Journal of Networks

ISSN 1796-2056
Volume 6, Number 5, May 2011

Contents

Special Issue: Nomadic Services and Applications

Guest Editors: Jason C. Hung

Foreword of Special Issue on "Nomadic Services and Applications"

Jason C. Hung
Department of Information Management, Overseas Chinese University, Taichung, Taiwan
E-mail: jhung@ocu.edu.tw

Nomadic Service has emerged rapidly as an exciting new paradigm that offers a challenging model of cyber-physical services and poses fascinating problems regarding distributed resource management, ranging from information sharing to cooperative computing. This special issue is intended to foster state-of-the-art research in the area of nomadic services and related applications, cloud computing technologies and services, including the topics of collaboration environment, implementation and execution of real-world architectures, and novel applications associated with this new paradigm. The published papers are expected to present high quality results for tackling problems arising from the ever-growing cloud computing technologies. The issue will serve as a landmark source for education, information, and reference to professors, researchers and graduate students interested in updating their knowledge about or active in cloud computing technologies and services. Nine papers are included in this special issue.

In the first paper entitled "Broadcasting Data Items with Time Constraints Based on On-Demand Multichannel in Ubiquitous Environments", Dr. Ding-Jung Chiang and Timothy K. Shih proposes mobile services with time constraints in ubiquitous computing environments based on on-demand multichannel broadcasting. This study provides on-line scheduling strategies for mobile data with timing constraint on multichannel broadcasting environments based on on-demand mode. The goal of this study ensures the most requests meeting their deadlines, which differs from previously proposed scheduling algorithms that aim to minimize the mean access time. The experimental results show the proposed algorithm, called on-demand broadcast program with time constraint (BPTC), outperforms miss rate for data timely delivery to mobile clients.

In order to illustrate a real-time hand gesture recognition system by using shape context matching and cost matrix, the second paper entitled "Real-time Hand Gesture Recognition by Shape Context Based Matching and Cost Matrix" by Prof. Lawrence Y. Deng, Jason C. Hung, Huan-Chao Keh, Lin Kun-Yi, Yi-Jen Liu, and Nan-Ching Huang tried to develop a perceptual interface for human-computer-interaction based on real-time hand gesture recognition. User could interact with computer program by performing body gesture instead of physical contact. The image of hand gesture was captured from CCD. The hand gesture image was transformed into proper instruction according to the shape information respectively.

Prof. Jiung-yao Huang, Huan-Chao Keh, Wai Shu-Shen, Ji-jen Wu, and Chung-Hsien Tsai purposed to enable users in a physical place to receive ubiquitous services from the environment while they communicate with each other unwittingly. In the third paper entitled "Research of Place-based 3D Augmented Community-Taking The 3D Virtual Campus as an Example", with the help of the augmented reality technique, on-the-spot member can visually sense the remote users by their representing avatars. To achieve this goal, the ambient communication environment is required to support message flow among the remote users and people on site. The context issues and context-awareness approaches of PDA community are fully discussed in the paper. Finally, the infrastructure of this PDA community is also presented along with preliminary result of the prototyping environment.

"Interactive E-diagnosis: An Efficient Scheme for Medical Diagnosis Supporting System" written by Rong-Chi Chang aims to design a prototype medical diagnostic support system, in which DICOM-based image analysis algorithms are utilized to develop an image browser and graphical user interface (GUI), allowing medical personnel to read X-ray, CT scan, MRI or other medical imaging files via a simple browser interface. The system also combines Internet functionality to provide remote medical diagnosis and patient data query features, all of which are conducive for doctors or medical personnel in different regions to cooperate and get hold of real-time test images and information of the patient. Follow-up studies can seek to incorporate a more convenient patient information entry and update interface into the system for automatic image analysis and information retrieval, where the data can be stored in the patient's personal information for up-to-date medical records.

In "A Web-based E-learning Platform for Physical Education", Prof. Chun-Hong Huang, Su-Li Chin, Li-Hua Hsin, Jason C. Hung, Yi-Pei Yu developed a Web-based E-learning Platform for physical education. The Platform provides sports related courseware which includes physical motions, exercise rules and first-aid treatment. The courseware is represented using digital multimedia materials which include video, 2D animation and 3D

doi:10.4304/jnw.6.5.687-688

virtual reality. The design concept of this project is based on ADDIE model with the five basic phases of analysis, design, development, implementation, and evaluation. Via the usage of this Web-based E-learning platform, user can learn the relative knowledge about sports at anytime and in everyplace. Authors hope to let players perform efficient self learning for sports skills, indirectly foster mutual help, cooperation, nice norms of law-abiding via the learning of exercise rules, and become skilled at accurate recreation knowledge and first-aid expertise.

Prof. Shih-Yang Yang, Ching-Ming Chao, Po-Zung Chen and Chu-Hao Sun proposed the ICspan algorithm to continue mining sequential patterns through an incremental approach and to acquire a more accurate mining result. In addition, due to the algorithm constraint in closed sequential patterns mining, the generation and records for sequential patterns will be reduced, leading to a decrease of memory usage and to an effective increase of execution efficiency. In the sixth paper entitled "Incremental Mining of Closed Sequential Patterns in Multiple Data Streams", the experiment results also support that ICspan algorithm will effectively reduce the sequential pattern records and consequently reducing the memory usage, while maintaining a sound mining efficiency under continuous data entries.

"Developing a Web-based and Competition-based Quiz Game Environment to Improve Student Motivation" written by Prof. Kuan-Cheng Lin, Ting-Kuan Wu and Yu-Bin Wang. In this study, authors develop a web-based and competition-based multi-player quiz game environment (WCMQGE) by incorporating DGBL into the web-based e-learning system. During the process of the game, the WCMQGE system can help players (students) review what they learned from school and use their knowledge to compete with their peers.

In "Using Active RFID to Realize Ubi-media System", Prof. Jason C. Hung purposed a human computer interface which is based on Active Radio frequency identification (Active RFID) technique to let human communicate with computer by analyzing the signal from tags. For retrieving those signals from tags, how to decreasing the noises created by surrounding environment and detecting useful information from variant signals are the most important. Author adopted a train procedure as pre-processing to categorize all of signals into two categories: noise and real data. After the real data is retrieved, we use "Music Director" and "DJ Scratch" as applications to let user play with computer.

In the ninth paper entitled "On the Design of A Contribution-based, Flexible Locality-Aware P2P Streaming Network", Yu-Wei Chan proposed the design of a two layered hybrid tree-push/mesh-pull overlay network with considering the problems of locality-aware of peers and incentive schemes. Prof. Chan proposed the schemes and algorithms to overcome the mismatch problem between the overlay network and the physical network, reduce the transmission latency, and provide a fully cooperative and reliable P2P streaming environment.

Jason C. Hung is an Associate Professor of Dept. of Information Management, Overseas Chinese University, Taiwan. His research interests include Multimedia Computing and Networking, Distance Learning, E-Commerce, and Agent Technology. From 1999 to date, he was a part time faculty of the Computer Science and Information Engineering Department at Tamkang University. Dr. Hung received his BS and MS degrees in Computer Science and Information Engineering from Tamkang University, in 1996 and 1998, respectively. He also received his Ph.D. in Computer Science and Information Engineering from Tamkang University in 2001. Dr. Hung participated in many international academic activities, including the organization of many international conferences. He is the founder and Workshop chair of International Workshop on Mobile Systems, E-commerce, and Agent Technology. He is also the Associate Editor of the International Journal of Distance Education Technologies, published by Idea Group Publishing, USA. The contact email is jhung@ocu.edu.tw.

Broadcasting Data Items with Time Constraints Based on On-Demand Multichannel in Ubiquitous Environments

Ding-Jung Chiang
Dept. of Computer Science and Information Engineering
Tamkang University
Dept. of Digital Multimedia Design
Technology and Science Institute of Northern Taiwan
Taipei, Taiwan, R.O.C
E-mail: djchiang@tsint.edu.tw

Chien-Liang Chen
Dept. of Tourism Information
Aletheia University
Taipei, Taiwan, R.O.C
clyde@cobra.ee.ntu.edu.tw

Chi-Yi Lin
Dept. of CSIE
Tamkang University
Taipei, Taiwan, R.O.C
chiyilin@mail.tku.edu.tw

Timothy K. Shih
Dept. of CSIE
National Central University
Taoyuan, Taiwan, R.O.C
timothykshih@gmail.com

***Abstract*—This study proposes mobile services with time constraints in ubiquitous computing environments based on on-demand multichannel broadcasting. For real-time mobile service, the need of mobile service for a real-time database management model is strong and providing real-time response to mobile transaction has became basic strategies in mobile data management. However, the resource constraints of ubiquitous computing systems make it difficult to satisfy timing requirements of supported strategies. This study provides on-line scheduling strategies for mobile data with timing constraint on multichannel broadcasting environments based on on-demand mode. The goal of this study ensures the most requests meeting their deadlines, which differs from previously proposed scheduling algorithms that aim to minimize the mean access time. The study presents a study of the performance of traditional real-time strategies and demonstrates traditional real-time algorithms which do not always perform the best in a mobile environment. The experimental results show the proposed algorithm, called on-demand broadcast program with time constraint (BPTC), outperforms miss rate for data timely delivery to mobile clients.**

***Index Terms*—ubiquitous computing; on-demand mode; time constraint**

I. INTRODUCTION

The advances in computer and communication technologies made possible the ubiquitous computing environment were clients equipped with portable devices can send and receive data any time and from any place. Due to the asymmetry in communication and the limit of wireless resources, data broadcast is widely employed as an effective method in delivering data to the mobile clients. For reasons like heterogeneous communication capabilities and variable quality of service offerings, we may need to divide a single wireless channel into multiple physical or logical channels. Thus, we need efficient algorithms for placing the broadcast data into these multiple channels so as to reduce the client access time.

A broadcasting mobile computing system based on on-demand mode consists of a number of mobile and fixed hosts. A fixed host (FH) is connected with each other via fixed high-speed wired network and constitutes the fixed part of the system. A mobile host (MH) is capable of connecting to the fixed network via a wireless link. Some of the fixed hosts, called mobile support stations (MSS), are augmented with a wireless interface to communicate with mobile hosts. The links between mobile computers and the MSSs can change dynamically.

In this study, the proposed method assumes that all fixed hosts act as mobile support stations (MSS). Each MSS has a database server which enforces strict data consistency. Each mobile host is associated with a coordinator MSS that coordinates the operations of the transactions submitted by that mobile host. Transaction execution descriptions are as the followings:

- Mobile host process (MHP): A mobile transaction exists at the generating mobile host.
- Fixed host process (FHP): A fixed transaction exists at the coordinator MSS of the generating host.
- Fixed cohort process (FCP): A process at fixed side deals with the required data between fixed and mobile host during transaction time.

Broadcast delivery based on on-demand mode [1][2] shown as figure 1 has been proposed and proven to be an efficient way of disseminating data to the mobile client population. This is due to the asymmetric nature of wireless communication, i.e., the downlink bandwidth is much higher than the uplink bandwidth. Associated with broadcast delivery is the problem of how to schedule the broadcasting of the requests to minimize the wait time of the clients. The wait time is also referred to as mean data access time, which is the average amount of time from the arrival of a request, to the time that the requested page

is broadcast. With broadcasting, the server can satisfy all pending requests on a data item simultaneously, thus, eliminating the potentially very large overhead of data requests, and saving both the wireless bandwidth and a mobile client's battery energy. Another feature is that it greatly increases the scalability of the broadcast system by keeping the server from being swamped with data requests.

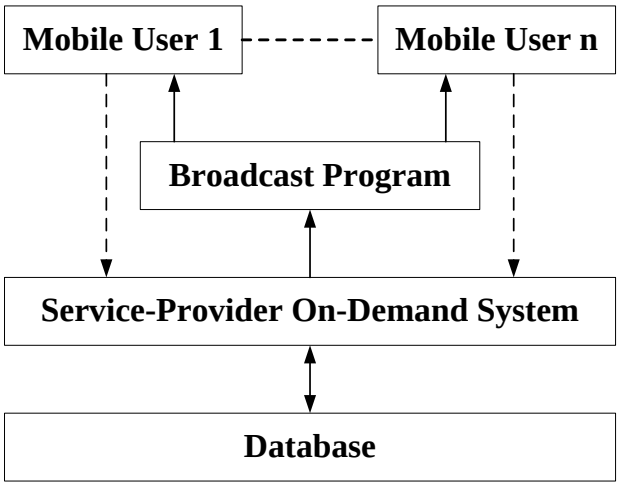


Figure 1. Broadcasting System Based on On-Demand Mode

With the rapid growth of time-constraint information services and business-oriented applications, there is an increasing demand to support quality of service (QoS) in mobile environments. In many situations, user requests are associated with time constraints as a measure of QoS. These constraints can be imposed either by the users or the applications. For example, the timing of buying/selling stocks for a stock holder is very crucial. If the stock information cannot reach a stock holder in time, the information might become useless. For another example, the information about traffic congestion that is caused by a traffic control should also reach a mobile client heading toward this direction timely. If a client receives such information early enough, the client is able to react accordingly to avoid the traffic jam. The value of the information would degrade significantly when the client gets closer to the spot of the control. With data broadcasting approach a broadcast server can serve many mobile clients simultaneously. Therefore, data broadcasting is usually adopted for disseminating data in mobile computing environments. Most of the related current research focuses on a data broadcasting approach, where the transmission of data is done without considering the data items with time constraints.

In this study, we present an on-line scheduling algorithm to maximize the total number of satisfied users in asymmetric communication environments with time requirements. This is achieved by means of dynamic adaptation of the broadcast program to the needs of the users, taking into account the bandwidth constraints inherent in asymmetric communication environments and the deadline requirements of the user requests. The goal of our research is to study scheduling strategies on mobile data with timing constraints. In such a ubiquitous environment, the goal is to determine how well the scheduling algorithms ensure that the database server does not miss deadlines, instead of minimizing wait time. There are several scheduling algorithms in mobile environments [3][4]. However, we demonstrate that traditional well-known algorithms do not always perform the best in a mobile environment, such as greedy and dynamic programming. We propose a model of scheduling mobile data with time constraint for a simulation analysis and also propose an efficient scheduling algorithm called scheduling priority of mobile data with time constraint (BPTC).

The rest of the paper is organized as follows. In section 2 we discuss related work for real-time strategies and mobile data dissemination. In section 3 we describe our model and propose the BPTC approach. In section 4 we present the experimental results and simulation environment. Finally, we conclude the paper in section 5.

II. RELATED WORK

Scheduling of transactions for real-time databases in a non-mobile environment is studied extensively in [5]. A real-time client/server model is considered in which the server assigns priorities to transactions based on several strategies, including Earliest Deadline First (EDF) and Least Slack (LS) first. As its name implies, for EDF the transaction with the earliest deadline is given the highest priority. For LS, the slack time is defined as: $d-\left(t+E-P\right)$, where d is the deadline, t is the current time, E is the execution time and P is the processor time used thus far. If the slack time is ≥ 0, it means that the transaction can meet its deadline if it executes without interference. The slack time indicates that how long a transaction can be delayed and still meet its deadline. The Least Slack LS differs from EDF because the priority of a transaction depends on the service time it has received. If a transaction is restarted, its priority will change. Simulation results show that the EDF is the best overall policy for real-time database systems in a non-mobile environment. However, when system loads are high, the LS and EDF strategies lose their advantage, even over FCFS, as most transactions are likely to miss their deadlines.

For mobile systems, the longest wait first LWF algorithm has been shown to outperform all other strategies at minimizing wait time [6]. In LWF, the sum of the total time that all pending requests for a data item have been waiting is calculated, and the data item with the largest total wait time is chosen to disseminate next. However, LWF has been recognized as expensive to implement. In [7] a strategy, called requests times wait (RxW), is presented for mobile systems that makes scheduling decisions based on the current state of a queue (instead of access probabilities). The RxW algorithm provides an estimate of the LWF algorithm by multiplying the number of pending requests for a data item times the longest request wait time. In general, the

performance of the approximate algorithms has been shown to be close to LWF.

There has been some research work to consider disseminating for mobile real-time systems [8]. A transfer protocol for organizing air disks for real-time applications, called Adaptive Information Dispersal Algorithm (AIDA), is presented in [9]. In this work, the data must be disseminated periodically to satisfy the timing constraints. The AIDA protocol considers fault-tolerance and the data items are allocated to the air disks to minimize the impact of intermittent failures by utilizing redundancy. AIDA guarantees a lower bound on the probability of meeting timing constraints. Similar work addressing fault-tolerant real-time air disks appears in [10]. In this work, the authors show that designing strategies for real-time air disks is related to pinwheel scheduling. The authors derived a pinwheel algebra, which utilizes rules that can be used to construct fault-tolerant real-time air disks.

Their works differ from our work because we assume that we schedule all data items with time constraints using adaptive algorithms under limited bandwidth to minimize miss rate. In the air disks model, the server periodically repeats a computed priority program, based on user access patterns. Deadline constraints have been integrated into the priority program in [3]. In order to minimize the total number of deadlines missed by making the most effective use of the available bandwidth, scheduling approach has to focus on critical factors such as access frequency, time constraint, and bandwidth requirements. In [11], scheduling mechanisms for disseminating data that are to minimize the delay incurred by insufficient channels, but it is reasonable that all clients are satisfied with an expected time to optimize average access time.

III. DISSEMINATING MOBILE DATA WITH TIME CONSTRAINTS

We now describe a framework to support disseminating mobile data with time constraint. In this section, the real-time scheduling problem, system architecture and solving mechanism are introduced.

A. System Architecture

The most common assumption about broadcasting is that there exists a single physical channel for the broadcasted data. There are, though, many scenarios where a server has access to multiple low-bandwidth physical channels which cannot be combined to form a single high-bandwidth channel. Possible reasons for the existence of multiple broadcast channels include application scalability, fault tolerance, reconfiguration of adjoining cells, heterogeneous client communication capabilities, and so forth [12][13].

In the following paragraphs, we give example scenarios of the aforementioned reasons.

- Scalability of system: Consider an application running on the server that needs to be scaled in order to support a larger number of clients. In this case, it may need to acquire additional physical channels. If these channels are in noncontiguous frequencies, they may have to be treated as separate channels when broadcasting data.
- Fault tolerance of system: Suppose that a transmitting station can have more than one server with a transmission capability, and let A, B and C be servers which broadcast data in three noncontiguous frequency ranges all in the same cell. If servers B and C crash, then frequencies assigned to them should be allocated to server A.
- Reconfiguration of adjoining cells in mobile environments: Suppose that there are two adjacent cells whose servers transmit in different frequency ranges and, at some point in time, we decide to ”merge” the two cells and use one of the two servers to serve the newly generated cell. Then, the frequency range of the other server should be migrated and added to the residual server. In this case also, the latter server gets multiple physical channels.
- Heterogeneous clients in mobile environments: The mobile clients may have heterogeneous communication capabilities, precluding the existence of a single high-speed transmission channel.

There are several concerns related to the exploitation of a multichannel broadcast system. The first consideration is related to the capability of the server to concurrently transmit in all channels; the second is related to the capability of the client to simultaneously listen to multiple channels and also perform instant ”hopping” among channels. Since the interest of this study is to focus on the time-constrained data placement problem, we will make simplifying assumptions, considering that the server is able to concurrently transmit to all channels, and the clients are able to listen simultaneously to all channels and perform instantaneous hopping from channel to channel. Moreover, we do not assume any kind of dependencies between broadcasted data. The two major issues in broadcast dissemination are how and what the server transmits and how the client retrieves.

The particular interests are the solutions that enable the mobile clients to get the disseminated data efficiently, that is, with short access latency and with minimum power expenditure. The former is quantified by the query access time, which is equal to the time elapsed between when the client starts seeking for an item until it gets it. The latter is quantified by the tuning time, which is the time the client spends actively listening to the broadcast channel. The access time is directly related to the size of the broadcast. On the other hand, providing information to the clients for selective auto tuning, that is, indexes, reduces the tuning time. However, including such information increases the overall size of the broadcast, which in turn increases the access time. The trade-off between these two performance measures is obvious. The focus of this study is the time-constrained data placement issues, so we assume that the clients have complete knowledge of the broadcast program. In other words, they know a priori the arrival time and channel of all broadcasted data items. Hence, we are only interested in

minimizing the client access time according to access probability of broadcasted data and optimizing miss rate of time constraints for broadcasted data so that we do not assume the existence of index packets in the broadcast.

Though, in order to make the broadcast program "predictable," we are interested in placement schemes, which guarantee that the broadcast is cyclic, that is, the schedule has a beginning and an end, and the inter arrival time between successive transmissions of an item is constant for all the broadcast cycles. Under the aforementioned assumptions, the problem addressed by this can be captured by the following definition:

Definition III.1 Given a number of identical broadcast channels and knowledge of the data item probabilities, we have to decide the contents of the multiple channels in order to reduce the average access time.

The descriptions of operation model in the mobile environment are as the following:

- Server side: We assume that there are K priority programs which server generates by algorithm we propose, each program denoted C_i, $1 \le i \le K$. A database is made up of N unit-sized items, denoted d_j, $1 \le j \le N$. Each item is disseminated on one of these programs, so program C_i disseminates N_i items, $1 \le i \le K$, $\sum_{i=1}^{K} N_i = N$. Each program cyclically disseminates its items. Time is slotted into units called ticks. The size of data item is fixed and equal to one tick. Each data item is denoted $d_i(id_i, t_i, p_i)$ by the following parameters[14]:
  - id_i: Identifier of data item.
  - t_i : Relative deadline, i.e. the maximum acceptable delay for its processing.
  - p_i: Access probability for d_i.

  Requests are for single item and assumed to be exponentially distributed.
- Client side: Each client can require one data item per request associated with a time constraint. When a client needs a data item, it first tunes in the hot spots to retrieve the contents of program. By the information provided cover area, the client can determine whether he can get the data item from the priority program. If the needed data item is in the priority data set, the client retrieves the desired data item. Otherwise, the client sends a request to the server via the uplink channel, and listens to the other program to retrieve the data pages.

B. Problem Formulation

We formulate our problem to make it a resolvable problem as follows. Given a number of data items N to be disseminated in multiple priority programs K. Each data item is associated with a time constraint. Every access of a client is only one data item. Expected delay, w_i, is the expected number of ticks a client must wait for the dissemination of data item d_i. Average expected delay is the number of ticks a client must wait for an average request and is computed as the sum of all expected delays, multiplied by their access probabilities:

$$Average\ Expected\ Delay(W) = \sum_{i=1}^{N} w_i p_i \qquad (1)$$

, where w_i is expected delay [15] and p_i is access probability for data item d_i respectively. With time constraints, a request for data item d_i has missed its deadline when timing fault (expected delay for data item d_i exceeds its time constraint $t_i < W$) occurred at some time slot. The miss rate of all data items is defined as follows:

$$Miss\ Rate = \sum_{j=1}^{K} \sum p_i \qquad (2)$$

Our goal is to disseminate all data items with time constraints over ubiquitous computing environments that minimize the miss rate.

C. Design of Algorithm BPTC

We provide an algorithm to generate a valid priority program so as to minimize miss-rate. If miss-rate is zero, let average access time minimized. We formulate our problem and make appropriate assumptions to make it a resolvable problem as follows.

Let each item contains two attributes: access probability and time constraint. Given a database D with its size $|D| = N$ and the number of channels = K, we aim to allocate each item in D into K channels, such that N items are cyclically disseminated in a service area, the miss rate can be written as:

$$\sum_{i=1}^{K} \left(\sum_{d_i \in C_i} p_i \right) \qquad (3)$$

Given an example using above the assumptions, we provide our algorithm (BPTC) based on dynamic programming which deals with optimal solutions [16]. Figure 2 illustrates the original overview according to example 1 without adjustment and exists with miss rate. Figure 3 illustrates the better data distribution and improved result (miss rate = 0) after adjustment by the proposed algorithm (BPTC).

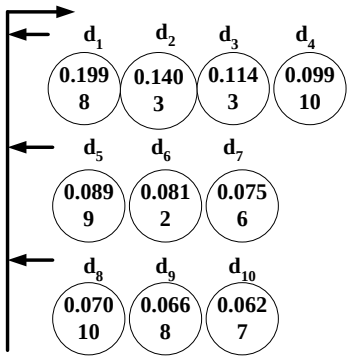
Example 1. Given $K=3$ *dissemination programs, we consider a set of* $N=10$ *data items with the following skewed access probabilities and time constraints:*

$\{d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, d_9, d_{10}\}$

d_1	d_2	d_3	d_4	d_5
0.199	0.140	0.114	0.099	0.089
8	3	3	10	9
d_6	d_7	d_8	d_9	d_{10}
0.081	0.075	0.070	0.066	0.062
2	6	10	8	7

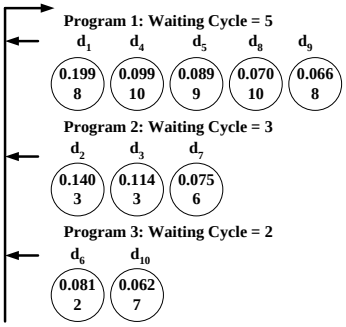
Algorithm 1 BPTC(int N, int K, float P, int T)

```
{N: number of items, K: number of channels}
{P: access probabilities, T: time constraints}
Require: N unit sized items ordered by popularity.
Ensure: K partitions to minimize miss-rate.
  Measure the priority of data items by data server;
  Partition_number = 1;
  while Partition_number < K do
    {Find the best point s to split in partition k}
    {Initialize the best split point for this partition as the first
    data item. If we find a better one subsequently, update
    the best split point.}
    {C_ij^d is computed as the expected delay of a data item in
    a channel of size j − i + 1}
    {Initialize the best solution as the one for the first
    partition. If we find a better one subsequently, update
    the best solution.}
    for each partition k with data items i to j do
      for (s = i; s ≤ j; s = s + 1) do
        if ((s = i) or (Local_change > C_ij^d)) then
          Local_S = s;
          Local_change = C_ij^d;
        end if
        if ((k=1)or(Global_change>Local_change)) then
          Global_change = Local_change;
          Global_S = Local_S;
          Best_part = k;
        end if
        Split partition Best_part at point Global_P;
        Partition_number = Partition_number + 1;
        for data items on each channel do
          Evaluate the priority of data items by data server;
          Minimize miss rate on each channel;
          Adjust data items from high level to low level
          according to priority value for data items;
        end for
      end for
    end for
  end while
```



For program 1, waiting cycle = 4
Time constraint of d_2 = 3 and d_3 = 3 < waiting cycle = 4
Miss rate = 0.140 + 0.114 = 0.254
For program 2, waiting cycle = 3
Time constraint of d_6 = 2 < waiting cycle = 3
Miss rate = 0.081
For program 3, waiting cycle = 3
Time constraint of d_8, d_9 and d_{10} > waiting cycle = 3
No miss rate
Total miss rate = 0.254 + 0.081 = 0.335

Figure 2. Disseminating 10 data items with time constraints to partition 3 disjoint subsets and the miss rate = 0.335.



Time constraint for each data item on its channel > Waiting cycle, total miss rate = 0

Figure 3. Using BPTC algorithm to adjust 3 priority programs and the miss rate = 0.

IV. EXPERIMENTAL RESULTS

A. Simulation Environment

In our simulation model, bandwidth is not explicitly modeled. Instead, similar to previous work [7], we use dissemination tick as a measure of time. The greatest advantage of this approach is that the results are not limited to any particular bandwidth and/or data item size. Rather, it aims to capture the fundamental characteristics of the systems.

The model simulates a one-hop wireless network. All data items are stored in a data server in a fixed location. Mobile clients need to send requests to the server via an uplink back-channel before the requested page can be disseminated. The arrival of requests generated by mobile clients follows a Poisson process and the inter-arrival time is exponential with mean λ. Each request has a request id, arrival time and deadline. For each page, a

queue is maintained to store the information about requests on the object. We assume the results produced after the deadlines are useless (firm deadlines), so all requests that have missed their deadlines are discarded. Mobile clients are responsible for re-sending requests when link errors occur. We also assume a deadline can capture the mobility of clients who are no longer able to receive the priority program. In our model, since newly generated data requests are sent to the server immediately, the request generating time is equal to the time the server receives it (assuming network delay is ignored). We also ignore the overheads of request processing at the server, because the main purpose of the model is to compare the scheduling power of various strategies. We assume requests generated by mobile clients are read only, and no update request is allowed.

Concurrency control issues are not our main concern, and thus, not considered. At each tick of the simulation clock, the following occurs. A simulated request generator generates requests with exponential inter-arrival time. The information about each request id, arrival-time and deadline is recorded. The request is then inserted to the corresponding queue. The server checks the deadlines of all the arrived requests, and discards those requests that have missed their deadlines. Then the server selects a page to disseminate by applying a scheduling strategy and starts to disseminate the selected page. All requests requesting the page are satisfied when the disseminated program is finished. A client can request multiple pages and a page can be requested by multiple mobile clients at a time. We assume that data demand probabilities p_i follow the $Zipf$ [17] distribution in which:

$$p_i = \frac{\left(1/i\right)^{\theta}}{\sum_{i=1}^{M}\left(1/i\right)^{\theta}}, \left(i = 1, 2, 3, ..., M\right)$$

where p_i represents the $i'th$ most popular page. The $Zipf$ distribution allows the pages requested to be skewed. Figure 2 shows the results of our simulation comparing the BPTC strategy to the other strategies for uniformly distributed deadlines.

Symbol	Description	Default	Range	Unit
DBSIZE	Total number of data pages stored in server	100	100-10000	pages
λ	Mean request arrival rate (exponential)	-	2-60	requests/tick
θ	Request skewness (Zipf)	1.0	0.0-1.0	-
$MinSlack$	Minimum slack time	1.0	-	ticks
$MaxSlack$	Maximum slack time	-	20-300	ticks
$\lambda_{deadline}$	Parameter of exponential deadline distribution	-	10-300	ticks

Table 1 Simulation Parameters

B. Simulation Parameters

We compare the BPTC approach with the algorithms described in Section 2: EDF, LSF, RxW, and LWF. Figure 4 illustrates the distribution of miss rate with the comparison from the aforementioned approach and the proposed method (BPTC). We only choose the scheduling algorithms to compare the results since we believe these scheduling algorithms better adapt to the dynamic changes of the intensity and distribution of system workloads. Figure 5 illustrates the comparison of miss number when the client number is increased and figure 6 illustrates the comparison of miss number when the channel number is varied. The scheduling (access probabilities, dissemination histories, etc.) off-line algorithms are not considered due to the fact that they are mainly for fairly stable systems. We implement the simulation model described in the previous section using C++. In each experiment, we run the simulation for 500 time units, and we use an average of 20 runs of each simulation as the final result. The Parameters used in this simulation are summarized in table 1. The default total number of data pages stored in the server, referred to as DBSIZE, is 100 pages. Client requests reach the system with exponential inter-arrival time with mean λ, and λ is varied in our simulation from 2 - 60. It is assumed that each data request requires 1 delivery tick to disseminate.

An open system model is used to simulate the system for extremely large, highly dynamic populations. Data access follows a skewed $Zipf$ distribution with parameter θ to control the skew. The minimum slack time is 10, with the maximum slack time ranging from 20 to 300. This variation in maximum slack time allows us to vary the tightness in the deadlines. In addition to a uniform distribution of deadlines, an exponential distribution is utilized with lambda ranging from 10 to 300. After doing a large number of experiments with various factors that affect the performance, we come up with an overall performance comparison between the previous algorithms and our scheduling algorithm BPTC in table 2. We grade the level of performance from 1 to 5. The higher the degree is, the better the performance is. On the contrary, overhead with higher degree shows that the algorithm gets more cost in table 2.

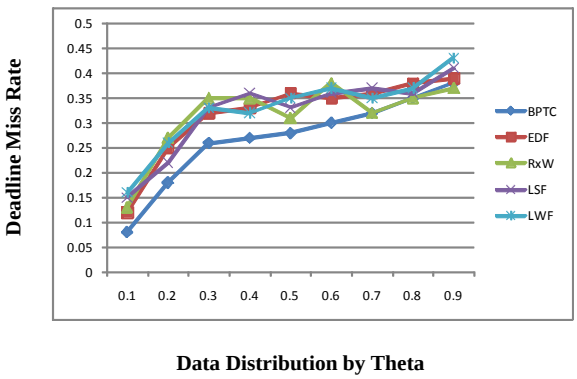


Figure 4. Deadline Miss Rate by Data Distribution

Figure 5. Deadline Miss Number by Client Number

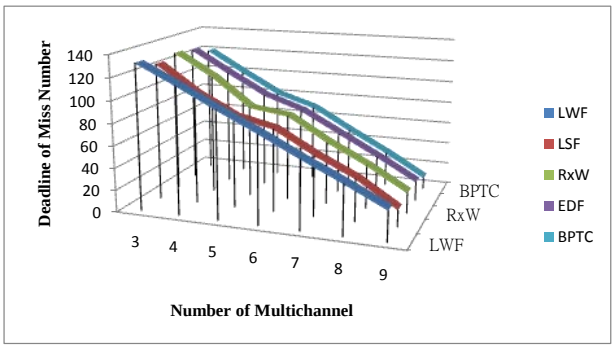


Figure 6. Deadline Miss Number by Channel Number

	DMR	ART	AS	Overhead
LWF	1	2	3	5
LSF	1	1	2	5
RxW	3	5	4	4
EDF	3	3	1	4
BPTC	5	5	4	3

DMR = Deadline Miss Rate
ART = Average Response Time
AS = Average Stretch

Table 2 Performance comparison of different scheduling algorithms

REFERENCES

[1] W. SUN, W. SHI, and B. SHI, "A cost-efficient scheduling algorithm of on-demand broadcasts," *Wireless Networks*, vol. 9, pp. 239–247, 2003.

[2] M. A. SHARAF and P. K. CHRYSANTHIS, "On-demand data broadcasting for mobile decision making," *Mobile Networks and Applications*, vol. 9, pp. 703–714, 2004.

[3] K. Prabhakara, K. A. Hua, and J. H. Oh., "Multi-level multi-channel air cache designs for broadcasting in a mobile environment." in *Proceeding of the 16th International Conference on Data Engineering.*, February 2000, pp. 167–176.

[4] Q. L. Hu, D. L. Lee, and W.-C. Lee., "Dynamic data delivery in wireless communication environments." In *Proceedings of International Workshop on Mobile Data Access*, November 1998, pp. 218–229.

[5] R. Abbott and H. Garcia-Molina, "Scheduling real-time transactions: A performance evaluation," *ACM Transactions on Database Systems*, vol. 17, pp. 513–560, 1992.

[6] J. W. Wong, "Broadcast delivery," *Proceedings of the IEEE*, vol. 76, no. 12, pp. 1566–1577, Dec. 1988.

[7] D. Aksoy and M. Franklin, "Scheduling for large-scale on-demand data broadcasting," in *Proceedings of the INFOCOM Conference*, March 1998, pp. 651–659.

[8] J. Fernandez-Conde and K. Ramamritham., "Adaptive dissemination of data in time-critical asymmetric communication environments." In *Proceedings of the 11th Euromicro Conference on Real-Time Systems*, 1999, pp. 195–203.

[9] A. Bestavros., "Aida-based real-time fault-tolerant broadcast disks." In *Proceedings of Real-Time Technology and Applications Symposium.*, 1996, pp. 49–58.

[10] S. Baruah and A. Bestavros., "Pinwheel scheduling for fault-tolerant broadcast disks in real-time database systems." in *Proceedings of the 13th International Conference on Data Engineering.*, April 1997, pp. 543–551.

[11] Y.-C. Chung, C.-C. Chen, and C. Lee., "Time constrained service on air." in *Proceedings of the 25th International Conference on Distributed Computing Systems*, June 06-10 2005, pp. 739–748.

[12] S. Acharya, R. Alonso, M. Franklin, and S. Zdonik., "Broadcast disks: data management for asymmetric communication environments." In *Proceeding of ACM SIGMOD*, March 1995, pp. 199–210.

[13] S. Acharya, M. Franklin, and S. Zdonik., "Dissemination-based data delivery using broadcast disks." *IEEE Personal Communications*, vol. 2, no. 6, pp. 50–60, December 1995.

[14] F. Cottet, J. Delacroix, C. Kaiser, and Z. Mammeri, "Scheduling in realtime systems," in *Scheduling in Real-Time Systems*. John Wiley and Sons Ltd, 2002.

[15] W. G. Yee and S. B. Navathe., "Efficient data access to multi-channel broadcast programs." in *Proceedings of the 12th International Conference on Information and Knowledge Management.*, 2003, pp. 153–160.

[16] T. H. Cormen, C. E. Leiserson, and R. L. Rivest., "Introduction to algorithms," in *Introduction to Algorithms*. The MIT, 1992.

[17] G.K.Zipf., "Human behaviour and the principle of the least effort," in *Proceedings of the 25th International Conference on Distributed Computing Systems*. Reading,MA: Addison-Wesley, 1949.

Ding-Jung Chiang is an instructor of Department of Digital Multimedia Design, Technology and Science Institute of Northern Taiwan, Taipei, Taiwan. He received the bachelor and master degrees in computer science and information engineering from Tamkang University, Taipei, Taiwan, in 1995 and 1998, respectively. He is currently a Ph.D. student in the Department of Computer Science and Information Engineering at the Tamkang University, Taiwan. His research interests include ubiquitous computing, embedded system and wireless multimedia.

Chien-Liang Chen received the Ph.D. degree in Electrical Engineering from National Taiwan University in 2009. He is an assistant professor of Tourism Information at Aletheia University, where he initially joined in February 2010. His current research interests include design and analysis of algorithms, wireless sensor networks and discrete

event systems.

Chi-Yi Lin received the B.S. and Ph.D. degrees in electrical engineering from National Taiwan University in 1995 and 2003, respectively. He was a visiting researcher at AT&T Labs-Research, New Jersey from August to December, 2000, an Assistant Researcher at the Telecommunication Labs, Chunghwa Telecom from 2003 to 2007, and a Postdoctoral Research Fellow at the Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology from 2007 to 2008. He joined the Department of Computer Science and Information Engineering, Tamkang University as an Assistant Professor in August, 2008. His research interests include fault tolerant distributed/mobile computing systems, RFID-based applications, and telecommunication systems.

Dr. Shih is a Professor of the CSIE department at National Central University, Taiwan. He was the Dean of College of Computer Science, Asia University, Taiwan and the Department Chair of the CSIE Department at Tamkang University, Taiwan. Dr. Shih is a Fellow of the Institution of Engineering and Technology (IET). In addition, he is a senior member of ACM and a senior member of IEEE. Dr. Shih also joined the Educational Activities Board of the Computer Society. His current research interests include Multimedia Computing and Distance Learning. Dr. Shih has edited many books and published over 440 papers and book chapters, as well as participated in many international academic activities, including the organization of more than 60 international conferences. He was the founder and co-editor-in-chief of the International Journal of Distance Education Technologies, published by the Idea Group Publishing, USA. Dr. Shih is an associate editor of the ACM Transactions on Internet Technology and an associate editor of the IEEE Transactions on Learning Technologies. He was also an associate editor of the IEEE Transactions on Multimedia. Dr. Shih has received many research awards, including research awards from National Science Council of Taiwan, IIAS research award from Germany, HSSS award from Greece, Brandon Hall award from USA, and several best paper awards from international conferences. Dr. Shih has been invited to give more than 30 keynote speeches and plenary talks in international conferences, as well as tutorials in IEEE ICME 2001 and 2006, and ACM Multimedia 2002 and 2007.

Real-time Hand Gesture Recognition by Shape Context Based Matching and Cost Matrix

Lawrence Y. Deng
Department of CSIE, St. John's University
Lawrence@mail.sju.edu.tw

Jason C. Hung
Department of Information Management
Overseas Chinese University
jhung@ocu.edu.tw

Huan-Chao Keh, Kun-Yi Lin, Yi-Jen Liu, and Nan-Ching Huang
Department of CSIE, Tamkang University
f08572@ms7.hinet.net

***Abstract*—How to recognize the shape gesture for new human-computer interface without controller required and bring entertainment, games industries and information appliances in new ways. In this paper, we would illustrate a real-time hand gesture recognition system by using shape context matching and cost matrix. The shape context is taken as a basis description for shape matching. It can be regarded as a global characterization descriptor to represent the distribution of points in a set with scale and rotation invariance. In this paper, we developed a perceptual interface for human-computer-interaction based on real-time hand gesture recognition. User could interact with computer program by performing body gesture instead of physical contact. The image of hand gesture was captured from CCD. The hand gesture image was transformed into proper instruction according to the shape information respectively. The instruction was transferred to an appropriate program to execute. The experience of our preliminary results shown the precision rates was up to 70% ~ 90%.**

***Index Terms*—Hand Gesture Recognition, Shape Matching, Cost Matrix and Human-Computer Interface**

I. INTRODUCTION

As technology rapidly advances, mouse and keyboard are not the only two necessities to control computer. Many substitutions for mouse and keyboard are being hard developed, such as speech recognition and gesture recognition. In this paper, the 'Virtual Petting Game' was taken as an example. We tried to present a novel interaction style on this game. By computer vision techniques, users could play game more intuitively and could interact with computer game without mouse and keyboard. This computer game system captured the image of user's hand gesture by video device. Users could hand down an order by performing hand gesture in front of a video camera. During playing with 'Virtual Pet' or applying this interaction style on other games, users would have exceptional experience. Through this mode of manipulation, users could be more active, explored more deeply, had more fun in the virtual world and felt that the game could be a part of real life.

We utilized computer-vision based approach in this study. Users didn't need to mark any colored sign on the hand or wear any glove or sensor. With only a single digital camera, the image of user's hand gesture was captured and analyzed. The proposed system of scenario contained following possible five steps (as shown in fig.1):

(1) The image was captured through CCD camera.
(2) The hand gesture image was segmented from the image.
(3) Identifying the hand gesture.
(4) Recognizing hand gesture and matching with proper instruction.
(5) Triggering the entertainment games.

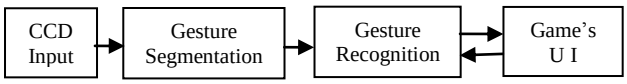


Figure 1. The scenario of real-time hand gesture recognition system

In this paper, we outlined the implementation procedures of real-time hand gesture recognition and then to interact with entertainment game. The related technologies and processes of hand gesture segmentation were illustrated in section II. The section III discussed the hand gesture matching based on shape context and cost matrix. The implementation considerations of the real time hand gesture human-computer interface system were discussed in section IV. Finally, we made a brief conclusion in section V.

II. RELATED TECHNOLOFIES OF HAND GESTURE SEGMENTATION

doi:10.4304/jnw.6.5.697-704

The first step is trying to segment the background image and eliminate impropriated pixel from the captured image initially. Then the hand region would be detected and segmented from the image. The contour of hand gesture would be described. The significant information would be extracted by shape context based approach. Using this information, we could search corresponding hand gesture in pre-defined gesture-base, such that user's hand gesture could be transformed into proper instruction to trigger correct action.

There were several interesting image processing issues for system implementation that included Background Separation, Color Tracking, Object Segmentation, Hand Extraction, Pattern Recognition, Hand Gesture Recognition, and Shape Context etc... We will discuss in further detail in the following sections.

The seperation of the Background and foreground

For the purpose of background subtraction, we could set up a reference image to describe the conformation for the background model [1]. And the background model would provide the upper limit and lower limit of each pixel's variation. It would be a foreground image if the pixel variation was exceeded those limit value obviously. It was also caused the pixel's variation value over the upper limit or lower limit if a worse background existed occasionally. We could solve this problem with an observation from background subtraction mask periodically.

Generally, the mask was a binary image that composed with black and white elements that provided foreground and background information individually. The foreground area was a large adjacent block of white area which was noticed in a user's interaction mode normally. It took about one corner or one to two block area of the image approximately. There were two kind of worse situations to be considered:

(1) When the situation of the shooting scene was worse, such as the lighting changed slowly, the decorated objects switched or the position of camera was unstable, etc. The mask would contain more small area of extension area for the expected block or image everywhere. This small area could be classified as an unnecessary noise that could be detected by the searching of small individual area with white image value as well.

(2) If the environment condition was improved thoroughly, such as the lighting was switched on/off or the camera was shielded by a large object, we also considered it was a stain without any small area of noise that the whole mask was consisted by a large white stain conditionally.

There were two measure methods for classifying of noisy signal:

(1) We defined a minimal foreground block; it would be a noisy signal if the area was smaller than that area.

(2) We defined a maximum foreground block of relative image; it would be an interference area if the area was bigger than that area.

Color Tracking

The purpose of color tracking was tracking a moving color area in a target image. It could input the result of extracted background with CAMSHIFT algorithm [2] or Kalman Filter Tracking algorithm [3] alternatively.

Hand Extraction

We had extracted a shape of hand with Canny Edge Detector. It was proposed by J. Canny for the edge detection algorithm. There was a grayscale image for the input, and came out a bi-level image that marked as detected edge with non-zero pixel.

Hand Gesture Recognition

There were three patterns defined for the hand gesture basically: The Static Hand Poses Gestures, The Simple Hand Path Gestures and The Staged Hand Path Gestures.

(1) The Static Hand Poses Gestures was a single hand pose in a space room invariably. It could be a one symbol in the sign language alphabet system probably.

(2) The Simple Hand Path Gestures was a hand pose that described a normal tracking route to compare with the primitive shape. It could estimate the characteristic coordinate of hand shape, such as the mass of the hand. And it also compared the sketch of produced data to a primitive shape simultaneously.

(3) The Staged Hand Path Gestures was a hybrid gesture that combined with the static hand poses gestures and the simple hand path gestures together. It similar to a vector based trajectories with the poly-lines certainly. Therefore, we could divide up the hand gesture to the control points of multi-lines with localized hand postures naturally.

The relative techniques of these three hand gestures were discussed as followings separately:

(1) The Static Hand Pose Recognition

We had applied the concept of shape context into the static hand pose recognition currently [4][5][6]. The shape context was a new shape descriptor that provided the correspondence recovery of shape information and the identification of shape-based object practically. A shape context of each point recorded the distributional status of its relative position. It described a shape configuration with a valuable, local descriptor generally. And the shape context simplified the recovery of correspondences of two known shape points mass effortlessly. The shape context introduced an exhaustive and comprehensive data that could measure the similarity of shape easily. And the descriptors of shape context could accept all shape deformation without any special landmarks or key points fundamentally.

(2) The Simple Hand Path Recognition

We could describe a tracking route of a moving hand with a projected image on a flat board, and compared it with a simple sketched shape, such as triangle, circle, square or ellipse. Thus we could identify the trajectory or path with the geometric property of the projected

image briefly. There were three mainly concepts for the recognition of hand path which shown as followings [7] [8]: (a) The identifier was depended on the geometry information principally. (b) To filter unnecessary shape in a specific standard with decision tree accordingly. (c) To differentiate the degrees of certainty of the identified shape with fuzzy logic algorithm every so often [16].

(3) The Hybrid hand path recognition

The method of the hybrid hand path recognition was based on the vector analysis and localized hand pose identification generally. We could describe the tracking and path by user's hand which was according to a group of limited control points that determined by the video capture rate and speed of hand pose. And it was not suitable for the vector analysis directly. Therefore, the key point should be identified in a plotted curve previously.

We could summarize a group of sector by a high curve rate was zero with a multi-sector searching basically. We collected these sectors with its geometry property with the algorithm of Douglas-Peucker approximation and a proper parameter precisely. And the vertex point of shown hand gesture would be verified as soon as it needed.

Shape Context

Shape context is a new shape descriptor presented by Serge Belongie and Jitendra Malik. They proposed this idea in their paper "Matching with Shape Contexts" in 2000[10]. The shape context describes the coarse arrangement of the shape with respect to a point inside or on the boundary of the shape. It can be used for measuring shape similarity and recovering point correspondences.

Microsoft Project Natal

Project Natal is the code name for a "controller-free gaming and entertainment experience" by Microsoft for the Xbox 360 video game platform [11]. It enables users to interact with the Xbox 360 without touching a game controller physically through a perceptual user interface using body gestures, spoken commands, or presented objects and images.

III HAND GESTURE MATCHING BASED ON SHAPE CONTEXT

The description of hand gesture was based on the Shape Context algorithm [4]. We also focused on the how to search the possible position of hand gesture and the technique of image comparison simultaneously. It was integrated with the algorithm that included the shape sampling, image shape calculation, calculation of the shape descriptors, and variation of the shape descriptors with cost matrix. The whole idea was input a hand video image from video camera, and positioned the sample matrix of hand gesture from the points of shape context. The following chapter would discuss more for the key process as well.

Shape Sampling

The first step of shape context analysis was translated the edge elements of image shape to a group of feature points with N value. These points could be inside or outside of the image shape simultaneously. Also, there would not be the key-point of the shape normally, such as an apex. We took these sample shapes with a roughly equal range generally. For example, we got a shape sample data from the digitized gesture image. Meanwhile, the sampling point of edge elements would be collected as figure 2. Then, we could calculate every point of shape context for the shape descriptors which would be used in the following analysis,

The 'C' would be the collection of all shape context point.

$$C = \{C_1, C_2, ...C_t\}, C_i \in R^2 \quad (1)$$

The t would be the total number of the shape context points. And the D2 would be a two-dimension matrix of $t \times t$:

$$D2_{(i,j)} = (C_i \cdot x - C_j \cdot x)^2 + (C_i \cdot y - C_j \cdot y)^2 \quad (2)$$

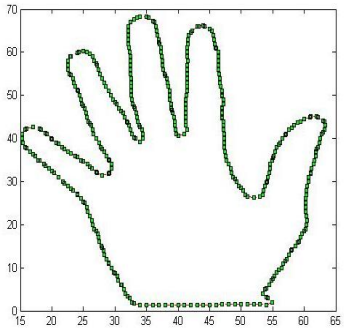


Figure 2. The image of Shape context point

We sampled the number of 'num_sample' points from the shape context with following algorithms:

```
while (Length(C) > num_sample) {
 //while collection C was bigger than num_sample, the loop would continued.
 [a,b]=min(D2);
 //a and b would be a vector of the row. The 'a' was the minimum value of each row in the D2 matrix, and 'b' was the number of row of the 'a' in that row of D2 matrix.
 [c,d]=min(a);
//The 'c' was the minimum value in the vector 'a'. And the 'd' was the index point of the minimum value in the vector 'a'.
 I=b[d];
//'I' was the number of row of the minimum value in the matrix D2.
 J=d;
//'J' was the number of column of the minimum value in the matrix D2. It deleted the Jth element from C.
//The Jth element was deleted from D2 matrix. It removed the Jth row from D2.
//The Jth column was deleted from D2 matrix.
```

}
//The 'C' was the point collection of the size of 'num_sample'. It also meant we collected a shape context after sampling with the number of 'num_sample' precisely (Refer to figure 3).

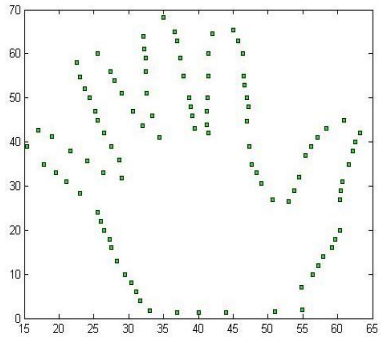


Figure 3. The image of sample points after shape context sampling (200 pixels)

The calculation of distance and angle within sample points

We took the coordinate of each sample point and other n-1 points relatively. And the data was stored into a distance array [r] and an angle array [theta] individually. The calculation of the distance array [r] was defined as following:

The 'r_array' was a $t \times t$ 2-dimension array, the t was the length of the collection of sample points C. And the 'r_array' stored the distance of every point and its next point orderly.

$$r\_array_{ij} = \sqrt{(c_i - c_j)^2} \qquad (3)$$

The calculation of angle array [theta] was defined as following: The 'theta_array' was a $t \times t$ 2-dimension array, the t was the length of the collection of sample points C. And the 'theta_arrayij' stored the tangent value of every two points and its next point orderly.

theta_arrayij = $\tan^{-1}$(Ci.y - Cj.y, Ci.x - Cj.x) *(4)*

Normalized Range

We define a 'dist' with a 'mean' function to calculate the average distance of every point. And the 'r_array_n' was a $t \times t$ 2-dimension array that stored the divided value of the distance of each points and its average distance precisely.

dist = mean(r)

The 'r_array_n' was a $t \times t$ matrix.

// t = Length(C), where r_array_nij = r_arrayij/dist

To divide the distance and the angle into equal parts

To classifing the distance into several range with a logarithm method, we resulted a 'nbins_r' points between 'r_inner' and 'r_outer', then stored these points into 'r_bin_edges' simultaneously. Therefore, the distance was divided into equal parts

```
r_bin_edges=logspace(log10(r_inner),log10(r_outer),nbins_r);
```

The 'r_array_q' and 'fz' were a $t \times t$ 2-dimension array. And we classified the distance of each points into 'nbins_r' sections.

```
r_array_qij = 0, fzij = 0
for (m = 0; m < nbins_r; m++)
    for (i = 0; i < t; i++)
    for (j = 0; j < t; j++)
        if (r_array_nij < r_bin_edges(m))
            r_array_qij++;
```

The point was a outer boundary if the r_array_n(i,j) was not in the range of r_bin_edges. Also, those were recorded by a 'fz' matrix particularly.

```
for (i = 0; i < t; i++)
   for (j = 0; j < t; j++)
      if (r_array_qij > 0) fzij = 1;
```

We set the opposite angle of every point to a range between 0 and 2*π . And the 'theta_array_2' was a $t \times t$ two-dimension array (where t = Length(C)). Therefore, the angle was divided into equal parts.

where

theta_array_2(i,j) = ((theta_array(i,j) mod 2*π)+2*π)

And the 'theta_array_q' was a $t \times t$ two-dimension array (where t = Length(C)).

theta_array_q(i,j) =
1+floor(theta_array_2(i,j)/(2*π/nbins_theta));

Calculation of Shape Descriptors

The shape descriptors recorded the characteristic of an image contour. The image contour was expressed by a series of discontinuous points with value of 'n'. Therefore, there were n-1 points left for the opposite position. And the recorded results could be rotated when the shape contour was rotated somehow. We could group these points by a symbol if the rotated relationship existed. Thus, an 'n' point would be represented by the 'n' symbols. This is how we could find the similarity between these shape contours so quickly.

We applied this comparison technique of shape contour to an existed image capture system. It also provided a related feedback with an area-based image capture method simultaneously. It would be have a high accuracy for the captured image that quite fit with a human thinking model at all. We could calculate the parameters of shape contour with following program where all elements of BH would set to zero for the initialization.

```
for (n = 0; n < sample; n++)
   for (i = 0; i < t; i++)
      for (j = 0; j < t; j++)
         if (fz(i,j) > 0)
            BH(n,theta_array_qij,r_array_qij)++;
```

The shape context would be represented by the shape information which described with the parameters of each sample point totally.

Cost Matrix

We could calculate the cost of every point from these sample points. Thus, the cost from i point to j point was equal to the Chi-squared similarity of the shape descriptors from i row to j row approximately.

$$C_{(i,j)} \equiv C(p_i, q_j) = \frac{1}{2}\sum_{K=1}^{K}\frac{[h_i(k)-h_j(k)]^2}{h_i(k)+h_j(k)} \quad (5)$$

Now we got all the matched points Ci,j from first shape to second shape. We also minimized the cost of all these pair points with one by one conditionally. It was a problem of the square assignment or the weighted bipartite matching. The complexity of time $O(N^3)$ was involved with Hungarian method probably. We set a slightly more efficient algorithm in this study [5]. We made a input matrix with Ci,j matrix instead, and the result was a permutation π(i), which was ΣiCi , and the π(i) was the minimum value.

IV. System Implementation

Generally speaking, the study of vision-based hand gesture recognition usually involved image capturing, gesture analyzing and the tolerability of ambiguous hand gesture. Our system captured the image of user's hand gesture by video device. The significant information would be extracted, such that user's hand gesture could be transformed into proper instruction. Finally, the instruction was transferred to an appropriate program to trigger correct action.

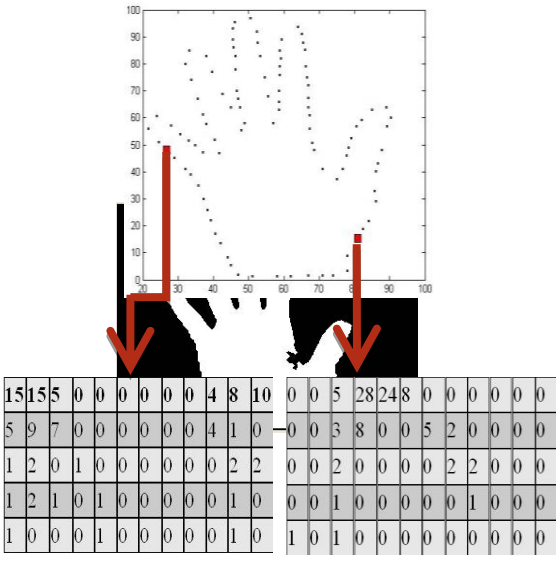


Figure 4. An example of the grabbed hand gesture

Figure 5. The binary image of the grabbed hand gesture

Definition of Hand Gesture

We had designed and implemented an interface between human and computer for the fully hand gesture system in this study. There were five model of hand gestures which defined as system control commands. The user could apply its command with hand gesture to replace any heavy equipment on the body.

The system would grab user's hand movement with video camera automatically. Then, it determined the color range of skin and transferred the specified color range to a binary data. And a smooth processing and a noisy elimination would be applied on those selected data simultaneously. Finally, the data would be compared in the cost matrix. The sequential procedures were described as followings:

Step 1: A completed hand was placed statically.

Step 2: The static hand gesture was grabbed and transferred to a binary data.

Hand Gesture Capturing

The experiment was setup in a 30 m2 square space with a 10~30 cm range, and the average brightness was 217 lm/m2. We used a CCD camera to grab the hand images for the recognizable gesture with 60 frames per minute. The resolution of these images was up to 640 x 480. (Refer to Fig 6)

Figure 6. The proposed interface for Hand Gesture Grabbing

Detection of Skin Color Range

We located the hand gesture area by the skin color detection in a normal and simple background simply. Unfortunately, there was much more complicated background actually. Therefore, we configured some other terms to search the possible gesture area precisely.

According to G. Kukharev, A. Novosielsk's theory, it was determined to a skin color while the value of YCbCr was fitted in with Y>80, 85>Cb<135, 135<Cr<180, where Y,Cb,Cr = [0, 255]. [9]

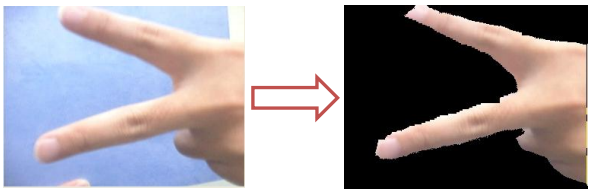


Figure 7. The image of Skin color range

After the detection of color skin, we could easy to identify most of the skin color with the filter actually (Refer to Figure 7).

It was only need some gray level to determine the threshold value of gesture recognition with its histogram on the processing, such as black and white image. It would be easily to translate a gray level image to a black and white image directly. The definition of the binary image was summarized as followings:

$g(x,y) = 1$, if $f(x,y) > T$

$g(x,y) = 0$, if $f(x,y) < T$

$f(x,y)$: the original image

T: the threshold value

$g(x,y)$: '1' means black. And '0' means white.

In the consideration of brightness of light and the factors of hardware, the threshold value was given in the skin color range with the adjudgment by YCbCr. And the threshold value were $T=45 < Y < 180$, $126 < Cb < 143$, and $122 < Cr < 130$.

Smooth Processing

There were some noisy found in the image process of digitizing of hand gesture. It was very common for the noise signal of a capture image from a normal video camera generally. This also made easier to wrong distinguish for the further binary processing. Therefore, we need to eliminate the noisy by the erosion and dilation process before the binary image processing which would made our digitizing image had more smooth and higher noisy ration concurrently.

Elimination of Noisy Signals

We fund some small skin color points around the hand gesture after the binary image processing with skin color detection. For the influence avoidance by the further procedures, we would apply the Opening operation in Morphology that could eliminate these minor noisy additionally.

The Opening operation was included erosion and dilation processing, that could narrow a binary image with erosion, then magnified that area image continuously. We also calculated the new binary image with mask operation after the skin color detection both for the erosion and dilation completely.

The Erosion: it determined the value of the position pixel P of the mask that was '1' or not. If it matched the value, the other 8 surrounding points would be determined repeatedly.

$P=P1 \cap P2 \cap P3 \cap P4 \cap P5 \cap P6 \cap P7 \cap P8$

The Dilation: It resembled to erosion. It would determine the value of the position pixel P of the mask that was '1' or not. If it matched the value, the other 8 surrounding points would be determined repeatedly.

$P=P1 \cup P2 \cup P3 \cup P4 \cup P5 \cup P6 \cup P7 \cup P8$

The binary skin color image would eliminate those minor noisy spot after the Opening operation (Refer to Fig 8 and 9).

Figure 8. The elimination of noisy signals with the algorithm: $S = (B \odot S) \oplus S$

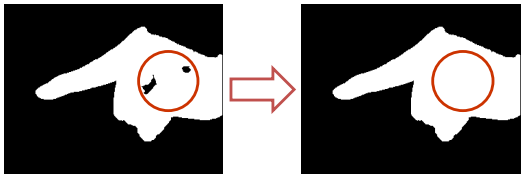


Figure 9. The elimination of noisy signals with the algorithm: $B \bullet S = (B \oplus S) \odot S$

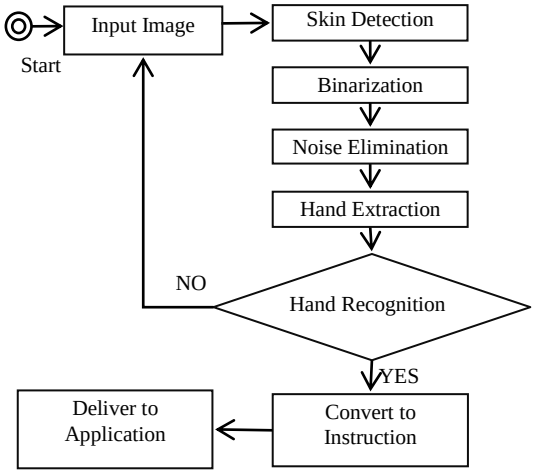


Figure 10. The system process flow

The implementation of the process flow was illustrated in figure 10. We used a video camera to shoot the images sequentially. The pixel quality of the images was 640 by 480 precisely, and the average luminance was 217cd/m2. The shooting speed was set to 0.5 frames per second actually. The first step, we started the program to select the interface of video to get hand gesture directly (refer to Fig 11).

The second step, we select the video image with 'Input Type', then pressed 'Start' to execute the program. And the CCD camera would grab the specified image frame automatically. It also displayed the hand motion and binary skin color range on the left area of the interface menu in the same time (refer to figure 12). There was some function key on the right area, such as 'Start', 'Stop' and 'Close'. Finally, the total count of shooting frames currently was displayed on the bottom of those function keys. Also, the analyzed result was shown on the right bottom area apparently.

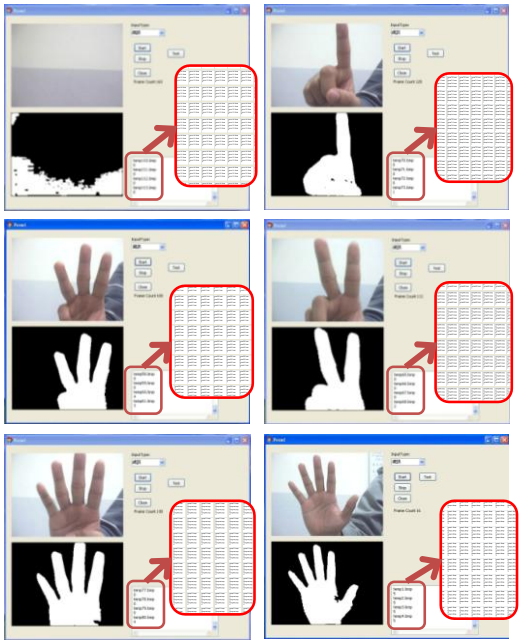


Fig 11. An interface of Hand Gesture

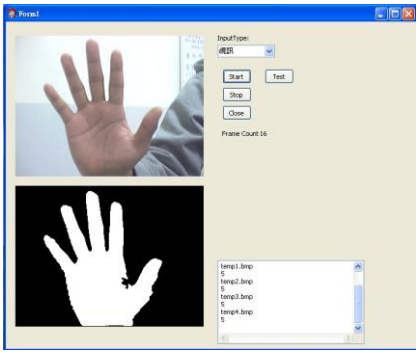


Figure 12. The Interface Menu that displayed the hand motion and the binary skin color range

The third step, the system would pass the analyzed data to the game system. The game system would response the suitable action according to the analyzed data previously. Therefore, the game system would execute a 'Head petting' motion if the hand gesture was analyzed as '3', and it executed 'Bone picking' motion if the hand gesture was analyzed as '4', same as the 'Feeding' motion for the '5' individually (refer to fig 13). We got the experimental data for 100 times from each hand gesture, and the accuracy was summarized as Table 1.

V. CONCLUSION

In this paper, we developed a perceptual interface for human-computer-interaction based on real-time hand gesture recognition. User could interact with computer program by performing body gesture instead of physical contact. We use a shape context based approach for matching hand gestures. We had proposed a simple and integrated method to recognize the hand gesture from a video image with several procedures, such as skin color detection, noisy signals elimination, comparison of hand gesture, and game command transferring. And it would trigger the control system of the virtual pet game with the previous transferred command consequently. We also solved a shield problem with a single video camera to avoid the influence factors such as the CCD position and the shooting angle. Finally, we had integrated with the game system for the amusements successfully.

TABLE I. THE ACCURACY RATE OF HAND GESTURES

Gesture Shape	1	2	3	4	5
Precision	70.8%	71.2%	75.0%	89.7%	81.7%
Correct Sample					
Error Sample					

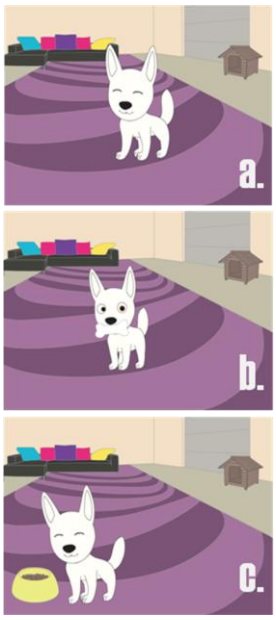


Figure 13. The Movement from Game Image (a) Head petting (b) Bone picking (c) Feeding

REFERENCES

[1] Larry S. Davis Thanarat Horprasert, David Harwood. : 'A statistical approach for real-time robust background subtraction and shadow detection', Technical report, Computer Vision Laboratory University of Maryland, 1999.

[2] Gary R. Bradski : 'Intel open source computer vision library overview', 2002.

[3] N. Liu and B.Lovell : 'MMX-Accelerated Real-Time Hand Tracking System', Proceedings of IVCNZ 2001. pp. 381-385.

[4] S. Belongie, J. Malik, and J. Puzicha : 'Shape context: A new descriptor for shape matching and object recognition', In NIPS, pages 831–837, 2000.

[5] S. Belongie, J. Malik, and J. Puzicha : 'Shape Matching and Object Recognition Using Shape Contexts', IEEE Trans. Pattern Analysis and Machine Intelligence, vol.24,no. 4, pp. 509-522,Apr. 2002.

[6] Eng-Jon Ong; Bowden, R.: 'A boosted classifier tree for hand shape detection', Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on 17-19 May 2004 Page(s):889 - 894.

[7] Manuel J. Fonseca Joaquim A. Jorge : 'A simple approach to recognize geometric shapes interactively', Technical report, Departamen to de Engenharia Informática, IST/UTL, 1999.

[8] Ajay Apte, Van Vo, and Takayuki Dan Kimura : 'Recognizing Multistroke Geometric Shapes: An Experimental Evaluation', In Proceedings of the ACM (UIST'93), pages 121{128, Atlanta, GA, 1993.

[9] Udo Ahlvers et al. : 'Model-Free Face Detection And Head Tracking With Morphological Hole Mapping', Germany 2005 , http://www.ee.bilkent.edu.tr/~signal/defevent/papers/cr1214.pdf.

[10] S. Belongie and J. Malik. : 'Matching with Shape Contexts', IEEE Workshop on Contentbased Access of Image and Video Libraries (CBAIVL-2000)

[11] http://en.wikipedia.org/wiki/Project_Natal

Lawrence Y. Deng Dr. Deng is an assistant professor in Department of Information Engineering of St. John's University, since 2002. He is also the Director of R&D in St. John's University since 2004. He received the B.S. degree in Computer Science Information Engineering from Tamkang University in 1997, and the M.S. and Ph.D. degrees in Information Engineering from the TamKang University, in 1999 and 2002, respectively. His current research interests are Distance Learning, Distributed Multimedia Presentation, Web Server Architecture, Floor Control Mechanism in communication and Multimedia Database System.

Jason C. Hung is an Associate Professor of Dept. of Information Management, Overseas Chinese University, Taiwan. His research interests include Multimedia Computing and Networking, Distance Learning, E-Commerce, and Agent Technology. From 1999 to date, he was a part time faculty of the Computer Science and Information Engineering Department at Tamkang University. Dr. Hung received his BS and MS degrees in Computer Science and Information Engineering from Tamkang University, in 1996 and 1998, respectively. He also received his Ph.D. in Computer Science and Information Engineering from Tamkang University in 2001.

Huan-Chao Keh Dr. Keh is a professor in Department of Computer Science and Information Engineering of Tamkang University. He is also a Dean of the office of Academic Affairs now. His current research interests contain Data Mining, Database System, Distributed Multimedia Presentation, Web Server Architecture, Mobile Agent and Multimedia Database System.

Lin Kun-Yi Mr. Lin is an P.H.D. candidate in Department of CSIE of TamKang University, since 2001. He received the B.S. and master degree in Computer Science Information Engineering from TamKang University in 2001. His current research interests are Distance Learning, Data Mining, Image Processing. The contact address of Mr. Lin is: Department of Computer Science and Information Engineering, Tamkang University, Tamshui, Taipei Hsien, 251, Taiwan, R. O. C.

Yi-Jen Liu Mr. Liu is an P.H.D. candidate in Department of CSIE of TamKang University, since 2002. He received the B.S. and master degree in Computer Science Information Engineering from TamKang University in 2001. His current research interests are Distance Learning, Ontology and Information Systems, Image Processing. The contact address of Mr. Liu is: Department of Computer Science and Information Engineering, Tamkang University, Tamshui, Taipei Hsien, 251, Taiwan, R. O. C.

Nan-Ching Huang Mr. Huang is an P.H.D. candidate in Department of CSIE of TamKang University, since 2007. His current research interests are Decision Tree, Data Mining, and Image Processing. The contact address of Mr. Lin is: Department of Computer Science and Information Engineering, Tamkang University, Tamshui, Taipei Hsien, 251, Taiwan, R. O. C.

Research of Place-based 3D Augmented Community-Taking The 3D Virtual Campus as an Example

Jiung-yao Huang[1], Huan-Chao Keh[2], Wai Shu-Shen[1], Ji-jen Wu[2], and Chung-Hsien Tsai[3]

[1]Dept. of Computer Science and Information Engineering, National Taipei University, San Shia, Taipei, Taiwan,

Email: jyhuang@mail.ntpu.edu.tw, waiss5409@gmail.com

[2]Department of Computer Science and Information Engineering, Tamkang University Tamsui 251, Taiwan,

Email: keh@cs.tku.edu.tw, wujj5770@yahoo.com.tw

[3]Dept. of Computer Science and Information Engineering, National Central University, Taoyuan County 32001, Taiwan,

Email: chtsai@csie.ncu.edu.tw

***Abstract*—Place-based virtual community is the trend of recent researches on pervasive computing. The purpose is to enable users in a physical place to receive ubiquitous services from the environment while they communicate with each other unwittingly. The paper further promotes this idea by allowing remote users to join such a virtual community as well as to interact with members on site and calls this type of community as the place-based 3D augmented (PDA) community. With the help of the augmented reality technique, on-the-spot member can visually sense the remote users by their representing avatars. To achieve this goal, the ambient communication environment is required to support message flow among the remote users and people on site. Besides, this environment should be able to discover context passing among members of this community to provide proper services. The context issues and context-awareness approaches of PDA community are fully discussed in the paper. Finally, the infrastructure of this PDA community is also presented along with preliminary result of the prototyping environment.**

***Index Terms*—Place-based 3D augmented (PDA) community, Pervasive computing, Context-awareness**

I. INTRODUCTION

Human beings are social animals, community activities will always be the focus of human activities. With the result that technical development would invariably comply with the development of human behavior. When the computer was invented, it launched the scenario of man-machine interaction. Comply with the emergence of the Internet, people can communicate with each other through the network in various styles and, as a result, a virtual community is formed on the networks platform. Similarly, while computer user interface progressed from text only input to graphical user interface; the creation of the Virtual Reality(VR) technology allows the interaction between human and computer to become more intuitive and coherent to human daily life's situations. And the ensuing Network Virtual Environment (NVE) further permits people to conduct social activities via VR technology through the network to form a large virtual community. NVE research aims to gather geographical distributed users in a shared virtual environment for the purpose of producing the simulated interaction activities within the virtual space. Via the networks, user's interactions inside this virtual environment are rapidly transmitting and spreading among players.

The similar development process also occurs in the transition of the mutual interaction relationship between human and computer. When the computer was invented, it can do nothing without specific command or data given by the human being. With the evolution of computing technology, researchers attempt to let computer sense user's needs via various sensors and autonomously provide appropriate services accordingly. The so-called Pervasive Computing technology is then emerged to progressively improve the interaction between humans and computers. However, such improvement is only limited to the interaction experiences between the individual and the environment. The individual's received services or perceived feelings cannot share with others. Nevertheless, with the increasing popularity of the wireless network (such as WiFi, 3G) infrastructure, together with the prolific sensors required by the pervasive environment to become quite cheaper than ever; the exchange and interaction among the users and equipments in the pervasive environment are rising much more dramatically with unusually expansive frequency. There are numerous virtual communities derived naturally from these interactions. However, such virtual community is randomly generated and does not exist in any fixed style. In 2002[1], Loke and Tuan first proposed the concept of LEC (Location-electronic-community)

doi:10.4304/jnw.6.5.705-712

based on virtual-community. They further enhanced the concept of LEC as a place-based notion on PBVC (Place-based virtual community) in 2008. The purpose is to enable members of the same virtual community to interact and exchange information in a specific regional environment. PBVC is extended from the previous applications of Location-based service to augment the former location specific limited services into the regional services. Hence, PBVC not only intends to provide the related services to a group of participants in the same physical space, most significantly, it implicitly implies the hinted meaning of multiple participation and interaction among people and equipments.

Both NVE and PBVC own the same purpose to supply users with an interactive platform through networks so as to enrich users' community experiences. And we can discover that, no matter if PBVC extended from LEC; or if NVE derived from VR; the current research trend is to integrate networked virtual reality with the physical world community and activities. Among proposed researches in the recent years, Cross reality [2] is the utmost major oncoming program, promoted by MIT Media Lab's Responsive Environments Group in July of 2009. The project theme is to integrate the reality world with the virtual environments based on the reality world's sensor device to provide the users with the vivid and impressive experiences crossing the virtual space and the reality world. Deployed in the specific environment with enormous sensors, Cross reality can allow avatars in the virtual environment to interact with the people existing in the physical world. Meanwhile, the fantastic experience of its hyperspace synchronization can exist in the virtual world and reality world simultaneously. Consequently, this is a fresh application of PBVC together with virtual reality. While the Cross reality stresses through sensors to create the integrated hyper reality space between the virtual world and the reality world, the reality experience vividly exists in the virtual world only but not the other side. In other worlds, the virtual world can depict the existence of user in the physical world with an avatar while the physical world can only use light or sound to imitate the presence of an avatar. Further, there is no direct user interaction between virtual and physical worlds. For the sake of further meeting the original intention of the pervasive computing environment development, that is, to explore the direct interaction between users and environment and then obtain appropriate services accordingly, this study enhances the cross reality concept to allow direct user interaction between virtual and physical world. Based upon NVE environment, this paper combines it with the Mobile Augmented Reality technology to create an overlapped community for user interaction between the physical space and the virtual space. Such a virtual-reality-integrated hyperspace community concept is then called place-based 3D augmented (PDA) community and it is defined as follows: "A place-based 3D augmented community is a group of people and avatars sharing the common interests and attributes of a physical place during a specific time. The virtual world is the clone of that physical place for remote users to interact with people on site. In other words, the community provides a cyber platform for the remote users and on-the-spot players to communicate and co-operate with one another over a common physical place."

In other words, PDA community is formed by user of the reality world wearing mobile device to interact with avatars, controlled by remote player, inside the virtual world. It is an indeed hyperspace community where interpersonal interaction activities occur in the reality space through virtual reality technology. Contrasting to Cross Reality which requires deploying enormous sensors to enable avatars in the virtual environment to interact with the people of the physical space, the difference between Cross Reality and PDA community is the user's experience. Cross Reality provides users' co-existence experience between virtual and physical world through massively deployed static sensors. Whereas PDA community allows direct co-operation experience exists between virtual world players who control avatars and physical space on-site users with mobile devices, such as wearable computers. Therefore, PDA community doesn't require costly implementation expenses to allow users in different geographical regions to perform the community activities at a specific reality space. In comparison with Cross reality, PDA community better meets the core concept of Pervasive Computing. PDA community stresses on the interaction between the virtual space and physical space and it allows different users under the various attribute spaces to communicate with one another by wireless networks and mobile equipments. It plays the architecture platform for users in this community to exchange interaction messages among them.

II. RELATED WORKS

Pervasive Computing has become the technological major trend in the 21st century. The earliest Pervasive Computing research prevailing popularity started from the Oxygen project [3] by MIT. Immediate following the Pervasive Computing Center [4] established by Denmark's Aarhus University; the Aura project [5] handled by Carnegie-Mellon University (CMU); the Websphere project [6] managed by IBM; the Cooltown plans [7] organized by HP; the ubiquitous network [8] progress from e-Japan to U-Japan in Japan; etc. The above mentioned prominent research projects belong to the successive conducted study. The major goal of Pervasive Computing is to help any users to obtain their desired service at any time, in any place.

Therefore, with the increased popularity of pervasive computing devices, the exchange and interaction among the users and equipments in the pervasive environment also are rising dramatically. There are extraordinarily numerous virtual communities derived naturally from these interactions. However, such virtual community is randomly generated and does not exist in any fixed style. In 2002[1], Loke and Tuan proposed the concept of Location-electronic-community (LEC) based on virtual-community. Four operators, which are union, intersection, restriction and overriding, are then defined to express the

relation among services. Furthermore, the composition scheme is adopted to explore the social activities within such virtual community. Moreover, LEC study utilized service-domains to discuss the concept of logical area, user types, mapping tables and service matching mechanisms in 2003[9]. In 2004[10], the ambient service was chosen by LEC to study the geographical boundaries of physical marketplaces scenario, and applied consumer buying behavior(CBB) as well as Consumer and Shopping Roles to research Task Boundaries and relevant "Service Classification Space". Above studies was all focus on various services of different geographical attributes among physical marketplaces. Various experiments were then conducted to verify the concept of LEC in 2005[11]. Finally, Place-based virtual community (PBVC)[12], which is the enhanced place-based notion of LEC, was defined in 2008 as following: "A place-based virtual community is a group of people, objects or agents, sharing common interests, attributes, and knowledge, that may share a common physical place at a specific time. Object can be a computational device or just a sensor, or a tag. The community is managed and operated by autonomous agents. Agent communications and co-operations are based on protocols, policies that give benefit to the members."

Aiming at the virtual communities activities for multiple users, in 2009, Beth Coleman [13] of MIT summarized the major spirits of Cross reality (XR) from its several prior projects to stress how to enable both users of the virtual world and physical world to share their individual users' experiences (UX) together. Consequently, the following three designed themes are introduced:

- The networked information must be available synchronously among all users.
- Users can more easily visualize data, particularly complex data, in 3D graphical form.
- Collaborative network tools support greater experiences of virtual presence (co-presence).

Above three design topics proposed by Beth Coleman, certainly, are the critical conditions to fulfill the integration of both virtual environment and physical reality in cross reality. Hence, XR emphasizes on allowing users of the physical space to experience the communication with the networked virtual environment through real world's equipments. And PDA community constructs a topological hyperspace to further enable users in a physical space smoothly co-operate with different users in another reality environment via the networked virtual world. Under such circumstances, PDA community does not only consider the User Experiences of man-machine but also more emphasizes on the User Experiences of human-machine-human. Therefore, PDA community further extends the above three research topics of cross reality to achieve distinct user experience in the hyperspace as:

1. Event synchronization between the virtual environment and the physical world.
2. The durable physical link (networking) between the virtual environment and the physical world.
3. Consistent user (Visualization) interface among the virtual environment and the physical world.

In the rest of this paper, services for above three research topics are addressed in chapter 3. The prototype of PDA community is then presented in chapter 4 and the evaluation of PDA community is followed in chapter 5.

III. SERVICES FOR THE HYPERSPACE

Since PDA community aims to provide the User-Experience of man-machine-man between mobile device users and desktop users under wireless network environment at a specific physical area, the message passing mechanism among man-machine-man has to be carefully designed. Based on the above mentioned research topics of PDA community, the study proposed the support service architecture as illustrated in Figure 1. Practically, PDA community uses various networks to interconnect user societies of physical space and virtual environment. Further, the limited computation power of mobile device and the fragile bandwidth of wireless network would influence the interactive performance of PDA community. Hence, PDA community needs Distributed Support mechanism to share the computation load of mobile device and, furthermore, to reduce the resource requirement of mobile system. On the other hand, this Distributed Support module also can improve the performance of Interaction between Desktop System and Mobile System. In addition, the locating support and mobility support modules in Figure 1 are two mechanisms to support mobile user navigating the physical space while interacting with other users in geographically different spaces co-exist in the same networked virtual environment.

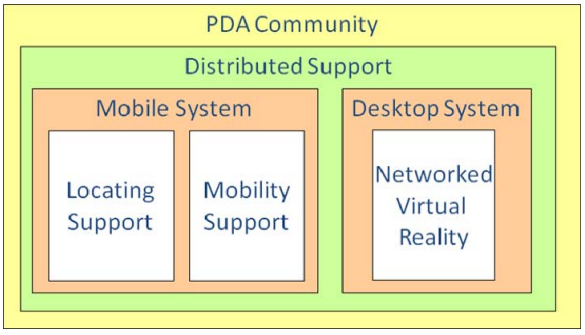


Figure 1. PDA community architecture

In order to achieve the co-experience within such hyperspace, PDA community must use the interactive message as the context to provide context aware services between mobile users and desktop players. There are several innate differences between both groups. Although prior researches had proposed context-aware framework and context-aware toolkit[14], different from these legacy approaches, the context-aware services of PDA community is built on top of distributed real-time interaction between virtual world and real world. Consequently, the three research topics of PDA community have to be further deliberated at the following sections. The event synchronization issue of the mobile device to achieve temporal and spatial consistency with

the desktop players is given first. The visual consistency issue of employing distributed computation technique to solve resource insufficient problem on the mobile device is then followed. Finally, the durable network issue of the stable connection for mobile device is given last.

3.1 Event Synchronization issue

Event Synchronization is the most critical issue to achieve interaction between virtual reality and physical space. The goal of PDA community is to make the users in the virtual and physical environment to have the experience of co-presence, and thus to achieve the effect of cooperation. Since all of the occurred events should have causal order relationship, Event Synchronization is the first issue needs to be solved to achieve the experience of co-presence and cooperation between two different worlds. This study focuses on the interactive behaviors between the physical world in which the mobile player exists and the virtual world which the mobile player joins. The most significant subject for this type of interactive behaviors is the synchronization of temporal and spatial events. For example, the conversion between the location of the mobile user in the physical space and his corresponded position in the virtual world is the typical issue. The mobile user in the physical space uses the positional sensor, such as GPS receiver, to acquire his location information and to navigate the virtual world accordingly. There are two major problems of this approach. First, the location data received from the positional sensor is a coordinate in the Geographical coordinate system while the position information of his simulated avatar inside the virtual world is in the Cartesian coordinate system. A mechanism is required to seamlessly and correctly translate data between these two coordinate systems. Secondly, although the virtual world is the clone of the physical space, it is impossible to achieve one-to-one duplication. Hence, except for the drift errors from the GPS receiver, the inconsistency between the physical space and the virtual world will cause the problem of spatial inconsistency significantly. Furthermore, the latency and instability of the wireless network will cause temporal discrepancy between the time when an event occurred in the physical space and the time of this event reflects to the virtual world. The inconsistency of the physical space and the virtual world will further magnify the temporal discrepancy when the mobile user is moving in the physical space. Hence, an event synchronization mechanism is required to solve the spatial and temporal inconsistency between two different interconnected worlds.

3.2 Durable Network issue

The Durable Network issue is mainly to discuss the wireless connection availability problem in the PDA community. The connection availability enables mobile users to join PDA community via wireless network in seamless way. The wireless network allows mobile players to freely move within Access point coverage while interacting with players of the virtual world. However, the wireless signal stability will affect the quality of user interaction between the physical space and the virtual world. Previous research [15] pointed out that the fragility of wireless signal leads to the connection availability problem. The problem was mainly caused by the interference of the environment or the lost link due to user's movement. Previous studies on fragile wireless signal all were focus on searching for the solution to the hardware network layer. The durable network study for PDA community does not focus on network technologies of any specific wireless network, but targets on how to handle the bad connection problem, such as disconnection or variable bandwidth, while the mobile user moving in the physical space and interacting with the virtual world.

3.3 Visual Consistency issue

Visual Consistency issue is resulted from inferior computing resource of the mobile device and insufficient bandwidth of the wireless network. The obvious benefit of the mobile devices is convenient to carry, so that display, power, and computation ability are limited in design accordingly. Due to the limitation of these factors, the performance of mobile device is the critical consideration in the PDA community. The sense of co-presence for the mobile player majorly comes from visually seeing the other players' co-existence in the same hyperspace. Hence, the major computation load of mobile device is to render the avatars of other players and other related visual information. To solve this problem, PDA community has to consider by means of distributed computing load sharing mechanism to reduce the rendering load on the mobile device. According to the technique of the Computer Graphics, such load sharing can be achieved by the visibility of objects, the level of detail of object appearance, and the realism of object animation[16].

The following chapter will focus on the above mentioned issues and utilize the context aware techniques to percept the temporal and spatial difference between the physical space and the virtual world, the bandwidth variation of the wireless network, and the visual state of the mobile users. It would then provide adaptive message transmission services to present co-presence interactive experience to the participants of this PDA community.

IV. THE PDA COMMUNITY ARCHITECTURE

In order to reach the service goal of allowing the mobile user to be co-present in the Networked Virtual Environment(NVE) in seamless way, the paper proposes a system architecture that is based on context-aware technique as illustrated in Figure 2. The system architecture includes Multi-user server, Context Management server, Remote client, and On-the-spot client. Multi-user server act as a lobby server and game server that manage users' account and exchange users' messages such as position, orientation, and text messages. Context Management server is a context-aware server to enable the mobile players of the physical space to seamlessly communicate with players of the virtual world. The mobile player in the physical space is called on-the-spot client thereafter. In essence, the Context

Management server plays the role as a data mediator between the Multi-user server and on-the-spot client. Therefore, from the perspective of Multi-user server, Context Management server can be regarded as a special-purpose desktop system to act as proxy server between on-the-spot client and remote client. As for the on-the-spot client standpoint, Context Management Server acts as the environment server of the virtual world.

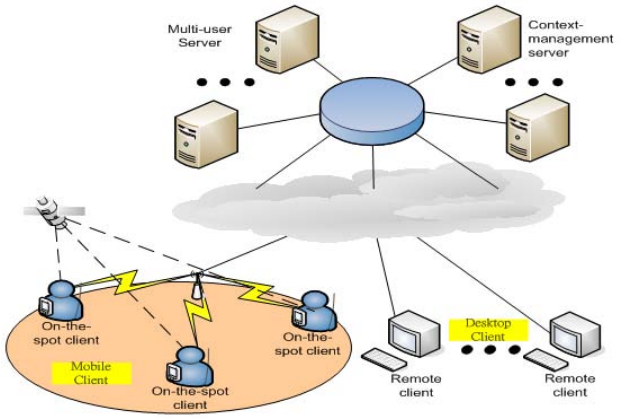


Figure 2. System architecture of the PDA community

As discussed in the previous chapter, the research goal of PBVC is to describe community service providing problem in a specific region whereas PDA community aims to elaborate real-time interactive service between a physical space and its duplicated virtual world. Furthermore, from another point of view, PDA community attempts to use the pervasive computing technique to solve a more complicated human-machine-human problem than Cross-Reality proposed. PDA community integrates Networked Virtual Environment and Mobile Augmented Reality as a whole unity to provide a fresh user-experience of traverse between reality and virtual worlds. In order to manage each on-the-spot client, the study is based on MiMAR[17] research together with context-aware techniques. As depicted in Figure 3, the presented PDA community system provides spatial consistency, temporal synchronization and balance data transmission schemes for Event Synchronization issue; connection availability service for Durable Network issue; and the last, computing load balance mechanism for Visual Consistency issue.

Figure 3. Services of PDA community environment

Since Context Management server plays a critical role of environment server in the PDA community, the context-aware technique learned from PGPS[18] research should be followed to adaptively provide services to on-the-spot client. That is, sensing context for the PDA community needs to be decided first; followed by extracting sensing data into feature data for user's behavior inference. Appropriate service then would be computed from inference result. Hence, the discussing of services for Event Synchronization issue, Durable Network issue, and Visual Consistency issue will be elaborated in the above mentioned pattern at the following sections.

4.1 Event Synchronization service

Event Synchronization service intends to solve the temporal and spatial difference problem between the physical space and the virtual world. It is expected that all of the interactive messages can be maintained as their proper causal order through Event Synchronization service. For example, due to the difference between two coordinate systems, on-the-spot client walking in front of a building can be misinterpreted by the Remote client as moving inside the building. To solve this problem, the Event Synchronization service must be aware of this situation and synchronize this interaction event accordingly. Under the variable bandwidth, Event Synchronization service views the interactive event message as sensing data and adapts data priority [19] mechanism to classify all event messages. The position, interaction, and voice messages are the highest priority, and then facial and text information, and the least is user's faced orientation information. Event Synchronization service can reduce the computing load of mobile device as well as avoid the influences of variable bandwidth by controlling the message priority. More importantly, after transmitting prioritized message, the system can infer the user's current state and provide the accurate service. For instance, there are two positional event messages transmitted within the PDA community, one is in Geographical coordinate system from the sensing location message of mobile device and the other is location message in Cartesian coordinate system of the virtual world. If the Geographic coordinate is received, Context Management server will then infer mobile user's location inside the virtual world using the corresponding relationship between the geographical marker of physical environment and the virtual marker of virtual space. On the contrary, if the Context Management server receives Cartesian coordinate value, the provided service will calculate the related Geographical position and orientation information for the mobile device to display represented avatar's proper image on the screen.

4.2 Durable Network Service

Durable Network service aims to solve fragile wireless signal problem. Durable Network service regards the available bandwidth as feature data, and detects user's linking status periodically to infer whether on-the-spot client is temporarily disconnected or broken link. If the user is disconnected temporarily, Context Management

server utilizes Dead Reckoning [Dead Reckoning] to simulate the corresponding position of the on-the-spot client and transfers this computed position to the virtual world. Besides, Durable Network service will manage the status of lost-link on-the-spot client in case that he is reconnected. In brief, Durable Network service mainly provides the relevant services regarding the connection availability problem.

4.3 Visual Consistency service

Visual Consistency service targets on reducing the computation load of mobile device. The dedicated mobile device of PDA community is SSD-based embedded system, which has inferior computing power. The most computational burden of such mobile device is to render all the PDA community users' information. Hence, the goal of Visual Consistency service is to utilize the human vision limitation to decrease the scale of resolution and the complexity of animation of the distant object to reduce the computation load of mobile device. Context Management server relies on location information of the on-the-spot client to calculate the number of objects and their respective resolutions that should appear within the field of view (FOV) of this client. By this way, Context Management can ensure visual consistency between on-the-spot client and remote players of the virtual world. For example, when the human being is moving, his vision to a distant object tends to be less perceptible. Hence, the study employs the depth perception [21] technique to infer the reasonable number of objects within users' FOV to reduce the number of objects that should be rendered. Further, by means of the level-of-detail (LOD)[22] technique to calculate the resolution of every visible object so as to reduce the computation load of mobile device.

V. EVALUATION

The original goal of Pervasive Computing is to automatically provide appropriate service to the users anytime, anywhere and anyplace. In order to verify the proposed PDA community architecture and the embedded context aware services, the study takes National Taipei University 3D virtual campus as an example to implement the prototyping system. First, the software architecture of the desktop system for the Remote client is shown in Figure 4. Desktop System applies the Virtools[23] software package to construct the user interface and its 3D real-time graph engine to render 3D virtual campus. Below it are Multi-user Manager, Skype Module, Event Manager, and Text Manager to individually implement multiuser interaction, voice chatting, event manager and text chatting services within the PDA community.

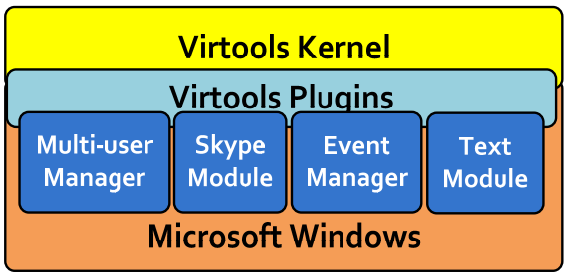


Figure 4. The functional architecture of Remote client

After the Remote client login PDA community system, Context Management Server will retrieve the aforementioned sensing context, such as bandwidth, position, orientation, …, and so on, of the on-the-spot client. The status of the on-the-spot clients will then be inferred to provide to multi-user server, so that Remote client can smoothly interact with their represented avatars, shown in figure 5.

Figure 5. The snapshot of Remote client's display

Furthermore, the 3D virtual campus project not only allows users to navigate the campus but also provides the voice chatting for interaction. This function is achieved by binding Skype client into the system. Figure 6(a) is the snapshot of Skype popup message when a player wants to call another one inside of the virtual campus. Figure 6(b) is the snapshot of the receiver when an incoming call is detected.

Figure 6. Skype voice chatting of the 3D virtual campus

The software architecture of the mobile device for the on-the-spot client is depicted in Figure 7. Mobile devices run on the Linux-based platform and take Dbus as the communication mechanism among the modularized software function. To facilitate the future expansion, the study further constructs a middleware layer on top of the Dbus layer based on IEEE 1516 standard[24]. All the functional modules exchange their messages and

communicate with Context Management Server via this middleware layer. The implemented functional modules include GPS & Compass Module for the mobile user's outdoor positioning, Event Manager to manage events of PDA community that are related to on-the-spot client, Display Module to realize co-present visual effect for on-the-spot client, and Skype Agent Module and Webcam Module to provide audio and video communication capabilities through Skype platform. Significantly, Webcam Module aims to improve the sense of telepresence of Remote client. That is, when on-the-spot client and Remote client communicate with each other via Skype, Remote client is able to see the live scene of whom on-the-spot client faces via the webcam of the mobile device.

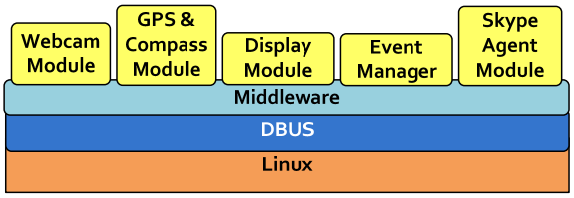


Figure 7. Function architecture of on-the-spot client

Further, to consider the computing power of the mobile device, Display module is implemented based upon the GTK graphical library to render avatars inside of the 3D virtual world. When On-the-spot client walks in the physical campus, the mobile device detects his motion via GPS and compass. This position and orientation information is then passed to the Context Management Server so as to interact with other users. Therefore, there would be two conditions need to be considered when On-the-spot client walks in the physical campus. First, when On-the-spot client meets another On-the-spot client, Display Module will reveal the player name on the screen. Second, while On-the-spot client walking in the physical campus, avatar controlled by other Remote client may appear in the neighborhood of his corresponding location inside of the 3D virtual campus. The corresponding image size of this avatar will be rendered on the screen of On-the-spot client depending upon the distance or orientation between the On-the-spot client and avatar.

VI. Conclusion

The paper elaborates new interaction model between physical communities and virtual communities and names this model as Place-based 3D(PDA) community. PDA community enables the users of physical space to directly interact with the users of virtual space via mobile device. In order to accomplish PDA community, the study exhaustively investigates three important issues faced by PDA community. The first is Event Synchronization issue during interacting between physical space and virtual environment. The second is to discuss the connection availability problem of durable wireless network issue. The third is to consider limited resource issue of the mobile device to have visual consistency result.

Focus on these above mentioned three issues, the study designs a Context Management server to sense and serve the mobile user based upon context aware technique. To further validates the presented system, an experimental PDA community was built on top of the existing 3D virtual campus. The preliminary experimental result shows that Context Management server not only plays the role of mediator between the physical space and the virtual world, and much more importantly, it but also acts as the event balancer in the PDA community with the increasingly growing of participants in the PDA community.

Although the presented PDA community is successfully verified through the 3D virtual campus project, more enhancements regarding the interactive experiences between the physical space and virtual world are required. For example, Asus EeePC 701 with Ubuntu Linux is used as the mobile device of the On-the-spot client at current implementation. The further study will actually design an embedded system as the mobile device. Therefore, the functional module of On-the-spot needs to be fine-tuned along with the embedded system. In addition, experiment shows that the accuracy of the position and orientation information will influence the sense of co-presence among users of PDA community. Consequently, a further study is required to probe the environment noises that affect the orientation sensor while the user changes his direction. These noises factors will lead to the real-time display error problem. The future work will set focus on employing the context-aware technique to solve the orientation discrepancy problem so as to promote the interactive experience within PDA community.

References

[1] S.W. Loke, Modelling Service-Providing Location-Based E- Communities and the Impact of User Mobility. In *Proceedings of the 4th International Conference on Distributed Communities on the Web* (DCW 2002), (eds) J. Plaice, P.G. Kropf, P. Schulthess, J. Slonim. Sydney, Australia, Springer-Verlag, LNCS 2468, pp. 266-277, 2002.

[2] J. Paradiso and J. Landay, "Guest Editors' Introduction: Cross-Reality Environments", *IEEE Pervasive Computing*, 8(3), pp. 14-15, 2009.

[3] http://oxygen.lcs.mit.edu/ [Accessed online at 02/19/2010]

[4] http://www.pervasive.dk/ [Accessed online at 02/19/2010]

[5] http://www-2.cs.cmu.edu/~aura/ [Accessed online at 02/19/2010]

[6] http://www-128.ibm.com/developerworks/web/library/wa-pvc/index.html [Accessed online at 02/19/2010]

[7] http://www.cooltownstudios.com/ [Accessed online at 02/19/2010]

[8] http://www.soumu.go.jp/menu_seisaku/ict/u-japan_en/index.html [Accessed online at 02/19/2010]

[9] T.T Naing, S.W. Loke, S. Krishnaswamy, A service-domain based approach to computing ambient services. ICSOC03, Technical Report, 2003.

[10] S. W. Loke, A. Zaslavshy, Integrated Ambient Services as Enhancement to Physical Marketplaces, In: *Proceedings of the 37th Hawaii International Conference on System Sciences*, 2004.

[11] S.W. Loke, S. Krishnaswamy, T.T. Naing, Service Domains for Ambient Services: Concept and Experimentation. Mobile Networks and Applications 10, pp. 395-404, 2005.

[12] T. Nguyen, S. W. Loke, T. Torabi, and H. Lu, Multi-agent Place-Based Virtual Communities for Pervasive Computing, in the 1st IEEE Workshop on Agent Technologies for Pervasive Communities (ATPC 2008), Hong Kong, 2008.
[13] B. Coleman, "Using Sensor Inputs to Affect Virtual and Real Environments," *IEEE Pervasive Computing*, 8(3), pp. 16-23, 2009.
[14] M. Baldauf and S. Dustdar. A survey on context-aware systems. Technical Report TUV-1841-2004-24, Technical University of Vienna, 2004.
[15] G.H. Forman, J. Zahorjan, The challenges of mobile computing, *IEEE Computer*, 27 (4), pp. 38-47, 1994.
[16] J. Clark, "Hierarchical geometric models for visible surface algorithms." *Communications of the ACM* 19(10) pp. 547-554, 1976.
[17] J.Y. Huang, M.C. Tung, and C.H. Tsai, "The Research of Multiplayer Mobile Augmented Reality (MiMAR) System and Its Application", in Qing Li and Timothy K. Shih (Ed.), *Ubiquitous Multimedia Computing*, Chapter 6, CRC Press, (ISBN 978-1-4200-9338-4) pp.153-166, Jan. 2010,.
[18] J.Y. Huang, C.H. Tsai, S.T. Huang, "PGPS: A Positioning Technique by Perceiving GPS Data", submitted to *IEEE Transitions on Mobile Computing*, Jan. 10, 2010.
[19] B.G.Witmer, M.J. Singer, Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and Virtual Environments*, 7, pp. 225-240, 1998.
[20] S. Singhal, M. Zyda, *Networked Virtual Environments Design and Implementation*. Addison-Wesley, 1999.
[21] J.P. Wann, S.K. Rushton, M. Mon-Williams, Natural problems for stereoscopic depth perception in virtual environments. *Vision Research*, 19, pp. 2731-2736, 1995.
[22] M.J. DeHaemer, M.J. Zyda: Simplification of objects rendered by polygonal approximations. *Computer & Graphics*, 15(2), pp. 175-184, 1991.
[23] Virtools, A.: Dassult Systemes Technology, http://www.virtools.com (Accessed online 03/09/2009)
[24] IEEE Standard for Modeling and Simulation [M and S], High Level Architecture [HLA] - Federate Interface Specification, IEEE Std 1516.1-2000, pp. i-467, 2001.

Jiung-yao Huang is a Professor and Chairman in Department of Computer Science and Information Engineering at National Taipei University. He received his PhD degree in Electrical and Computer Engineering from the University of Massachusetts at Amherst in 1993. His research interests include pervasive computing, augmented reality, computer graphics, networked virtual reality, and multimedia system. He has been a member of IEEE and ACM since 1990, and joined the Society for Computer Simulation in 1996.

Huan-Chao Keh received his BS degree in horticulture science from National Chung-Hsing University, Taiwan, in 1980, and his MS and PhD degrees both in computer science from Oregon State University in 1985 and 1991 respectively. He is currently a professor in the Department of Computer Science and Information Engineering at Tamkang University, Taiwan. His research interests include multimedia database systems, data mining, distance learning, and mobile computing.

Wai Shu-Shen is an Assistant Professor and Chairman in Department of Computer Science and Information Engineering at National Taipei University. He has a long time work experience on Modeling and Simulation. His specialty is computer assist teaching field which focus in the Distance learning.

Ji-jen Wu is a PhD candidate in the Department of Computer Science and Information Engineering at Tamkang University, Taiwan. He has more than 10 years work experience in the Joint Operation training center, His specialty are the analytic and interactive war-game.

Chung-Hsien Tsai, MS, received the MS in Telecommunication management from Polytechnic Institute of NYU in 2001. He is currently a PhD candidate in the Department of Computer Science and Information Engineering at National Central University, Taiwan. His current research interests include Networked Virtual Environment, Context-Awareness, Mobile Augmented Reality, and Navigation technology.

Interactive E-diagnosis: An Efficient Scheme for Medical Diagnosis Support System

Rong-Chi Chang
Department of Digital Media Design, Asia University, Taichung, Taiwan, R.O.C.
roger@asia.edu.tw

***Abstract*—With advancements in information technology, computer applications have now been largely implemented in health care, where medical support system or expert system is serving as a second opinion for medical personnel, underscoring its importance. The present study aims to design a prototype medical diagnostic support system, in which DICOM-based image analysis algorithms are utilized to develop an image browser and graphical user interface (GUI), allowing medical personnel to read X-ray, CT scan, MRI or other medical imaging files via a simple browser interface. Practitioners may also configure parameters for the intensity slicing and intensity analysis based on their medical expertise and experience to automatically generate converted imaging results for use as diagnostic references. Furthermore, the integration of GUI and network programming technology facilitates patient consultation and discussions among doctors in different areas, making this medical diagnosis system a telemedicine application, as well as an e-learning tool.**

***Index Terms*—Medical Diagnosis Support System; DICOM; Medical Image; Image Processing; Intensity Slicing**

I. INTRODUCTION

The current approach to medical treatments relies on apparent signs of serious symptoms or ailments from the patient to determine the course of treatment. However, when obvious symptoms or discomfort occur and the patient seeks medical attention, the illness has often progressed to a later stage. If it is possible to have an early diagnosis for the causes of the disease, simple medical procedure or preventive measures can be applied for proper and efficient treatments. For this reason, the use of effective medical diagnosis support system can assist practitioners in the diagnosis by providing relevant information, in turn enhancing the timely treatment of ailments.

With advances in medical engineering and technology, computer aided diagnosis system is now widely implemented in hospitals, where it is designed to complement the shortcomings of medical systems to provide practitioners with more pathological information on the patients for better understanding of patients' symptoms. Computer-aided diagnosis system takes many forms, e.g. X-rays, computed tomography (CT) scan or magnetic resonance imaging (MRI). The reports, data and files can be stored on a network medical database. Using a picture archiving and retrieval system (PACS), medical personnel (doctors or nurses) can access and retrieve these data to determine the course of treatment based on patient symptoms.

A picture archiving and communication system (PACS) consists of image and data acquisition, storage, and display subsystems integrated by various digital networks. It can be as simple as a modality connected to a display workstation with small image database, or as complex as a total hospital image management system. Now, most clinical PACS developed as open architecture systems are following the DICOM standard in image communication, image format and image management [2]. The image distribution and display inside radiology departments or hospitals mostly use DICOM services, e.g. Storage, Query/Retrieval, Printing, etc., and these standardized services greatly and efficiently improve the interpretabilities among different manufactures' PACS components [1].

Results of medical exams, e.g. X-rays, CT scan, or MRI, are stored in the commonly used DICOM medical image format. The image content is then interpreted and analyzed by a doctor for prescribing further medical treatment. DICOM medical images are high-resolution grayscale images. Doctors can use the image browser to review a single image or the subtle variations among a series of continuous images, and interpret the patient's symptoms based on observation and medical experience. The interpretation of these medical images requires the use of high-resolution display as the output device to obtain the best imaging results. Differences in grayscale images are sometimes too subtle, resulting in interpretation difficulties or misjudgment. Therefore, if parameter settings can be specified to display particular marked information according to the interpretation needs of the doctor, it serves as a good supporting reference during the interpretation process.

For the reasons above, this study aims to design a Medical Diagnosis Supporting System using image processing technology that can interpret DICOM file format and process image content conversion. An interactive interface allows doctors to read the medical images and specify parameters for the image files and image regions under observation according to different diagnostic needs. Then, through automatic interpretation and comparison, the regions of interest (ROI) can be

doi:10.4304/jnw.6.5.713-720

displayed in different colors to provide references and suggestions for doctors to perform medical interpretation. In addition, the functional integration of a web browser with video features can be embedded on a website, where the interface facilitates remote medical diagnosis or medical image transmission between different hospitals to achieve more efficient and convenient patient diagnosis. The paper is organized as follows: Section II describes the related work by our project. Section III introduces our prototype system with its user-friendly interface. The experimental evaluation and results are reported in Section IV. Finally, Section V concludes this paper and lists future works.

II. Related Work

Medical diagnosis is in essence a cognitive process of complex and ambiguous nature. Drawing from the computational model of artificial neural networks, it is shown that medical decision support system (MDSS) has great potential and is worthy of further developments. The main purpose of MDSS is to help doctors in the diagnosis process, presumably through displaying medical records for practitioners to analyze a patient's susceptibility to a certain disease. Yan *et al*. [3] developed a multilayer perceptron-based decision support system for heart disease diagnosis. The system consists of 40 input variables in the input layer which are encoded and divided into four groups. The results show that the proposed MLP-based decision support system can achieve very high diagnostic accuracy, evidencing its value in making clinical decisions for heart disease. Ogiela *et al*. [4] introduced the application of structured artificial intelligence, in particular semantic reasoning mechanism, to develop an intelligent medical information system. The main purpose of the system is to use graphs and data for readers to quickly understand the implication of medical test images.

Image processing [5, 6] plays a critical role in computer science, through which digital images are processed to enable detection of information imperceptible to the human eye. Several studies have exploited image processing techniques, e.g. color conversion [7], brightness enhancement [8, 9], color level variations [10, 11], in hopes of providing supporting medical diagnostic information for medical practitioners.

Digital medical imaging technologies provide powerful tools for diagnosis, treatment and surgery, acting as the cornerstone to modern medicine and health care [12, 13]. Digital Imaging and Communications in Medicine (DICOM) is a standard for handling, storing, printing, and transmitting information in medical imaging. It includes a file format definition and a network communications protocol [15]. The communication protocol is an application protocol that uses TCP/IP to communicate between systems.

Digital images are generated by a wide variety of radiological hardware. Each device collects data, which are then encoded and stored electronically in DICOM format [16]. This is a universal file type, developed to facilitate data exchange between hardware, irrespective of manufacturer. DICOM files store a large amount of data and usually need to be viewed on dedicated workstations but may be transferred electronically to other computers where they can be displayed provided appropriate DICOM viewing software is installed. DICOM files can easily be converted to a variety of image formats and edited before use in teaching and publications.

Hu *et al*. [17] used medical image processing techniques to develop a clinical diagnosis and treatment support platform, where it implements an interactive image processing technology to process DICOM format images, such as image smoothing, sharpening, histogram processing, pseudo-color processing, segmentation, reading, local amplification and measurement functions, targeting patients' CT image viewing to provide flexible and accurate reference data. In remote areas with limited resources, web-based decision support systems can pool together a network of doctors to jointly diagnose and treat patients [18, 19].

III. The Proposed Methods

This study endeavors to develop a prototype medical diagnostic support system using image processing technology and DICOM file attributes. Through file information acquisition and analysis and parameter setting, the system produces indiscernible contents for a series of medical images that would otherwise gone undetected by the human eye to assist doctors in medical diagnosis. This study proposes a number of data analysis approaches, as detailed in the following sections.

A. System Design and Implementation

This section explains the design architecture of the system. Figure 1 shows the flow chart of the DICOM image analysis. A DICOM image contains a patient's medical test images and personal information. According to variant needs, the user may input one or more DICOM image files; and once the images are processed and patient data modules acquired, they can be stored in XML format in the system database, wherein the DICOM header information and image attributes are used for image enhancement and sequence image analysis.

Intensity slicing is an image processing technique that enables image conversion and visual-effect display based on statistics collected for different threshold settings, where the converted images serve as aids in the diagnosis. In order to provide medical personnel with the convenience of observation, a user interface for intensity slicing and analysis is designed with the following features:

(1) The threshold values for intensity slicing can be divided according to the signal change statistically calculated from the images or based on the user-defined levels and segments; every set of threshold value is represented by one color level.
(2) Color level is defined by the user prior to analysis according to the patient's pathological features; different definition value can be input for different medical image analyses.

(3) Continuous time-sequence images are displayed in 2D planar images at different time points to demonstrate signal changes for medical practitioners to use as diagnostic reference.
(4) The statistically-obtained image information is used to generate charts through numerical statistics.
(5) The resulting information after numerical analysis and induction of multiple test images allows the generation of various charts and graphs for different needs, providing plentiful support information for diagnosis.

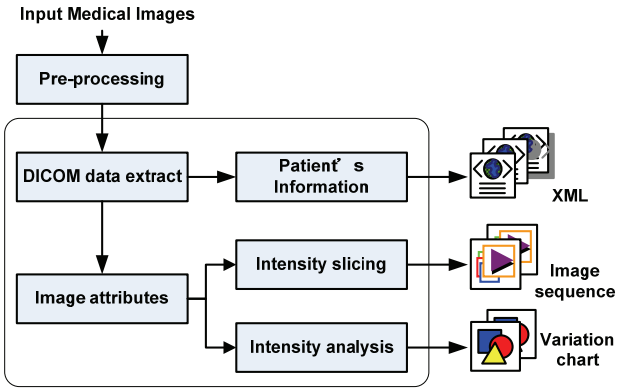


Figure 1. Flow chart of medical image recognition.

B. DICOM medial image data extract

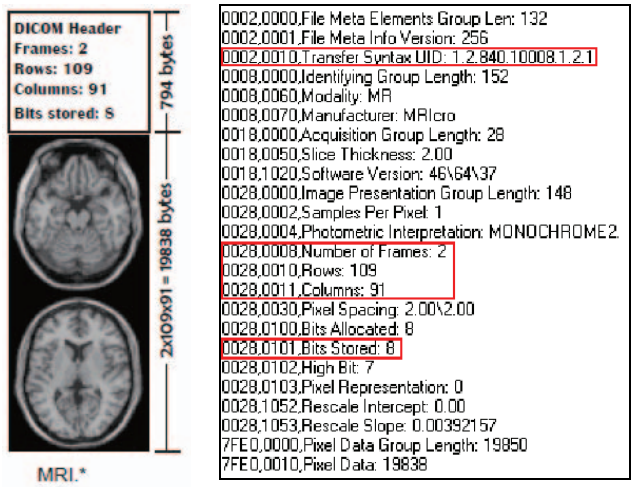
DICOM content can include many different types of data, such as the following:

- Patient administration information
- Waveforms
- Images
- Slices of 3-D volumes
- Video segments
- Diagnostic reports
- Graphics
- Text annotations

DICOM content also contains standard attributes and private attributes. Standard attributes are defined and published in the DICOM standard. Private attributes are defined by and specific to private organizations, such as manufacturers and other enterprises. The DICOM data dictionary provides the definitions for DICOM standard and private attributes.

Each DICOM file has a header containing amongst other items, patient demographic information, acquisition parameters, referrer, practitioner and operator identifiers and image dimensions. The remaining portion of the DICOM file contains the image data (e.g. Figure 2). Because they often contain multiple high-resolution images, DICOM files tend to be large and are frequently compressed before storage and transfer.

Figure 2 shows an example DICOM image and its file header information. In the Hospital Information System (HIS), different types of services call for different types of data elements. During the image processing, we need to acquire the image resolution (row × column), color space, image size, gray level image or color image, among other information. A part of the image content recorded in the DICOM file contains tags recording the image structure information, such as the content bracketed in red in Figure 2 (b). Table I presents important data elements related to image structure when reading the DICOM files.



(a) (b)

Figure 2. Example of the DICOM image file, (a) pictures and (b) their data element information.

Table I.
THE TAG DESCRIPTION OF THE DICOM IMAGE

Data elements	Description
0002,0010, Transfer Syntax UID	Sequence of bits during the read, and whether there is a value representation field
0028,0008, Number of Frames	Number of frames in the DICOM file
0028,0010, Rows	Number of rows in the image (number of pixels per row)
0028,0011, Columns	Number of columns in the image (number of pixels per column)
0028,0101 ,Bits Stored	How many bits is used for storage for each pixel
7FE0,0010 ,Pixel Data	Physical part of the image, the intensity of whole image.

According to DICOM image tag records, we convert the DICOM file into an image that can be easily read on regular displays and, at the same time, acquire other relevant information in the DICOM file, e.g. date of examination, patient information, to be stored in XML format for follow-up data transmission and exchange. The present study exploits image processing to develop a medical diagnostic support system using a self-designed DICOM file reading interface. We propose numerous operating modes for improved analysis of medical images that can serve as valuable references during the diagnosis for practitioners. Figure 3 shows the system interface architecture. In addition to the function of reading DICOM files, the browser has other features such as color conversion. Related features are presented in Figure 4 and Table II.

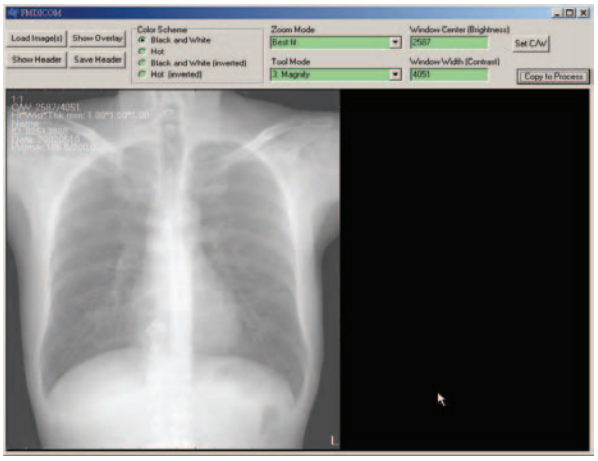


Figure 3. The DICOM image browser of e-diagnosis system.

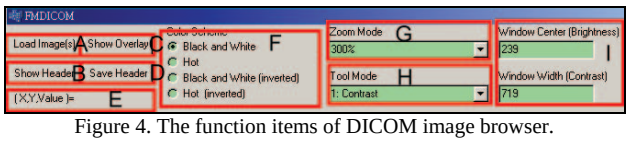


Figure 4. The function items of DICOM image browser.

Table II.
THE FUNCTION DESCRIPTION OF DICOM IMAGE BROWSER

Item	Function name	Description
A	Load Image(s)	Loading of the DICOM files in need of analysis
B	Show Header	Reading of DICOM header information
C	Show Overlay	Whether to display simple tags on the image
D	Save Header	Storage of DICOM header information
E	(X,Y, Value)	Display of the raw data from the reading
F	Color Scheme	Color models include ■ Black & White ■ Hot ■ Black & White (inverted) ■ Hot (inverted)
G	Zoom mode	Image zoom in / out, local amplification
H	Tool mode	View functions include ■ Contrast ■ Brightness ■ Magnify ■ Local enlarge
I	Window Center (Brightness)/ Window Width (Contrast)	Reading and configuration of window level and window width

C. Intensity slicing

The technique of intensity slicing (sometimes called density slicing) and color coding is one of the simplest examples of pseudo-color image processing.

We assume that each image is 3D function consisting space coordinates (x, y) versus signal intensity. In this 3D space coordinates is a slicing plane interlaced with the image. Suppose 0 ~ L－1 represents the intensity of gray level image (signal intensity), then the origin l_0 denotes pure black, i.e. f (x, y) = 0, whereas the maximal point l_{L-1} represents pure white, i.e. f (x, y) = L-1.

If anything above the slicing plane is defined as color C_1, and anything below is defined as another color C_2, then the entire image is divided into two colors; signals greater than the slicing plane will be converted to color C_1, while signals below will be converted to color C_2.

In this study, pixel brightness of the medical image determines the signal intensity. Figure 5 shows an MRI image of a knee-joint, where the grayscale value is set to be between 0 to 255, with the slicing plane set at 128 to divide the image into two different color areas.

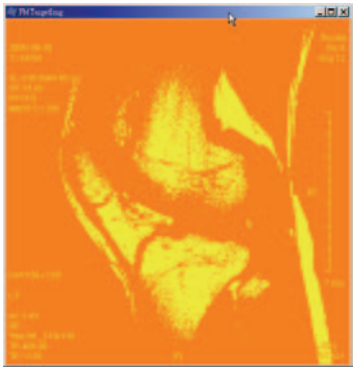


Figure 5. MRI of a knee.

We further extend the concept of slicing plane. Assume M number of slicing planes l_1, l_2, l_3,…, l_M are perpendicular to the X-axis (gray value) (0 <M <L-1), and these M number of planes can divide the gray level into $M+1$ intervals to represent planes V_1, V_2, …,V_{P+1}, then the gray level can be mapped to the defined color using formula (1)

$$\text{if } f(x,y) \in V_k \text{, then } f(x,y) = C_k \qquad (1)$$

where C_k denotes the defined color of the k^{th} intensity interval V_k (areas divided from the two slicing planes $l=k-1$ and $l=k$). Figure 6 demonstrates the use of different slicing plane intensity to display different colors of an MRI image of a local lumbar spine.

Here, intensity slicing technology for medical image analysis is implemented, where doctors are given the option to perform intensity slicing on the area of interest for each of the images in the same series to designate different colors according to signal intensity. The resulting images are merged into a series of images for practitioners to play back with video equipment, and they can choose to browse single image or play the entire series of images in sequence.

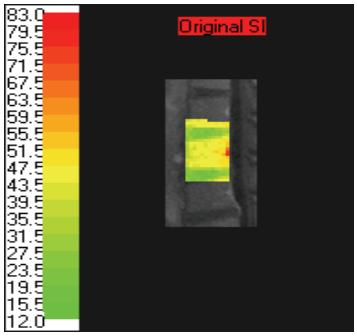


Figure 6. Example of intensity slicing of a lumbar spine's MRI image.

D. Intensity analysis

In addition to performing intensity slicing on the DICOM image files for visual enhancements, numerical analysis is also analyzed for the entire image. Take MRI for instance, a series of images are taken at a fixed position during the procedure; once the imaging is completed, doctors will then interpret the images according to the captures at different time points. If 4D images (2D plane plus image intensity and time) can be statistically or numerically calculated to reflect the image variations, then it can assist doctors in a way that is different from the conventional visual analysis of images.

We analyze pixel data of the DICOM data elements (7FE0, 0010) to represent different brightness intensities with different colors. Figure 3 shows an example of the brightness intensity analysis of the pixel data of an entire DICOM image, where pixels of similar intensity are close to one another in terms of color distribution. The area of analysis can be adjusted through varying the unit of observation, e.g. 1×1、2×2、3×3,…,*n*×*n* pixels. The smaller the units, the higher the accuracy, but the computational complexity also increases.

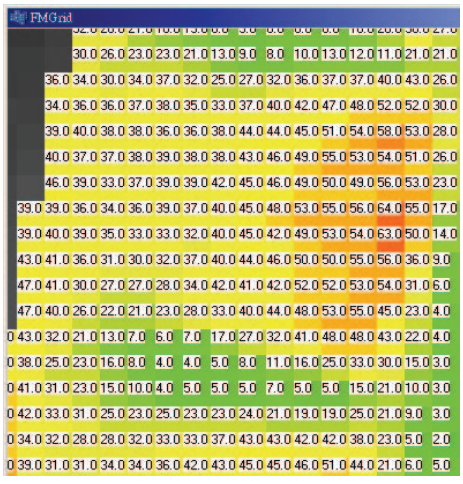


Figure 7. An example of the brightness intensity by analyzing the pixel data of an entire DICOM image.

In order to facilitate the observation of changes in image pixel values, we calculate the image intensity variations at each time point based on the preset analysis unit. Calculation of the variations includes the original SI value, slope value and peak value (%) changes. By utilizing these changes, a time point (X axis) versus area intensity variations (Y axis) line chart can be drawn for practitioners to use as diagnostic reference.

Among them, the original SI value is used to represent the original area signal intensity of the area, where the analytic data are acquired from reading the raw data value. Figure 8 shows the changes in original SI values for a series of CT image files, where the X-axis represents time change and Y-axis the brightness change. In the present study, the analysis values at three time points are taken, coupled with intensity slicing technology, to map the three analytic values onto an image using different colors to highlight the changes in (e.g. body tissue) signal intensity after an injection of test drugs (e.g. radio contrast agent).

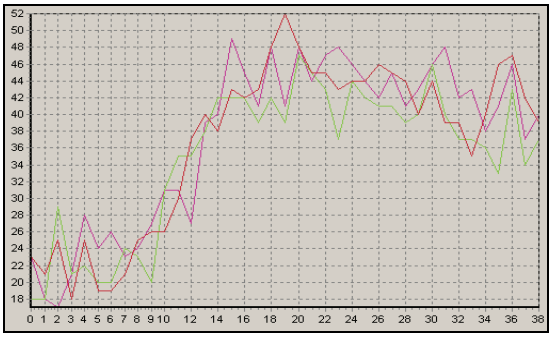


Figure 8. Variation chart of the original SI values from the image file.

In this study, the slope value is used to analyze an image between two neighboring time points, which can be seen as the signal difference of an image area between two consecutive time units. This analysis can help practitioners get a better grasp of the tissue of interest; for example, the higher the blood vessel density and tissue permeability would result in greater signal differences.

The slope value is defined as the changes in intensity between two neighboring time points. For instance, if the signal intensity is Y_1 at time point X_1 and Y_2 at time point X_2 for area *R*, then the slope for X_2 is calculated as:

$$S_R = \frac{(Y_2 - Y_1)}{(X_2 - X_1)} \tag{2}$$

Peak (%) value is defined as the signal difference between a particular time point and the initial time point. Assume the signal density is Y_0 at time point X_0 and Y_3 at time point X_3, then the peak (%) value at X_3 is calculated as follows.

$$P_R = \frac{(Y_3 - Y_0)}{Y_0} \tag{3}$$

Figure 9 shows a line graph for the peak value changes from the brightness analysis, where the X-axis represents the time point of observation, and the Y-axis the difference in variation. From the line chart, the differences between the intensity slicing areas at different time points can be observed. Drastic changes in the line graph imply more in-depth observation by the practitioner is demanded.

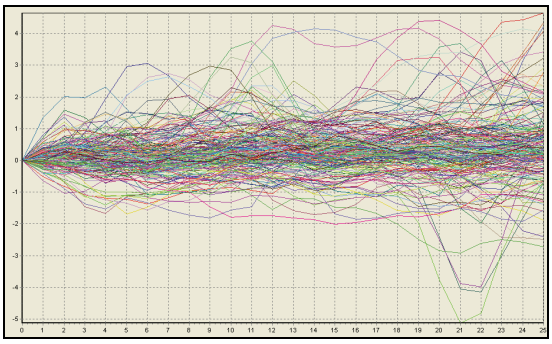


Figure 9. A line graph for the peak value form the image file.

IV. IMPLEMENTATION AND RESULTS

This section integrates the proposed strategies and the research methods to convert the algorithms into a user interface for medical personnel to load image files for analysis and observation. The following describes the results of practical application.

A. *Results of Intensity slicing*

Figure 11 (a) is an MRI image of a knee. We observed that the signal obtained for the same organs and tissues do not exhibit contrasting differences (pixel brightness of the image). Considering the naked eye is not able to pick up subtle signal differences, we intensified the original image by intensity slicing to convert the grayscale images that are difficult for observation into enhanced color images that can be subject to color change.

We strengthened the DICOM image browser functions by adding a color definition interface for users to assign levels for the observed image, and the system would in turn automatically assign the color distribution based on the defined levels. As shown in Figure 10, users can also define their own color to observe the color changes of the intensified area; they can also set up parameters for different examination type or body position to be saved as defaults for different examination scenarios for direct use in the future.

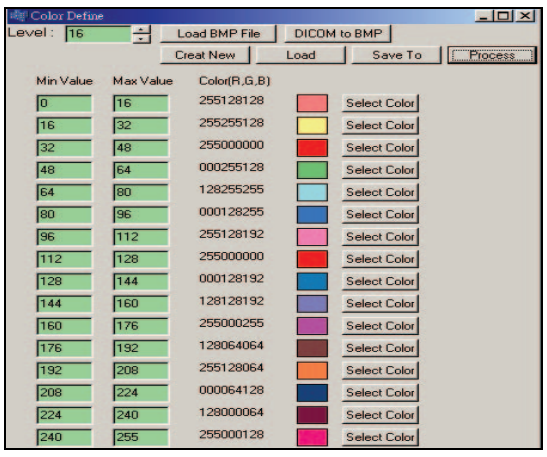
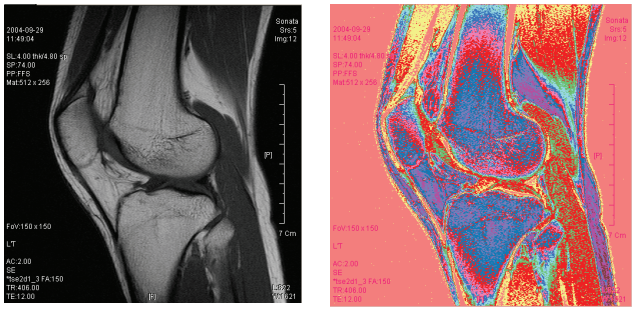


Figure 10. Color definition interface.

Figure 11 (b) demonstrates the results of a 16-color intensity slicing of a knee MRI image. As can be seen in the resulting image, grayscale differences that are difficult to be distinguished by the human eyes are now differentiated in the color image, which can efficiently facilitate the observation and diagnosis by medical practitioners.



(a) (b)

Figure 11. A color intensity slicing result for a knee MRI image, (a) MRI image of a knee; (b) A 16-color intensity slicing result.

B. *Numerical analysis and curve diagram*

Medical examinations usually produce a series of image files, and these files display variations due to different shooting times or reactions to drug injection. Through the analysis formulas in the previous section, we integrated the parameter definition for the line chart of original SI, slope, and peak (%) value, thereby allowing users to set up parameters to observe the sequential changes of the images.

Figure 12 shows the interface for numerical analysis, where parameter settings can be adjusted, as detailed as follows.

(A) Intensity slicing parameter setting
Provides the parameter setting for the line graphs of original SI, slope, and peak (%) to intensity slicing.

(B) Auto playback mode
Sets up the playback mode and delay time (ms) for the intensity slicing browser; the intensity slicing results can be saved in different image formats.

(C) Axis scroll setup
The axis represents the changes in brightness intensity in a given period of time. By scrolling along the axis, tissue sections at different time points can be observed, akin to watching a video in auto playback. This system provides four sets of axis scrolls for browsing the intensity slicing at four different time points for comparison, as presented in Figure 13.

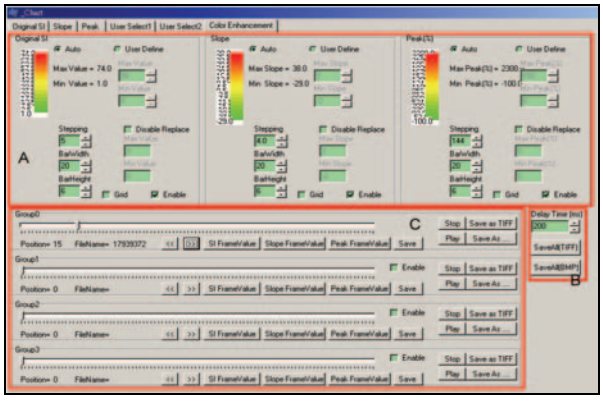


Figure 12. The user interface of Intensity Slicing tool.

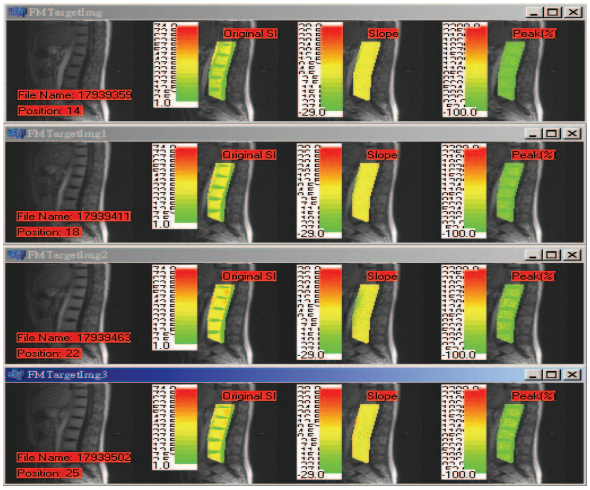


Figure 13. Brightness intensity of section images at four different time points.

In addition, users can choose to observe the entire area or a selected area. The system automatically generates a line graph (Figure 14) according to the area selected by the user. With the auto playback mode, image variations can be readily observed.

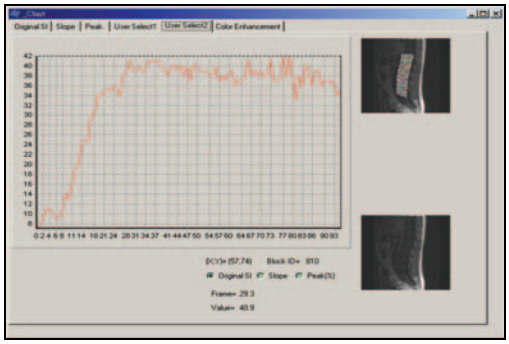


Figure 14. Intensity slicing line chart.

C. Component-based E-diagnosis Web Viewer

We use component software technology to develop a web-based image processing and display component to visualize and manipulate various DICOM images in a Web browser.

This web-based e-diagnosis system uses module-based design framework with multi-thread processing and multimedia content presentation capabilities, integrating the methods and algorithms in Section III to process DICOM image data, including static medical image display module, dynamic medical image display module, intensity slicing and analysis module and other functions, as shown in detail in Figure 15. The system execution results are presented in a web browser for user operation and manipulation.

A GUI is designed for the web browser, called the Web Viewer, containing functions such as the DICOM diagnosis system, organ image registration system, DICOM browser and patient database access. In terms of image reading and operation, users can select to perform color-level or color conversion, zoom in/out, image reorientation and size measurement through a simple mouse click. They can also carry out intensity slicing and analysis of sequence-images, or directly access previously stored information for reading and diagnosis. Figure 17 presents the Web Viewer interface in detail.

A function menu is placed on the left of the Web Viewer screen, while the top provides operation options, with the middle area displaying medical image or video, in which multiple images can be viewed and analyzed simultaneously. In addition, the system is equipped with a remote control module, permitting practitioners from different hospitals or regions to query electronic patient medical records or read the image sequence in the database via the web browser user interface upon authentication and authorization. Medical personnel at different hospitals can send and receive operating instructions to each other on the platform for online cooperation and telemedicine education.

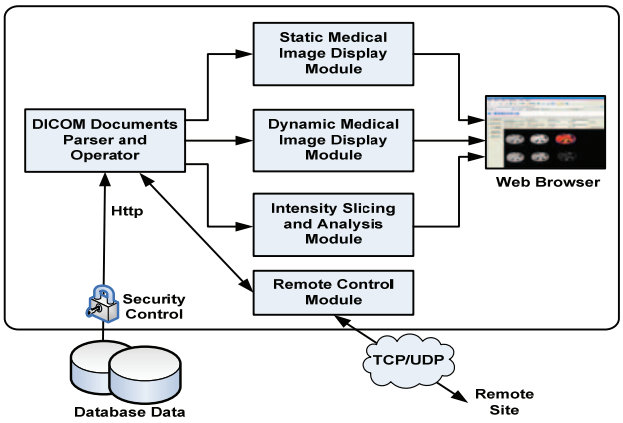


Figure 15. Architecture of the e-diagnosis module.

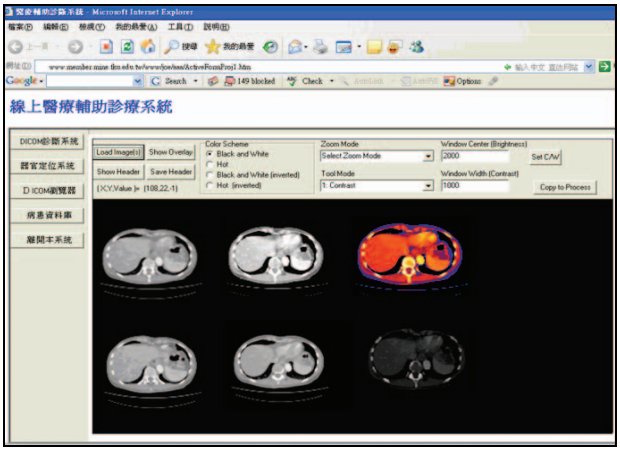


Figure 16. Screenshot of the E-diagnosis Web Viewer.

V. CONCLUSIONS

Medical practitioners need to diagnose numerous patients and analyze test results for the course of treatment, coupled with the progress and innovation in medical engineering, speed and accuracy are more important than ever. Commonly seen X-rays, CT scans or MRI tests often call for experienced doctors to read relevant data and produce a diagnosis and treatment recommendations

With our DICOM-based interactive e-diagnosis system prototype, doctors are provided with reference information to facilitate the diagnostic process. Using

image processing and analysis, the DICOM files are converted and acquired, allowing the user (doctor) to set up parameters or perform intensity slicing and analysis through the e-diagnosis system to obtain image information or label specific regions of interest (ROI) for computation to generate charts depicting image variations.

The system also combines Internet functionality to provide remote medical diagnosis and patient data query features, all of which are conducive for doctors or medical personnel in different regions to cooperate and get hold of real-time test images and information of the patient. Follow-up studies can seek to incorporate a more convenient patient information entry and update interface into the system for automatic image analysis and information retrieval, where the data can be stored in the patient's personal information for up-to-date medical records.

ACKNOWLEDGMENTS

This work was supported in part by the Asia University and National Science Council of the Republic of China under contract 98-NSC-03 and 99-2410-H-468-022-.

REFERENCES

[1] Rampado O, Garelli E, Zatteri R, et al, "Patient Dose Evaluation by Means of DICOM Images for a Direct Radiography System," *Radiologia Medica*, 2008; Vol. 113, No. 8, pp. 1219-1228.

[2] DICOM Standard, http://medical.nema.org/dicom/2009

[3] Hongmei Yan, Yingtao Jiang, Jun Zheng, Chenglin Peng, Qinghui Li, "A Multilayer Perceptron-Based Medical Decision Support System for Heart Disease Diagnosis," *Expert Systems with Applications*, Vol. 30, 2006, pp. 272–281.

[4] Lidia Ogiela, Ryszard Tadeusiewicz, Marek R. Ogiela, " Cognitive Techniquesin Medical Information Systems," *Computers in Biology and Medicine*, Vol. 38, 2008, pp. 501 – 507.

[5] Zheng NN, You QB, Meng GF, et al, "50 Years of Image Processing and Pattern Recognition in China," *IEEE Intelligent systems*, 2008, Vol. 23, No. 6, pp. 33-41.

[6] R. Tadeusiewicz, M.R. Ogiela, "Medical Image Understanding Technology," *Springer*, Berlin, Heidelberg, 2004.

[7] Chung-Ming Wang, Yao-Hsien Huang, "A Novel Color Transfer Algorithm for Image Sequences," *Journal of Information Science and Engineering*, 2004 ,Vol. 20, pp. 1039 -1056

[8] Lee E, Fahimian BP, Iancu CV, "Radiation Dose Reduction and Image Enhancement in Biological Imaging Through Equally-Sloped Tomography," *Journal of Structural Biology*, 2008, Vol. 164, No. 2, pp. 221-227.

[9] Wahed MES, "Image Enhancement using Second Generation Wavelet Super Resolution," *International Journal of Physical Sciences*, 2007, Vol. 2, No. 6, pp. 149-158.

[10] Spetsieris K, Zygourakis K, Mantzaris NV, "A Novel Assay Based on Fluorescence Microscopy and Image Processing for Determining Phenotypic Distributions of Rod-Shaped Bacteria," *Biotechnology and bioengineering*, 2009, Vol. 102, No. 2, pp. 598-615.

[11] Rodriguez J, Martin MT, Herraez J, et al, "Three-Dimensional Image Orientation Through Only One Rotation Applied to Image Processing in Engineering" , *Applied Optics*, 2008, Vol. 47, No. 35, pp. 6631-6637.

[12] Weidong Cai, Dagan Feng, Roger Fulton, "Web-Based Digital Medical Images," *IEEE Computer Graphics and Applications*, 2001, pp.44-47.

[13] M. Sonka, J.M. Fitzpatrick (Eds.), Handbook of Medical Imaging: Vol. 2. Medical Image Processing and Analysis, *SPIE* Press, Vol. PM80, Bellingham, USA, 2000.

[14] Prabhakar B, Reddy MR, "HVS Scheme for DICOM Image Compression: Design and Comparative Performance Evaluation," *European Journal of Radiology*, 2007, Vol. 63, No. 1, pp. 128-135.

[15] Lebozec CD, Bourquard K, Fabiani B, et al, "Virtual Blades for Routine Diagnosis. The Importance of Using Computing Standards HL7 and DICOM," *Annales de pathologie*, 2008, Vol. 28, No. 1, pp. S109- S112.

[16] Prabhakar B, Reddy MR, "HVS Scheme for DICOM Image Compression: Design and Comparative Performance Evaluation," *European Journal of Radiology*, 2007, Vol. 63, No. 1, pp. 128-135.

[17] Zhanli Hu, Hairong Zheng, Jianbao Gui,"A Novel Interactive Image Processing Approach for DICOM Medical Image Data," in *2nd International Conference on Biomedical Engineering and Informatics, BMEI '09*, 2009. pp. 1-4.

[18] Shusaku Tsumoto, "Web Based Medical Decision Support System: Application of Internet to Telemedicine," *2003 Symposium on Applications and the Internet Workshops*, 2003, pp. 288-293.

[19] Rainer Anzböcka, Schahram Dustdar, "Modeling and Implementing Medical Web Services," *Data & Knowledge Engineering*, Vol. 55, No. 2, 2005, pp. 203-236.

Rong-Chi Chang is an Assistant Professor of Digital Media Design at Asia University, Taiwan. He specializes in image restoration and also works as the program coordinator of applied computer science in the Multimedia Computing Laboratory. He earned both M.S. and Ph.D. degrees from Tamkang University in Computer Science and Engineering. His research focuses on multimedia computing, interactive media technologies and image processing.

A Web-based E-learning Platform for Physical Education

Chun-Hong Huang[1]

[1]Dept. of Computer Information and Network Engineering, Lunghwa University of Science and Technology, Taoyuan, Taiwan
E-mail: ch.huang@mail.lhu.edu.tw

Su-Li Chin[2] , Li-Hua Hsin[3], Jason C. Hung[4], Yi-Pei Yu[5]

[2]Dept. of Physical Education, Tamkang University, Taipei, Taiwan
E-mail: Csunny@mail.tku.edu.tw

[3]Physical Education Office, Lunghwa University of Science and Technology, Taoyuan, Taiwan
E-mail: hsinlihua@mail.lhu.edu.tw

[4]Department of Information Technology, Overseas Chinese University, Taichung, Taiwan
Email: jhung@ocu.edu.tw

[5]Dept. of Computer Elertrical Engineering, Lunghwa University of Science and Technology, Taoyuan, Taiwan
E-mail: lucky760729@hotmail.com

Abstract- The major purpose of this paper is to develop a Web-based E-learning Platform for physical education. The Platform provides sports related courseware which includes physical motions, exercise rules and first-aid treatment. The courseware is represented using digital multimedia materials which include video, 2D animation and 3D virtual reality. Courseware within digital multimedia materials not only can increase the learning efficient but also inspires students' strong interest in learning, especially in the area of Physical Education. The design concept of our project is based on ADDIE model with the five basic phases of analysis, design, development, implementation, and evaluation. Via the usage of this Web-based E-learning platform, user can learn the relative knowledge about sports at anytime and in everyplace. We hope to let players perform efficient self learning for sports skills, indirectly foster mutual help, cooperation, nice norms of law-abiding via the learning of exercise rules, and become skilled at accurate recreation knowledge and first-aid expertise. Moreover, coaches can use the system as a teaching facility to mitigate loading on teaching.

Index Terms- **Digital Multimedia, E-learning, ADDIE, sports skill, exercise rules, first-aid**

I. INTRODUCTION

Duo to the invention and quickly development of internet, the accumulated knowledge and information has violent development. And people have more varied methods to learn everything in everywhere at anytime. E-Learning is a trend of education, it can assistant teacher and reduces the loading of teaching. Students can develop the relative knowledge or skills through experience in virtual laboratories and simulated environments0. However, the most existent E-Learning systems focus on general subjects such as linguistics, mathematics, management, or science. The E-learning should try to lay more stress on physical education since sports activities are placed importance on our daily lives progressively and can strengthen someone's mind and body.

In our work, we aim to develop an E-learning platform for physical education which integrates the different courseware within multimedia factors such as 2D/3D animation, and digital video. Since Physical education has the distinguishing characteristics to other areas such as philology, management, business, and etc.., the applications of computer multimedia are suitable for E-learning on physical education fairly. One of our issues is to focus on the training of basic kinematical movements, via the usage of the videos which represent the basic kinematical movements user can learn the sports skill. Since the implementation of the sports morality of is based on cognition, the player should know and understand the rules of exercise. In order to increase the understanding of exercise rule and the cognition of sports morality, we fabricate the courseware of exercise rule with the form of 2D animation. Therefore, another import issue of our works is to learn the rules of different sports to form sports morality of players and avoid the foul trouble in competition. No one can promise nothing to happen when a player is in a competition game or in practicing, so our platform also supply the courseware of first-aid which can be represented by 3D VRML.

II. RELATIVE WORKS

ADDIE [2] [3] model is the step of content design which is based on Instruction System Design (ISD). It usually applies to map out the course, courseware design and instruction. It also can support to the design and development of system. Based on the concept of ADDIE model, most policy of instruction can be considered. In our jobs, the instructional policy for physical education applies the courseware within multimedia contents to instruction.

Using instructional media such as video, picture or animation is better than only the text. In the area of physical education, video is a good media to supply the instructional activities. The video has the capability to

doi:10.4304/jnw.6.5.721-727

display or replay the sports action or key motion loop and loop. Teacher or coach can explain the action to the students directly through the media repeating the key motion or difficulty slowly. Student can learn the course by himself via voice guidance and sample action demoing within the video scene. When students learn the sports skill, to view the action skill repeatedly and supply the diversiform learning method are very important [4] [5][6] [7]. Du to the specific features of media, courseware which is created based on computer multimedia can satisfy the demands of teaching and learning. Therefore, a learning application which integrates multimedia elements to apply to sports can supply the sense effect and increase learners' learning motivation and desire. Multimedia courseware also plays an assistant role to help teacher and give the diversiform instructional policies to students when students get out of class. As the description in [8], instructional actives under the environment of multimedia technology can stimulate students' interest and mobilize their enthusiasm in learning. Students are faced to computer that is a magic world with text, sound, pictures, animation, and video, even or virtual reality. It is the new form within the vivid graphics and a lot of resources of digital media factors which cause students' interesting in it. It not only activates the classroom atmosphere, inspires students' strong interest in learning, but also mobilizes the students to participate actively. Katz, L believed that the integration with the strong capability, plasticity, and the information management policy of computer can produce the distant prospect to the combination with scientific technology and physical education. Knudson, D., and Kluka indicated that the action demonstration in the media is an efficient and appreciative guidance for physical motion learning. It can strengthen learners' realization, cognition and study efficient.

III. DESIGN OF PLATFORM AND COURSEWARE

A. *Design concept*

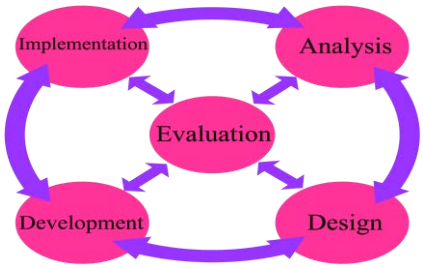


Figure 1. The concept of ADDIE model.

The procedure of platform development and courseware construction is based on the concept of ADDIE model as figure 1 illustrates. Basically, ADDIE model includes five phases which may reference mutually and depend on each other in some phases. It includes several works as the following:

(1) Analysis:

In this phase, we should consider two major tasks as the following:

(1-1) Demand Analysis:

The task carries out the analysis of demand to specific subject of sports teaching and learning. Learner, learning environment, resource, the goal, and system framework need to be analyzed. Our goal is to let the players can get the sports skill, understand the exercise rule which may establish their morality with nice norms of law-abiding directly, know the processing of first-aid via a web learning platform which supplies the serial courseware of physical education. Coach also can use the platform as a teaching facility to mitigate loading on teaching. For our demand in this project, we need the specialists of digital contend, art designer, programmer, professors of physical education. The achievements and literatures are also needed to investigate and study in order to improve the demand and create the new ideals.

(1-2) Content Analysis:

In this task, the work focuses on the kernel of courseware, correction and the suitability of course content. Besides to carry out the content analysis of network data, scientific or technical literature, and teaching media, it also needs to invite the specialists of sports to analyze the contents, and institutes the teaching framework and learning policy. The delivery options which will be included in the Sports E-learning platform also should be considered. The platform must display teaching contents with professional, interesting and funny way to attach players and coaches to use.

(2) Design:

To represent the logicality of courseware clearly and systematically, it is necessary to draw the script with matching up the Human-Computer Interface. The tasks would be considered as following:

(1-1) The design of platform architecture and the production of courseware will be carried out in this phase.

(1-2) The framework of the courseware and the represented way of course should be designed.

(1-3) The mechanism of interaction and feedback evaluation of the application should be designed.

(1-4) Considering the media factors such as video, image, voice, text, animation and effect that will be used and integrated for courseware and platform.

(3) Development:

Depending on the output of Design and Implementation, a programmer starts to develop the platform which can integrate with the teaching material. The developer of courseware also starts to create the course. Since the courseware will be integrated with the platform in the final, the programmer and developer of courseware should make a well communication with each other and need to understand the demand between

them.

(4) Implementation:

In the phase of implementation, the plans have to draw up. The plans should consider the timeline of implementation and the procedures for training at least. The procedures of formative evolution for the learning efficient of student should be established. Since the final system is developed based on needs, testing and modification while utilizing a prototype system with members of the target audience, it is necessary to construct the procedure of summative evolution for the final products which include courseware and platform.

(5) Evaluation:

During the period of testing and modifying for system, learners and coaches start to use the system, and make the evaluation for system. They may propose their suggestions that will be the reference to modify the system and the courseware. Then, the task carries out the testing and evaluation by specialists whose research areas are in Human-Computer Interface, educational technology, and physical education. It tries to find the problems, and makes modification by carrying out the entire evaluation to system. In the next, the correction of courses needs to certify. The formative and summative evolution would be implemented and finished in this phase. Finally, the correct course will apply to the activities of instructions.

B. Design the Courseware

According the demand analysis, we fabricated three different multimedia courseware including sports skill, exercise rule, and fist-aid treatment. All of the courseware can be viewed via the Sports E-learning platform that integrates different digital contents of video, voice, animation, and 3D VRML model.

B-1. Courseware Design of Sports Skill

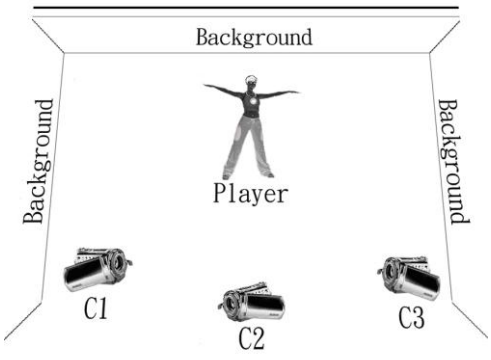


Figure 2. The cameras are placed in the studio.

For the courseware of sports skills, they are displayed with the video and can be viewed with three different shots. We shot the serial courseware includes kickboxing, tennis, badminton, Chinese martial art, basketball, and physical fitness and etc.. Each sports type includes basic and advanced actions, and several of the types have the continuous motions. Basically, the degree of the courseware is from easy to hard. To display the Multi views of a sports action, three cameras are set to shoot the action synchronously in three different orientations and directed to the player as shown as figure 2. To adapt to the need of different sports, three cameras can be located in suitable places.

B-2. Courseware Design of Exercise Rule

The courseware of the exercise rule also be design and fabricate in our work. In order to increase the learning interesting of users and develop player's cognition for exercise rule, the learning contents of the exercise rule should be displayed with the easy, vivid and vigorous way. By the above demand of request, Flash animation is considered to use to represent the exercise rule. As the indication [9][10], something in instruction using animation should be watched out. The arrangement or the decoration of multimedia material contents will result in Cognitive Overload of learning[11]. Using multimedia instructional material, both high density text and complexity graphic will influence visual attention. Since students trend to rely more heavily on narration with low text density and relevant animation, the instructional design in animation should integrate the multimedia material that includes clear narration using voice, text with low density, and concise graphic.

B-3. Courseware Design of first-aid CPR

In recent year, a very popular area of computing is Virtual Reality. The purpose of Virtual Reality is to describe method of interaction and simulation with the 3D environment. VR can be thought of Human-Computer Interface to 3D simulation model which allows the user to enter, interact and grope for the real world that concerning to the system [12] [13]. In our project, we create a 3D VRML model for CPR of first-aid treatment that can be delivered and view on the web. VRML is the Virtual Reality Modeling Language, a standard file format with small file size to display 3D models on the web. User just only needs to install a free VRML plugin application that plugs in web browser for viewing the 3D model.

C. Design the framework of Platform

According the investigation, the most professors whose area is in physical education consider that the traditional instructional method of sports should be improved. The Computer Assisted Learning System can help them on the work of instruction. And the digital content that at least integrates the video and speech is a good choice to help them to teach the players to learn the sports skill in the developed era.

Regarding with the notion of these professors, to construct a web-based Sports E-learning platform which can exhibit some types of courseware of physical education will be a meaning task. The framework for Sports E-Learning platform on physical education as the figure 3 shows.

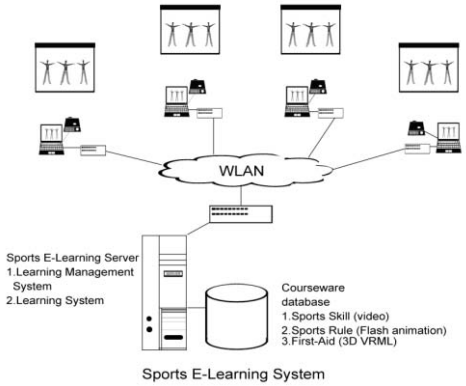


Figure 3. The framework for Sports E-Learning platform on physical education

IV. DEVELOPMENT OF PLATFORM AND COURSEWARE

The works of development include the fabrication of courseware and platform development. Courseware fabrication is the time-consuming job, since the pre-work of courseware that is presented with multiple-view video includes the shooting, editing, post-production, and needs to evaluate the discrimination between good and bad. And both the animation and VR models which represent the exercise rule and first-aid treatment also need to spend much time to create the teaching materials. Next is the description of the fabrication of different courseware.

A. Fabrication of Courseware

A-1. Courseware of Sports skill

For sports skill learning, each action is shot in three different orientations synchronously, and we use the video editing application to edit the video as figure 4 shows. The video editing application can be used to extract the suitable clips which need to be edited further in the next. The effect with slow motion is blended into the clips. Then, the clip with the slow motion effect should be connected within to its relative clip. We asked the specialists and coaches to examine the correction of actions. According to their sufficient and professional knowledge, the actions are discriminated into standard actions and non-standard actions. All of the actions will be reprocessed with combining the speech guidance which indicates the key points or the mistakes for sports skill. No matter standard or non-standard actions will be stored into the database. Users can observe the action with the three different angle views via the user interface. With the comparison between the correct and incorrect actions, students may nose out the mistake of the non-standard action or the key skill of standard action. When they practice the skill or in sports tournament, they can beware of the mistake.

Figure 4. The video editing application is used to edit the film and integrate with the voice of action guidance.

The narrative guidance will be integrated with the video, since the voice is also an important factor for instruction. Therefore, the voice guidance of action from specialists is recorded and integrated with the video. Finally, all of the videos will be transformed into the format of Flash FLV which supports the streaming capability on the web. Now, the amount of sports action in our sports database is above 150 including standard and non-standard actions, and we still increase and make up the more sports actions continuously.

A-2. Courseware of Exercise Rule

The courseware of exercise rule is created by Flash animation development tool. It is a popular development tool for fabricating 2D animation, and supplying the interactive capability via the action script language as shown as figure 5. In each course of exercise rule, the narration of speech also matches up the animation. With the mechanism of interaction, user can interact with the course content and choice the items that may interest them more. Therefore, the idea of the courseware design considers the instructional method with double delivering. Figure 6 illustrates the Flash animation that represents the rules of athletics sport.

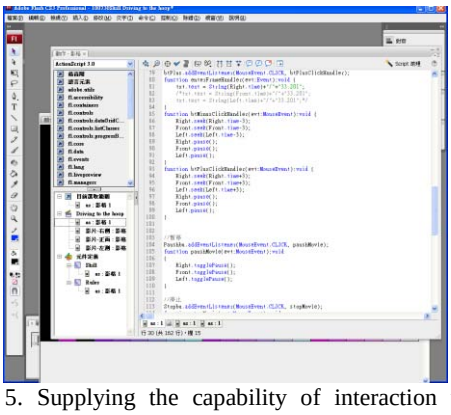


Figure 5. Supplying the capability of interaction via the programming of Action script language

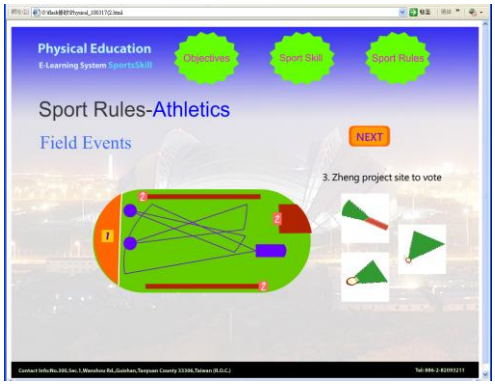


Figure 6. Exercise rule is represented by Flash animation.

A-3. Courseware of first-aid treatment

In fabrication of first-aid, we used the 3D development tool- 3D MAX studio which is a famous software application in developing the 3D models and scenario. However, the file size of the most 3D models is large since they need the heavily computing with the great quality data to render the 3D graphic. To save this problem, we transform the 3D MAX model into 3D VRML text file that is a standard file format with small file size to display 3D models on the web. In our project, we create a fist-aid course material for CPR firstly, and we will create the other fist-aid processing instructional material such as triangular bandage and tube bandage. Figure 7 shows the creative job using 3D MAX studio, and Figure 8 shows the content of CPR with 3D VRML model.

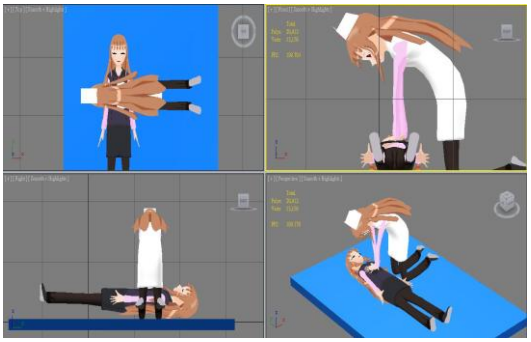


Figure 7. The fabrication in 3D MAX studio.

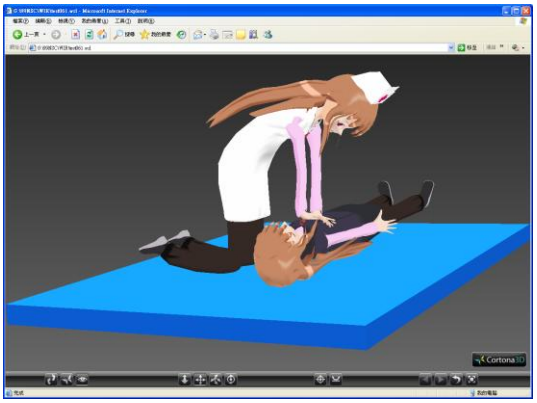


Figure 8. The course content of CPR with 3D VRML model.

B. Platform Development

In platform development, the development tools we used including PHP developer, Adobe Dreamwerver, and Flash action script to develop the learning platform.

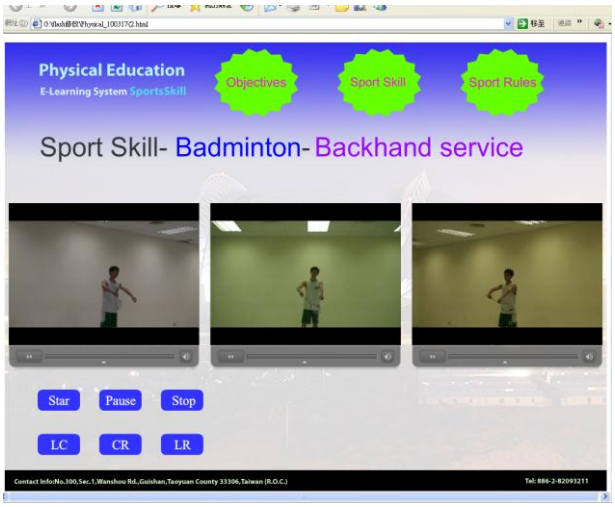


Figure 9(a). A specific action with three different angle views is displayed via three media players simultaneously.

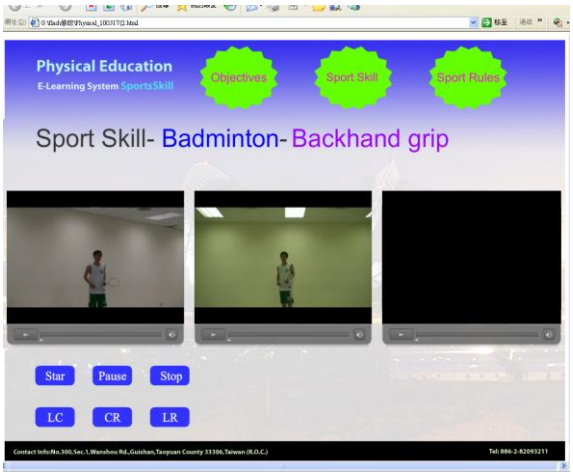


Figure 9(b). A specific action with two different angle views is displayed via two media players simultaneously.

The media players that plug in system screen display the same action with different angle views as shown as figure 9(a)(b). A user can interact with the system via this interface. One action which has three different orientations can be displayed in three or two media players simultaneously. And user also can control any one of them certainly. The instruction for the sports skill with the diversiform and repeated way yields twice the result with half the effort. Therefore, player can learn the sports action with the repeated clip via the system since the film integrates the effects such as slow-motion effect and voice of guidance.

At present, the platform and courseware had been demonstrated to 6 professors whose area is in physical education and 253 undergraduate students, and most of

them thought that it is helpful in learning phase of some specific sports. Some of them suggested increasing one view that is located on the top of the player. According to their concluding comments, the platform is feasible on E-learning for Physical education. However, we must emphasize that the system just only plays the role as auxiliary. The teachers or coaches are still the most important guide roles on learning sports knowledge.

V. CONCLUSION

In this paper, we developed an E-learning platform for sports which is based on the concept of ADDIE model and fabricated the courseware. They are represented matching up with the different multimedia digital contents and factors such as video, voice, animation and 3D VRML model. Users can view the one sports action clip that can be watched with the different orientation views. Users can only control the one buttons of any media player via the user interface, and any media player will display the synchronic action as the other media players. For students, all the courseware that our system supplies is not only used to learn or to improve the sports skill but also help them to establish the cognition of sports morality. And they can get the skill and knowledge of first-aid treatment that may help themselves or other ones to salve their life when some accidentals occur.

Here, we do not emphasize on the complex technique for system development, but on the contrary we use the easy and ripe technique to fabricate the courseware and develop the Physical education E-learning platform. We hope that E-Learning on Physical education can be attached great importance since people pay close attention to their body healthiness and many sports activities have became the professional job.

Now, our platform only supplies the multimedia courseware for instruction. To achieve the better learning efficiency, we will develop the Learning management system and Assessment system that will integrate with the existent learning system in the future. And we will also demonstrate the system to the more professors and students of physical education and general area. And getting more suggestions from them to verify the degree of feasibility, and improve the system and course contents.

ACKNOWLEDGMENT

The authors would like to thank the National Science Council of Republic of China, Taiwan for financially supporting this research under Contract No. NSC 99-2410-H-262-014.

REFERENCES

[1] Harris and Krousgrill (2008), "Distance Education: New Technologies and New Directions" Proceedings of the IEEE, Volume: 96, Issue: 6, 2008 , pp. 917 – 930.

[2] Molenda, M, (2003) The ADDIE model, Encyclopedia of Educational Technology, ABC-CLIO

[3] LaMaster, K., Williams, E., & Knop, N. "Technology implementation: Let's do it." Journal of Physical Education, Recreation, and Dance, 69(9), 1998. pp.12-15.

[4] Katz, L., "The role of interactive video, multimedia and technology in Physical Education: Towards the years 2000." Processing of the International conference on computer Application in Sport and Physical Education, 1992, pp.22-31.

[5] Knudson, D., PhD and Kluka, D., PhD.- 2708 -Impact of Vision and Vision Training on Sport Performance, Reprinted with permission from the Journal of Physical Education, Recreation and Dance ,Vol.68, Iss.4, April 1997, pp.17-24.

[6] Bonnie S. Mohnsen, "Using Technology in Physical Education (7th edition). Bonnie's Fitware Inc.

[7] Morrison, G. R., Ross, S. M., & Kemp, J. E. (2001). "Designing effective instruction (4rd ed.)." New York: John Wiley & Sons, Inc.

[8] Zhihua Tan; Song Li, "Multimedia Technology in Physical Education", International Symposium on Computer Network and Multimedia Technology, 2009, pp.1-4.

[9] S. M. Alessi and S. R. Trollip, Multimedia for Learning: Methods and Development (Boston: Allyn and Bacon, 2001).

[10] Eric Wiebe and Leonard Annetta, "Influences on Visual Attentional Distribution in Multimedia Instruction," Journal of Educational Multimedia and Hypermedia, 2008, pp.259-278.

[11] R. E. Mayer and R. Moreno, "Nine Ways to Reduce Cognitive Load in Multimedia Learning," Educational Psychologist, 2003, pp.43-52.

[12] Maria Roussou. Virtual reality and interactive theaters: Learning by doing and learning through play: an exploration of interactivity in virtual environments for children. Computers in Entertainment (CIE). 2(1): 10 (2004)

[13] Mike Fraser, Tony Glover, Ivan Vaghi, Steve Benford, Chris Greenhalgh, Jon Hindmarsh, and Christian Heath. Revealing the realities of collaborative VR. In Proceedings of the third international conference on Collaborative virtual environments, 2000, pp.29-37.

Biographical notes:

Chun-Hong Huang is an Assistant Professor of the Department of Computer Information and Network Engineering at Lunghwa University of Science and Technology from 2007 to date. He received his PhD in Computer Science and Information Engineering from Tamkang University in 2007. His current research interests are in the areas of Multimedia Processing, E-learning, 3D Information Analysis and Retrieval.

Su-Li Chin is an Associate Professor of the Department of Physical Education, Tamkang University, Taiwan. She was the player of the national basketball team, and retired from the national team in 1991. She received her Master degree in Graduate Institute of Coaching Science from National Taiwan Sport University in 2001. Her research interests include Sports Biomechanics and Sports Skill Analysis.

Li-Hua Hsin is a Lecturer of the Physical Education Office at Lunghwa University of Science and Technology from 1998 to date. She received the Master degree in Graduate Institute of Coaching Science from Chinese Culture University in 1998. His current research interests and specialties include the areas of Athletics、Dancesport、Prevention and Treatment of Injuries, and Sports Medicine.

Jason C. Hung is an Associate Professor of the Department of Information Management, Overseas Chinese University, Taiwan. His research interests include Multimedia Computing and Networking, Distance Learning, E-commerce and Agent Technology. He received his PhD in Computer Science and Information Engineering from Tamkang University in 2001.

Yi-Pei Yu is a Master student at Lunghwa University of Science and Technology, Department of Electrical Engineering. Her current research interests are in the areas of E-learning, Virtual Reality and Multimedia Processing. Her contact email is lucky760729@hotmail.com

Incremental Mining of Closed Sequential Patterns in Multiple Data Streams

Shih-Yang Yang
Department of Media Art, Kang-Ning Junior College of Medical Care and Management, Taipei, Taiwan 114, R.O.C.
Email: shihyang@knjc.edu.tw

Ching-Ming Chao
Department of Computer Science and Information Management, Soochow University, Taipei, Taiwan 100, R.O.C.
Email: chao@csim.scu.edu.tw

Po-Zung Chen and Chu-Hao Sun
Department of Computer Science and Information Engineering Tamkang University, Tamsui, Taiwan 25137, R.O.C.
Email: pozung@mail.tku.edu.tw, 894190320@s94.tku.edu.tw

***Abstract*—Sequential pattern mining searches for the relative sequence of events, allowing users to make predictions on discovered sequential patterns. Due to drastically advanced information technology over recent years, data have rapidly changed, growth in data amount has exploded and real-time demand is increasing, leading to the data stream environment. Data in this environment cannot be fully stored and ineptitude in traditional mining techniques has led to the emergence of data stream mining technology. Multiple data streams are a branch of the data stream environment. The MILE algorithm cannot preserve previously mined sequential patterns when new data are entered because of the concept of one-time fashion mining. To address this problem, we propose the ICspan algorithm to continue mining sequential patterns through an incremental approach and to acquire a more accurate mining result. In addition, due to the algorithm constraint in closed sequential patterns mining, the generation and records for sequential patterns will be reduced, leading to a decrease of memory usage and to an effective increase of execution efficiency.**

***Index Terms*—Multiple Data Streams, Data Stream Mining, Sequential Pattern Mining, Incremental Mining**

I. Introduction

Due to widespread use of the network, increasing hardware processing speed, and expanding disk storage capacity, the amount of data stored on the computer and network is expanding enormously. As such, discovering useful knowledge from large amounts of data is of particular importance for businesses and relevant practitioners. The objective of data mining is merely to discover knowledge from large amounts of data, which enables the businesses to understand and predict user behaviors so as to expand business opportunities.

Various data mining techniques have been developed to meet different needs, while sequential pattern mining is one of the techniques. Sequential pattern mining is to discover sequential patterns that frequently occur in time sequence or specific order. Through analysis on the state change of sequences, we can make predictions on future states. For example, physicians can discover the evolution process of diseases in terms of patients' medical records in order to prevent and cure diseases sooner.

Traditionally, data are first stored in databases or integrated into data warehouses and are then provided for data mining. However, the advancement of computer and network technology leads to the emergence of the data stream environment, which is data rapid growing, information fast changing, and real-time demand highly enhancing. In such an environment, stream data are different from traditional data in which they are changing rapidly, massively or possibly infinitely, and fail to be completely stored. Therefore, traditional data mining techniques were ineffective. As a result, new approaches to data stream mining have been studied, in which [1, 2, 3, 4] proposed techniques for mining sequential patterns over data streams.

In recent years, an increasing number of emerging applications start off to gear towards monitoring multiple data streams in order to perform more advanced analysis. In the intensive care unit (ICU), for example, medical devices used on a patient will generate multiple data streams. By analyzing these stream data, doctors can receive a more insightful understanding of the patient's physiological changes. Therefore, mining in multiple data streams is one of the latest research issues for data mining.

Among the studies on sequential pattern mining in multiple data streams, Oates and Cohen [5] proposed the MSDD (Multi-Stream Dependency Detection) algorithm to search from different data streams for the rules of particular events taking place at a fixed time frame. Chen et al. [6] held that the rules found by MSDD merely account for an exception of sequential patterns and therefore proposed the MILE (Mining in Multiple Streams) algorithm to search for complete sequential patterns. However, MILE uses one-time fashion for mining multiple data streams at a certain time period. When new stream data are input, therefore, MILE only mines new data rather than integrates new mining results with old ones to generate more accurate sequential patterns.

doi:10.4304/jnw.6.5.728-735

In view of the aforementioned problems, this paper proposes ICspan (Incremental Mining of Closed Sequential Patterns in Multiple Data Streams) algorithm for mining sequential patterns through an incremental approach, in order to provide a more accurate mining result and to reduce special and time consumption in the mining process. Nonetheless, we have the following difficulties to overcome during the study process. First, the data amount in multiple data streams environment is more enormously relative to a single data stream, while the more sequential patterns will also been generated. Under the limitation for memory capacity, an effective processing for unnecessary sequential patterns is required. As a result, in the data stream environment, data could not be stored indefinitely, and to assure continuous data entry, preserve the accuracy of sequential pattern mining results.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the SPAMDAS method, which covers data sampling and incremental mining. Section 4 evaluates and compares the performance of the ICspan algorithm. Section 5 concludes this paper.

II. RELATED WORK

Due to the rising data stream environment in recent years, many studies have started to improve traditional data mining algorithm to accommodate data stream environment. Mining techniques for data stream environment have been put forward in Association Rule, Classification and Clustering, as well as sequential pattern mining.

First of all, the eISeq method proposed by Chang and Lee [1] will be discussed. This method could catch the latest change for the sequential data stream within a short time frame; while through a decaying mechanism to gracefully discarding the non-useful old data.

In the same year, Chen et al. [6] propose the MILE algorithm to implement weaknesses in the two papers [5, 7], which was used to process multiple data streams within a time cycle. In the time-series data stream environment, input data are considered a series of consecutive and time ordering token. Tokens include stream identification number, timepoint, and values, while each data stream is composed by tokens from the identical stream identification number. For example, in the stock market, stream identification number may serve as a stock identification, while in the medical treatment; stream identification number could represent the model number of a medical instrument. In addition, all tokens for multiple data streams at the same timepoint are all synchronously recorded, and the sequential pattern mining across data stream will be conducted in response to these multiple data streams in synchronized records. While MILE is a reformed method based on PefixSpan, enhancing efficiency for algorithm through recording already generated sequential patterns and expediting search by a hash method. However, the algorithm records generated sequences based on the fundamental framework of PrefixSpan, accordingly resulting in a high consumption of memory. In addition, due to MILE is a one-time fashion mining algorithm, previously mining results could not be preserved.

Subsequently, Raissi et al. [4] proposed SPEED (Sequential Patterns Efficient Extraction in Data Streams) algorithm to search for the maximal sequential pattern in the data stream. This algorithm maintains frequent sequential patterns based on the new data structure in addition to augmenting prompt decaying strategies, allowing users to search for the maximal sequential pattern at an arbitrary time interval in any given time.

Nonetheless, Ho et al. [3] proposed IncSPAM (Incremental Mining of Sequential Patterns using SPAM) algorithm which differs from earlier mentioned methods in which it applies incremental method for mining sequential patterns from the data stream. In this algorithm, bitmap representation is used for calculating the supports of sequential patterns, in addition to implementation through the concept of CBASW (Customer Bit-Vector Array with Sliding Window), in order to raise the supports for calculating sequential patterns and the speed of sequential pattern ordering. Furthermore, to process the false positive problems from out-of-date data, the concept of decay-rate proposed by Chang and Lee [8] will be improved and used to determine the importance of data through a decay mechanism, allowing algorithms to accommodate to the concept drift problems frequently seen for data stream mining.

III. THE SPAMDAS METHOD

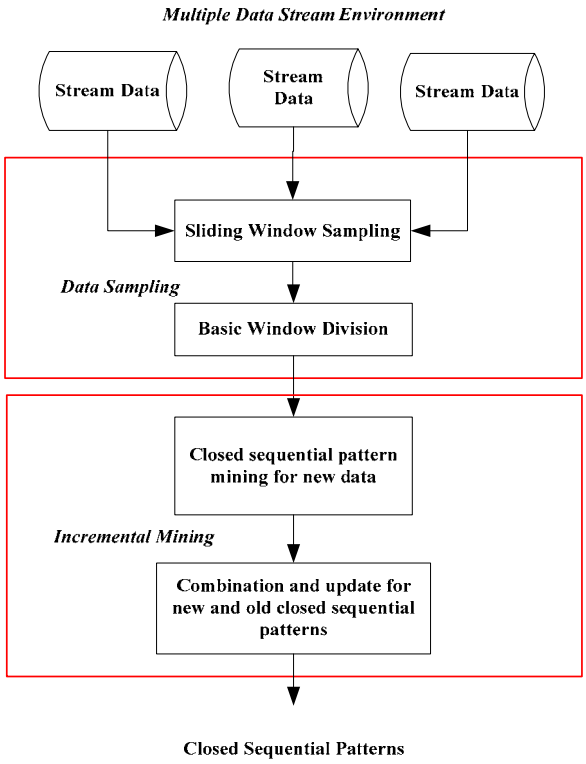


Figure 1. Overall process of SPAMDAS.

From the previous section for literature review, we have discovered the relevant studies on sequential pattern

mining in multiple data streams could not preserve the results of the previously mined sequential pattern, leading to a less accurate mining result. To approach such a problem, this paper proposes SPAMDAS (Sequential Pattern Mining in Multiple Data Streams) method to solve this problem. SPAMDAS mainly consist of two phases in data sampling and incremental mining. First, in the data sampling phase, a sliding window mode samples from stream data and to divide this sample data into units in lieu of the basic window. Followed by the incremental mining phase, the different basic windows in the sliding window are matched to search for closed sequential patterns, and further combine and update on the new and old sequential patterns to generate more accurate mining results of sequential patterns. Figure 1 shows the overall process of SPAMDAS.

A. Data Sampling

As referred in Figure 2, in the data sampling phase, we adopt time-sensitive sliding window to proceed with non-overlapping sampling on input stream data. In addition, we obtain mined periodical sequential patterns, limiting frequent sequential patterns with time interval, less than the fixed time intervals. Therefore, we divide the data within the sliding window according to the scale of time interval into several identical length window (these windows are called basic windows). Then we will be able to match the sequential data from these basic windows to discover a certain period of frequent sequential patterns.

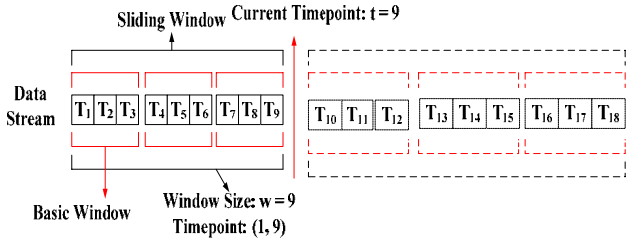


Figure 2. Example of Sliding Window Sampling.

Figure 2 shows sampling from one sliding window, whereas T1, T2, … , T9 in the figure represents a timepoint respectively. In this figure, sliding window sample data on every 9 timepoint, then displace another 9 timepoints to be next data sampling. We take three time intervals as the periodical length for sequential pattern while dividing the timepoint of sliding window in three equal lengths, in anticipation of discovering frequent sequential pattern through matching different basic windows. Consequently, in the process of mining a sliding window, we first take a basic window as transactions of one customer, while the data at the same timepoint will be treated as transactional data, as showed in Figure 3. In Figure 2, the three lattices, T1, T2, and T3 correspond with three timepoints, T1, T2, and T3 in Figure 3, while the center left S1, S2, and S3 respectively represents a different stream. Each timepoint T in the basic window is a set composed of input values from different streams, equal to the previous set we took as a transaction of one customer. In addition, the three timepoints in the basic window represent the time ordering. With these orderings joined with data at the timepoint, we regarded the basic window as transactions of one customer. For example, in Figure 3, the sequence data in the basic window represent as<(11,33,1) (55,86,81) (7,8,22)>. Thereupon through matching the sequence data of different basic windows in the same sliding window, frequent sequence patterns will be searched. We will set the minimum support = 100%, and therefore the frequent sequence pattern discovered in this sliding window will show <(11,33) (22)>.

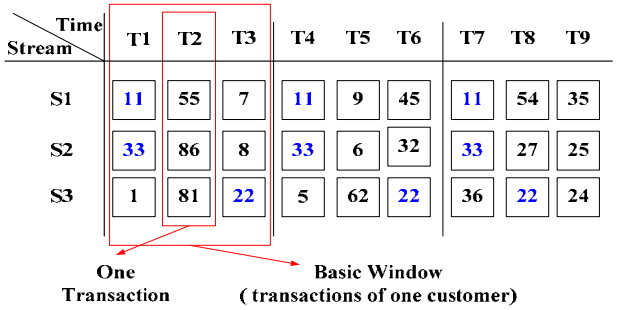


Figure 3. A Basic Window Model.

B. Incremental Mining

In addition to sampling through sliding window models, mining accurate sequential pattern in multiple data streams environments will tackle the stream data which cannot be preserved permanently, leading to a loss of accuracy. As such, the concept of an incremental mining is introduced. We could take the previously sampled data as the previous database (represented as SW), and on that occasion of a new entry from a sliding window data sampling (represented as SW'), SW' will be processed with sequential pattern mining in an incremental approach in order to be integrated with the result of SW. Nevertheless, due to the massive data for multiple data streams and to avoid excessive branches and nodes generated by the lexicographical sequence tree in the process of sequential pattern mining, which could ultimately lead to problems in memory space consumption and time for searching the lexicographical sequence tree. Also, we have integrated concept of the mining closed sequence pattern, and concurrently using a hash table to index closed sequential patterns and to enhance efficiency for establishment and maintaining of the lexicographical sequence tree.

Definition 1 (The support of sequential pattern): Have α be a sequential pattern. The support of α is equal to the transactions of including α divided by all of transactions.

Definition 2 (Frequent sequential pattern): Have α be a sequential pattern. α is a frequent sequential pattern if and merely if the support of α is larger or equal to minimum support.

Definition 3 (Semi-frequent sequential pattern): Have α be a sequential pattern. α is semi-frequent sequential pattern if and merely if the support of α is smaller then minimum support and larger or equal to buffer ratio r*minimum support.

Definition 4 (Frequent closed sequential pattern): Have α be a sequential pattern. α is a frequent closed sequential pattern if and merely if α is a frequent sequential pattern

and there exists no proper super sequential pattern which has the same support as α.

Definition 5 (Semi-frequent closed sequential pattern): Have α be a sequential pattern. α is a semi-frequent closed sequential pattern if and merely if α is a semi-frequent sequential pattern and there exists no proper super sequential pattern which has the same support as α.

In the incrementally update sequential patterns, processing will occur in new entry data, influencing the original sequential pattern mining result, and possibly leading to the following situations:

1. Frequent closed sequential pattern remains the same.
2. Frequent closed sequential pattern becomes a semi-frequent closed sequential pattern.
3. Semi-frequent closed sequential pattern remains the same.
4. Semi-frequent closed sequential pattern becomes a frequent closed sequential pattern.
5. Frequent closed sequential pattern becomes a non-frequent closed sequential pattern.
6. Semi-frequent closed sequential pattern becomes a non-frequent closed sequential pattern.
7. New data bring on new item.
8. Non-frequent closed sequential pattern becomes a frequent closed sequential pattern.
9. Non-frequent closed sequential pattern becomes a semi-frequent closed sequential pattern.

From Situation 1 to 4, due to our preservation of sequential pattern information from the lexicographical sequence tree, we could modify the preserved information directly for any update in new data. In Situation 5 and 6, the support of sequential pattern is less than the support of semi-frequent sequential pattern, will delete the sequential pattern information from the lexicographical sequence tree. In Situation 7 where new data bringing new entry, in other words, the sequence data non-exist in the lexicographical sequence tree, consequently will find out the new data when scanned. In Situation 8, we notice that for a non-frequent closed sequential pattern to become a frequent closed sequential pattern, all of the previously existing sub-sequential patterns of this sequential pattern must be a frequent sequential pattern and without any ultra sequential pattern. Simply put, if one of the sub-sequential pattern is non-frequent in the lexicographical sequence tree then this sequential pattern does exist. Situation 9 is similar to Situation 8 whereas for a non-frequent closed sequential pattern to become a semi-frequent closed sequential pattern, all of the previously existing sub-sequential patterns in the lexicographical sequence tree must be frequent or semi-frequent closed sequential patterns without containing any super sequence. Otherwise, the sequence will not be accounted for a semi-frequent closed sequential pattern.

Figure 4 shows the parameter definitions for the incremental closed sequential pattern mining algorithm (as known as ICspan algorithm) as in the following:

- SW' is the data for the current sliding window, whereas SW is the data of the previous sliding window.
- r is the buffer ratio(0 r 1), used for determining the number of semi-frequent closed sequential patterns.
- min_sup is the minimum support.
- F is the frequent closed sequential pattern in SW, whereas F' is the newly found frequent closed sequential pattern.
- SF is the semi-frequent closed sequential pattern found in SW, whereas SF' is the newly discovered semi-frequent closed sequential pattern.
- $sup_{sw}(p)$ is the support for sequential pattern p in SW.
- $sup_{sw'}(p)$ is the support for sequential pattern p in SW'.
- Sup(p) is the support for sequential pattern p.
- $sup_{append}(p)$ is the support for sequential pattern p in the sequential set of all appended items or itemsets.

Algorithm: ICspan

Input: SW', SW, min_sup, r, F, SF

Output: F and SF

1. F' ←ψ; SF' ←ψ;
2. Scan SW' to find 1-item sequential patterns and add into F' or SF';
3. **For each** 1-item sequential pattern in F' or SF' **do**
4. transform into closed sequential patterns and add into F' or SF';
5. **For each** sequential pattern p in F or SF **do**
6. **If** $sup_{sw}(p) + sup_{sw'}(p)$ min_sup **then**
7. **If** p is a closed sequential pattern **then** add p into F';
8. **If** $sup_{append}(p)$ (1-r)*min_sup **then**
9. transform into closed sequential patterns and add into F' or SF';
10. **If** $(sup_{sw}(p) + sup_{sw'}(p) \leqq$ min_sup) **and** $(sup_{sw}(p) + sup_{sw'}(p) \geqq$ r*min_sup) **then**
11. **If** p is a closed sequential pattern **then** add p into SF';
12. F←F'; SF←SF';
13. **return;**

Figure 4. Incremental closed sequential pattern mining algorithm.

The steps for ICspan algorithm are divided into two sections. The first section is to identify a new sequential pattern (line 1-4). This section is used to find the closed sequential pattern from the new data and therefore the sequential pattern with length = 1 will be projected. The second section is an integration of the new and old sequential patterns (line 5-13). This section projects on the new patterns found in Section 1 with the previously recorded patterns in order to find the combined sequential patterns derived from the new and old items. The following are the detailed steps to ICspan algorithm:

Step 1 (line 1): Set F' and SF' as empty sets.

Step 2 (line 2): Scan the data sampling from the sliding window, find a sequential pattern with length = 1 and support larger or equal to min_sup or r*min_sup. Add the result to F' or SF' respectively.

Step 3 (line 3-4): Project the all of the new sequential patterns in F' or SF' with length = 1, then transform into the closed sequential patterns.

Step 4 (line 5-11): Previously recorded frequent or semi-frequent sequential patterns are updated and projected then transformed into closed sequential patterns.

Step 4.1 (line 6): Determine if the support for sequential pattern p in SW added with support in SW' is larger or equal to the minimum support. If the support is determined to true then it becomes a frequent sequential pattern and executes steps in 4.2 and 4.3.

Step 4.2 (line 7): Determine if sequential pattern p includes other sequential patterns or is included under other sequential pattern. Update to assure recorded sequential patterns are frequent closed sequential patterns.

Step 4.3 (line 8-9): Determine if the support for sequential pattern p in all appended items or item sets in the sequential pattern set is larger or equal to (1-r)*min_sup. If the support is determined to be true, then it is possible that a non-frequent sequential pattern has becomes a frequent sequential pattern, which requires projection in order to generate closed sequential patterns.

Step 4.4 (line 10-11): Determine if the support for sequential pattern p in SW added with support in SW' is smaller minimum support and larger or equal to the r*min_sup. If the support is determined to true then this sequential pattern p is a semi-frequent sequential pattern and requires update to assure the recorded sequential patterns are all semi-frequent closed sequential patterns.

Step 5 (line 12): Store F' back to F; store SF' back to SF.

Step 6 (line 13): Output F and SF.

The following example will be used to explain the ICspan algorithm. First, we will set min_sup = 100% and r = 0.5. Figure 5 will show data sampling from the first execution of ICspan. Step 1 will set the two empty sets, F' and SF', to store the newly found sequential pattern. Step 2 searches for new data sampling from the sliding window which are frequent or semi-frequent sequential patterns with length = 1. The frequent sequential patterns found in the example of figure 5 are <A>,<a> and <3>，the semi-frequent sequential patterns found in the example are <f>,<d> and <6>, then add each of the found sequential patterns to F' or SF'. Step 3 will project the all of new sequential patterns in F' or SF' with length = 1 and transform into frequent closed sequential patterns <(A, a)(3)> and semi-frequent closed sequential patterns <(6)(3)>, <(A, a)(F)>, and<(A, a)(d, 3)>. Due to lack of previously mined sequential patterns in Step 4, the step will go straight to Step 5 to update F and SF. Lastly, Step 6 output F and SF. Table I shows the result of the first ICspan execution.

Time / Stream	T1	T2	T3	T4	T5	T6	T7	T8	T9
S1	A	C	F	A	E	B	A	F	D
S2	a	b	d	a	c	f	a	e	d
S3	1	2	3	6	3	4	5	6	3

Figure 5. Data sampling of the first ICspan execution.

TABLE I.

RESULT OF FIRST ICSPAN EXECUTION.

Frequent closed sequential pattern F	<(A, a)(3)>
Semi-frequent closed sequential pattern SF	<(6)(3)>、<(A, a)(F)>、<(A, a)(d, 3)>

Time / Stream	T10	T11	T12	T13	T14	T15	T16	T17	T18
S1	C	A	E	B	D	B	A	F	D
S2	f	e	b	a	d	f	c	e	a
S3	4	3	6	1	3	2	3	6	5

Figure 6. Data sampling of second ICspan execution.

Figure 6 shows data sampling of the second ICspan execution. Step 1-2 will find frequent sequential patterns <3> and semi-frequent sequential patterns <A>, <6>, <a>, <D>, <f> and <e> with length = 1 from figure 6, and then add <3> to F' while adding the remaining sequential patterns to SF'. Step 3 will project the all of new sequential patterns in F' or SF' with length = 1 and transform into semi-frequent closed sequential patterns <(A, 3)(6)>. Table II shows the results for F' and SF' to Step 3. Step 4 combines new and old sequential patterns. First, Step 4.1 determines the support for sequential patterns in Table I added with support for the frequent or semi-frequent sequential patterns from Table II is larger or equal to min_sup. If the condition is in accordance to Step 4.2 will proceed projecting and transformed into frequent closed sequential patterns <3>. Furthermore, since the support of these sequential patterns for all appended sequential sets in Step 4.3 is smaller then (1-r)*min_sup, the projection will not take place to transform the new frequent or semi-frequent closed sequential patterns. Table III shows the result for F' and SF' to Step 4.3. Next, Step 4.4 will be determined if the support for sequential pattern p in SW added with support in SW' is smaller minimum support and larger or equal to the r*min_sup, and this step will add <A>、<a>、<6>、<(A, a)(3)>to SF'. Step 5 will store the last result of F' and SF' back to F and SF, then output F and SF through Step 6. Table IV shows the final output result.

TABLE II.

RESULTS F' AND SF' FOR SECOND EXECUTION TO STEP 3.

Frequent closed sequential pattern F'	< 3>
Semi-frequent closed sequential pattern SF'	<a>、<D>、<f>、<e>、<(A, 3)(6)>

TABLE III.

RESULTS F' AND SF' FOR SECOND EXECUTION TO STEP 4.3.

Frequent closed sequential pattern F'	<3>
Semi-frequent closed sequential pattern SF'	<a>、<D>、<f>、<e>、<(A, 3)(6)>

TABLE IV.

FINAL OUTPUT RESULTS.

Frequent closed sequential pattern F	<3>
Semi-frequent closed sequential pattern SF	<A>、<a>、<6>、<(A, a)(3)>

IV. PERFORMANCE EVALUATION

A. Experimental Environment and Data

The program has been written by C++ STL in the Visual Studio 2005 compiling environment. All experiments have been conducted in the experimental environment described in Table V. We used the data generator from SQL Server 2005 to generate synthetic datasheets a Uniform distribution and Gaussian distribution. Table VI describes the generated parameters of the synthetic data. For example, S9T200V4 is expressed by 9 stream numbers and 200 data at a timepoint, with the value range for each data stream as much as 4 times more from the multiple data streams environment.

TABLE V.

EXPERIMENTAL ENVIRONMENT.

Item	Specification
Processor	Pentium4 3.00GHz
Memory	1 GB
Hard Drive	80 GB
O.S	WINDOWS XP
Programming Language	C++ STL

TABLE VI.

GENERATED PARAMETERS FROM SYNTHETIC DATA.

Parameter	Parameter Description
S	Number of Data Stream
T	Number of Timepoint
V	Value of Stream Data

B. Comparison Between ICspan and MILE

In multiple data stream environments, MILE algorithm emphasizes mining sequential patterns from periodical data through recoding generated sequential patterns and using a hash method to shorten the time to mine sequential patterns. Therefore the process does not involve an incremental mining. In addition, due to the massive amount of generated data from multiple data streams, the recorded sequential patterns are also relatively increasing as a result. On the contrary, ICspan algorithm not only conducts incremental mining but also reduces the number of recorded sequential patterns.

(1) Accuracy

From the mining results of sequential patterns in an identical data environment, if the new sequential pattern β includes the previous sequential pattern α [9], and then it means β has a longer and more accurate length than that of α. Through comparing the derivations of mining sequential pattern set from ICspan and MILE, we will observe that all the sequential pattern sets of MILE corresponding to those derived from ICspan. For this reason, we use the following formula to calculate their accuracy:

$$\sum_{i=1}^{N} p_i \quad (1)$$

$$\sum_{i=1}^{N} p_i' \quad (2)$$

N is the corresponding sequential pattern number for both algorithms in individual mining result, whereas p_i $(1 \le i \le N)$ represents the sequential pattern length of each MILE mining, and p_i' $(1 \le i \le N)$ represents the corresponding sequential pattern length for the sequential pattern set derived from each ICspan algorithm. By adding the total length of sequential patterns, the higher the value means better accuracy.

Figure 7 shows the comparison of accuracy for ICspan and MILE. The vertical axis from the figure indicates the total amount of the sequential pattern length whereas the horizontal axis indicates the minimum support is represented by percentage. The MILE line represents the total amount of sequential pattern lengths found by MILE, whereas ICspan line represents the sequential pattern set derived from ICspan with the sequential pattern lengths corresponding to MILE. Figure 7 shows under various minimum supports, the total amount of sequential pattern lengths found by ICspan is larger than that of MILE. Since ICspan integrates the results of the new and old sequential pattern mining by using a incremental method, the mining result of ICspan will be more accurate.

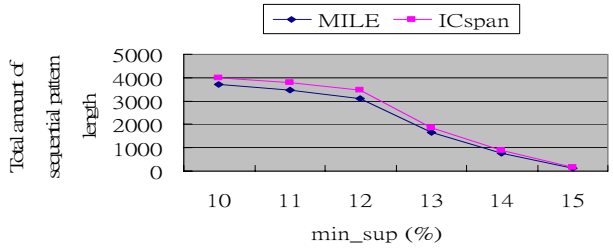


Figure 7. Comparison of Accuracy.

(2) Number of Sequential Patterns

Figure 8 shows the comparison of the number of generated sequential patterns for MILE and ICspan in S9T2000V4 environment. The min_sup from the figure indicates the minimum support represented by percentage, while sequential patterns represent the number of generated sequential patterns. In order to compare with MILE, the buffering parameter r is set to 1 for ICspan algorithm. Under various minimum supports, the recorded sequential pattern amount for ICspan is much less than that of MILE. The reason for the difference lies on ICspan looking for closed sequential patterns only, leading to the reduction of many undesired sequential patterns.

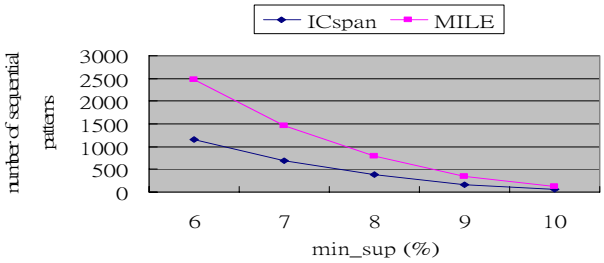


Figure 8. Comparison of the Number of Sequential Patterns.

C. Analysis on ICspan Algorithm

The experiment from the second part emphasizes on the ICsapn algorithm in different environments and parameter testing by applying different conditions such as data sizes, basic window lengths and data distribution to interpret the executing efficacy of the algorithm.

(1) Data Size

Figure 9 and 10 show the three different multiple data streams environments in S9T20000V4, S9T2000V4 and S9T200V4, compared with the execution time and memory usage for ICspan. The vertical axis in Figure 9 is the executing time in units of seconds, whereas the vertical axis in Figure 13 is the maximum usage for memory in units of kilobytes (KB). As we observe from Figure 9 and 103, when the minimum support for ICspan is lower than 7%, the execution efficacy is worse. Nonetheless, the execution efficacy to maintain minimum support is satisfying. The reason could result from the number of sequential patterns meeting the conditions, which is excessively enormous, extending the time to transform closed sequential patterns.

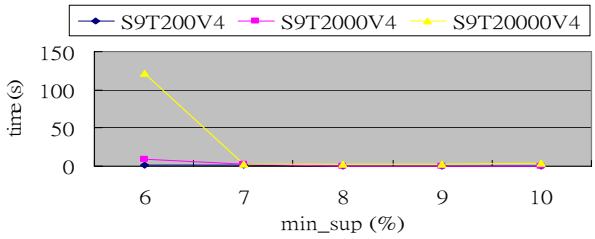


Figure 9. Execution Time for Various Data Size.

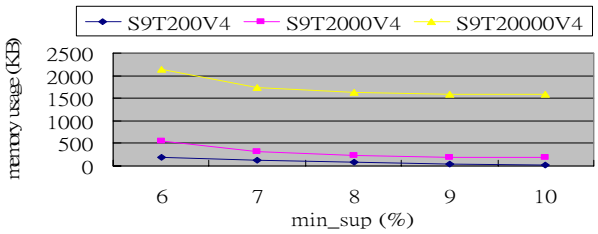


Figure 10. Memory Usage for Various Data Size.

(2)Basic window length

Figure 11 shows the comparison of the ICspan execution time for various base window lengths in the multiple data streams environment in S9T2000V4. The line BW = 3 indicates the basic window length is equal to 3, whereas the same indication applies to the other lines. Figure 11 indicates the execution time for BW = 3 is smaller than that of BW = 5 and for BW = 7, whereas BW = 5 has execution time slightly higher than BW = 7 at the minimum support of 8%. As the base window length increases for the former, the sequential patterns also increase the mining time; while due to the change of the basic window length for the latter, shortening the time to transform from sequential data to closed sequential patterns.

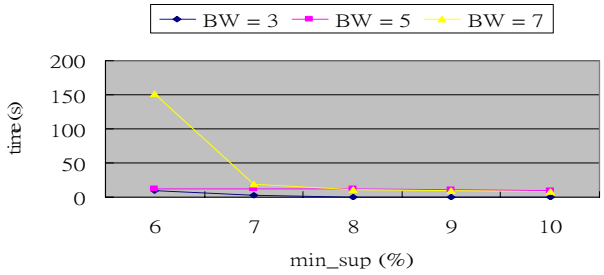


Figure 11. Basic Window Length.

(3) Data distribution

Figure 12 shows the comparison of the execution time in Gaussian distribution and Uniform distribution for executing ICspan, in the multiple data streams environment in S9T2000V4. The figure indicates that the execution time is acceptable when the minimum support is larger than or equal to 25%; while the execution time for Gaussian distribution will climb rapidly following a continuously declining minimum support, resulting from the variation in probability for Gaussian distribution.

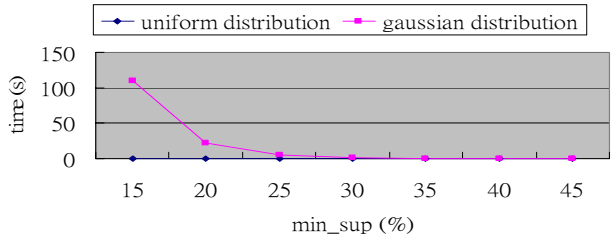


Figure 12. Comparison of Different Data Distribution.

V. CONCLUSION

Most studies favor the analysis on long-term data rather than short-term data because short-term data is easily interfered by instantaneous events. Long-term data have less possibility to be interfered, and instead it is easier to find out the periodicity of occurring events, allowing people to be readily interpreting the periodical changes to events. For this reason, in order to find out the sequential pattern mining resulting from long-term data, we have put forward this ICspan algorithm, applying an incremental mining method to conduct a direct sequential mining to the new data, and then to integrate with former mining results for generating new sequential patterns. In addition, we have implemented the method of closed sequential pattern mining to reduce the number of sequential pattern records. This concept will reduce a waste of memory space in increasing sequential patterns. Finally, the ICspan algorithm is collaborated with sliding window mode to obtain a comprehensive sampling of all historical data, resulting in more accurate sequential patterns for reference and decision-making to all users. The experiment results also support that the ICspan algorithm will effectively reduce the sequential pattern records and consequently reducing the memory usage, while maintaining a sound mining efficiency under continuous data entries.

ACKNOWLEDGMENT

The authors would like to express their appreciation for the financial support from the National Science Council of Republic of China under Project No. NSC 98-2221-E-031-003.

REFERENCES

[1] Chang, J.H. and Lee, W.S., "Efficient Mining Method for Retrieving Sequential Patterns over Online Data Streams," Journal of Information Science, Vol. 31, Issue 5, pp. 420-432 (2005).
[2] Ezeife, C.I. and Monwar, M., "SSM: A Frequent Sequential Data Stream Patterns Miner," Pro-ceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining, Hawaii, USA, pp. 120-126 (2007).
[3] Ho, C.C., Li, H.F., Kuo, F.F., and Lee, S.Y., "Incremental Mining of Sequential Patterns over a Stream Sliding Window," Proceedings of the 6th IEEE International Conference on Data Mining, Hong Kong, China, pp. 677-681 (2006).
[4] Raissi, C., Poncelet, P. and Teisseire, M., "SPEED: Mining Maximal Sequential Patterns over Data Streams," Proceedings of the 3rd In-ternational IEEE Conference Intelligent Systems, Varna, Bulgaria, pp.546-552 (2006).
[5] Oates, T. and Cohen, P.R., "Searching for Structure in Multiple Streams of Data," Proceedings of the 13th International Conference on Machine Learning, Bari, Italy, pp. 346-354 (1996).
[6] Chen, G., Wu, X., and Zhu, X., "Sequential Pattern Mining in Multiple Streams," Pro-ceedings of the 5th IEEE International Conference on Data Mining, Houston, USA, pp. 27-30 (2005).
[7] Das, G., Lin, K.-I., Mannila, H., Renganathan, G., and Smyth, P., "Rule Discovery from Time Series," Proceedings of the 4th International Conference of Knowledge Discovery and Data Mining, New York, USA, pp. 16-22 (1998).
[8] Chang, J. and Lee, W., "Decaying Obsolete Information in Finding Recent Frequent Itemsets over Data Stream," IEICE Transaction on Information and Systems, Vol. 87, No. 6, pp. 1588-1592 (2004).
[9] Yang, J., Wang, W., Yu, PS., and Han, J., "Mining Long Sequential Patterns in a Noisy Environment," Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data, New York, USA, pp. 406-417 (2002).

Shih-Yang Yang received his Ph.D. degree in Computer Science and Information Engineering from Tamkang University, Taiwan, in January 2008. Since January 2008, he is an Associate Professor with the Department of Media Art at Kang-Ning Junior College of Medical Care and Management (Taipei, Taiwan). His research interests include parallel & distributed systems, web technology, and multimedia.

Ching-Ming Chao received his Ph.D. degree in Computer Science from The University of Iowa, Iowa City, Iowa, U.S.A. in 1990. From 1990 to 1992, he was an assistant professor in the Department of Computing Sciences at the University of Scranton, Scranton, Pennsylvania, U.S.A. He joined the faculty of the Department of Computer and Information Science at Soochow University in 1992 as an associate professor. From 1992 to 1996, he served as the department chair. Since 2003, he has been a professor in Department of Computer Science and Information Management at Soochow University. His research interests include data mining, data warehousing, database, and web technology.

Po-Zung Chen received his Ph.D. degree in Computer Science from the University of Iowa in December 1989. From November 1989 to May 1990, he was a visiting Assistant Professor at Michigan Technological University (Houghton, Michigan). Since August 1990, he is an Associate Professor with the Department of Computer Science and Information Engineering at Tamkang University (Taipei, Taiwan). His research interests include object-oriented distributed programming, parallel & distributed systems, and simulation & modeling.

Chu-Hao Sun is currently a candidate of Ph.D. student in department of Computer Science and Information Engineering in Tamkang University (Taipei, Taiwan). He received his B.E. and M.E. degree from the same university in 1995 and 1998. His research interests include database management system, parallel processing, web technology, and data mining.

Developing a Web-based and Competition-based Quiz Game Environment to Improve Student Motivation

Kuan-Cheng Lin*, Ting-Kuan Wu, Yu-Bin Wang
Department of Information Management, National Chung Hsing University,
250, Kuo Kuang Rd., Taichung 402, Taiwan
*kclin@nchu.edu.tw

***Abstract*--Game-Based learning (DGBL) has become an attractive pedagogy in recent years because it can effectively improve the intrinsic motivation of learning. Besides, the development of web-based e-learning systems can help students to carry mastery learning out. In this study, we develop a web-based and competition-based multi-player quiz game environment (WCMQGE) by incorporating DGBL into the web-based e-learning system. During the process of the game, the WCMQGE system can help players (students) review what they learned from school and use their knowledge to compete with their peers. Moreover, the WCMQGE provide three item generation algorithms to increase the randomness of quiz items and enhance the interest of students. Furthermore, the WCMQGE system also provides the learning management module for the students and teachers to review the learning outcome of the students.**

Index Term: **e-learning, digital game-based learning, mastery learning**

I. INTRODUCTION

A. Background

Computer games are a growing part of our society, the value of game industry which includes the selling of games and other related media such us magazines and internet communities, has increased continuously. According to research, game industry is worth multi-billion dollars and its revenue and influence on related industries are not less than those of Hollywood [1]. It shows that more and more people are attracted by the games' characteristics: fantasy, challenge, and curiosity, and more and more people choose video games as their entertainment so that the impact of games becomes significant. Now there is increasing interest in researching why games attract so many people and what powerful supports can games provide to learning activities.

On the other hand, because of the development of internet technology and portable devices, people are able to access more types of media and highly-functioned web applications without time limitation and place restriction. In this context, e-learning is more and more an efficient way for leaning because what one of the main purposes of e-leaning is to give learners opportunities to learn anytime and anywhere. To explain it in detail, a digital learning information system based on web technology has several characteristics which help learning [2]:

(1) Learning anytime and anywhere without installing certain client.

(2) Large and easy-to-link community.

(3) Promoting visualization and operation, and linking visual and symbolic representation: computer software can visualize abstract symbols and concepts which are hard to be presented in real life.

(4) Providing an immersive learning environment: multimedia learning materials provide interactive learning and simulated leaning environments.

Basis on the above characteristics, researchers believe combining web-based digital games with learning processes is an effective way to enhance learners' motivation thus improve outcome.

Here are some cases of digital game-based learning. In Chin-Tau Wang's Joyce, the system provided a board game-style leaning environment. To win, players must reach their goal sooner than their opponents, so players would eager to enrich their skills and knowledge in order to quickly respond to given questions [3].

On the other hand, M. Minović, M. Milovanović, D. Starcevic, and M. Jovanović designed an adventure game-style learning system, this system helped educators use leaning materials construct a fictional world, in which players had to use their skills to complete phase specific target. Every time a goal was completed, players would be rewarded, so this system helped players learn by providing simulated learning environments and giving them motivation [4].

Apart from academic studies, now there are many practical applications of game-based learning systems, such as the U.S. Army (Americans Army: Operations, http://www. Americasarmy.com) in which players learn how to accomplish their task as an American Army in the virtual battlefields created by the game, and projects sponsored by other organizations [5].

B. Research motivation

Motivation is one of the most important factors in a learning process, and it directly affects the effectiveness of learning. In order to enhance positive emotions of students [6], a common way is to combine game elements in the learning environments. An example is that

doi:10.4304/jnw.6.5.736-742

Zong-Bin Guo promoted the development of moral cognition by making students participate in various activities [7].

Generally speaking, repeating review and practice account for a large proportion in a learning process, and this is usually done by writing assessment. After all, not all of the skills and knowledge can be instructed by implementing, especially the basic knowledge of every academic subjects. For example, reciting English vocabularies of English learning, and arithmetic of mathematic. The problem is, the process of this type of review is often monotonous, and it difficultly put learners into independent thinking since learners do not have to gain basic knowledge by observation and explanation. Therefore, learners would be easily distracted and the efficiency would be reduced.

When the learning process is routine and boring, a game-like learning environment can effectively make this activity enjoyable because a game brings learners sense of curiosity, challenging, and interactivity by providing unpredictable outcome, feedback, and appropriate complexity [8]. It's believed that if using games to improve this fact, that is, learners are interested in this reviewing process when they can feel an optimal level of uncertainty and interact with others by competition and cooperation during the writing assessment, learners' outcome will be improved.

In this paper, we design a digital game-based learning system and learners are able to compete with others in quiz games. Furthermore, to increase playfulness, we design several types of questions and let the system to use them randomly. On the other hand, the system will collect games' data and analyze learners' learning history which allows learners to view their learning outcomes.

II. Literature Review

In this paper, group-competing game is the primary method for interaction between learners. Therefore, we'll concentrate on "competitive learning", "cooperative learning", and "game-based learning".

A. Competitive learning

D. W. Johnson and R. T. Johnson proposed the definition of competition [9]:

(1) The only purpose of competition is to defeat opponents to victory, regardless of the shape of "victory".

(2) In a competition, only a few people can win, and be rewarded.

For instance, a school only provides a few numbers of scholarships to the best performance in-school students, intending to create a competing environment to give students motivation to study hard.

Whether competitive learning really increases the learning motivation, and then promoting learning, some researchers hold positive views. Stipek considered that under the premise of that the learning objectives are clear, simple, and not time-consumed to reach, compared to cooperative learning, competition learning resulted in significantly better outcomes, and the whole learning process took a shorter time [10].

However, some researchers hold a negative attitude. Hamachek found that stress from competition may reduce the trust between peers, affecting the cohesion and creativity of learning groups [11]. On the other hand, Aimeur and Frasson proposed that competition can improve the motivation of capable learners, but not applicable to incapable learners [12].

In order to moderate the side-effects of competition, D.W. Johnson and R.T. Johnson put forward several proposals[13]:

(1) Emphasize the fun of learning is more important than winning or losing.

(2) Use competition between groups rather than between individuals.

(3) Understand the relationship between competition and cooperation, and maximize the pros and minimize the cons of competition by cooperation.

(4) During the competition, instructor should pay attention to three levels of (instructor and learners, learners and learners, learners and learning materials) interaction.

(5) Maximize the winners.

(6) Respect others' learning condition.

(7) View the importance of quantity and quality equally when evaluating students' achievement.

B. Cooperative learning

There are often a few winners in a competing learning environment, and this fact will result in significant backward position for learners who seldom win the competition. To buffer the impact of competition, group cooperation is a way in which learners focus on helping their teammates win instead of defeating their opponents, attempting to create a more moderate competitive conditions, but with the learning motivation the competition brings. A point the instructor should know is that all teams' capabilities should not differ much to avoid excessive influencing the essence of "victory" [14].

The fact is that cooperation accounts for the most important role in the learning process. In the cooperative learning process, learners exchange information, generate new ideas to simplify the problem, and solve the problem. In addition, instructor becomes a learning mentor partner who assists learners to construct their own knowledge, rather than passive convey knowledge to learners [15].

The effect of cooperative learning has been well recognized, Attle and Baker pointed out that cooperative learning helps learners generate new ideas, and competition will be objective and motivation for learners [16]. If combining the heterogeneous learning methods mentioned above, using group competitive learning, it often results in achievement beyond expectations. Bulut said during to process of cooperative learning, team members support and communicate with each other to enhance the trust between team members, this help them develop good relationships [17].

However, cooperative learning also possesses risks, Ballantine and Larres proposed that, if a group lack of task management, time management, communication and coordination, it will result in low efficiency of group

learning, worse than studying individually [18]; Hong-Jia, Zhang tried to use cooperative game to improve students' motivation in math class and he found that the competency members actively participated in group learning while the incompetency members seemed not engaged in leaning activity. Therefore, instructor should always pay attention to learning conditions in each group, and give the necessary coordination [19].

C. Game-based learning

It's an effective way to improve learners' motivation by combining gameplay in the learning process. A. Martens, H. Diener, and S. Malo pointed out that the game creates a virtual environment, allowing the players to play a completely different role from real life and into the game as part of the game. In addition, the game gives every players a common task (usually simple and can be reached in short time), and the players will be rewarded when they finish their tasks [20]. In order to achieve the tasks, the players must cooperate or compete with other players. And this novel experience, community participation, and incentive of reward, are the source of game motivation.

From the point of view of combination of learning, a good game-based learning system must maintain the users a high participation rate of the game and learning. To achieve this, Prensky and Marc proposed several principles of designing game-based learning [21]:

(1) The game itself provides enough contents so users will not feel bored playing it.

(2) The system makes users think themselves as the "players" rather than "students" or "trainees" when they use the system.

(3) The users want to play with game until they complete their objectives, and continuing pursuing their next goal.

(4) The users can feel their knowledge have grown because of investment of their time in games.

(5) The users can increase their chance of winning by the experience and the knowledge gain from the game.

According to Kiili's study, in the process of learning, to reach their objectives, learners will be in a very involved, focused state, called "Flow". In this state, the learners can very efficiently gain their knowledge and enhance their skills from the learning experience [15].

To promote Flow, two conditions must be satisfied. First, there must be challenges, and learners can use their skills to overcome this challenge, if the objectives are too difficult or too simple to achieve, learners will not be in the state of flow. Therefore, a successful game-based learning system must take the users' prior knowledge and skills into account, and give them appropriate challenges to help users construct new knowledge and skills based on their prior knowledge and skills, shown as fig. 1. However, although a well-designed game actually makes learner feel enthusiastic and improves learner's outcome, the game's effectiveness varies depends on how many time the objectives cost, learning environment, and the age of the learners [22]. A simple game is fit for short-term and single study, and a sophisticated game is suit for long-term and cooperative study.

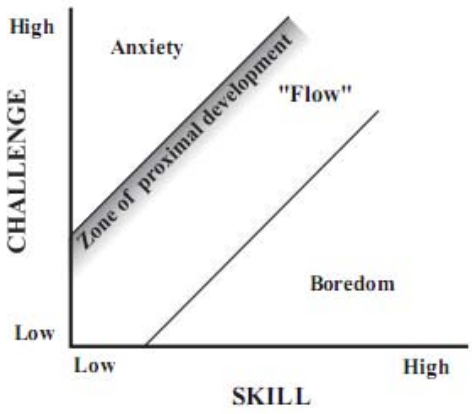


Figure 1. Three channel model of flow [15].

III. DEVELOPMENT OF THE WEB-BASED AND COMPETITION-BASED MULTI-PLAYERS QUIZ GAME ENVIRONMENT

In this section, we describe the development of the web-based and competition-based multi-players quiz game environment. The WCMQGE system includes: quiz game module, item generation module and learning management module. In addition, the design of the WCMQGE system has three main considerations as follows. First, we design various types of quiz used in the games to increase the randomness of items and enhance the interest of players. Second, we set staged, short-term goals for the players to give them simple objects which are expectedly achievable. Third, the system will reward players with special titles and enhancement of capacity of their avatar if they complete each staged goal. On the other hand, the system allows players to use anonymous in games to reduce the negative impacts of competition. Moreover, the WCMQGE system will adjust the parameters of quiz game, such as difficulty of quizzes, based on players' learning situation.

Fig. 2 shows a screenshot of the main page of WCMQGE. On the left side are the function list and the status of the user and on the ride side is the course list. The user needs to login first and enter a course in order to use the functions. After the user does that, he or she can see his or her name and what course he or she has entered on the left-bellow side.

Figure 2. Main page

A. *Quiz game module*

The main method of game is to compete in the quiz tests. In a game, every players receive several different questions simultaneously, then one of players chooses one question to "attack" another at the other side (here it means to make opponents do the same questions which attacker chosen) until one of the groups wins.

According to the principle of encouraging players to actively participate in game and balance, we generally put attackers in a more predominant position. On the other hand, to avoid some occasional scenarios that may influence the balance among players, the injuries that the players get in the game process would be calculated considering several conditions:

(1) If the player chooses the wrong answer of the question, he or she will get injured and the value of injury is based on the difficulty of the question. The more difficult the question is, the fewer injuries the player gets.

(2) There are more skills can be used in attacking, meaning that attackers have more strategies used to win the game.

(3) If the attacker chooses the right answer of a question when the defender does not, the defender will get more injuries when the question is difficult. However, the attacker will get more injuries if the attacker chooses the wrong answer when the defender chooses the right one.

In addition to above conditions, there are some limitations to make the game more balanced:

(1) The attacker cannot attack the same player continuously more than two times.

(2) Only a pair of players can compete by the same question at a time, others cannot intervene in the competition.

Figure 3. Quiz game module

Fig. 3 shows a screenshot of the process of quiz game. Section 1 shows the players' information such as nickname, HP, MP, and what they are doing. Section 2 places quizzes used to be the main tool to play the game, player can choose one quiz, answer it, and target one opponent to attack it, then the player who is attacked should do the same quiz. After both sides choose answer, the system will check both answers and distract the players' HP basis on what they answer and how difficult the quiz is.

Figure 4. Game Module (choosing skills)

Section 3 contains a textbox which shows some information of the game. In addition, the players can use skills, which are designed to improve the game's playfulness, before they answer their question. If the players successfully attack their opponents, they will be influenced by the skill chosen by the other side.

In fig. 4, section 1 shows the skill list in a game. The players can choose one skill to use before they answer the question.

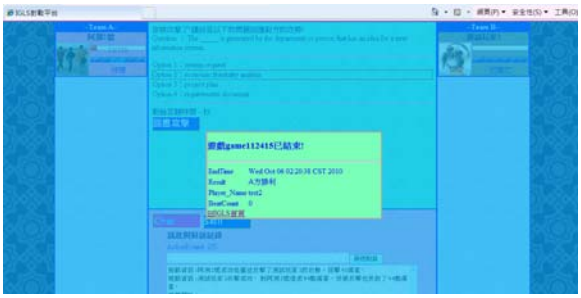


Figure 5. Quiz game module (ending)

Fig. 5 shows the screen shot of the game which is end. A panel on which shows the result of the competition and some information of this game.

B. *Item generation module*

The WCMQGE system provides three item generation algorithms to increase the randomness of quiz items and enhance the interest of students. The three algorithms are described in the following:

a. Multiple Choice With Random Options:

The system randomly chooses tree incorrect choices and one correct choice (there are more than three incorrect choices and one correct choice in database.) and permutes them by randomness.

Example 1:

Question Description	Which is a prime number?
Correct choices	2, 3, 5
Incorrect choices	4, 9, 15, 21, 27

When this question is used in a game, it may look like below:

Question: Which is a prime number?
Options: (A) 9 (B) 3 (C) 4 (D) 21

b. Match Items

The system randomly chooses a word to substitute the variable defined in the question, and the match description will be the correct answer.

Example 2:
Question Description: What's the definition of [v]?
Variables and matched descriptions:

When this question is used in a game, it may look like below:

Variable	Matched Description
Planck constant	$h = 6.626\ 069\ 3(11) \times 10^{-34}\ \mathrm{J \cdot s}$
Dirac constant	$\hbar \equiv \frac{h}{2\pi} = 1.054\ 571\ 68(18) \times 10^{-34}\ \mathrm{J \cdot s},$
gravitational constant	$F = G\frac{m_1 m_2}{r^2}.$
Avogadro constant	$N_A = (6.022\,141\,79 \pm 0.000\,000\,30) \times 10^{23}\ \mathrm{mol}^{-1}$

Question: What's the definition of Planck constant?
Option:

(A)	$h = 6.626\ 069\ 3(11) \times 10^{-34}\ \mathrm{J \cdot s}$
(B)	$\hbar \equiv \frac{h}{2\pi} = 1.054\ 571\ 68(18) \times 10^{-34}\ \mathrm{J \cdot s},$
(C)	$F = G\frac{m_1 m_2}{r^2}.$
(D)	$N_A = (6.022\,141\,79 \pm 0.000\,000\,30) \times 10^{23}\ \mathrm{mol}^{-1}$

c. Random Argument Computational Problem:

The system randomly generates arguments in a specific range to substitute the variables in the function defined with question, then compute the answer based on the function. When learners give the answer and the system will check whether the answer matches what it has computed.

Example 3:

Question	What's square measure of the triangle with the bottom [v1], height [v2]?
Function	[v1] * [v2] / 2
Arguments	v1[3, 10], v2[5, 15]

When this question is used in a game, it may look like below:

Question: What's square measure of the triangle with the bottom 3, height 14?
Option: (A) 10 (B) 24 (C) 75 (D) 21

Figure 6. Item generation module

Fig. 6 shows a screenshot of the item generation module. On the above side of the figure is the section where a question can be added or edited. On the bellow side of the figure is the section where the existed questions are listed and the user can choose one of the questions to edit. The format of the questions is described before, users should decide what type of question to add or edit and input the parameters needed.

C. Learning management module

Users can view their game records which are presented as correct answer rates, and the information provided change depends on users' role. As a learner, the user can view his or her progress by the correct answer rate of the questions the user has answered which are sorted by chapters in a certain curriculum. As an instructor, the user can view all the students' correct answer rate in a certain curriculum so that the instructor will know which student needs more assistance on his or her learning process.

Figure 7. Learning management module (Student)

Figure 8. Learning management module (Instructor)

Fig. 7 is what a student sees when reviewing records. In section 1, the system lists how many quizzes, which are sorted by chapter, the student has done, and section 2 shows the ratio of the correct answers to the total answers.

On the other hand, fig. 8 shows what an instructor sees when reviewing records. In section 1, the system lists the students who attend the class and their ratio of the correct answers to the total answers, and section 2 shows the same as in fig. 7.

IV. Conclusion and Future Research

In this study, we have developed a web-based and competition-based multi-player quiz game environment (WCMQGE) by incorporating DGBL into the web-based e-learning system. The WCMQGE system includes three main modules: quiz game module, item generation module and learning management module. During the process of the game, the WCMQGE system can help players (students) review what they learned from school and use their knowledge to compete with their peers. Moreover, the WCMQGE provide three item generation algorithms: Multiple Choice with Random Options, Match Items and Random Argument Computational Problem to increase the randomness of quiz items and enhance the interest of students. Furthermore, the WCMQGE system also provides the learning management module for the students and teachers to review the learning outcome of the students.

In the future, we will apply the WCMQGE system to the practical teaching and learning. The pedagogical experiments will be designed to assess the improvement of student motivation and learning effectiveness.

References

[1] K. Squire, "Game-based learning: Present and future state of the field," Doctor, ADL Co-Lab, University of Wisconsin-Madison, Madison, 2005.

[2] J.-r. Sohng, "The Application of the Problem-based Learning Strategy in Game-based Learning—A Case Study with Black-Faced Spoonbill Ecology Simulation Game," Master, National University of Tainan, Tainan, 2007.

[3] C.-T. Wang, "Exploring the Design of Computer Supported Question-and-Answering Competitions Exercise Game," Master, Computer Science and Information Engineering, National Central University, Taoyuan, 2001.

[4] M. Minović, et al., "Knowledge Modeling for Educational Games," Lecture Notes in Computer Science, vol. 5736, pp. 156-165, 2009.

[5] D. W. Shaffer, et al., "Video games and the future of learning " University of Wisconsin-Madison and Academic Advanced Distributed Learning Co-Laboratory2004.

[6] A. M. O'Donnell, et al., Educational Psychology: Reflection for Action, 1 ed. Taipei: Yeh Yeh, 2009.

[7] Z.-B. Guo, "The Research of Elementary School Students' Development of Moral Cognition in Learning Game," Doctor, Department of Civic Education and Leadership, National Taiwan Normal University, Taipei, 2010.

[8] M. Ebnera and A. Holzinger, "Successful implementation of user-centered game based learning in higher education: An example from civil engineering.," Computers & Education, vol. 49, pp. 873-890, 2007.

[9] R. T. Johnson and D. W. Johnson, Learning together and alone: Cooperation, Competition and Individualization, 4th ed. Massachusetts: Allyn & Bacon, 1994.

[10] D. J. Stipek, Motivation to Learn: From Theory to Practice, 2 ed. Massachusetts: Allyn & Bacon, 1992.

[11] D. E. Hamachek, Psychology in Teaching, Learning, and Growth. Massachusetts: Allyn & Bacon, 1985.

[12] E. Aimeur and C. Frasson, "Analyzing a new learning strategy according to different knowledge levels," Computers & Education, vol. 27, pp. 115-127, 1996.

[13] S.-W. Jhuang, "Design of cooperative and competitive online math game," Master, National Taipei University of Education, Taipei, 2003.

[14] M. Prensky, "Digital Game-Based Learning," Computers in Entertainment, vol. 1, pp. 21-21, 2003.

[15] K. Kiili, "Digital game-based learning: Towards an experiential gaming model," The Internet and Higher Education, vol. 8, pp. 13-24, 2005.

[16] S. Attle and B. Baker, "Cooperative Learning in a Competitive Environment: Classroom Applications," International Journal of Teaching and Learning in Higher Education, vol. 19, pp. 77-83, 2007.

[17] S. Bulut, "A cross-cultural study on the usage of cooperative learning techniques in graduate level education in five different countries," Revista Latinoamericana de Psicología, vol. 42, pp. 111-118, 2010.

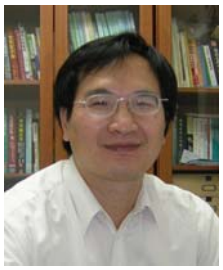
[18] J. Ballantine and P. M. Larres, "Cooperative learning: a pedagogy to improve students' generic skills?," Education + Training, vol. 49, pp. 126-137, 2007.

[19] Z. Hong-Jia. Using Competitive and Cooperative Learning in the Teaching of Elementary Mathematics. Available: http://163.20.178.5/groth/92/math.doc

[20] A. Martens, et al., "Game-based learning with computers: learning, simulations, and games," in Transactions on edutainment I, ed: Springer-Verlag, 2008, pp. 172-190.

[21] M. Prensky, Digital Game-Based Learning. New York: McGraw-Hill, 2004.

[22] M. Papastergiou, "Digital Game-Based Learning in high school Computer Science education: Impact on educational effectiveness and student motivation," Computers & Education, vol. 52, pp. 1-12, 2009.



Kuan-Cheng Lin was born in Taiwan on September 13, 1964. He received a BS in chemistry from National Taiwan University in 1988 and a PhD in applied mathematics from the National Chung-Hsing University in 2000. From 2000 to 2006, he was an assistant professor with the department of information management at the Northern Taiwan Institute of Science and Technology, Taipei, Taiwan. From 2006 to 2008, he was an assistant professor with the department of management information systems at National Chung-Hsing University, Taichung, Taiwan. Since 2008, he has been an assistant professor with the department of management information systems at National Chung-Hsing University, Taichung, Taiwan. His current research interests include affective computing, intelligent tutoring system and data mining.

Ting-Kuan Wu was born in Kaohsiung, Taiwan, R.O.C. on June twenty-sixth. Now he is an undergraduate student of management and information system department in National Chung Hsing University (NCHU), Taichung, Taiwan. He is the chief administrator of NCHU male dormitory's network, and he had worked in internship as a developer to build a student performance management system in Daojian High School of Commerce, Taipei, Taiwan and a volunteer serve in Sinlau Hospital for one and half month, from July 5th, 2009 to August 25th, 2009. His primary research interests are system architecture and integrating information systems with the operating process of any organizations.

Yu-bin Wang was born in Tainan, Taiwan, R.O.C on May eighth. Now he is an undergraduate student of management and information system department in Nation Chung-Hsing University (NCHU), Taichung, Taiwan. He is the administrator of school male dormitory's network, and he had worked in internship as an experimental educator in junior care class, Tainan, Taiwan. His primary research interests are computer science, optimizing system in algorithm, and contact system from each type of user by psychology.

Using Active RFID to Realize Ubi-media System

Jason C. Hung
Department of Information Management
Overseas Chinese University
jhung@ocu.edu.tw

***Abstract*—This paper proposes a human computer interface which is based on Active Radio frequency identification (Active RFID) technique to let human communicate with computer by analyzing the signal from tags. For retrieving those signals from tags, how to decreasing the noises created by surrounding environment and detecting useful information from variant signals are the most important. In our proposed method, we adopt a train procedure as pre-processing to categorize all of signals into two categories: noise and real data. After the real data is retrieved, we use "Music Director" and "DJ Scratch" as applications to let user play with computer. The experimental results show the proposed method can analyze the signals from tags successfully with high detection rate.**

***Index Terms*— RFID; Human Computer Interface; Received Signal Strength Indication**

I. INTRODUCTION

In the past few years, Radio Frequency IDentification (RFID) techniques is often used to identify object's identity such as warehousing, pet tracking, car anti-thief, payment system[3] and access controlling system[4]. In some research topics of local sensing, an object position can be estimated by analyzing the signal sent by tags to corresponding readers. However, due to the radio frequency is easy effect by obstructions such as walls, tables and metal objects, so there are also many researchers emphasis on how to improve the obstructing problems [1,2,5,6].

In this paper, we integrate the RFID technique and the concept of designing human computer interface to proposed a system which break through the limitation of above systems [1,2,5,6] and let user play with computer successfully. At the core of our proposed algorithm, we adopt a procedure of signal identifying by analyze the distance between identification tag and reader. The problem of signal obstructing is also improved by using our proposed signal categorizing procedure. But, due to the price of active-RFID device is expensive, so our goal is only using limited active-RFID devices to accomplish this system. In our implementation, we only are using one active-reader and some active-tags. For improving the categorizing rate, we add another reference tags.

"A Dream music director system" is a real application of our proposed system. We let user play as director of a musical group. The characteristic of this system is that whole music group only has one role: the director. User needs to hold the active-tag in their hand to act the pre-defined posture for controlling the music group. During the user change their posture, our system will detect the variance of signal then play the corresponding music file. By the way, user needs to finish an initial process which is used to determine the acceptable range of signal before they play. Different style of music can be generated and played by using our system.

The advantage of our proposed algorithm is that we can only using the estimation result of distance measurement to identify the posture of user successfully. Those categorized meaningful signal can be a guidance to let user control their computer. The proposed algorithm can easy be use in any kind of application.

This paper is organized as following. All of related works are discussed in section 2. The main methodology of the proposed system is described in section 3. The experimental results and conclusion is shown in section 4 and 5 separately.

II. RELATED WORKS

In this section, we will talk about how the RFID system works and reviewing some related researches. The main concept of how to implement a user-friendly interface will also be discussed in this section.

2.1 Active-RFID and Passive-RFID

There are four major devices constituting a RFID system: Reader, Tag, Client and Server. Figure 1 shows the flow chart of RFID system. The connection between client and reader is build by using USB port or RS232. There are two categories of RFID device: Active-RFID and Passive-RFID. The work flow of Active-RFID device is described as follows. In the beginning, tags will send radio frequency to reader continuously. When the reader received those radio frequencies, it will translate them into some useful signals by using local device and show the results on the client PC by middleware. After that, the client will send the tag's Identification or the message from tag to the sever side and then compare with the information in database. Since the comparison process is completed, server will response the request to the client. The work flow of Passive-RFID is almost same with Active-RFID. The only difference between Active and Passive is that user needs to put those tags near to reader. Those Passive-tags will work if and only if they are in the active range of reader.

doi:10.4304/jnw.6.5.743-749

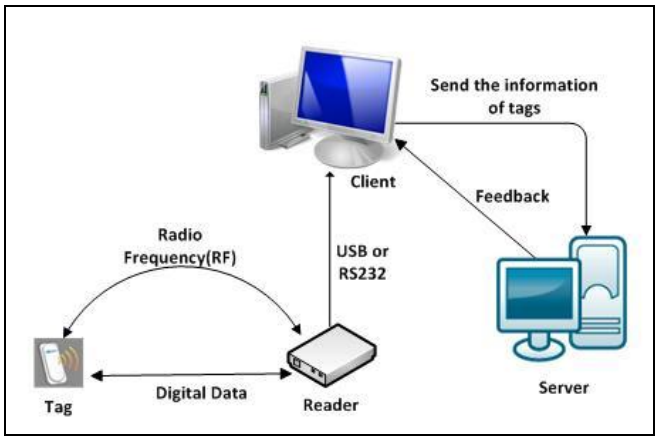


Figure1. Work flow of RFID system.

2.2. Related works of locate sensing and object tracking by using RFID technique.

2.2.1 SpotON[6] :

Authors have created SpotON to investigate ad-hoc location sensing, a exible alternative to infrastructure-centric location systems. SpotON tags use received radio signal strength information as an inter-tag distance estimator.

This paper designed and built hardware that will serve as object location tags, part of a project called SpotON. SpotON tags use received radio signal strength information (RSSI) as a sensor measurement for estimating inter-tag distance. Using many collocated nodes, the measured positional accuracy can be improved through algorithmic techniques and erroneous distance measurements caused by signal attenuation (e.g. by metal objects in the area) can be automatically factored out.

2.2.2 LANDMARK[1]:

A location sensing prototype system that uses Radio Frequency Identification (RFID) technology for locating objects inside buildings. The major advantage of LANDMARC is that it improves the overall accuracy of locating objects by utilizing the concept of reference tags. Based on experimental analysis, we demonstrate that active RFID is a viable and cost-effective candidate for indoor location sensing.

LANDARC (Location Identification based on Dynamic Active RFID Calibration) system employs the idea of having extra fixed location reference tags to help location calibration.

These reference tags serve as reference points in the system (like landmarks in our daily life).

the simplest way to find the nearest reference tag to the tracking tag is to use the coordinate of the reference tag with the smallest Euclidian-distance value as the unknown tag's coordinate. When use k nearest reference tags' coordinates to locate one unknown tag, we call it k-nearest neighbor algorithm.

2.2.3 LEMT[7] :

Present a novel algorithm, known as Location Estimation using Model Trees (LEMT), to reconstruct a radio map by using real-time signal-strength readings received at the reference points. This algorithm can take real-time signal-strength values at each time point into account and make use of the dependency between the estimated locations and reference points.

2.3. Discussion of Human Computer Interface Design

Tovi et al. propose a 3D user interface [8] which is base on the concept of human computer interaction. Fingerprint-based techniques consist of two phases: an offline training phase and an online localization phase In the offline phase, a radio map is built by tabulating the RSS measurements received from signal transmitters at predefined locations in the area of interest. In the online localization phase, the real-time RSS samples received from signal transmitters are used to search the radio map to estimate a user's current location based on the learned model.

The human computer interaction technique also can be used in learning. Bravo et al. [9] propose a system which is using ubiquitous computing technology. In this paper, authors use passive-RFID reader and tags to let user interact with computer. They also design two user interfaces for both teacher and students.

III. SYSTEM ARCHITECTURE AND EXPERIMENT PROCESS

3.1. System architecture as music director

Due to the signals will easy effect by humidity, obstructing problems and noise created by air, we gather the statistics from variance of signals and RSSI values to determine whether the tag is moving or not. The proposed system will record every signal during each time slot. After that, all of signals will be categorized according to those computed statistics. Hence, the proposed system is not easy effect by surrounding environment. The identification rate more precise, the output music plays smoother.

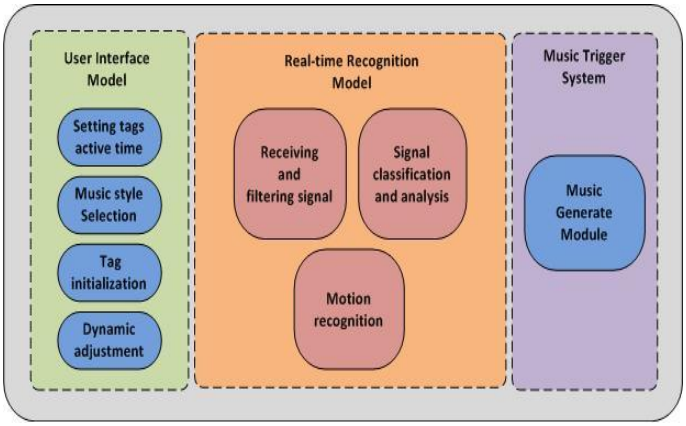


Figure2. Work flow of RFID system.

3.1.1. User Interface Model

According to the proposed system is used as a human computer interface based on RFID techniques, so how to let user easy to use is the most important part during the implement process. Figure 2 shows all of components included in our proposed system. We provide four

functionalities for user to tuning our system in different environment:

1. Setting tags active time.
2. Music style selection.
3. Tag initialization.
4. Dynamic adjustment.

3.1.2. Real-time Recognition Model

This part is not only the core of our propose system but also a most complicate module. The real-time recognition model consist three major parts:

1. Receiving and filtering signal: this functionality is used to receive signal and filter out the noise create by surrounding environment. For implementing this part, we modified the API which is provided by factory owner and make it more compatible for different kinds of environment.
2. Signal identification and analysis: This functionality is used to identify the useful information from all of received signals. We will classify received signals according to the active range of each reader.
3. Motion recognition: Since the useful information is identified in previous step, we can use those obtained information to recognize user's posture and trigger the player.

3.1.3. Music Trigger System

For playing different music simultaneously and reducing the computation cost of system, we use DirectSound proposed by Microsoft as base unit to build up our music trigger system. Table 1 shows the compatibility of DirectSound.

Table1. Wave format

File format	File Size	Number of sound track	Sample frequency
WAV	8bit	Single channel and Signal track	8KHz、11KHz
	16bit	Double channel and Stereo sound	22KHz、44KHz

3.2. Experiment process

The proposed system use active-RFID devices as input to let user command the computer. User can control computer by acting some pre-defined posture. Figure 3 shows the flow chart of our system. There are several functionalities consist our system and described as follows.

3.2.1. Initialization

Before user use our system to direct the computer play music, user have to trigger the initialize button for initializing our system. In this stage, the system will analyze twenty to thirty signals to get the maximum and minimum strength value of each tag. This step will raise up the identification rate.

3.2.2.

3.2.3. Data collecting and integration

Due to the signals sent by RFID tags are easy effect by surrounding environment, all of received signals in a time slot must be recorded and identified. Since the useful information is identified, we will collect them into a reference point set to help the system to decide if the music can be played or not. Those reference points also can be viewed as interaction trigger points.

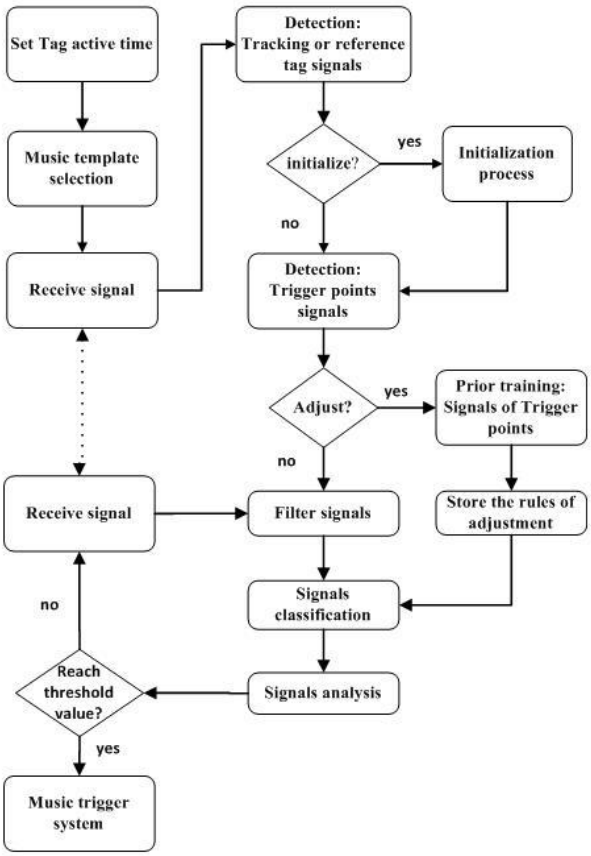


Figure3. Flow chart of the proposed Music trigger system

(1) Trigger point design

Table 2 shows an example of trigger point design. We put the reader on the table and the distance between user and reader is 40cm. The frequency of tag is 0.6 sec. User hold only a tag and acts the pre-defined posture.

Table2. An example of trigger point design.

Trigger points	**Description of action**	**RSSI Range(0-255)**	**Sample signal**
Ref. point-1	Tag is held at right-hand side. User's right hand needs is at front of neck and unbends toward right.	130－140	100
Ref. point-2	User swings their right hand to the top of reader.	150－160	100

(2). Design of adding reference tag to the trigger point

Because of the number of human postures will be limited by the number of tags used in our system, so we add a reference tag to let the system receive more signals and identify more postures. Table 3 shows the definition of trigger point which added a reference tag. All of settings are same as previous design.

Table 3. The definition of trigger point-adding reference tag

Trigger points	Description of action	RSSI(0-255)	Sample signal
Ref. point-1	Tag_A is held at right-hand side. User's right hand needs is at front of neck and unbends toward right.	Tag_A：： 140-150	100
	User held Tag_B at left hand side and swing to the top of reader.	Tag_B：： 160-170	
Ref. point-2	Right hand swing back to the top of reader.	Tag_A：： 160-170	100
	Left hand held Tag_B and unbend toward left.	Tag_B：： 140-150	
Ref. point-3	Right hand held Tag_A and unbend toward right.	Tag_A：： 130-140	100
	Left hand held tide Tag_B and stop at point-2.	Tag_B：： 90-110	

IV DJ SCRATCH SYSTEM

In this system, we proposed a virtual desk via active RFID tags and Readers. The setting process was shown in Figure 4. Figure 5 shows the flow chart of DJ Scratch system.

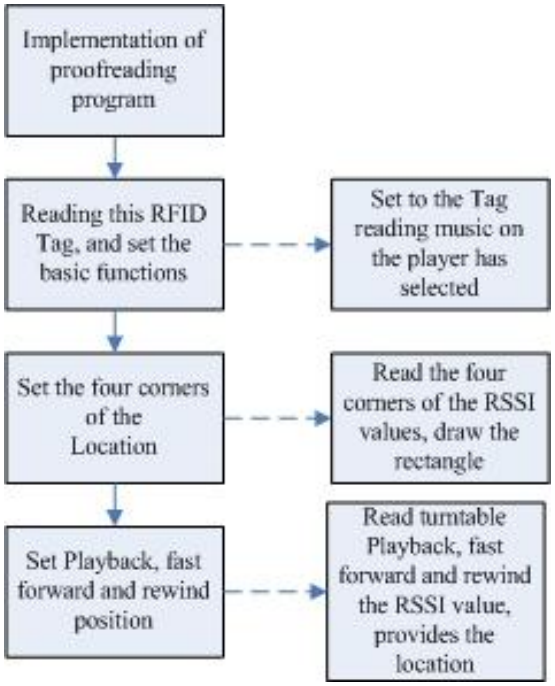


Figure4. Setting process of DJ Scratch system

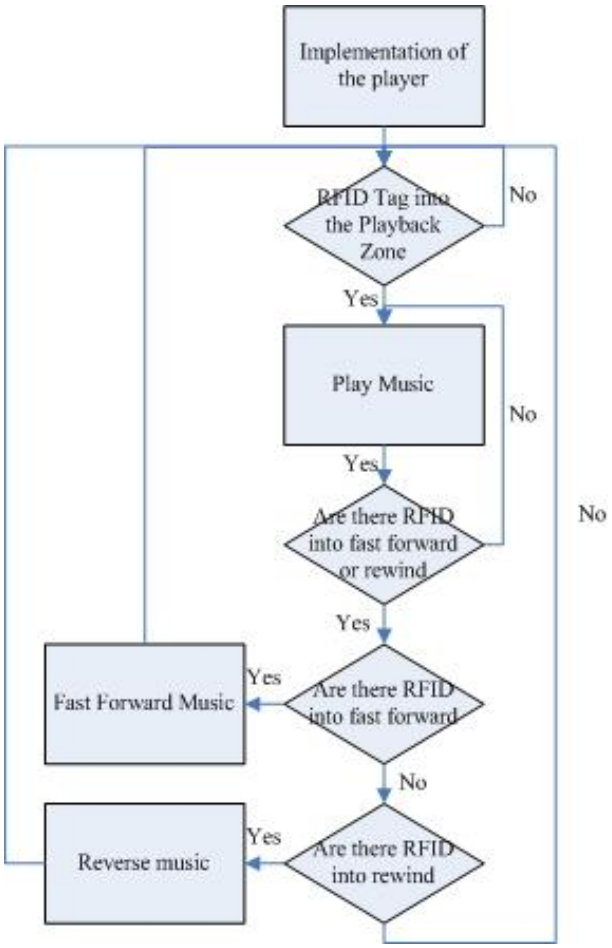


Figure5. Flow chart of the proposed DJ Scratch system

4.1 Execution process

Step1. To execute the adjustment or proofreading

program first, then to set a RFID Reader in the left-top side and another one in the left-down side.

Step2. To set the Virtual Desktop via the left-top and left-down Readers. Accompanied music will be set via the first active RFID tag positioned. Relative position will be decided according to the RSSI value. Figure 6 illustrated the situation.

Step3. Putting the tag from left-top, left-down, right-up to right-down sequentially is to virtualize a square desktop.

Step4. The scratch position will be set when we put the active Tag on the default place like figure 7.

Step5.The scratch zone will be decided according to the scratching movement. Figure 8 illustrates this action.

Step6. Setting the reverse action based on the reverse movement like the figure 8.

Step7. After the setting processes, music will be played in scratch special effective sound.

Step8. Putting the active RFID tag in the scratch position and playing the music.

Step9. Putting the active RFID tag in scratch zone and play the special effective music based on changing the playing speed.

Step10. The music will be played back based on the reverse movement.

Figure 9 illustrated the positioning processes and the following equations are the basic theory. Variable r1 and variable r2 in (1) and (2) are derived from r, x, and y.

$$(r1)^2 = x^2 + y^2 \quad (1)$$

$$(r2)^2 = x^2 + (r + y)^2 \quad (2)$$

$$y = ((r1) - (r2) + r2) / 2r \quad (3)$$

$$x = \sqrt{(r2 - y2)} \quad (4)$$

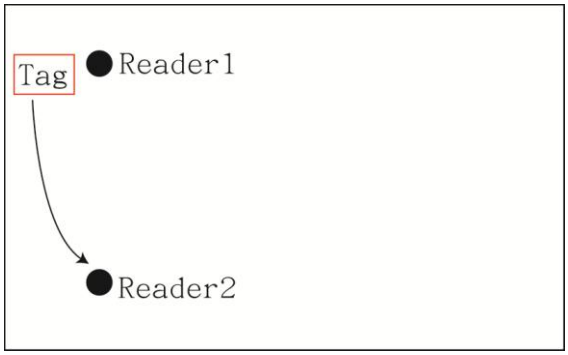


Figure6. Setting the Virtual Desktop

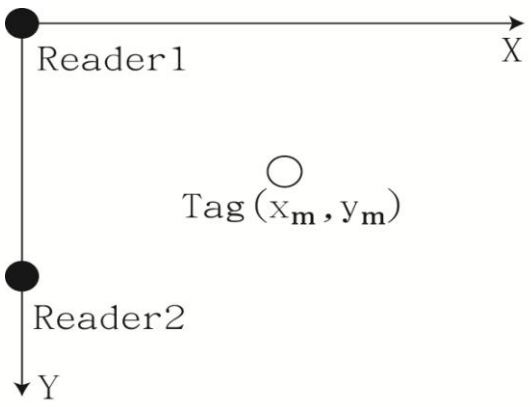


Figure7. Setting Scratch Position

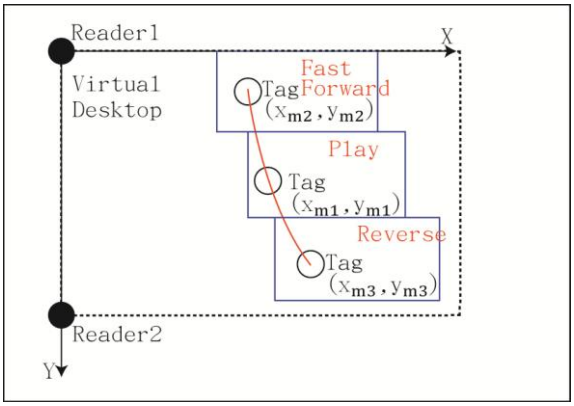


Figure8. Scratch Zone Setting

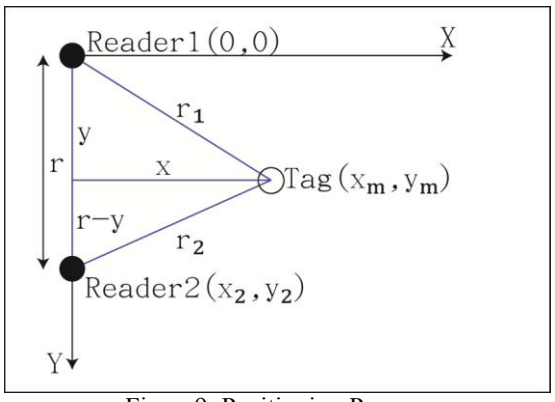


Figure9. Positioning Process

V. EXPERIMENTAL RESULTS

In this section, we provide two experimental results for two different experiments: without and with prior training.

5.1. Analysis of only using tracking tag without prior training

In this experiment, all of settings, environments and used device are same as mentioned in section 3.2.2. The proposed system will average the strength of received signals (about four signals) every three seconds. The total number of signals is 100 times. As shown in figure 4, the identification rate (accuracy) at Ref. point-1 is 0.08. The miss detection rate (error) is 0.2. The field of exception

means that the system cannot determine out the human action. As you can see, the system cannot identify human action correctly without prior training.

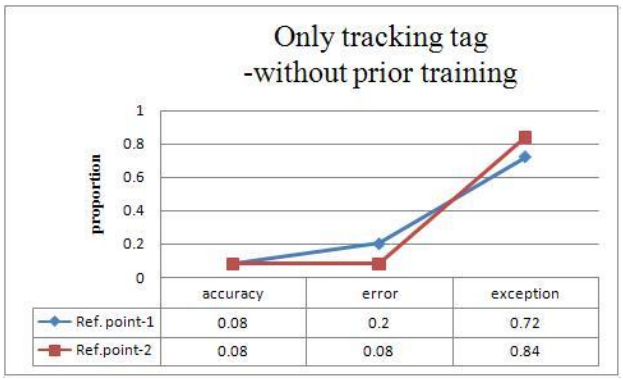


Figure10. Using single tag and without prior training.

Figure 11 shows the another analysis which at the situation of adding a reference tag but still without prior training.

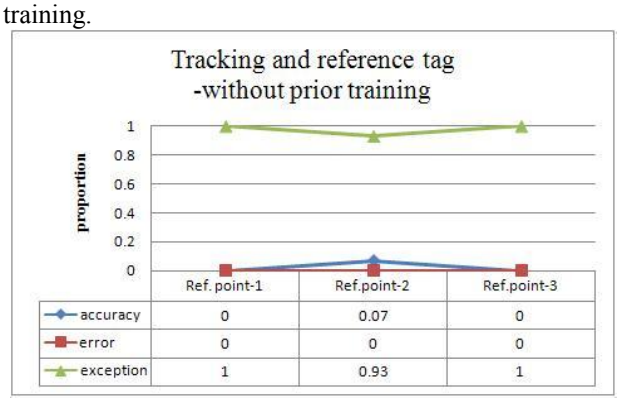


Figure 11. Adding reference tag but without prior training.

5.2. Analysis of experiments with prior training : using adjustment rules

Figure 10 and 11 show that the system is unstable because the user did not trigger the initial button to training our system. Even though the received signals are the same, the proposed system still cannot recognize human actions. Hence, in this paper, a pre-processing step is adopted in our system to improve the accuracy in identification process.

Figure 12 and 13 shows the results after we adopt some adjustment rules in pre-processing step. They are quite different with Figure 10 and 11. The identification rate is improved and the exception rate is reduced.

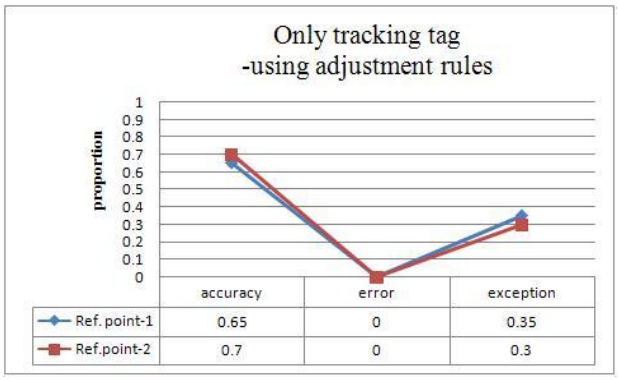


Figure 12. Using single tag and adjustment rules

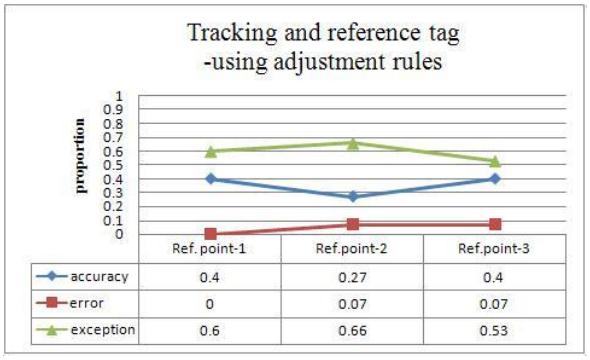


Figure 13. Add reference tag and both with adjustment rules.

VI. CONCLUSION

In this paper, we propose a new system which is based on Active RFID techniques and conform to the main concept of human computer interaction. Designing and adding a pre-processing mechanism before user interacts with computer can increase the correct rate of trigger events and reduce the incidents of wrong identification. “Music Director” and “DJ Scratch” are applications to let user play with computer. Our main contribution is to realize an Ubi-media system based on using the Active RFID suite. In the future, we will try to build an environment that utilizes more readers or tags enhance the lack of the number of interaction and improve our user interface.

ACKNOWLEDGEMENT

This work was partially supported by National Science Council, Taiwan: NSC 98-2622-E-240-001-CC3 and NSC 99-2221-E-240-003-.

REFERENCES

[1] L.M. Li, Y. Liu, Y.C. Lau, and A.P Patil, “LANDMARC: indoor location sensing using active RFID,” Wireless Networks, Vol. 10(6), pp. 701-710, 2004.

[2] D. Hahnel, W Burgard, D. Fox, K. Fishkin, and M. Philipose, “Mapping and localization with RFID technology,” Proceeding of IEEE International Conference on Robotics and Automation(ICRA 2004), April 26-May 1, 2004.

[3] S.L. Garfinkel, A.Juels, R. Pappu, "RFID privacy: an overview of problems and proposed solutions," IEEE Security and Privacy, Vol.3(3), pp. 34-43, May-June 2005.
[4] D.H. Wilson, "Assistive Intelligent Enviroments for Automatic Health Monitoring," Robotics Institude, Carnegie Mellon University, Sep. 2005.
[5] A.Malekpour, T.C. Ling, W.C. Lim. "Location Determination Using Radio Frequency RSSI and Deterministic Algorithm," Communcation Networks and Services Research conference(CNSR 2008), pp.488-495, 2008.
[6] J. Hightower, C. Vakili, G. Borriello, and R. Want, "Design and calibration of the spoton ad-hoc location sensing system," 2001.
[7] Jie Yin, Qiang Yang and Lionel M. Ni "Learning Adaptive Temporal Radio Maps for Signal-Strength-Based Location Estimation" Hong Kong University of Science and Technology, IEEE TRANSACTIONS ON MOBILE COMPUTING, VOL. 7, NO. 7, JULY 2008.
[8] Tovi Grossman, Daniel Wigdor, and Ravin Balakrishman, "Multi-Finger Gestural Interaction with 3D Volumetric Displays," Proceedings of the 17th annual ACM symposium on User interface software and technology 2004, pp. 61-70, Santa Fe, NM, USA, Oct. 24-27, 2004.
[9] José Bravo, Ramón Hervás, and Gabriel Chavira, "Ubiquitous Computing in Classrom: An Approach through Identification Process," Journal of Universal Computer Science, vol. 11, no. 9 (2005), pp.1494-1504.

Jason C. Hung is an Associate Professor of Dept. of Information Management, Overseas Chinese University, Taiwan. His research interests include Multimedia Computing and Networking, Distance Learning, E-Commerce, and Agent Technology. From 1999 to date, he was a part time faculty of the Computer Science and Information Engineering Department at Tamkang University. Dr. Hung received his BS and MS degrees in Computer Science and Information Engineering from Tamkang University, in 1996 and 1998, respectively. He also received his Ph.D. in Computer Science and Information Engineering from Tamkang University in 2001. Dr. Hung participated in many international academic activities, including the organization of many international conferences. He is the founder and Workshop chair of International Workshop on Mobile Systems, E-commerce, and Agent Technology. He is also the Associate Editor of the International Journal of Distance Education Technologies, published by Idea Group Publishing, USA. The contact email is jhung@ocu.edu.tw

On the Design of A Contribution-based, Flexible Locality-Aware P2P Streaming Network

Yu-Wei Chan
Department of Information Management, Chung Chou Institute of Technology, Yuanlin, Taiwan, R.O.C.
ywchan@dragon.ccut.edu.tw

***Abstract*—P2P live streaming networks recently become an emerging research topic. In such multimedia steaming networks, autonomous users cooperate with each other to provide a self-organizing, distributed, scalable, and cost-efficient transmission environment. Due to the mismatch problem between the overlay network and the physical network, dynamics of peers, and peers' self-interested, low start-up delays, full cooperation, and stable playback quality can not be guaranteed. Therefore, how to incentivize the participanting peers to have the willings to provide more resources and how a group of geographically neighboring peers cooperate with each other to achieve better streaming performance are all important issues. In this paper, we propose the design of a two layered hybrid tree-push/mesh-pull overlay network with considering the problems of locality-aware of peers and incentive schemes. We propose the schemes and algorithms to overcome the mismatch problem between the overlay network and the physical network, reduce the transmission latency, and provide a fully cooperative, and reliable P2P streaming environement.**

***Index Terms*—live streaming, P2P streaming, locality-aware, incentive mechanism**

I. INTRODUCTION

In the past few years, with the success of peer-to-peer (P2P) file sharing technology, a large number of P2P live streaming systems have become a reality in the Internet [1][2][3][4][5]. In these systems, a large number of peers self-organize into streaming overlays such that the peers can exchange media chunks with each other and construct a self-organizing, scalable, and efficient live media content distribution environment. However, in the currently proposed P2P streaming systems, some peers in an overlay are selected as the streaming suppliers by the topology and the maintenance schemes which would affect the common performance metrics, such as data-stream delivery efficiency and perceived quality of streaming. A proper overlay for peer-to-peer streaming can keep stable suppliers, shorten transmission delays and balance the load of peers. Therefore, how to form an overlay to overcome the uncertainty factors such as reliability of live media streaming quality, non-guaranteed communication efficiency, limited upload capacity, dynamic of suppliers and so on thus becomes challenging issues.

Currently, according to the media streaming delivery architectures, P2P live media streaming systems can be classified into two major approaches which are tree-push and mesh-pull, respectively. Some early P2P streaming applications [6][7][8] proposed the single tree based structure as their communication architecture for media data delivery. While the single tree based structure is simple and efficient in terms of delay optimization, it is vulnerable to network dynamics due to its rigid structure. Moreover, the failure or leave of a node near the streaming source may cause data outage in all its descendants and the upload bandwidth of the leaf peers of the tree cannot be utilized. Therefore, existing approaches slice the stream into multiple sub streams and render the structure to build multiple trees [9] to utilize the leaf peers' upload bandwidth resources fully. Although the multiple-tree structures improve resilience, it is more complex to maintain and construct.

In order to achieve better resilience, some data-driven P2P streaming systems [2][3] use gossip-based unstructured framework to construct the mesh structures where each peer independently selects neighboring peers and exchanges data with them. The mesh structure achieves inherent robustness for highly dynamic, high churn rate of P2P environment [10]. In the mesh-based method, peers exchange live media streaming sessions with each other. A peer has to pull data to avoid significant redundancies. Although the mesh-based systems are more robust, they often suffer from longer start-up delays, higher traffic control overhead and playback time lags of large peers. Therefore, recently, some systems of hybrid push-pull architecture were proposed [12][13][14].

In this system, to overcome the problems of start-up playback delays, and construct a more reliable cooperative streaming environment, we propose a two-layered contribution-based cooperation, flexible locality-aware overlay network which addresses the contribution-level strategies for stimulating peers' cooperation and uses the group concepts to construct locality-aware groups. By exploiting the resource of geographically neighboring peers with low transmission delay, the delivery efficiency and perceived quality can be constantly satisfied in our system.

In the proposed 2-layered overlay, peers are clustered into locality groups based on the transmission delays of peers. In each locality group, peers form an overlay mesh for exchanging the media sessions with other peers of the same group. Once the overlay mesh is formed, the group head will be selected among peers of the same group by using the designed scheme which named the optimal

doi:10.4304/jnw.6.5.750-756

stable peers identification scheme. Each group head acts as the stable peer (or super peer) of the mesh who is responsible for exchanging the media chunks among groups. At the same time, the stable peer is also the *collector* who collects the requests of media chunks of other peers in the same group except the stable peer and transmits these requests to other stable peers for exchanging the chunks. Therefore, a key question to answer is: *"how to select the stable peer in a locality-aware group and how a group of neighboring peers should cooperate with each other to achieve better streaming performance?"*

The main contributions of our proposed system are summarized as follows.

(1) We propose a novel P2P streaming architecture which contains the overlay construction and maintenance schemes to adjust the size of the overlay dynamically according to the dynamics of peers, and the contribution-based cooperation scheme to stimulate peers' cooperation for achieving better streaming performance.
(2) We consider the importance of peers' locality in the system. We propose some schemes such as the overlay construction scheme, the membership management scheme, the overlay maintenance scheme, the streaming scheduling scheme, and the contribution-based cooperation scheme, we successfully reduce the source-to-end delivery latency and the start-up delay. Also, we increase the reliability of streaming delivery paths of our system.

The remainder of this paper is organized as follows. In section II, some related works are given. In section III, we propose the system model. Then, we propose the system schemes in detail in section IV. Section V gives the discussions of the feasibility of our system. Section VI concludes this paper.

II. RELATED WORK

Recently, many overlay schemes have been proposed in the literature for efficient peer-to-peer streaming. The goals of these schemes are to assure that the data streaming delivery efficiency and the received quality metrics can be constantly satisfied. They can be classified into tree-based [6][7][9] and mesh-based structures [1][2][10][11].

In spite of the proposed single or multiple tree-based systems, these systems create and maintain an efficient tree overlay, mimicking a multicast tree structure. However, these structures are sensitive to node dynamics and needed to adjust the topology frequently that may cause worse streaming quality. Moreover, Due to the streaming of high bit rate, the tree-based structure is not suitable properly because it does not take the heterogeneity of peers into account.

Mesh-based P2P streaming systems have enjoyed a number of successfully deployments to date, such as DONet/CoolStreaming [2], PPLive [4], PPStream[5] and others [3][11]. The major advantages of mesh-based systems are the simple design principle and inherent robustness, particularly desirable for the highly dynamic, high-churn P2P environment. Bullet [11] is a scalable and distributed algorithm used for constructing high-bandwidth streaming overlay. In Bullet, nodes can self-organize into an overlay tree to transmit the disjoint data sets and retrieve the missed parts simultaneously. DONet/CoolStreaming [2] is a data-driven overlay network for live media streaming. By employing a gossip protocol, peers can periodically exchanges the availabilities of data blocks for retrieving yet unavailable data and supplying available data. However, the streaming quality can not be guaranteed.

Recently, researchers focus on the new class of hybrid push-pull architecture for P2P live media streaming. In [13], the GridMedia system adopts the gossip-based approach for the construction of the overlay and uses a push-pull mechanism to not only keep robust with the peers' churn behavior but also reduce the latency of peers efficiently. In [12], authors proposed a mTreebone framework which was a hybrid tree-push/mesh-pull design that leverages both overlays. The main focus of this research work is on the effective use of stable nodes. In [20], authors propose a two-layered locality-aware P2P streaming framework which considers the problem of how to exploit the bandwidth resource of physically neighboring peers.

Generally, peers in the system are autonomous and rational such that they want to maximize their utilities or profits with minimum contribution cost. Hence, if no proper incentive mechanism exists, the selfish characteristics of peers will result in the situations of free-riding and the tragedy of the commons.

Therefore, to address the unique challenges in providing incentives in P2P live streaming, some different incentive mechanisms [15][16][17][18][19] are presented. In [15], a payment-based incentive mechanism was proposed, where peers pay points for receiving data and earn points by forwarding data to others. A distributed incentive mechanisms on mesh-pull P2P live streaming systems wa proposed [16]. In [17][18], a game theoretic framework is presented for incentives in P2P systems. In [19], authors propose a coalition-based framework to enable peers to join the coalitions so as to achieve cooperation.

III. SYSTEM MODEL

Based on the concept of super-peer network [21] and hybrid tree-push/mesh-pull structure [12], we propose a two-layered architecture which is shown in Fig. 1. From Fig. 1, we can realize that the proposed architecture which consists of the tree and mesh structures. The tree structure is constructed by connecting the super-peers or the stable peers whose resources are more stable and more sufficient. Besides, peers of the same overlay connect with each other to construct the mesh structure. In the system, peers are clustered into locality-aware groups (or meshes) with bounded size whose value was defined based our previous research works [20].These locality-aware groups form the bottom-layer of this overlay.

In each locality-aware group, one of the peers will be selected as the group head which we called stable peer (or super-peer) according to the designed schemes which will be described in the following sections. With the help of the stable peers, the media chunks in a locality-group can be rapidly exchanged with other group members. In our proposed system, we assume that the maximum number of stable peers should not exceed one but exceed zero in a locality-aware group. Therefore, we must select one proper peer to be the stable peer in each group. These stable peers are connected with each other to form the top layer of this overlay and are also interconnected as a multicast tree rooted at the streaming source.

In the proposed system, we assume that peers are classified as the stable peers and normal peers, respectively. We also assume that when a new incoming peer joins the overlay, its played role will be the normal peer. If one of the normal peers wants to be promoted to become the stable peer, the designed optimal stable peer identification scheme will be performed.

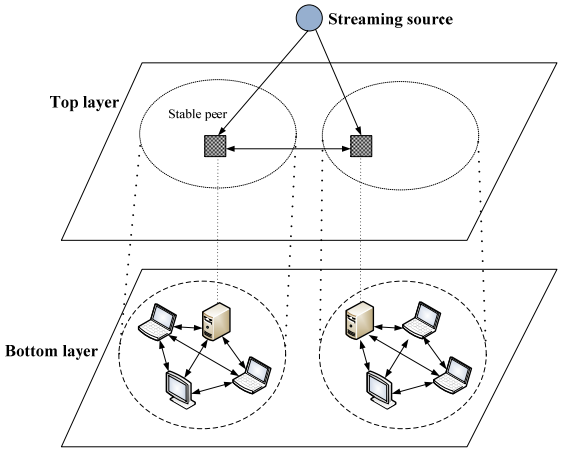


Figure 1. Overview of system architecture

As a streaming session starts, the corresponding overlay would be constructed by the streaming source according to the specifications of the session. The new incoming peers join a proper locality-aware group in the overlay by using the overlay construction scheme. Streaming data from the streaming source are ditsributed along with the multicast tree by continuous requests and relays. The clustered peers in a locality-aware group are managed by the membership management scheme for cooperative streaming. To keep sufficient and stable suppliers for streaming, the overlay maintenance scheme for overlay maintenance would be performed on groups if the number of peers in a group is over its bounded size or less than a threshold, respectively. The ability to grow or shrink the number of groups in an overlay makes the proposed overlay flexible and scalable. In the following sections, we will sequentially describe the system schemes in details.

In addition, since our framework is a hybrid tree-push/mesh pull overlay. The core of the framework is the construction of the tree-based backbone which was constructed mainly by stable peers. Therefore, how to encourage the stable peers stay longer and contribute more resources instead of reducing the probability of churn behavior? How to stabilize the treebone structure in a highly dynamic environment? How to incentivize the normal peers to become the stable peers? These above described problems are all we considered in this research.

IV. SYSTEM COMPONENTS

Our system mainly consists of five major schemes which are shown in Fig. 2.

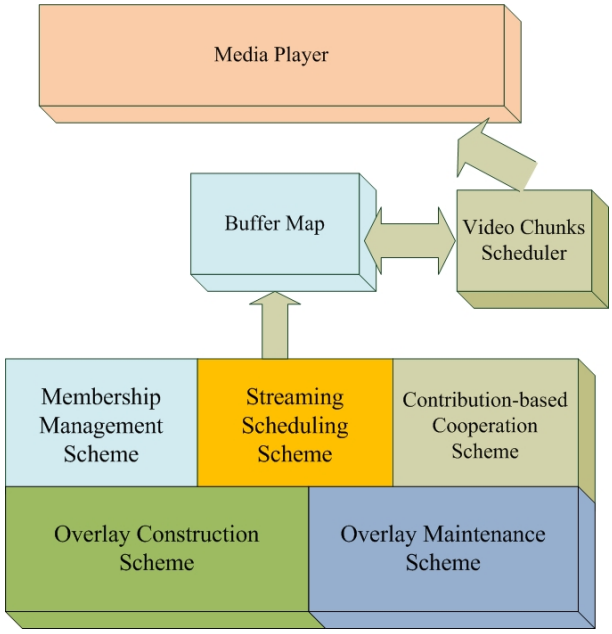


Figure 2. System components

The main functions of these schemes are listed as follows.

1. The overlay construction scheme: this scheme is used to help the participating peers with locating themselves into proper locality groups and organizing them into two-layered overlay.
2. The membership management scheme: this scheme is used to help peers with organizing the membership of peers in locality groups.
3. The overlay maintenance scheme: this scheme is also used to dynamically adjust and optimize the overlay structures when some stable peers of the tree-based backbone leave and crash suddenly.
4. The streaming scheduling scheme: this scheme determines the delivery of streaming data chunks and reconciles the tree and the mesh overlays, so as to fully utilize their potential.
5. The contribution-based cooperation scheme: this scheme provides some incentive cooperation mechanisms for cooperative streaming between the stable peer and the normal peer. They include the stable peers identification mechanism, the differentiated service-based incentive mechanism, and the peers' resource reciprocation mechanism.

A. The overlay construction scheme

This scheme mainly consists of two mechanisms, they are the groups-formation mechanism, and the peer locating mechanism, respectively.

(1) Groups-formation mechanism

Firstly, we define the locality-aware group in the system. In reality, every peer can have a large number of geographically neighboring peers with large intra-group upload and download bandiwdths. Therefore, these peers are connected with each other to form locality-aware groups. For peers in the same locality-aware group, the transmission delays among them are less than or equal to the predefined value according to the rate of a streaming session. In this paper, the predefined value is set between $l/2$ and l based on these research works [20], where l is the tolerable delivery latency of the streaming rate of a media chunk.

In addition, with respect to the groups formation process, we mainly refer to our previous works [20] to set the size (number of peers) of a group that is bounded by $[k, (3k - 1)]$, except for the locality group where the streaming source belongs is bounded by $[1, (3k - 1)]$, where $k \geq 1$. If the size of a locality-aware group is equal to $3k - 1$, we say that the group is full. When a peer joins a full group, it will cause this group spilt into 2 smaller groups. If the size of a group except for the streaming source belongs is less than k due to some peers leave, this group will be merged with other groups such that the size of the merged group is larger than $3k$. If no such locality group is available, the merge processes will not performed until such a group is available or when the size of the group is greater than or equal to k again.

After the formation of locality-aware groups, one of the normal peers who contributes more upload and download bandwidths will be selected as the group head (or the stable peer). The stable peers which are interconnected with each other to construct the tree-based backbone rooted by the streraming source. In this system, we assuem that the track server is the streaming server. The track server records the essential information of the divided media chunks and the corrresponding overlays which were considered as the metadata of the streaming data.

Since our designed system adopts the hybrid tree-push/mesh-pull data delivery mechanism, data are delivered with two stages. In the first stage, data are pushed from the streaming source to stable peers through the tree-based backbone. In the second stage, through the stable peers, data are delivered to the normal peers or other stable peers who require the media chunks they need. Through the designed two-stage data delivery framework, the transmission performance can be enhanced significantly. Since the data are pushed by the tree-based backbone in the first stage, the stable peers attached to the backbone generally will contribute more upload bandwidth to provide more efficient and resilient transmission. In a case if a peer who receives the stream has been disrupted, the peer may pull the missed media chunks through the locality-aware groups such that its received streaming quality can be guaranteed.

In addition, since the data delivery mechanism of meshes uses the P2P swarm-like content delivery framework. In this kind of framework, a peer downloads from multiple suppliers in parallel. Each peer uses the gossip protocol to find neighbors. Furthermore, each peer keeps a "buffer map" indicates the media chunks that it has currently buffered and can share with others, and they exchange their buffer maps with each other frequently. In addition, each peer maintains a window of interest, which is the set of sequence of packets that the peer is interested in requesting at the current time.

(2) The peer locating mechanism

When a new incoming peer wants to join the streaming overlay, it will call the peer locating mechanism to join a proper locality-aware group. Therefore, the algorithm of the peer locating mechanism is shown below.

Algorithm 1 Peer locating mechanism

For a new incoming peer p_i:
Step1. Obtain the stable peers who have the interested chunks from the track server.
Step2. Calculate and sort the average network delays betwee p_i and the connected stable peers in decreasing order n times (n is the predefined value).
Step3. Select and join the group with the smallest network delays between p_i and the corresponding stable peer.
Step4. If the selected group is full, peer p_i will select the group with the second smallest network delays to join.
Step5. Repeat steps from step2 to step4 until peer p_i has joined the overlay successfully.

B. The membership management scheme

The membership management scheme is used to organize the peers in a locality group in this system. Through this scheme, the peers in a locality group can collaborate for streaming data. In this paper, with respect to the membership management scheme , we mainly refer to our previous works [20]. In this scheme, the stable peer of a group serves as the group head to handle the join and leave operations of normal peers, monitor the status of normal peers, manage and broadcast the metadata information of group and media chunks.

In our system, a member cache is used to store the information of members in a locality group. For each joined locality group, a peer maintains the corresponding member cache. The information stored in the member cache consists of four fields which are type, network address, contributor rank and subset. The type field specifies the role of a member. The network address field is used to record the network address of a member. The contributor rank field is used to record the rank among all contributors. The rank is used to recover the failure of the stable peer and for the split scheme. The subset field specifies the subsets of members. During the join and leave procedures of peers, the information of the members is recorded in the member cache of the stable

peer. For monitoring the status of peers, a stable peer receives the keep-alive messages from its members constantly to assure that they are alive. If a stable peer does not receive the keep-alive messages from a member in a period of time, it will drop the information of that member.

To manage a contributor in the member cache, the stable peer will set the type field of the corresponding entry as the contributor with a contributor rank. The contributor rank is a unique stamp (ex. the global timestamp when receiving informing messages) generated by the stable peer. When a contributor lacks of the streaming data in its data cache, it will inform the stable peer. The stable peer will set the contributor as a free-rider by setting the type field of the corresponding entry in the member cache. Based on the management of contributors, a stable peer periodically updates the information of contributors to each member. With the proposed membership management, a set of contributors act as the streaming suppliers are called active contributors. Corresponding to the rate of a session, a peer obtains the streaming data from the data caches of active contributors based on the collected bandwidth. If the status of an active contributor is changed (ex. leave, lower bandwidth and so on), a peer would seek one non-active contributor in its member cache to replace this active contributor.

C. The overlay maintenance scheme

To keep sufficient and stable suppliers for streaming data and ensure that the overlay can adjust itself with the dynamics of peers, the split and merge schemes will be performed on locality-aware groups if the number of peers in a group is over its bounded size or less than a threshold, respectively. In our overlay, a stable peer periodically checks the size of its group and performs the split/merge schemes if needed.

(1) Overlay split scheme

When the size of a locality-aware group is larger than $3k - 1$, the split scheme would be triggered to split this group into two smaller groups. The following algorithm is the procedure of the split scheme.

Algorithm 2 Overlay split scheme

For a locality-aware group g_i:
Step1. The stable peer m_i of a group g_i who selects the contributor c_j in its member cache as the stable peer of a new locality group.
Step2. Peer c_j claims itself as the stable peer m_j of a new locality group g_j by informing the track server and other stable peers who connected with it.
Step3. Decide the members of the new group according to the selection criteria defined by [20].
Step4. The stable peer m_i creates a *split list*, and broadcasts the split list to all its members in g_i.
Step5. When a member received this split list, it will decide whether to move to the new group or stay in the original group.

(2) Overlay merge scheme

In addition, to maintain sufficient and available resources in each locality group, a locality group would perform the merge scheme when the size of the locality group is smaller than the predefined threshold k. Assume that the size of a locality group g_i is smaller than the predefined threshold k. The stable peer m_i of g_i first queries the stable peer m_s of its source group g_s to obtain the size of g_s. There are possible situations which may happen during the overlay merge process, they are listed as follows:

(1) Case1: If the size of the group g_s is less than $3k$ after performing the merge processes with group g_i, all members in g_i would join the group g_s and the stable peer m_i of g_i would act as the contributor in g_s.
(2) Case2: If Case 1 is not satisfied, the stable peer m_i will find a locality group g_j whose size is less than $3k$ after merging with the group g_i. If such a group g_j is found, all members in g_i would join g_j and the stable peer m_i would act as a contributor in g_j.
(3) Case 3: If such a g_j is not available, the stable peer m_i will find a group g_k whose size is less than $3k$ after merging with g_i. If such a group g_k is found, all the members in g_k would join g_i and the stable peer m_k of g_k, would act as a contributor in g_i.
(4) If none of cases specified above is satisfied, the merge process will repeated from case 1 to case 3 until one of the cases is satisfied or the size of g_i is greater than or equal to k.

D. The contribution-based cooperation scheme

In this section, we will propose a two-phase contribution-based cooperation scheme which overcomes the problems of stable peers identification and incentive cooperation strategies of peers. At the first phase of the scheme, we propose a novel method to identify the optimal stable peers. In the second phase, we propose a method concept to stimulate normal peers' contribution in the meshes.

(1) Optimal stable peers identification mechanism

With respect to the optimal stable peer identification, the research [12] proposed a method by using a node's age to identify the optimal stable nodes. In this method, nodes with higher age tend to stay longer. Although the research proposed method to identify stable nodes, we consider that there exist some problems. For examples, the proposed method [12] does not consider the situation that stable peers who stay longer do not mean that they will contribute more resources. Therefore, we present the following algorithm 3 to summarize the process of the optimal stable peer identification.

The expected contributed level represents peer i's commitment level of the expected contribution of upload bandwidth promised to the connected matched peers. A higher value of the expected contribution level represents that the peers can provide more stable multimedia quality. In addition, in this push-pull transmission scenario, stable peers generally contribute more upload bandwidth than normal peers. But all the peers enjoy the same video playback quality. Therefore, in [16], layered video can be

used to provide incentives for one normal peer to become a stable peer. Through the differentiated service, normal peers only download a basic-layer video to receive the basic video playback quality; however, super peers download multiple-layer video to enjoy an enhanced video quality.

Algorithm 3 Stable peers identification mechanism

Step1. Determine the Expected Contribution Level (ECL) of each peer based on the peers' commitment level of the expected contribution of upload bandwidth promised to the connected matched peers.

Step2. Determine the Expected Service Time (EST) of each peer. Through the calculations, we can obtain each peer's age and predict the peers' staying time.

Step3. Define the optimal size ratio of amounts between stable and normal peers based on the obtained results of the above two steps.

The notations in (1) are specified in Table 1.

TABLE 1. NOTATIONS OF PARAMETERS OF THE EQUATION

parameter	Parameter Description
t_{0i}	the joining time of new incoming peer *i*
SW_i	peer *i*'s size of window of interest
$ECLi(t)$	peer *i*'s expected contribution level function
$RU_i(t)$	peer *i*'s contributed upload bandwidth function
PUC_i	peer *i*'s physical upload capacity

Equation (1) is used to calculate the ECL value.

$$ECL_i(t) = \frac{\sum_{t=t_{0i}}^{t=t_{0i}+SW_i} RU_i(t)}{\frac{SW_i}{PUC_i}}, 0 \le ECL_i(t) \le 1 \quad (1)$$

(2) Incentive cooperation strategies

Next, we will discuss about the incentive cooperation strategies among normal peers in the meshes. We plan to propose a coalition-based incentive resource exchange mechanism based on the strategic bargaining game. In [19], authors also propose a coalition-based resource reciprocation strategy for P2P multimedia streaming. In this research, a new framework where each peer creates a coalition of matched peers with which it can exchange resources was proposed. This research work adopts the proportional bargaining solution to negotiate the upload bandwidth among the matched peers. However, in our approach, the P2P streaming environment was considered as a decentralized market model. Each peer in the system is a market agent. Since each peer will form a coalition with its matched peers, peers in the coalition can transact with media chunks with each other through virtual monetary exchanges. Peers in the coalition can negotiate with each other to achieve the equilibrium prices and the appropriate demanded upload bandwidth. In our system, we assume that the peers are rational such that they want to maximize their utility. Since the P2P streaming system is a highly dynamic, self-organizing, and strict time-constraints environment, how the peers can form coalitions, and how to formulate the cooperative behavior of peers with the Nash Bargaining Game (NBS) [19] are our research works in the future.

V. DISCUSSION

In this section, we will give some discussions about the feasibility of our system on different aspects according to these described issues. The first one issue is the rate of streaming session. When the session's rate increases, the load of upload bandwidth of peers would become heavier. Since the limited capacity of bandwidth and policy on each end host, relaying the streaming data becomes problematic. Thus, a dedicated multicast architecture for peer-to-peer streaming is considered to be more viable. Considering the high rate session, the provision of streaming data with constant acceptable performance may be difficult due to the limited upload bandwidth of end hosts.

Comparing with the tree-based multicast architecture, a peer can gather resources from multiple suppliers to easily achieve the acceptable performance in our system. On the other hand, when the rate of streaming session is relatively low, the amount of forwarded data streams would be minor to end users so as to release those limitations specified above. On this condition, the benefit of this paradigm can be exploited to shorten the streaming delivery paths by the escaped factor of overlay size as the simulation results show.

A benefit of typical multicast architecture can be demonstrated that the participanting peers only concern about their directly connected suppliers, not the delivery paths from the multicast source. The peer-to-peer streaming system applies this kind of architecture to keep the consistency of streaming delivery paths. Since each peer is responsible for transmitting the data streams for its suppliers and customers, this will produce long latency due to the predecessors on those paths. Although some optimization schemes of these systems such as AnySee [3], [2] or the properties of these structured overlay networks effectively decrease the delivery latency and relay hops from the streaming source, they have frequent recovery times of failure paths due to the dynamics of peers.

Finally, we assume that if the source-to-end delivery delay is an important factor to evaluate the efficiency of a streaming system. Our system exhibits the flexibility in a small-scale system by the compact streaming delivery paths.

VI. CONCLUSION

In this paper, we have presented a peer-to-peer streaming system based on a contribution-based cooperative, and flexible locality-aware overlay network. Based on the contribution-aware incentive mechanism,

the optimal stable peer can be identified as well as the contribution degree of stable peers.

In addition, by exploiting the surrounding neighbors of peers with low communication delays, our overlay is constructed to match the underlying network topology. Based on the group concept, the resources of peers in our overlay can be fully utilized. Based on the properties of flexibility and locality-awareness of our system, the peers would benefit from sufficient, stable and efficient suppliers in the joined locality groups for streaming.

With respect to the future works, we plan to do the simulations in the setting environment to prove that our proposed methods are sound. Meanwhile, we will develop our system prototype based on the design to compare our simulation and experimental results with other systems. Furthermore, we will investigate more efficient and reliable transmission schemes such as network coding and layered coding to achieve high-throughput, scalable and robust peer-to-peer streaming environments.

REFERENCES

[1] C. Diot, B. N. Levine, B. lyles, H. Kassem, and D. Balensiefen, "Deployment Issues for the IP Multicast Service and Architecture," *IEEE Network*, vol. 14, no. 1, January/February 2000, pp. 78-88.

[2] X. Zhang, J. Liu, B. Li, and T.P. Yum, "Coolstreaming/DONet: A Data-Driven Overlay Network for P2P Live Media Streaming," in *Proc. INFOCOM*, 2005.

[3] X. Liao, H. Jin, Y. Liu, L. M. Ni, and D. Deng, "AnySee: Peer-to-peer live streaming," in *Proc. INFOCOM*, 2006.

[4] *PPLive*, http://www.pplive.com/.

[5] *PPStream*, http://www.ppstream.com/.

[6] Y.Chu, S. G. Rao, S. Seshan, and H. J. Zhang, "A Case for End System Multicast," *IEEE Journal of Selected Areas in Networking*, vol. 20, no. 8, October 2002, pp. 1456-1471.

[7] S. Banerjee, B. Bhattacharjee, and C. Kommareddy, "Scalable application layer multicast," in *ACM SIGCOMM*, 2002.

[8] J. Liu, S. G. Rao, B. Li, and H. Zhang. "Opportunities and challenges of peer-to-peer internet video broadcast," in *Proc. IEEE*, 2007.

[9] M. Castro, P. Druschel, A Kermarrec, A. Nandi, A.Rowstron, and A.Singh, "SplitStream: high-bandwidth multicast in cooperative environments," in *ACM SOSP*, 2003.

[10] X. Hei, Y. Liu, and K. Ross, "IPTV over P2P streaming networks: the mesh-pull approach," *IEEE Communications Magazine*, 2008.

[11] D. Kostic, A. Rodriguez, J. Albrecht, and A. Vahdat, "Bullet: high bandwidth data dissemination using an overlay mesh," in *Proc. ACM SOSP*, 2003.

[12] F. Wang, Y. Xiong, and J. Liu, "mTreebone: A Hybrid Tree/Mesh Overlay for Application-Layer Live Video Multicast," in *Proc. IEEE Int. Conf. Distributed Computing Systems (ICDCS)*, 2007.

[13] M. Zhang, L. Zhao, Y. Tang, J. Luo, and S. Yang, "Large-scale live media streaming over peer-to-peer networks through global internet," in *Proc. ACM P2PMMS*, 2005.

[14] Z. Li, Y. Yu, X. Hei, and D. H. K. Tsang, "Towards low-redundancy push-pull P2P live streaming," in *Proc.of ICST QShine*, Hong Kong, July 2008.

[15] G.Tan and S. A. Jarvis, "A payment-based incentive and service differentiation mechanism for peer-to-peer streaming broadcast," in *Proc. Int. Workshop Quality of Service (IWQoS)*, Jun. 2006.

[16] J. Liu, Y.Shen, S. Panwar, K. Ross, and Y. Wang, "Using layered video to provide incentives in P2P live streaming," in *Proc. ACM Special Interest Group on Data Communication*, Aug. 2007.

[17] C. Buragohain, D. Agrawal, and S.Suri, "A game-theoretic framework for incentives in P2P systems," in *Proc. Int. Conf. Peer-to-Peer Computing*, Sep. 2003.

[18] W. S. Lin, H. V. Zhao, and K.J. Liu, "A game theoretic framework for incentive-based peer-to-peer live-streaming social networks," in *Proc. of IEEE Int. Conf. Acoustic, Speech, and Signal Processing (ICASSP)*, 2008.

[19] H. Park and M. van der Schaar, "Coalition-Based Resource Reciprocation Strategies for P2P Multimedia Broadcasting," *IEEE Transaction on Broadcasting*, Sep. 2008.

[20] Yu-Wei Chan, Chih-Han Lai, and Yeh-Ching Chung, "A Construction of Peer-to-Peer Streaming System Based on Flexible Locality-Aware Overlay Networks," *Innternational Journal of Pervasive Computing and Communications*, Vol. 6, No. 1, January-March 2010, pp. 104-124.

[21] Beverly Yang, Hector Garcia-Molina, "Designing a Super-Peer Network," *in Proc. of 19th International Conference on Data Engineering (ICDE'03)*, 2003.

Yu-Wei Chan received his B.S. and M.S. degrees in Information Engineering from TamKang University, Taiwan, in 1997 and 2001, respectively. He is a lecturer in the Department of Information Management, Chung Chou Institute of Technology, Taiwan. His current research interests include peer-to-peer computing, peer-to-peer streaming and game-theoretic resource allocation.

Research and Implementation of Three HTTPS Attacks

Kefei Cheng
College of Computer Science
Chongqing University of Posts and Telecommunications
Chongqing, China
chengkf@cqupt.edu.cn

Tingqiang Jia, Meng Gao
College of Computer Science
Chongqing University of Posts and Telecommunications
Chongqing, China
jiatingqiang@gmail.com

***Abstract*—With the rapid development of network applications, the issues of Network transmission security become very important. Therefore, SSL protocol is more and more widely used in a variety of network services. But the SSL protocol itself is not perfect, in practice, there are also problems. For the deficiencies of endpoint authentication in the SSL handshake process, the paper analyzes two kinds of defects existing in the SSL hand-shake process. Firstly, handshake process, in the first stage of the SSL connection, using plaintexts, existing the possibility of being monitored and tampered. Secondly, SSL deployment of the actual application. Because of considering the factors about the performance of the network connection, that usually uses the way of switch connection based on HTTP protocol. In response to these deficiencies, this thesis adopts the two ways of forged certificates and converting the data stream from HTTPS to HTTP to attack them. In addition, a new attack mode against the data stream of HTTPS is designed and implemented. Experiments show that the above three methods cause significant security risks to HTTPS communications. Therefore, taking a static ARP table, enhanced certificate mechanism and mutual authentication of three different measures are proposed to enhance network security in the paper. It is shown that three ways can relative effectively defense against attacks on HTTPS in the experiments.**

***Index Terms*—SSL, HTTPS, Man in the Middle Attack, Session Hijacking**

I. INTRODUCTION

With the development of e-commerce and cloud computing, SSL protocol is more and more widely used in all kinds of network services. SSL protocol by providing end to end authentication, message encryption, message integrity check and other security mechanisms protects the security of the communication process. For example, Yahoo ensures the security of the mail account through SSL, to protect the safety of the user e-mail account. Amazon shields the user transaction account and transaction security by it. In recent years, due to the development of cloud computing, the connection security between the client and the cloud is also an extremely important issue. December 2009, VeriSign announced, VeriSign will provide security and authentication services cloud-based computing for Microsoft Windows Azure platform. And Microsoft will use VeriSign SSL Certificate and VeriSign code signing certificate, to ensure the security of the cloud-based computing services and applications that being developed and deployed on the Windows Azure platform. Because, in the using of cloud computing model, the users` all computing resources are stored in the cloud, so network connection is essential for you to normally work. Therefore, if the server ends were safe enough, the security of network transmission would become very important. The SSL protocol is widely embedded in the client browser currently, the server-side also is relatively easy to deploy and implement, and the SSL protocol itself has good security features. At present, most bases of the cloud security applications are using the SSL protocol, many current hacker communities study SSL, so that the security issues of the SSL protocol are more and more concerned.

But SSL is not perfect in the practical application, there still exists the possibility of middle attack. 2003, in the literature[1], Peter Burkholder studied the defects of SSL handshake and verified the possibility of SSL attack, using the way of hijacked session of the typical middle attack to deceive the client to achieve the attack. In the year of 2009, Michael Howard, in the paper[2], showed to carry out SSL attacks by the use of tools based on the Webmitm, which is a tool for SSL attack, and eventually received the data after decryption. The paper[3]shows, in the International security conference in 2009, Moxie brought forward that the HTTPS to the HTTP transfer connections in the practical application would have security problems, causing link substitution attack, but also the client has not traditional security warning signal, so there is a serious hazard. In the literature[4], basing on the analysis of SSL attacks, put forth to improve the way of the browser interaction design to help users find the presence of attack, thereby enhancing safety. In the paper[5],according to the feature about SSL attacks often based on LAN-deceived, proposed ARP deception defense tool designed by their own to enhance LAN security, this is also obtained better results by the comparing experimental. However, there is a problem that this method is not for the SSL protocol itself. In the paper [6] being similar to [4], improved the

doi:10.4304/jnw.6.5.757-764

security of the parties through upgrading the browser interaction interface.

This paper is on the basis of analyzing the defects of the SSL handshake, mainly do the following work:

First of all, the attack experiments were carried out in three ways. The first method is by use of the characteristics of plaintext in the handshake phase and the defects of client certificating authentication, hijacks sessions through ARP deception and DNS spoofing within the LAN, then the attacker fakes server-side certificate by a proxy way to attack, he can decrypt all the https data stream. The second one is that using the deployment defects of SSL in practical application, after by ARP tricked, forwarding all client traffic, and then monitoring HTTP messages, through altering the https address in return HTTP packets so as to achieve that the client is still working on the http connection, and the attacker maintains https connection with the remote server. An attacker sends directly plaintexts to the client in the way of http by decrypting encrypted data, so he can access all communication data. The last one, SSL intercepted procedures working on the embedded device are designed and implemented, the device is accessed the network in the series way, processing the passing SSL traffic, which mainly achieved to tamper the certificate. Because it uses the defects of the client certificate authentication, so once the client completes the recognition of the certificate, the equipment can also complete the decryption of the passing SSL data.

In the experiments, it was analyzed and compared to the specific course, dangers, feasibility of the several attack methods above, and carried out the corresponding packet user testing, so as to get the probability of success of various attack methods.

What`s more, based on the analysis on the attack, it was put forward to three kinds of defensive methods. They were static ARP table, EVSSL certificate and two-way certificate authentication, and analyzed the advantages and disadvantages of various defense methods. The results showed that, EVSSL was a more effectively preventing SSL attack way in practical applications.

II. HTTPS HANDSHAKE AND DRAWBACKS

A. Basic SSL Handshake

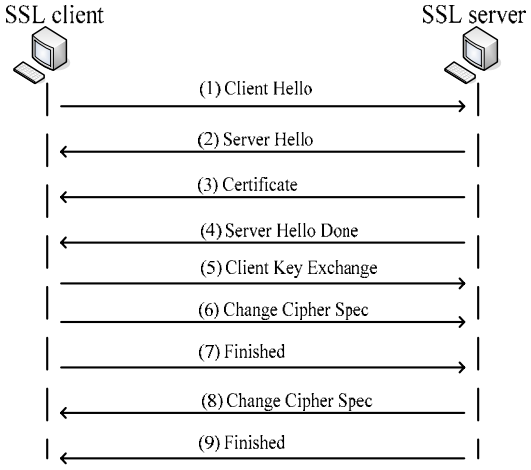


Figure 1. SSL handshake proess

The first version of SSL Protocol is SSL3.0, because the wide range of applications, it is included in the Internet standards by the IETF, and it becomes the TLS. Currently, the version of SSL protocol widely used is TLS (SSL3.1), SSL handshake process is clearly defined in RFC2246[7], as is shown in the Fig. 1 above.

1) Client sends ClientHello: The client initiates SSL request message at first to show the beginning of an SSL handshake by sending ClientHello message. ClientHello packet includes the packet length, SSL version, a random number generated by the customer's key, sessionID, all the cryptographic suite which the client browser can support, compression method which may be adopted.

2) Server sends ServerHello: The server receives the ClientHello message sent by the client, then returns the ServerHello to the client basing on the message, which includes SSL version, the SSL version the server can highestly support, the random number key generated by the server, sessionID set by server, the communication encryption package which is selected from all encryption package of ClientHello by server, compression method which may be adopted.

3) Server sends its certificate: The server sends its own Authentication certificate chain, which probably includes server itself and the higher certificate. The certificate includes its version, serial number, signature algorithm, the information of the certificate issuer, effective time, information of the certificate's subject, the public key, extended domain, the signed value of the all certificate.

4) Server sends ServerHelloDone: The server shows that its all authentication informations have been sent to the client completely.

5) Client sends ClientKeyExchange: After the client passes the authentication of the certificate that is received from the server, the client will generate a pre-master key, and makes use of the server certificate's public key to encrypt the pre-master key, then the encrypted pre-master will be sent to the server in the form of client key exchange packets.

6) Client sends ChangeCipherSpec: The client shows that the next communication packets will be encrypted by the consultative suite of encryption key, and the client generates the master key used in this communication through the pre-master key and the information exchanged previously.

7) Client sends Finished: handshake finish message. The client show that the finished message in handshake phase, the main contents of which are the hash value of all previous the handshake messages and are sent after encrypted through the communications master key generated by (6).

8) Server sends ChangeCipherSpec: The server shows that the next communication message will be encrypted through the negotiation key suite. The server generates the master key of this communication through pre-master key and the previous exchanged information.

9) Server sends Finished: The server shows that the finished message in handshake phase, the main contents of which are the hash value of all the previous handshake messages(including the clients' finished message),and are sent after encrypted through the master key of communication generate by (7).

After these steps, the communication messages of the two sides are both encrypted by means of the consultation symmetric key, therefore, all subsequent data streams are ciphertext flow, so that this ensures the security of communications content.

B. Anlyasis of the drawbacks

1) Certificate authantication

The data content is plaintext in SSL handshake, so attacker can obtain data packet by monitoring and tampering to it. It makes the hijacking attack to SSL session possible. Client authenticates the server by step 3 which transfers server certificate, but the certificate is transmitted in plaintext which can easily be intercepted and tampered. When the attacker intercepts and obtains the server certificate, it can forge its own fake certificate, and then fill the information with server's, inserting its own public key into certificate, signature and send the forged certificate. Once users accept the fake certificate, it means that the attack is successful. The attacker can decrypt the premaster key that client sends, there is no secret to attacker. Therefore, steps 3 has defect in the SSL handshake.

2) HTTPS connection initiates by HTTP

According to analysis of users' habits and the practical applications of HTTPS, HTTPS request will be initiated by the following two ways:

a) *Users' habits*

When user accesses HTTPS sites by web browser, it usually types directly the URL without https head in the address bar, such as: www.xxx.com. If there is no protocol head in URL, the browser will use the HTTP protocol to connect the site. When client initiates HTTP connections, but the server is the HTTPS site, it will return messages of HTTP redirection. The content in this packet contains the actual HTTPS address, such as https://www.xxx.com. The client receives this packet, and browser will be re-launched to initiate HTTPS connection. Compared with directly typing https://www.xxx.com in client browser, this redirection will be no difference.

b) Application in practice

With HTTP connection, there are some buttons on the page to initiate HTTPS connections. For example, when you want log in personal account of E-mail, you will click on the submit button to transmit your ID and password. When you click on the button, the client initiates HTTPS connections to protect confidentiality of personal information.

Case a) is because of the habits of users, they don't pay attention to the difference between http and https in the URL. Case b) is because of consideration of the overhead of SSL handshake, only the important information has encrypted, not the whole data in the connection. In general, websites don't use HTTPS connection in the whole process, because HTTPS connection is usually 2 to 100 times slower than HTTP connection[8]. Therefore, the submission of confidential information (such as ID, password) is by HTTPS connection, and other services are still by HTTP connection. In this way, the delivery of HTTPS URL is in the HTTP message content, while the inherent insecurity of HTTP protocol, so it results in security vulnerabilities.

III. IMPLEMENTATION OF ATTACK AND RESULT

A. Implementation of attack

1) Implementation of the first method

Because of data transmission in plaintext in SSL handshake, the attacker can forward data by ARP Spoofing between the client and server. The attacker intercepts HTTPS requests of client, and connects with the server itself. When the server sends the certificate for authentication, the attacker can forge certificate which is a self-signed, and then sends it to the client. Users tend to choose to accept the forged certificate, so that the attacker has successfully set up SSL communication link. The attack has following steps:

a) Conducting ARP Spoofing, it can make the attacker between the server and client. Then it can forward packets between the server and client.

b) Carrying out DNS Spoofing, and listening to the port 443. After that, the attacker can easily establish connection with the client.

When receiving the SSL request from the client, the attacker accepts the TCP request and initiates SSL handshake with the server.

c) After successful connecting with the server, the attacker forges its self-signed certificate, replacing the public key with its own. And then accepting the client's SSL request, the fake certificate has sent to the client.

d) When the client receives the fake certificate, the browser will have an alert dialog which tells the user that the certificate is suspicious. Most users tend to accept the certificate, and it means the attack is successful. The client and attacker have established SSL connection.

e) Because of successful connecting with the client and the server respectively, the attacker can receive encrypted data from one part, decrypt it by the key of one part, encrypt it with the key of the other part, and send to the other. Hence, all the data between the client and the server is hijacked by attacker.

2) Implementation of the second method

Similarly, the attacker can forward data and monitor the HTTP packet by ARP Spoofing between the client and the server. When the packet content sent from the server have HTTPS address for redirection, the attacker tampers the content, and sends to the client. At the same time, the attacker initiates HTTPS handshake with the server. If the attacker receives the HTTP request from the client which receives the tampered content, it decrypts the data in the ciphertext receiving from the server, and sends it to the client. Because of HTTP connection between the client and attacker, there is no alert dialog in the browser of the client. So this method is better deceptive. As the Fig. 2 and 3 below shows, and the attack has following steps:

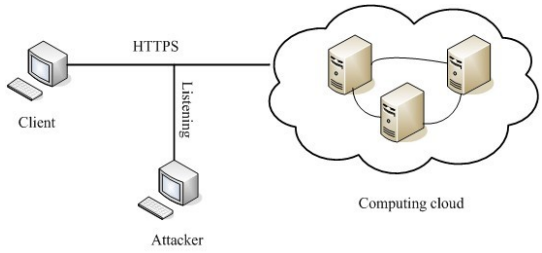


Figure 2. Attacker listen to the connection.

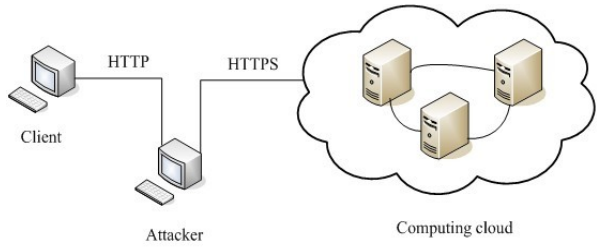


Figure 3. The attack 2 in the second method.

a) Conducting ARP Spoofing, it can make the attacker between the server and client. Then it can forward packets between the server and client.

b) The attacker monitors the HTTP data between the server and client.

c) When the attacker receives the data content from the server has <a href="https://...">, replacing it with <a href="http://...">. If the head of HTTP packet has "location:https://…", replacing it with "location:http://…", making a record about the addresses which have been tampered before.

d) If the attacker receives the HTTP request to the address which is in the record, it connects with the server by HTTPS.

e) After establishing the HTTP connection with the client, the attacker decrypts the ciphertext of HTTP from the server, and sends the plaintext of HTTP to the client. In the browser of the client, there is no difference with the normal HTTPS connection, except the URL. In normal way, the URL begins with https. But after attacking, it begins with http.

3) Experimental result

Experiments are conducted in the 100M Ethernet, and the browser of the client is IE 6.0. The client machine runs on the windows xp, whose IP address is 192.168.1.2. The attacking machine runs on the Linux whose kernel is 2.6.11, and IP address is 192.168.1.3.

a) The result of first method

A program is designed to carry out ARP Spoofing and DNS Spoofing. With the OpenSSL library, two modules are designed to tamper certificate and forward SSL data. The result is below:

Figure4. Alert dialog after attack

In the browser of the client, this alert dialog will appear, the Fig. 4 above showing. If clicking on the yes, then the attack is successful. We can obtain all the information in the communication between the client and server.

b) The result of second method

The same program with that above is used to carry out ARP Spoofing and DNS Spoofing. With the OpenSSL library and sslstrip of Moixe [9], the attack is designed and verified. The result is shown in the Fig. 5 and 6 below:

Figure 5. Normal HTTPS connection

Figure 6. Connection after attacking

There are two distinct differences between two pictures. The one is that normal URL begins with https, but after attacking it is http; the other one is that normal connection will have a lock icon under the browser, but after attacking there is not.

B. *Implementation of the thrid attack*

The principle of this attack is similar to the first attack method, they is by means of defects of a Certificate of the SSL handshake. Method 1 is only used in the same LAN with only one switch connected, or ARP deception can not be implemented, so the data stream of the client can not be forwarded. An experiment is designed to run SSL attack program on the embedded ARM development board, which has two Ethernet interfaces, being cascaded in the network, and does not be configured IP address in running, so that this does not change the local network topology. This board has two Ethernet interfaces eth0 and eth1, the

role of the basic forwarding procedure is as follows: the eth0 receives datagrams with no treatment, sending directly them from the eth1. The eth1 receives datagrams in the same way, sending directly them from eth0. When this basic forward program runs on the equipment, then this can achieve the function of forwarding completely. The device itself has not IP address, when being accessed networks, it will not affect the topology. The specific topology is shown by Fig. 7 below.

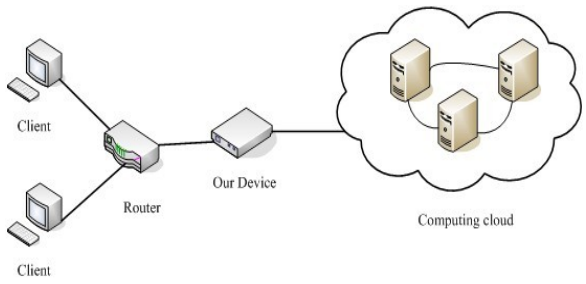


Figure 7. specific topology.

As is shown above, the device is cascaded with the exit of the router, so you can forward all the datagrams of the network.

In the process of forwarding the program, judge whether the datagrams are the SSL packets, if so then further processing, or transmitting directly. Take the connection of HTTPS for example below, HTTPS port number is 443, so the forwarding program can determine whether those are the HTTPS packets only by the 443 port.

When the HTTPS datagrams are received, it can be determined whether that are SSL handshake or SSL data messages, basing on the value of the SSL header fields. Here, the program only deals with the tampering of the SSL messages in the handshake phase.

According to the analysis of the principle of the SSL handshake certification, the process in a single handle HTTPS session is as follows:

1. Clienthello messages, record the random number, sessionID.

2. Serverhello messages, record the random number, sessionID and key package.

3. Certificate messages, tamper the last one of the issuer of the certificate messages to the '/ 0', and generate a pair of public and private key, use the generated public key to replace the public key of the certificate, and use the public key to sign for the whole certificate.

4. Client key exchange messages, decrypt messages by the generated private key, receive negotiation pre-master key, then the generated public key is used to encrypt.

5. Client finish messages, because of getting the pre-master key, it can generate the primary communication key by means of previously recorded information and the pre-master key. Decrypt the finish messages by the master key, generate the value of the hash by means of the actual communication packets having been tampered, and then the master key is used to encrypt.

6. Server finish messages, as is shown in the 5 above, the hash of the server finish messages is generated, as according to the stipulated rules of the agreement, but its generation need to the record information after distorted, and the client finish messages having been generated together, then encrypt by the master key.

The communication after the treatment above, because of the certificate information being tampered, the client will also have the warning dialog of the certificate authentication. Once the user confirms the authentication to the certificate, it is normal connection between the server-side program and the client browser, and the data can be transmitted by that. We use the generated master key to identify the SSL data messages transfered, and then obtain the content of the plaintext by decrypting.

a) The results of the third method are as follows:

The forward procedure is designed by the libpcap and libnet library. Because there are not only the functions relating to SSL in the OpenSSL library, but there also are a large number of cryptography API, with their help, you can easily implement encryption, decryption and the operation of hash calculation. And so we use the OpenSSL library to implement the processing procedures of the SSL packets.

The experimental platform is the development board with the Intel ixp435 network processor, which has four LAN interfaces, and a WAN port, only one LAN and a WAN port are used in the experiment. And Linux kernel version 2.6.16, the client environment is the same with the previous experiment.

The same structure with the above one is adopted in the experiment, the device is stringed to the local switch export.

The results are the same with the first method of attack, as the Fig. 8 shows below.

Figure 8. Alert dialog after attack

Similarly, the client browser will pop out the certificate warning dialog to prompt the authenticity of the client certificate. If the customer clicks "Yes", then the attack would be successful, and then all data packets of the communication can be gained by decryption.

b) Disscusion

Because the device itself, accessed in series in the network, will forward the data stream of all the machines coming from the LAN, then the operation load of the equipment may be very big, so the speed of the whole communication may be affected. Especially, in the case of a large number of hosts in the LAN, this phenomenon may be more significant. Moreover, the forward program designed by the libcap and libnet library, as it mainly working in the user layer, so that will also have a greater impact on the performance.

C. Analysis and comparison about three methods

1) T he preparetion for attacks

The similarity between the first and second method is both of them must use the ARP Spoofing. The third method does not depend on it. With the development of Antivirus Software, many of them can detect the ARP Spoofing, and even some can deal with it in special method, except the third method.

2) Protocol with each side

In first method, the attacker establishes both SSL connections with the client and the server. In the second method, the attacker establishes SSL connection with the server but HTTP connection with the client. But in the third method, it doesn't establish connection with each side. The device just modifies the certificate and the handshake message for the following communication. The connection after attack is the same to the normal connection. The device just forwards the SSL encrypted data, if needing, it can decrypt the encrypted data.

3) Analysis of attack

The first and third method monitors the HTTPS request, after tampering the certificate, there is a dialog in the browser. If users accept the fake certificate, all the following are the same as the normal connection. On the contrary, the second one is different. It monitors the HTTP streams, and tampers redirection of https address. After attack, the browser of the client will have no alert, because it is actually a HTTP connection. Though it seems perfect, it has two distinct differences with the normal connection. One is the head of URL, and the other is the lock icon on the browser. If users are aware, the attack fails.

4) Harmfulness

In the first and third method, the success depends on the vigilance of the users. Only when the user clicks on the yes button, the attack can be successful. But in the second method, because of no alert dialog, users can't discern easily. Especially, a few of E-mail system only conduct HTTPS connection to send ID and password, after verifying them quickly redirect to the HTTP connection. In this case, it seems hardly note the existence of attack.

In the test experiment, 30 people through the client computer verify the harmfulness. The statistic data is shown in the TABLE I below:

TABLE I. REASULT OF ATTACK.

Result	Number	
	successful	***failed***
Method 1	20	10
Method 2	30	0
Method 3	21	9

Experimental results show that the number of the success of the second method is more than the first and third method, it is more harmful.

5) Difficulty

Method 1 and Method 2 can be completed by programming in the local machine, so the implementation does not require attached factors. But method 3 will need the programs to be run on the development board, the runtime library need to be migrated to the ixp435 platform, and it's relatively more complicated to implement. However, given the host owning the firewall with the ARP client protection function, so the third method will be completely free to consider the impact of such factors, but once the ARP deception fails, there is not only the warning of the ARP attack in the client, but also the user may sense to attack there. Then the software will prompt the user to take appropriate measures, such as prompting the user to use a static ARP table, and sometimes the first and second methods may also not succeed.

In the test experiment, 30 people use the client computer which has been installed special firewall that can defeat ARP Spoofing. The statistic data is shown in the TABLE II below.

TABLE II. REASULT OF ATTACK

Result	Number	
	successful	***failed***
Method 1	15	15
Method 2	18	12
Method 3	22	8

Experimental results show that the success of the third method is more than the first and second method. Because the first and second method depend both on the ARP spoofing, so their results seem similar in this test.

IV. PROTECTION SCHEME

A. Static ARP

In the first and second method, at the beginning of attack, the need for ARP Spoofing is in order to deceive the client and the server for forwarding packets, so effectively preventing ARP deception is essential to prevent from HTTPS attacks. For most users, manually configuring a static ARP table is the most effective way [10]. Because IP address and MAC address in the static ARP table are fixed pairs, so it can prevent that fraud correspondence is added into the ARP table by ARP Spoofing. The experiment is conducted on a machine with windows xp sp3 configuring a static ARP table, the ARP deception fails. It means the static ARP table can effectively prevent the ARP cheating. But the disadvantage of this method is that it is not flexible to the large network, and changes need to be adjusted according to the topology. And the static ARP table can not work effectively in the third method.

B. EVSSL certificate

EVSSL certificate is a new stringent authentication standard of SSL certificate, which is the world's leading digital certificate authority and main browser developers to formulate. The browser can identify security EVSSL then display green in the address bar, so ordinary users can be confident that the visiting website is the entity that strictly authenticated by the authority. All digital certificate authority that issuing EVSSL certificate must

follow the uniform criteria to carry out strict authentication, while the browser can identify EVSSL certificates makes the address bar turn green. For the first and third method of attack, browsers will have an alert dialog to tell the users; for the second method of attack, because the connection between the client and attacker is HTTP connection, the address bar will not turn green, it also can't check specific security information from the browser, so users can realize the connecting problems and conduct timely defense. Disadvantage of this approach is that only the newer versions of browsers can support EVSSL features, the general browser does not support it. In addition, it depends on the user's own safety awareness whether the color of the address bar can be noted.

Using two machines as the client machines, attacking in the second method, people involved in the experiment is divided into two groups, each group has 20 people. The client machine of first group is without the browser of EVSSL supported, and the second group uses the browser that supports EVSSL on the machine. The experimental results are in the figure 9 below. In comparison, it indicates that because of the EVSSL certificate, the success rate of attack decreases by 34%, effectively enhancing the security.

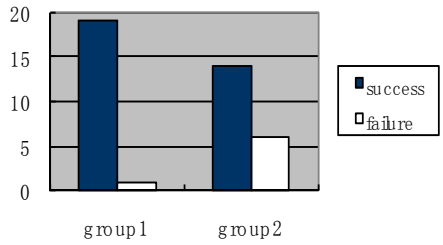


Figure 9. Comparison between two groups.

C. Two-way authentication

In RFC2246, a full SSL handshake includes not only the server authentication, but also the client authentication. But in applications, there is just server authentication, and the two-way authentication is conducted only when the server requires. This can not only speed up the network service, but also reduce the overhead. If the server is set to require authentication to client certificate, client must send an authorized certificate to the server. SSL connection can be successfully established after both sides are certified. In the first and second attack, because attacker can not obtain a legitimate certificate, so it is difficult to establish normal HTTPS connection with server. In the third attack, we can also design program to modify the client certificate. But the server will detect the fake certificate, so the connection will be terminated by the side of server.

In the local area network, we build a HTTPS site of self-configuration which requires client certificate, then attack in the second method. Experiment result show that attacker fails to establish HTTPS connection with the server. Disadvantages of this approach are that two-way certificate authentication requires the client having a legal certificate issued by the authority, the management of client certificate will be complicate. In addition, two-way authentication will have more encryption and decryption operations. The speed of connection will slow down and overhead of server will increase.

V. CONCLUSION

Experiments show that three methods of attack on the HTTPS session are feasible. In normal, connection speed of HTTPS services is 2-100 times slower than normal HTTP connection, users will not be aware of attacks even if the delay caused by the change of link. Because users usually don't care about the alert in the browser, when attacking in the first and third method, the majority of users will tend to accept a warning certificate even if the alert dialog; when attacking in second method, the user will not be aware because the little difference between the normal and attacking pages. In our experiment, using EVSSL certificate and two-way authentication are both effectively to avoid the attack. Configuring a static ARP table can avoid attack in first and second method How prevent man-in-the-middle attacks on HTTPS session more effective is the next focus of our study.

ACKNOWLEDGMENT

This paper is sponsored by the major science and technology programs of the information industry of Chongqing key technologies R&D programs -Smartphone R&D and industrialization (CSTC2010AB2003), Science Foundation of Chongqing Municipal Education Commission (KJ20090511), and the Cloud Computing project of College of Compute Science, Chongqing University of Posts and Telecommunications (JK-Y-2010005). We thank them for their support, and we are also grateful to the reviewers for fruitful comments.

REFERENCES

[1] Peter Burkholder, "SSL Man-in-the-Middle Attacks", SANS Institue InfoSec Reading, 2003.

[2] Michael Howard, "Man-in-the-Middle Attack to the HTTPS Protocol", IEEE computer society, 2009, pp.78-81.

[3] Marlingspike Moixe, "New Tricks For Defeating SSL in Practice", BlackHat Conference, USA(2009).

[4] Haidong Xia and Jose Carlos Brustolonl, "Hardening Web Browsers Against Man-in-the-Middle and Eavesdropping Attacks". Proc. 14th Int'l Conf. World Wide Web(IW3C2), ACM Press, 2005, pp. 489-498.

[5] Thawatchai Chomsiri, "HTTPS Hacking Protection", 21st International Conference on Adcanced Information Networking and Applications Workshops(AINAW), 2007.

[6] Andre Adelsbach and ebatian Gajek, "Visual Spoofinging of SSL Protected Web Sites and Effective Countermeasures", Proceedings of the 1st Information Security Practice and Experience Conference, Singapore, 11-14 April, 2005.

[7] T.Dierks and C.Allen, The TLS Protocl, IETF RFC 2246, 1999; www.ietf.org/rfc/rfc2246.txt.

[8] S.Thomas. SSL and TLS Essentials. New york: Wiley Computing Publishing, 2004.

[9] Moxie,Ssltrip.http://www.thoughtcrime.org/software/sslstrip/.

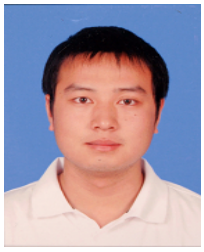
[10] Somnuk puangpronpitag and Narongrit Masusai, "An Efficient and Feasible Solution to ARP Spoofing Problem", IEEE Electronics, Computer, Telecommunications and Information Technology (ECTI) Association conference, Thailand, 2009.

Kefei Cheng Chongqing, China, 1974.3. Received Master Degree in computer application from Chongqing University of Posts and Telecommunications(2000), Ph.D in engineering in the field of computer software and theory at Chongqing University(2005).

He is working in the department of Computer Science and Technology, Chongqing University of Posts and Telecommunications from 2000. He has published more than 20 papers in various conferences and journals, and directed two funded projects on networking security. His current research interests are in the areas of computer network security and high performance computing.

Dr. Cheng, is a member of China Computer Federation(CCF).

Tingqiang Jia Henan province, China, 1985.2.Received Bachelor Degree in information management and system from Xinyang Normal University(2009). Now, he is a master student at Chongqing University of Posts and Telecommunications, who mainly study the security of compute network.



Meng Gao Hubei province, China, 1986.9.Received Bachelor Degree in information security from Hubei university of police(2008). Now, he is a master student at Chongqing University of Posts and Telecommunications, who mainly study the security of compute network and embedded systems.

Multi-tier Grid Routing to Mobile Sink in Large-scale Wireless Sensor Networks

Zujue Chen and Shaoqing Liu
Computer Science and Communication Engineering, Jiangsu University
Zhenjiang Jiangsu, China
Email: chenzujue@126.com, king8411twins@yeah.net

Jun Huang
School of Information Technology and Engineering (SITE), University of Ottawa,
Ottawa, Ontario, Canada
Email: steedhuang@hotmail.com

Abstract — **Improving the efficiency of data dissemination algorithms and protocols to mobile sink remains as an interesting research and engineering issue, especial for large-scale wireless sensor network. As the node energy and resources are limited, theses protocols should meet energy-efficiency, low delay and high delivery ratio requirements. Although an energy-efficient dissemination tree (d-tree) can be constructed with sink mobility, the delayed handoff could lead to suboptimal routing trees for a substantial amount of time as the network grows larger, and also the efficiency heavily relied on the location of the sink, thus it may offset the load balance resulting from hierarchical tree structure. In this paper, we propose an energy efficient routing protocol MGRP (Multi-tier Grid Routing Protocol) which introduces a special hybrid multi-tier structure for data dissemination. MGRP divides the observation areas into square grids, could be different size. Within each grid, we form an optimized cluster which transmits reliable data to its higher tier cluster head, the uppermost cluster head from neighbor grids further negotiate to construct the data d-tree from which the mobile sink can access and send query. Through intensive simulation in a given mobile sink experiment, MGRP performs better in terms of energy consumption and average delay respectively compared to previous protocols such as TTDD (grid-based), SEAD (tree-based) and COSEN (chain-based), under practical conditions, where sensors may die out, but with initial large size.**

***Index Terms*—Wireless sensor network; routing protocol; energy efficiency; sink mobility**

I. INTRODUCTION

Wireless sensor network is made up of a set of sensor nodes and sink nodes which has significant application in weather forecasting, military target tracking, medical monitoring, and environmental detection [1][2][3]. Sensor nodes transmit the data through multiple hops to the sink which is located far from the target. However, sensor nodes are limited by the lifetime of battery, as such, how to effectively save the energy of battery and prolong the network lifetime has been the important research and engineering issue in wireless sensor network.

This paper studies the problems of hierarchical data dissemination with high-speed moving sink in large scale sensor networks. We define a source as a stimulus that generates an interest event over a vast field. This application prefers the occasion in geological exploration where earth crust status associated with earth quake and/or potential gas and oil basin is probed by low-flying aircraft or mobile vehicle. The sensors attaching to the surface of earth crust can detect the soil types and vibration patterns, aggregate the data and report to the mobile base station. In a battlefield the sensors can also be used for the detection of land mine, thus enables the safe removal of landmines in former war zones, reducing the risk of soldiers involved in the removal process. Similarly the same technology can also be utilized to provide an early warning system for flood prone area where access is difficult or expensive; we have seen such applications both in China, Canada and USA. This paper is trying to extract our experience on interesting common topic from various projects, and share with our research community.

Sensor nodes have many modules of which the communication module consumes the most electricity. The cost of communication is closely related to the way of routing protocol is set up. If routing protocol operates efficiently, the energy that every node consumes would be reduced. Therefore, it's desirable that the routing protocols should resolve issues like energy efficiency, load balance, minimum delay, etc. Especially, energy efficiency is the key issue for keeping a longer network lifetime.

Various energy efficient protocols have been proposed such that the energy for data transmission can be evenly consumed in the whole network to extend the network lifetime, under fixed sink situations. In [4], Zen et al proposed GREES-L and GREES-M methods, which combine geographic routing and energy efficient routing techniques and take into account the realistic lossy wireless channel condition and the renewal capability of environmental energy supply when making routing decisions. Both GREES-L and GREES-M exhibit graceful degradation on end-to-end delay, without compromising the end-to-end throughput performance. In [5], Koutsonikolas et al present HGMR (Hierarchical

doi:10.4304/jnw.6.5.765-773

Geographic Multicast Routing), a new location-based multicast protocol that seamlessly incorporates the key design concepts of GMR [6] and HRPM [7] and optimizes them for wireless sensor networks by providing both forwarding efficiency (energy efficiency) as well as scalability to large networks. In [8], Cheng et al analyze a network model which has Query and Response packets travel different routes to address the problem of efficient data collection in wireless sensor networks and propose an efficient Query-Based Data Collection Scheme (QBDCS). In order to minimize the energy consumption and packet delivery latency, QBDCS chooses the optimal time to send the Query packet and tailors the routing mechanism for partial sensor nodes forwarding packets with minimum energy consumption and delivery latency.

In [9] Lou et al propose a routing protocol suitable for networks with one mobile sink. The sink visits certain anchor points in the network area and remains still while collecting data at each one of them. The sink samples the global energy consumption of all nodes while stationed in an anchor point. It uses this data to create power consumption profiles and calculate the optimal resting time at each anchor point. Another approach that optimizes the sink's trajectory is presented in [10]. Ma et al assume a mobile sink that initially follows a linear trajectory and collects network information from the sensors. The sink uses this information to break its trajectory into separate line segments that are closer to the network nodes, thus minimizing the data propagation cost. However, collecting network knowledge incurs a significant overhead on the sensor nodes.

In this paper, we propose a Multi-tier Grid Routing Protocol (MGRP) which uses a special hybrid multi-tier structure to effectively restrict the transmission distance between sensing grids and enhance the scalability between adjacent grids thus avoid energy consumption growing exponentially. The lower tier is divided to observation squares where energy efficient cluster head (CH) is formed. One sensor node is elected as a CH in every grid based on the maximum residual energy. Among all CHs a dissemination tree (d-tree) is further constructed for higher tier based on gradient broadcast [11] where the mobile sink have access to the root and send queries. The distributed transmission from these hierarchical tiers plus trees structure could effectively reduce the average delay for each round and balance energy consumption over the whole network.

The rest paper is organized as follows: Section 2 reviewed and discussed some previous related works that motivate our research; Section 3 describes the network model and explains energy consumption problems; Section 4 presents the implementation of our MGRP protocol in detail; Section 5 studies on the simulation process and analysis of the results obtained. Section 6 concludes this paper and point out some future work.

II. Related Prior Works

Most of the protocols can be classified as either data-centric or hierarchical or location-based. In this paper, we only review three classic routing protocols which contributed to the design of our work.

Two-Tier Data Dissemination (TTDD) [12] mainly resolves situation of multiple mobility sink nodes. It solved excessive energy consumption issues due to the global flood caused by direct diffusion. In TTDD, the source node which was defined as the origin of the incident built an entire a grid into the environment when the sink is defined as the terminal to the observers. When the event of source is triggered, the source node will send this information from its nearest dissemination node, which is called the immediate dissemination node, and thus the other ordinary nodes which locates near the dissemination area will continue to relay this message until the entire network is aware of this incident, so that all the dissemination nodes of the grid could record the information of this event sent from the source node. As the sink should locate within a grid or boundary of any two grids, there must be a dissemination node around the sink. When the mobile sink node needs information, it only need to do the local Flooding within the local grid and obtained acquisition of data in order to save the energy. A reverse transmission path is built back to the sinks based on data trace after the sink sent queries, so the overhead for controlling the transmission and keeping track of the sink is limited to the local grid. TTDD carry most traffic that fix on the virtual grid and caused intensive energy consumption on the dissemination path.

The SEAD (Scalable Energy-efficient Asynchronous Dissemination) [13] protocol, provides a recursive algorithm to search for the minimum energy transferring trees, as well as the conduct of energy conservation and management for the mobile sink. The sink doesn’t need to report the current location to the delivery tree. As the sink moves, no new access nodes is chosen until the hop count between the access node and sink exceeds a threshold which allows tradeoff to be made between the energy consumed to reconstructing the tree and the path delay. In the midway of transferring data to the sink node SEAD will establish a temporary path and will minimize the energy consumed on the way to the root of tree. The dissemination tree (d-tree) is established from a source to different access nodes. However, the delayed handoff in SEAD could lead to suboptimal routing trees which also consumes so much energy to complete the tree construction meanwhile offsets the load balancing effect resulting from sink mobility, especially in the large scale sensor network where event frequently happens.

COSEN [14] is a two-tier hierarchical pure chain-based routing protocol in which path calculation and cluster setup are both carried out by sensor nodes themselves. COSEN builds a path based on two-tier chain structure and operates in two phases: chain formation and data transmission. For chain formation, each low-level chain is formed with fixed length based on greedy algorithm. The chain leader is elected based on maximum residual energy in each chain. All low-level leaders then connect into a high-level chain and one low-level leader is elected as a high-level chain leader the same way as low-level chain does. In data transmission, each sensor

nodes sends the fused data via their respective low-level leaders and the high-level leader, toward the sink which is the root of entire path. Although COSEN can alleviate the transmission delay and energy consumption compared to TTDD and SEAD, it still introduces a lot of redundant transmission paths, especially for those nodes which are nearest to the sink but would detour their fused data toward the farther leaders.

In our work, we propose a hybrid Multi-tier grid based routing protocols which combines the merits (chain, tree and grid which we call cluster) of above three mentioned protocols, and it will get better performance than TTDD, SEAD and COSEN both in energy efficiency and average delay.

III. PRELIMINARIES

Let's consider a network of N sensor nodes randomly distributed over the observation area of two-dimensional space to periodically collect data. Some assumptions are made according to the network model:

1) The sink node locates near the outer side of the square region. Sensors are stationary and sink is mobile.

2) All nodes are homogeneous and have the same capabilities.

3) Links are symmetric. A node can compute the approximate distance to another node based on the received signal strength, if the transmitting power is given.

4) All nodes are aware of their geographical coordinates.

The first order transmitting and receiving radio model is used in this paper. The model consists of transmitter circuit, amplifier circuit and receiver circuit. Both the free space (d^2 power loss) and the multi-path fading (d^4 power loss) channel models are used in the model, depending on the distance d between the transmitter and receiver. The energy spent for transmission of a k-bit packet over distance d is:

$$E_{Tx}(k,d) = E_{Tx\_elec}(k) + E_{Tx\_amp}(k,d) = \begin{cases} E_{elec} \times k + \varepsilon_{fs} \times k \times d^2, & d < d_0 \\ E_{elec} \times k + \varepsilon_{mp} \times k \times d^4, & d \geq d_0 \end{cases} \quad (1)$$

and to receive this message, the radio spends the energy:

$$E_{Rx}(k) = E_{Rx\_elec}(k) = E_{elec} \times k \quad (2)$$

Where $E_{TX}(k,d)$ denotes the energy consumption for transmission, $E_{RX}(k)$ denotes the energy consumption for reception, E_{RX_elec} and E_{TX_elec} denotes power of transmitter and receiver circuit respectively. E_{TX_amp} is the energy consumed by amplifier. A sensor node also consumes E_{DA} (nJ/bit/signal) amount of energy for data aggregation.

IV. HIERARCHICAL GRID ROUTING

A. Grid Encoding

For most sensor networks, the distance between neighboring nodes largely affects the energy consumption. The farther are the neighboring nodes, the higher energy consumption grows exponentially. The MGRP proposed in this paper utilizes multi-tier grid to restrict the neighboring transmission distance between cluster head in each sensing grids. The coordinate space is completely logical and bears no relation to any physical coordinate system. The entire sensing area is dynamically partitioned according to preset requirements.

In order to distinguish between different tiles of grids, we encode a tile to binary code. Each tile will be given an address as addr (tile binary) by which a tile can be identified. The addr (tile binary) is a bit stream formed by 0 or 1 according to classification of different grid and is defined as follows. The length of the address is decided by the hierarchy it takes which is defined as $\log_2 4^n$ (n>1). Suppose Tier Level n (TLn) is equipped with 4^n grids. For top TL0, it can have any grids according to preset requirements, so the address is denoted as $G_{i,j}$. TL1 is taken to have 4^1 grids, so the length of address is $\log_2 4^1=2$, as for the TL2 is $\log_2 4^2=4$, etc. We can see from the code framework that the underlying code contains the belongingness relative to the upper tier in a recursive way. For an a grid from TL3 with a length of $\log_2 4^3=6$, e.g. addr(001011), we dissect it into three element as Fig.1.

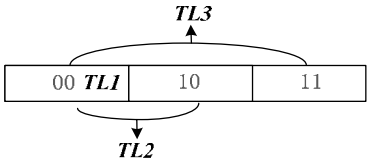


Fig.1. Example dissection of ***addr****(001011)*.

As can be seen from the Fig.1, a tile encoding of TL3 covers position information of TL1 and TL2. Let's map TL3 and TL2 to TL1 and thus a tier with 3 embedded tiers make up of a whole tile mirror chart from which we can see more clearly details of the relationship between different tiers. Meanwhile, our approach of addressing can also estimate the location of other adjacent grids from the local tier. From Fig.2 addr(001011)'s projections we can see three different grids embraced in one field, they are mutually dependent.

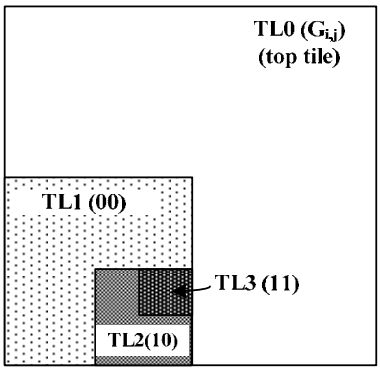


Fig.2. Example of ***addr****(001011)*'s projections.

Next, we discuss the coding sequence of tier address. Since a grid can be divided into two segments, namely, the vertical segment and horizontal segment, we define the order as vertical first and horizontal after. For the vertical division, there will be segment 0 and segment 1 left, respectively denoting the left half and right half of the plane; for the horizontal division, there will be 1 and

0, respectively denoting the upper half and lower half of the same plane. According to the division sequence, we can proceed to divide the left segment into next two pieces of upper half and lower half which we can in turn encode 00 and 01. The first bit 0 represents the left segment, while the second bit represents further division sequence. For a tier with 4^1 grids, each grid from the lower left corner respectively names 00,01,11,10 in a clockwise direction, so are other tier grids named. For the TL2, based on the fineness division of TL1, e.g. 01, it can proceed to be divided into 0100,0101,0111,0110. The only difference between these four addresses is last two bits. For TL3, the sequence of naming is applied by analogy as is shown in Fig.3. From which we can see 0101 is further divided into 010100,010101,010111, 010110.

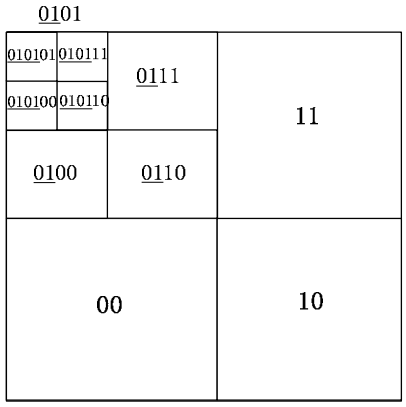


Fig.3. Naming sequence for tiles of different floors.

First, the two dimensional space is divided into square grids. Fig.4 shows a 3×3 grid with side length of *a*. Each grid is surrounded by 8 grids so the transmission distance of any two nodes in the neighboring grids should be less than d_0. To be able of location awareness, each node is equipped with GPS device from which it can read its current location. Each grid is given an ID which can tell the predefined coordinates. The group of nodes in a grid is denoted by $G_{i,j}$ as illustrated in Fig. 4.

Fig.5 shows part of hierarchical aggregation structure of MGRP of 3 tiers. There are multiple hops between adjacent tiers. Each local tile from the lower tier grids transmits data to the cluster head where upper tier grids locate.

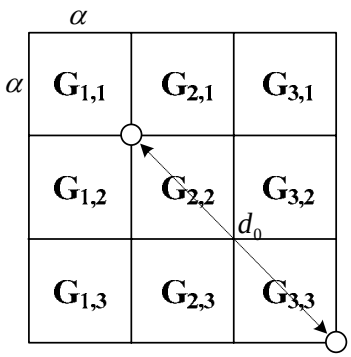


Fig.4. Grid Division

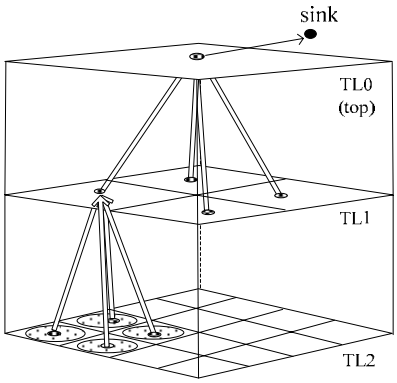


Fig.5. Hierarchical aggregation structure.

B. Cluster head election

To maintain the data aggregation in each grid, it is necessary to elect a gateway in each grid. Each node in the local grid computes its probability of being a cluster head and arbitrates automatically. A node initially sets its probability to become cluster head in the mirror tile

$$CH_{prob} = C_{prob} \frac{E_{residual}}{E_{\max}} \quad (3)$$

where $E_{residual}$ is the estimated residual energy of the node, E_{max} is a reference maximum energy, and C_{prob} is a small constant fraction used to limit the number of initial cluster head announcements. After that, the cluster head will send announcements and waits for the final affirmation from the local grid. If it has not received any affirmation, it elects itself to become a cluster head. Then the cluster head sets the timer and prepare for next iteration when the time expires. For most applications, it is possible to make nodes except cluster head in the local grid enter sleep mode after each election.

C. Data Transmission

After inter-grid aggregation, the cluster head further send the data to upper cluster of next grade, thus a hierarchical data aggregation tree is gradually built, from the stand point of sink. When a CH detects the interesting event, it will propagate to all its neighboring CHs using simple flooding. Thus, all the CHs know where the events happened. When the mobile sink sends the query to its CH of local grid, the CH receives the query packet and initiates the construction of d-tree.

In order to reduce the broadcast storm to a certain degree, the forwarding zone is defined with the locations of interested grid and sink's local grid. Let $G_{2,2}$ and $G_{4,4}$ be the interested grid and sink's local grid respectively (see Fig. 6). A larger rectangular forwarding zone is formed based on the diagonal line across the square from upper left $G_{2,2}$ to lower right $G_{4,4}$. We must ensure that the diagonal line always starts from the furthest vertex of interested grid to the corresponding furthest vertex of sink grid. Thus sufficient redundancy can be saved to avoid contention and collision.

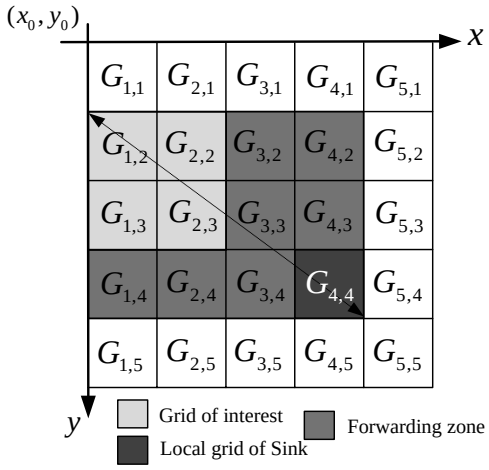


Fig.6. Definition of forwarding zone

Then the sink will flood the query packet to the Immediate Cluster Head (ICH) in local grid of interest. The query packet includes the fields of the location of sink and destination of CH that covers the grid of interest. In MGRP, the ICH of sink will be the root in the forwarding zone. Each CH constructs the d-tree and so the d-tree is restricted to the forwarding zone. The d-tree starts from the ICH to the destination CH of interest grid. Some CHs outside the forwarding zone also receive the packets but do not forward it.

MGRP build and maintain a cost field in the forwarding zone, providing each CH the direction to forward the data. Each CH keeps the cost for forwarding the packet from itself to the sink. Each CH will compare its cost with the sender so the data will follow the direction of descending cost to the sink. Since receiver decides whether it should forward a packet, the routing table can be removed and a packet can simply travel through whichever working nodes to the sink. When a CH forwarding a packet, it inserts its own cost in the packet, only the neighboring nodes with the smaller cost will continue forwarding otherwise they will discard it, means that the path is abandoned. After the sink confirms the grid of interest, the ICH broadcasts the query packet with cost field. The other CHs rebroadcast the packet until it reaches the CHs of interested grids.

The cost field is an artifact created by the sink and nodes farther to the sink have greater cost. The sink first builds a cost field by propagating the advertisement (ADV) packet announcing the cost of 0. The cost is defined as minimum energy overhead to forward a packet from this CH to the sink over a path. The initial cost of each CH is ∞. As illustrated in Fig. 7, the costs of CH M and N are initially C_M and C_N respectively. After M receives the ADV of N, M computes its link cost $L_{N,M}$ to N and adds it to cost of N, the result is $C_N+L_{N,M}$. If $C_N+L_{N,M}<C_M$, it means M should update its cost to $C_N+L_{N,M}$ and continue to rebuild the cost field for neighboring CHs, otherwise M discards the ADV packet and maintains the current link.

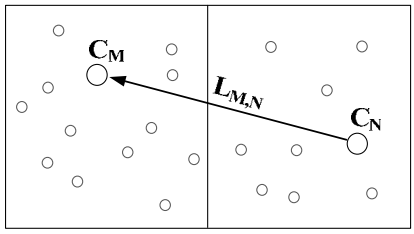


Fig.7. Cost update of CH

Then, the d-tree can be constructed step by step between each grid. If any CH receives any ADV packets, it will choose the parent CH on its minimum cost path. When the CH receives multiple copies from different upstream CHs, it searches its cache for signatures of recently forwarded packets to avoid duplicates. When a new CH i is built and receives a downstream CH j, it checks that j may be one of its potential parents, so it becomes connected to the d-tree and makes j as its parent. When CH i is already connected to the d-tree and whose parent is j, receives a messages from downstream j', then i compares the cost through different parent and finally selected a minimum cost path by j'. In this case, i changes its parent in the d-tree from j to j'. Fig. 8 shows an example of d-tree in forwarding zone. The CHs in light-colored spare grids which do not connect to the d- tree are called Substitute CH (SCH) and which connect to the d-tree called Relaying CH (RCH). The SCHs will be awakened when the energy of RCH is below E_r.

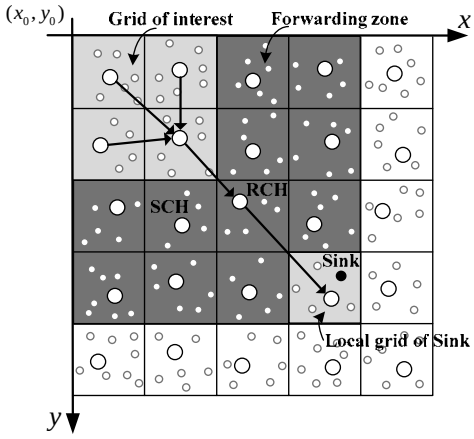


Fig.8. Spanning tree in forwarding zone

D. Sink Mobility

The MGRP is designed for mobile sink, so how the sink keeps continuously receiving updated data from the source is discussed in this section.

Since handling mobiles sinks brings new challenge to large scale sensor network research, many researches have been provided in order to maintain the data transmission during the process of sink mobility. TTDD exploits local flooding within local grid to find immediate dissemination node, however it does not optimize the path from source to the sink. Also, TTDD frequently resumes entire grid construction when the network energy status deteriorates sharply. CODE (a COordination-based data Dissemination protocol for wireless sEnsor networks) [15] protocol considers energy efficiency. When a source

detects an event, it generates a data announcement and sends it to all coordination nodes using simple flooding. Then the sink sends the query to the senor nodes along the backward path. However, when the sink moves to other gird, the sink just resend the query to the sensor nodes along a different path which is frequently updated by CODE.

According to the above analysis, in order to reduce energy consumption and minimize the cost of path reconstruction, taking advantage of our new chain-tree hybrid structure, we further introduce interior and exterior grid routing mechanisms to achieve efficient data transmission.

In interior grid, the sink selects the nearest sensor node as Primary Agent (PA) from itself and includes the location of primary agent in its queries. PA is used to find the ICH for sink and transmit data to the sink whose radio radius is limited. We assume that the radio radius of agent is smaller than the side length of grid, otherwise energy efficiency of agent decreases and sinks uses CH continuously even if the sink is moving. The sink is ready to send a query packet to the source if the PA finds the ICH within the radius of PA. In the process of sink mobility, the Immediate Agent (IA) is selected in neighboring nodes of sink to forward the data to PA when the sink leaves the radius of PA (see Fig. 9). When the sink is moving out of the radius of IA, it selects its new IA from its neighboring nodes and sends updating messages to old IA and PA, so that the future data is stopped forwarding from old IA to new IA. The new IA transfers data continuously from sink to PA until it finds new CH. Then the new IA changes to PA and the detected new CH becomes new ICH meanwhile existing ICH changes to CH. Thereafter, data transmission is carried out only between new ICH and new PA. The new ICH and new PA keeps sending update messages to the upstream CHs until the sink's ICH does not have sinks or downstream agents requesting data for a period of time, so that data are no longer forwarded to this gird.

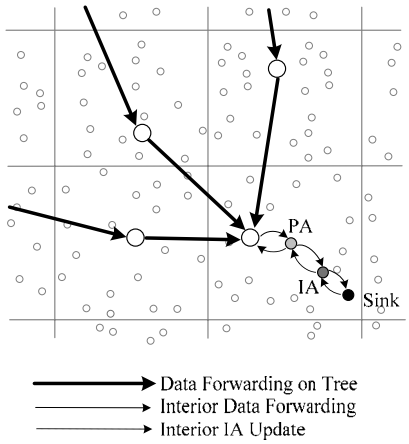


Fig.9. Interior grid routing

In the case when the sink moves out of a grid size, it picks a new PA and floods a polling message locally to discover new CHs that might be closer and checks its current location to know which grid it is locating periodically. The sink sends a message to PA along the old path to stop data transmission. As shown in Fig. 10, the mobile sink receives polling reply and chooses the closest CH1. Then CH1 broadcast a message to neighboring CHs. If the neighboring CH belongs to the forwarding root CH of the d-tree, it will send a message to the CH1 with its location and cost field. When node CHr belongs to the root of d-tree, it will send a message to CH1 which can further chooses CHr as its parent of the d-tree. In addition, we can also refer CH1 as the temporary root of the old d-tree and the former root CHr changes to normal relaying CH of the old path. This mechanism can reduce the overhead of rebuilding of d-tree in the forwarding zone.

If the mobile sink moves out of three hops from the branch CH of the d-tree which means that CH1 can't find any neighboring CH that belongs to root of the d-tree, the mobile sink will rediscover the new path by rebuilding the d-tree, so that devious path can be avoided and more energy can be saved.

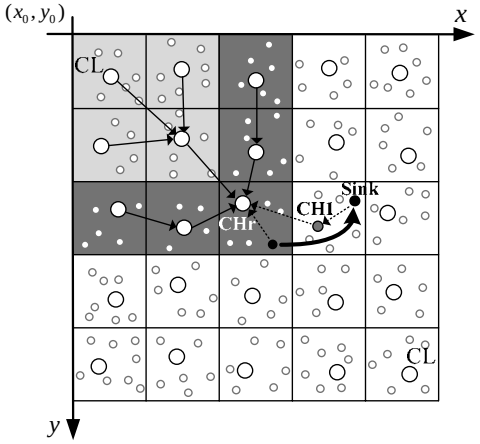


Fig.10. Exterior grid routing

V. PERFORMANCE EVALUATION

We implemented MGRP in OMNeT++ [16] simulator and the underlying MAC is 802.11 DCF. The environment in basic experiment is as follows. The entire sensor field is 2000×2000 square meters; the side length d of grid is r/22 (r=100m); the sink mobility follows the standard random waypoint model. The transmitting, receiving and idle energy consumption rates of a sensor node are set to 0.660W, 0.395W and 0.035W. Each sensor is initially equipped with 10 joules power and the simulation lasts for 1000 sec. We choose two main metrics to analyze the performance of MGRP; namely energy consumption and average delay, compare it to TTDD, SEAD and COSEN. Energy consumption is defined as the total energy consumed of our system which includes communication, computation, aggregation and sensing dissipation. The idle energy is not counted since it does not indicate the efficiency of data delivery. Average delay is defined as the average time between the time a source transmits a packet and the time the sink receives it.

A. *Impact of Network Size*

We first experiment these protocols assuming the sink moves at an average speed of 20m/s. Fig. 11 shows the graph of energy consumption versus the network size which ranged from 100 nodes to 500 nodes. Protocols other than MGRP consume more energy because of higher cost for tree construction, especially the initial sensor flooding causes a lot of energy. SEAD has a minimum update rate enforced to detect failures which may not be suitable for large network size. As the sink moves, the replica nodes have to feed different access node thus the maintenance cost of d-tree increases correspondently. Moreover, access nodes in SEAD keep tracks of the current sink and changes when the sink moves out of grid size, therefore a branch to the new access point is generated and a lot of energy is consumed for replica search phase. In TTDD, dissemination nodes which don't participate in routing process still consume energy for data update. COSEN used a two-tier structure for data dissemination. However, it could not reach energy efficiency when the network grows larger especially when the high-level leaders are far away from sink, thus causing a lot of energy for data transmission. MGRP demonstrates better energy consumption than the mentioned protocols because of hierarchical hybrid construction of both low level tier and high level branch nodes of the tree based on a flexible grid structure matched with up-to-date grid utility, which are used to confine the final transmission distance.

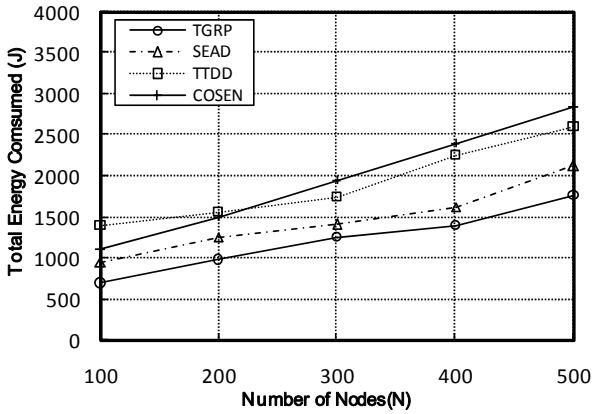


Fig.11. Energy consumption with different number of nodes

The forwarding zone contributes further in energy efficiency. MGRP uses confined forwarding zone and tries to connect to the original root with PA and IA, thus the total energy consumption can be saved for d-tree reconstruction. The multi-tier grid routing process also ensures the full utilization of nodes in each grid. Moreover, the mobile sink will try to connect to the original path based on local flooding until the temporary root CH is found and the restructuring overhead can be greatly reduced.

Fig. 12 shows the average delay versus the network size. As the number of nodes increases, the average delay of MGRP is lower than the other protocols. Since end-to-end delay and hop count in SEAD are considered for the new access node in large scale network, so the average delay is proportional to the number of single broadcasts needed to propagate a message. TTDD uses local flooding to reach the immediate dissemination node. When the sink moves, it has to generate a lot of broadcast storm along the trajectory of sink, so its dissemination path tends to be longer than the other protocols. COSEN uses non-flexible d-tree structure which can not adjust to the topology change, as such; when the sink moves out of a grid size, it has to reconnect the path between the high-level leader and sink thus costing more delay. However in the hierarchical structure of MGRP, each node in local grid can efficiently send data to CH even within lowest tier grid which significantly decreases the delay. Moreover, the branch nodes of d-tree are selected based on the optimized CH, so the network size has little influence on average delay.

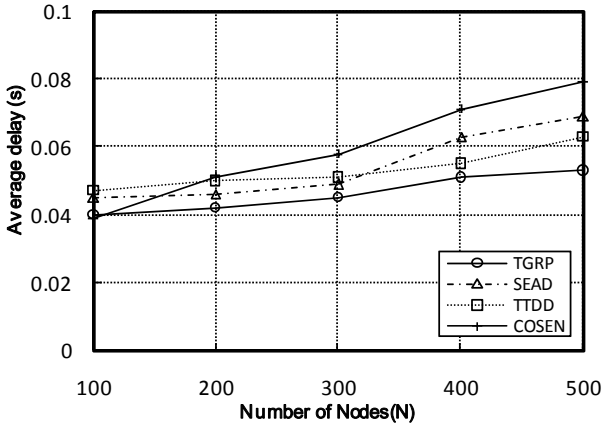


Fig.12. Average delay with different number of nodes

Fig. 13 shows the success ratio in packet delivery as the network size changes. MGRP disseminates most of the data successfully because it uses multi-tier grid routing method to make the all grids operable. In SEAD, more replica nodes will be selected as the nodes increase, more trade-offs will be made between the path delay and restructuring overhead thus causing the packet collision. In TTDD, it dissemination path tends to longer than the other protocols. With the nodes increase, it is inevitable to avoid packet collision along the path. In COSEN, the two-tier structure is not optimal when the nodes are large, because more control packets will be sent along the high-level chain and thus caused more packet collision due to devious paths toward the sink.

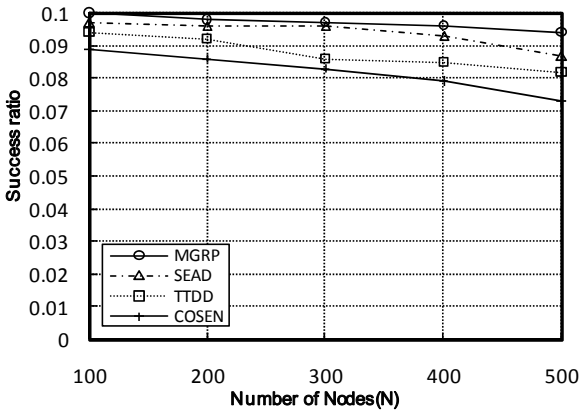


Fig.13. Success ratio with different number of nodes

B. Impact of sink mobility

Fig. 14 shows the total energy consumption versus different sink speeds. The number of nodes is set to 400. We compare the routing protocols changing the speed of sink from 0~30m/s. MGRP still outperforms the other three, because MGRP constructs the d-tree based on the optimized cluster head of each grid, thus the energy for readjusting the d-tree is reduced substantially.

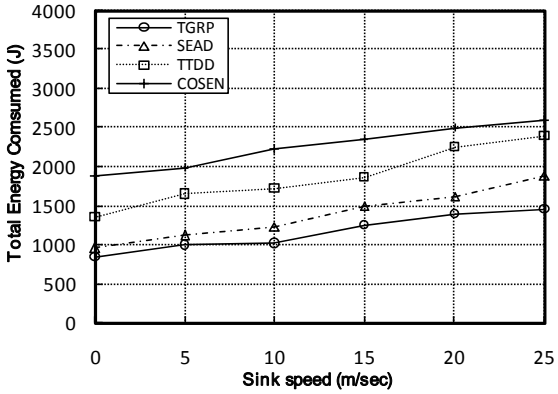


Fig.14. Energy consumption with different sink speed

Fig. 15 shows that the average delay of MGRP is lower than the other protocols both in low speed and high speed of sink when the number of nodes is set to 400. The reason is that in SEAD when the sink moves at higher speed, more cross nodes can be found at the neighboring nodes of sink and replica placement runs frequently in order to reconstructing the d-tree. Therefore configurable trade-off is not efficient at high speed of sink. In TTDD, when the sink moves faster, the devious transmission path frequently happens and it has to rebuild a new multi-hop path between the immediate dissemination nodes and the mobile sink. In COSEN, as the sink moves, the high-level leader remains connected to the sink via intermediate relay nodes. When the sink moves faster, the number of intermediate relay nodes increases causing more average delay along the path between the sink and high-level leader. However in MGRP, the depth of the d-tree is only relevant to the interest region in forwarding zone thus saved quite a bit delay. It also elects each branch nodes of d-tree based on maximum residual energy to ensure the efficiency of data transmission under changing situations.

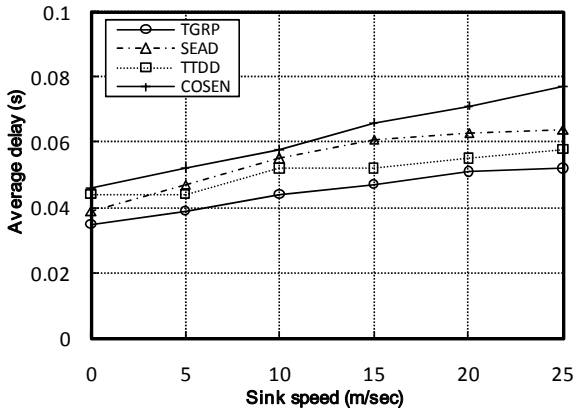


Fig.15. Average delay with different sink speed

Fig. 16 shows the success ratio in packet delivery as the sink's speed changes. MGRP transfers successfully most of the packets even to the high-speed sink. The first reason is that MGRP uses forwarding zone to turn off the grids that don't participate in the routing process. The second is the MGRP uses hierarchical routing scheme to ensure the continuity of data transmission with high-speed sink. TTDD frequently renews the entire path to mobile sink therefore increases the connection loss ratio. Moreover, the election of optimal cluster head can also reduce the success ratio. In SEAD, as the sink moves faster, more data update are disseminated along the d-tree meanwhile many existing access nodes needs to be replaced which still causes transmission interruption. In COSEN, the long relay path between the sink and high-level leader can not be avoided thus more broadcast needs to maintain this path which also increases packet collision.

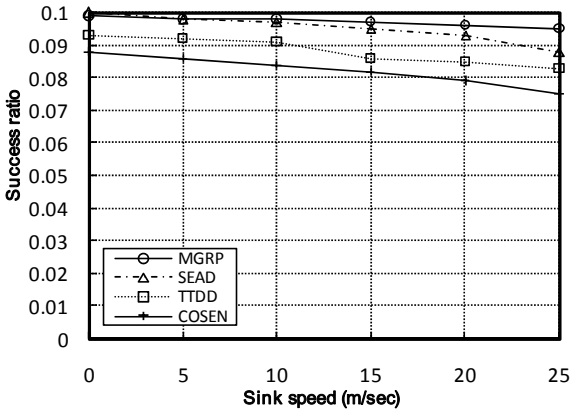


Fig.16. Success ratio with different sink speed

In summary, MGRP has higher scalability than it's precede protocols TTDD, SEAD and COSEN, MGRP makes use of the advantage of each protocols, and combined them in a hybrid way, getting rid of their disadvantages respectively. With more sensor nodes, MGRP has significantly energy efficiency than the other protocols even with high-speed sink mobility.

VI. CONCLUSIONS

In this paper we proposed MGRP, a hierarchical hybrid grid routing protocol for large scale sensor network. The MGRP constructs a multi-tier structure where each member of global d-tree is elected based on residual energy of local chain in each grid. The d-tree member periodically updating the cost field to make adjustment of optimized d-tree, so other Cluster Headers in grid can find the minimum cost path to the tree. In addition, we utilize the forwarding zone to restrict the flooding region and also reserve spare grid in case of transmission interruption. When the mobile sink moves to other grid, the presented approach searches the neighboring nodes and elected a temporary root of d-tree for adjustment. The basic idea is enable the CH to connect to the old path without restructuring of entire d-tree. The simulation results show that MGRP outperforms the other protocols such as TTDD, SEAD and COSEN, in terms of the energy efficiency and average delay while

delivering most of the data successfully to the high-speed mobile users, especially for the sensors of large scale.

The work remains to be done in future are: investigating the optimum cluster size, the melting down point of related cluster head election algorithm, the implantation complexity of the grid density control algorithms, the up limit of d-tree updating frequency associated to the mobile speed. We are grateful for the supporting received from both industry and academic circles.

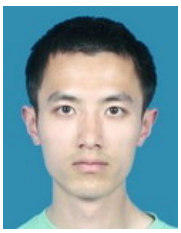
ACKNOWLEDGEMENTS

This work is supported by National High Technology Research and Development Program (863 Program) (2006AA10Z258); Zhenjiang Municipal Social Development and Technology Support Projects (SH2009002). Ottawa University and GenieView Inc partnership program (Auto21-F303-2010).

REFERENCES

[1] I.F. Akyildiz and M.C. Vuran. Wireless Sensor Networks. Advanced Texts In Communications And Networking, p.480, 2010.
[2] D. Pompili, T. Melodia and I.F. Akyildiz. Three-dimensional and two-dimensional deployment analysis for underwater acoustic sensor networks. Ad Hoc Networks. 7(4):778-790, 2009.
[3] L. Li and T. Kunz. Cooperative Node Localization for Tactical Wireless Sensor Networks. Proceedings of IEEE MILCOM'07, 2007.
[4] K. Zeng, K. Ren, W. Lou and P.J. Moran. Energy aware efficient geographic routing in lossy wireless sensor networks with environmental energy supply. Wireless Networks, 15(1):39-51. 2009.
[5] D. Koutsonikolas, S. Das, Y. C. Hu, and I. Stojmenovic. Hierarchical Geographic Multicast Routing for Wireless Sensor Networks. Wireless Networks, 16(2):449-466, 2007.
[6] J.A. Sanchez, P.M. Ruiz and I. Stojmenovic. GMR: geographic multicast routing for wireless sensor networks. Proceedings of IEEE Sensor and Ad Hoc Communications and Networks (SECON '06), pp. 20–29. 2006.
[7] S. M. Das, H. Pucha, and Y. C. Hu. Distributed Hashing for Scalable Multicast in Wireless Ad Hoc Networks. IEEE TPDS, (19):347-362, 2008.
[8] L. Cheng, Y. Chen, C. Chen and J. Ma. Query-based data collection in wireless sensor networks with mobile sinks. Proceedings of the 2009 International Conference on Wireless Communications and Mobile Computing: Connecting the World Wirelessly, pp.1157-1162, 2009.
[9] J. Luo, J. Panchard, M. Piorkowski, M. Grossglauser and J.-P. Hubaux. Mobiroute: Routing towards a mobile sink for improving lifetime in sensor networks. IEEE International Conference on Distributed Computing in Sensor Networks (DCOSS), 4026:480-497, 2006.
[10] M. Ma and Y. Yang. Sencar: An energy efficient data gathering mechanism for large scale multihop sensor networks, in: IEEE International Conference on Distributed Computing in Sensor Networks (DCOSS), 4026:498-513, 2006.
[11] F Ye, G Zhong, S Lu and L Zhang. Gradient broadcast: A robust data delivery protocol for large scale sensor networks. Wireless Networks, 11(3):285-298, 2005.
[12] H. Luo, F. Ye, J. Cheng, S. Lu, and L. Zhang. Ttdd: Two-tier data dissemination in large-scale wireless sensor networks. Wireless Networks Journal (WINET), 11:161–175, 2005.
[13] H. Kim, T. Abdelzaher, and W. Kwon. Minimum-Energy Asynchronous Dissemination to Mobile Sinks in Wireless Sensor Networks. ACM Conference on Embedded Networked Sensor Systems (SenSys), 2003.
[14] N. Tabassum, Q.E.K. Mamum and Y. Urano. COSEN: a chain oriented sensor network for efficient data collection. Third International Conference on Information Technology. New Generations (ITNG 2006), pp.262-267, 2006.
[15] H.L. Xuan and S. Lee. A coordination-based data dissemination protocol for wireless sensor networks. Proceedings of the Sensor Networks and Information Processing Conference, pp. 13-18, 2004.
[16] Varga, A. OMNeT++ Discrete Event Simulation System User Manual - Version 3.2. Technical University of Budapest/Hungary. 2005.

Zujue Chen was born in Shanghai, China. He received B.E. degree in Automobile Engineering from Jiangsu College of Technology, China, in 1978. He finished advanced studies in College of Computer Technology from HeFei University of Technology and College of Computer Technology from Beijing University of Posts and Telecommunications, in 1982 and 1985, respectively. Since then, he has been on faculty of School of Computer Science and Telecom-munication engineering in Jiangsu University, China, where he is currently Professor and Dean of Communications Engineering division. His research interests include embedded computing, wireless sensor network, intelligent instruments, monitoring and controlling system, remote wireless monitoring system.



Shaoqing Liu was born in Tianjin, China. He received B.E. degree in Communication Engineering from Jiangsu University. In 2010, He received his Master Degree in communication and information systems from Jiangsu University. His technical interest includes ad hoc wireless and sensor network, distributed computing and geographic information systems.

Jun Huang was born in Shanghai, China. He received his doctor's degree in 1992 from a Joint Ph.D program between Southeast University China and Concordia University Canada. He worked at a number of places including Bell Canada, Lockheed Martin USA, Carleton University and Ottawa University. He is currently Professor of School of Computer Science and Communication Engineering in Jiangsu University China. His current research interest includes wireless sensor network, network switching for automotive electronics and artificial intelligence for field robotic control.

ElGamal Digital Signature Algorithm of Adding a Random Number

Xiaofei Li, Xuanjing Shen and Haipeng Chen
College of Computer Science and Technology, Jilin University, Changchun, China
Email: xiaofei09@mails.jlu.edu.cn, {xjshen, chenhp}@jlu.edu.cn

***Abstract*—As for the problem that ElGamal digital signature scheme's security is constantly being challenged and increasingly becomes increasingly serious, an improved ElGamal digital signature algorithm is proposed. As the original ElGamal algorithm has its own security disadvantages that only one random number is used, in order to improve its security, the scheme presented in this paper improved this demerit by adding a random number to the original one and increasing difficulty of deciphering key. The security of the improved signature scheme is the same with the ElGamal signature scheme which is based on the difficult computable nature of discrete logarithm over finite fields. Then issues about how to increase the complexity between the random number and the key by adding a random number is discussed. Last, we analyzed the improved signature scheme from the following two aspects: security complexity and time complexity. The analysis showed that the safety of the improved signature scheme was higher than that of the original one, and the improved one has a smaller time complexity.**

***Index Terms*—ElGamal type;digital signature; random number; program improvement**

I. INTRODUCTION

Digital signature is one of the main application for public key crytography. At present, the wider application of digital signature systems are RSA signature scheme[1] and ElGamal-type signature scheme[2], such as the Schnorr signature[3], DSA signature[4]. Although the hash function can avoid some attacks, however,if the system parameters are inappropriate, there is security risk[22]. Therefore, this is necessary to how to choose digital signature in study of application examples' design.

In 1985, ElGamal raised one of the two most important digital signature scheme,which is the ElGamal digital signature system[2]. It's security is based on the difficult computable nature of discrete logarithm over finite fields. The signature scheme with message recovery has many obvious advantages: a shorter signature for short message, and the shorter produced verification. If this scheme is applied to group signature, it has the difficult computable nature of discrete logarithm over finite fields and the advantage of message recovery in the digital signature[5-6].

As a widely used digital signature algorithm, ElGamal algorithm has the security on the discrete logarithm problem[16].Generalized ElGamal-type signature scheme has stronger security than the original ElGamal algorithm through improving the original ElGamal-type signature scheme[7] and increasing the number of hidden parameters. This paper presents an improved ElGamal signature scheme based on generalized ElGamal digital signature scheme, and the safety of the improved signature scheme was higher than that of the original one. Not only dose it broaden the scope of its application, but can be used for the signature of the group users[8]. Group signature, as a branch of a subject oriented group cryptography, was proposed by Chaum and Van Heyst in 1991. This improved ElGamal algorithm makes ElGamal Digital Signature have a more extensive application in the fields of authentication and e-commerce system[9-13].

The rest of this paper is organized as follows: Sect. 2 gives a background on ELGamal algorithm and presents the research work related to digital signature in general, and mainly analyzes various attack means of ELGamal digital signature algorithm. The improved idea of ElGamal digital signature algorithm is given in Sect. 3. In Sect. 4, we introduce the specific improved ElGamal digital signature algorithm, including the specific process steps and implementation of improved algorithm. In Sect. 5 gives the performance analysis of the improved algorithm , and compares the performance of the original algorithm, it's security has significantly improved. Finally, the drawn conclusions and the planned future work are discussed in Sect. 6.

II. BACKGROUND AND RELATED WORK

This section provides an introduction to ElGamal digital signature algorithm on how it to be proposed and developed. Though research into this subject has been taking an important part, its security and efficiency are attacked gradually. This paper mainly analyzed the performance of ElGamal algorithm, and the security risks existing in the traditional ElGamal digital signature algorithm are discussed from four aspects.

A. Background

In 1985, ElGamal proposed a cryptosystem based on the discrete logarithm problem which can be used for both data encryption and digital signature [2]. Since it was put forward, the new cryptosystem aroused widespread interested in the password academic field

National Natural Science grant (No.60773098);
Scientific and Technological Development Foundation of Jilin Province(No.20080317).

doi:10.4304/jnw.6.5.774-782

because of its good cryptographic properties, which is the second most prominent signature scheme after the RSA signature scheme. Although the study of ElGamal's digital signature rose many years ago, it still has many problems with application in practice, especially the problems of its safety and efficiency. With the rapid development of e-commerce, the research on this project will be crucial.

The ElGamal digital signature scheme is non-deterministic like ElGamal public-key cryptosystem. Namely, the same clear message has different signatures due to the different parameters selected randomly, and each signature only corresponds to a random number [14], which has brought a great hidden danger to the security of ElGamal digital signature scheme [15-17].

In terms of the ElGamal digital signature scheme, the algorithm's security depends on the security of the private key x. Once the private key x is intercepted by the hacker, the entire digital signature algorithm is accessible to anyone, and no security exists at all [16]. Therefore, the primary target of the attacker is the private key x. Therefore, the primary target of the attacker is the private key x. The following are the common ways of attack, though in which the author compared the traditional ElGamal digital signature scheme with the ElGamal digital signature scheme after adding a random number, then analyzed and verified its security that is improved, it turns out that the private key x and random number k are unknown to the attacker.

B. Analysis of ElGamal Digital Signature Algorithm Security

In general, the following are the main ways of attack: a direct hack on the private key, arbitrary forged signature attack, substitution attack according to known signatures, the assault in homomorphism (using the same random number k), the assault in homomorphism (use the relevant random number k1, k2, k3,).

1) A direct hack on the private key

Since the attacker's primary target is the private key x, obviously the direct attack on the private key x is the most direct methods of attack.

The first thing we need to determine is: ElGamal digital signature scheme is based on the discrete logarithm problem. At present, there is no feasible solution to this problem. Then the attacker can only use the simplest method of exhaustion to calculate the value of the private key x compulsorily, and use the equation $\alpha^x \equiv \beta \bmod p$ to work out the solution set of private key x through a large number of operations. Then test each values of these x to determine which is the private key used by signers that need to get another signature. And calculate the value of the random number k using the equation (1) in signature equation. Then test it though the equation (2).

$$\delta = (m - x\gamma)k^{-1} \bmod (p-1) \quad (1)$$

$$\gamma = \alpha^k \bmod p \quad (2)$$

In the course of the hack, the hackers apparently need to solve logarithm for a time, then test the solution sets obtained from solving logarithm, and test of each solution all need to go through an inverse element and exponentiation. In addition, the middle data which are produced in the operations will be greater, so there is the need for corresponding processing algorithms and adequate storage space. The complexity of time and space are very high.

2) Arbitrary forged signature attack

The goal that hackers attack algorithm is to find the private key x, but their fundamental goal is to forge the signature using the private key x, then sign the selected file of the attacker or sign the message. As the calculation of the equation and verification of the signature equation are public, then the attacker may forge signature directly through a public key and digital signature and verification equation without knowing the private key x.

The attacker get the public key p, α and β, as well as the equation which need to be used for signing and verifying from the public sources. To forge a digital signature (γ, δ). An attacker can first select a value of γ, and through γ and the known public key to calculate the value of δ using signature equation or verification equation ,thus a digital signature is forged completly.

If we use equation (1) in the signature equation to calculate the value of δ, then relate to the unknown private key x and the value of the random number k will be involved. The two calculations will encounter discrete logarithm problem. Therefore, only consider using the verification equation to calculate the value of δ is the only way can be considered. Moreover, if the value of δ is calculated through the verification equation, then it is clear that this forged digital signature can certainly through validation.

Calculate the value of δ using the following validation equation:

$$\beta^\gamma \gamma^\delta \equiv \alpha^m (\bmod p) \quad (3)$$

Where, p, α and β are public key, m and γ are selected by hacker. Then what we can obtain through equation (3):

$$\beta^\gamma \gamma^\delta \equiv \alpha^m (\bmod p)$$
$$\Rightarrow \gamma^\delta \equiv \alpha^m \beta^{-\gamma} (\bmod p)$$

It is clear that the process of solving the index δ must face the discrete logarithm problem.

3) Substitution attack according to known signatures

Set (γ, δ) is the signature of m, if γ is reversible, then set k'= ek + n, where e, n is any two integers, and satisfy:

$$k' \in Z_{p-1}$$

$$r' = r^e \alpha_n \bmod p$$
$$\delta' = \delta k \gamma^{e-1} \alpha_n (tk+n)^{-1} \bmod (p-1)$$

So, (γ', δ') is the signature of m'.where:

$$m' = \gamma^{e-1} \alpha_n m \bmod (p-1)$$

Because (γ, δ) is a signature of m, there is $m = (\delta k + x\gamma) \bmod (p-1)$. If γ is reversible, there also is:

$$x = \gamma^{-1}(m - \delta k) \bmod (p-1)$$

From signature equation can obtain:

$$m' = \delta' k' + x\gamma' \bmod (p-1)$$

δ', k', x, γ' are substituted into the above equation, get:

$$m' = \alpha^n \gamma^{e-1} m \bmod (p-1)$$

In this way, it also takes the attacker a long time to wait for the documents or information available after the digital signature has been forged. Take the long file into account, in order to reduce the computational times, the value of m obtained by using the open one-way function is used to sign. Thus the attacker can also consider attacking on the one-way function, so that make the selected file generate the summary corresponding with the value of m '. This relates to another algorithm of attack. At the same time attacked the two algorithms, the degree of difficulty is surely bigger than any method of attack by only attacking the ElGamal digital signature algorithm.

This method of attack allows an attacker to get a lot of legitimate digital signature, however, as the signatures associated with the value of m', it is difficult to be practical.

4) The assault in homomorphism

There are two cases:

a) Using the same random number k

Among the parameters used in the ElGamal digital signature algorithm, the value of random number K is confidential in spite of the private key x. k is randomly generated and are discarded after each signature, it seems there is no attack value, in fact, there are significant security vulnerabilities.

The random number k is only used once,because if the same value of k is used to sign on two or more signature files or messages,then the attacker can calculate the value of k by using the signature two times ,and thus get indirectly the value of private key x indirectly.

Using the same random number k, same γ in two signatures from (2) can be introduced. A hacker could use these two signatures which are (γ_1, δ_1) and (γ_2, δ_2), files or messages named m1 and m2. We can get group equation through (1).

$$\begin{cases} \delta_1 k + x\gamma_1 = m_1 \\ \delta_2 k + x\gamma_1 = m_2 \end{cases} \quad (4)$$

According to this group equation, the hacker can easily work out the value of k: $k = (m_2 - m_1)(\delta_2 - \delta_1)^{-1}$. Then substitute the calculated the value of k into any equation in the group equations to calculate the value of private key x, and then attacks will be successful.

In view of the shortcoming of random number k, so ElGamal digital signature algorithm requires to use the different value of k every time. As long as the signers strictly do this, the attacker can not use this simple and effective way to attack successfully.

For example, when the value of k with one signature is equal to the value of k with another signature, it is said that three random numbers k1, k2 and k3 satisfy:

$$k_3 = k_1 + k_2 \quad (5)$$

Obviously: $\gamma_3 = \gamma_1 \gamma_2$. To calculate the value of the private key x:

$$x = \left(\delta_1 \delta_2 m_3 - \delta_2 \delta_3 m_1 - \delta_1 \delta_3 m_2\right) \left(\delta_1 \delta_2 \gamma_1 \gamma_2 - \delta_2 \delta_3 \gamma_1 - \delta_1 \delta_3 \gamma_2\right)^{-1} \bmod (p-1) \quad (6)$$

To improve security, h (m) is required instead of m in the signature equation, so, ElGamal digital signature algorithm signature equation is:

$$\delta = (h(m) - x\gamma) k^{-1} \bmod (p-1)$$

According to the above analysis on ElGamal Digital Signature Algorithm, due to the discrete logarithm problem has yet no possible solution has been worked out yet, so the ElGamal type digital signature algorithm based on the question has high security keys. Before the discrete logarithm problem is effectively resolved, any direct attack on the keys, the computational needs are staggering.

According to the signature and the verify equation, the signer needs to complete a signature through a power operation and an inverse operation, and the verifier needs three power operations. Given that sign documents or information for the sign may be more, but the verifier is the different user; this conditions which the computational complexity of verifier is higher than the computational complexity of signers.

In terms of the current computing power, it is not difficult for the traditional ElGamal digital signature to compute. But insecurity caused by the random numbers has posed a great threat. An attacker can easily use the link among random numbers to access the value of private key k after the uncomplicated calculation. Analysis of the selected contact is the most simple two links; I believe there are other links making a threat to the security algorithm. Signer can of course find out the value of these random numbers to avoid using them to ensure the security of the algorithm. But this also brought another problem. Since the selection of random numbers is greatly reduced, and each random number can only be

used one time, and then discarded, which has seriously affected the life of the algorithm. Especially with network communication increased due to the development of the Internet, the need for digital signatures in more places, and the signer has to replace the value of the key x from time to time in order to ensure the safety of the signature algorithm, or even replace the public key α . The replacement of the public key needed to be re-issued through public channels. Therefore, some customers maybe fail to access to the updated information and still used the no-consistent public key for authentication leading to the erroneous conclusion. As Internet users and the need to sign a document or a piece of information increased, Authenticator to the public key of the update delays will become more apparent because of this updating. We can also consider sending the public key and signature together. In fact, this is the public key as a part of the signature, which increases the length of the signature and the occupier of the amount of network resources in the transmission process. The error probability will increase resulting from increased data transmission. So, there will be more digital signature retransmission due to the error in transfer process.

Thus, in the ElGamal digital signature algorithm, the insecure random number constitutes a very large threat to its security. At the same time, measures taken to ensure the safety of the algorithm will lead to a series of problems, and the further extension of the algorithm has a significant restriction.

To improve the algorithm's security, we have improved the ElGamal digital signature algorithm by adding a random number and strengthening the link between the random number and the private key to make it more difficult to decipher.

III. IMPROVING METHODOLOGY

According to the analysis of two attacking methods on random number, it was found that the hacker can easily calculate the value of random numbers or the value of the key by calculation of a random number if a signer uses the insecure random numbers. This generally resulted from that it is easier to hack the random number than hack the key or too intimate relationship between the random number and the key. So, for the vulnerability of random number vulnerability in the ElGamal digital signature algorithm and too simple link between the random number and the key, in this paper, improved program was proposed to enhance the security of the algorithm, which can make the link between the random number and the key more complicated.

There are two signature equations of the ElGamal digital signature algorithm, shown as follow:

$$\gamma = \alpha^k \bmod p \tag{7}$$

$$\delta = (m - x\gamma)k^{-1} \bmod (p-1) \tag{8}$$

Public key β is calculated by the follow equation:

$$\beta = \alpha^x \bmod p \tag{9}$$

Compared with (7), (9) is the same as (7) in form, with β generated through the private key x and as a public key. In (7) γ is generated through the random number k and as a part of the signature. Then we can regard these three equations as a signature equation, but what is calculated as a public key instead of as a signature. If we take β as part of the signature, γ as a public key, corresponding to x as a random number, k as the private key, and the equation (8) is changed to follow:

$$\delta = (m - k\beta)x^{-1} \bmod (p-1) \tag{10}$$

Verification equation is changed as follow:

$$\alpha^m = \gamma^\beta \beta^\delta \bmod p \tag{11}$$

This result of replacement is also an ElGamal digital signature algorithm. It can be seen that there is no essential difference between the random number k and the private key x. They are in different positions only because (8) is different. Then we can consider adding such a random number, and a corresponding increase in a form such as the type (9) of the equation to the signature equation, namely:

$$\lambda = \alpha^t \bmod p \tag{12}$$

Accordingly, it is needed to make the appropriate changes to the signature equation and the verify equation. The signature equation in (8) is changed as follow:

$$m = (x\gamma + k\lambda + t\delta) \bmod (p-1) \tag{13}$$

The authentication equation is changed as follow:

$$\alpha^m = \beta^\gamma \gamma^\lambda \lambda^\delta \bmod p \tag{14}$$

In this way, the linkage between the random number and the private key x is established based on (13). As for (13), the values of x, k and t are to be identified. If the hackers obtained a random number value successfully and wanted to continue hacking, this is clearly more difficult than that for the original algorithm. Improved the above scheme further, signature equation of the new digital signature algorithm's signature equation can be obtained as follow:

$$\beta = \alpha^x \bmod p \tag{15}$$

$$\gamma = \alpha^k \bmod p \tag{16}$$

$$\lambda = \alpha^t \bmod p \tag{17}$$

$$m = (x\gamma + k\lambda + t\delta) \bmod (p-1) \tag{18}$$

Verification equation can be drawn as follow:

$$\alpha^m = \beta^\gamma \gamma^\lambda \lambda^\delta \bmod p \tag{19}$$

IV. IMPROVED DIGITAL SIGNATURE ALGORITHM

The difference between the improved algorithm and the original ElGamal digital signature algorithm is mainly reflected in adding a random number aiming to make the

original algorithm more complicated and more difficult to decipher. The specific algorithm is as follow:

Step 1: A large prime number p is produced by system, α is a generator of Zp*, x($1<x<\varphi(p)$) is the signer's private key, the corresponding signature public key β can be calculated by $\beta=\alpha^x \bmod p$, and opened to the public key.

Step2: Two different random numbers t and k are randomly selected by system. Where t, k and x must be co-prime and there is inverse. γ and λ are calculated by the $\gamma=\alpha^k \bmod p$, $\lambda=\alpha^t \bmod p$ and retain γ and λ.

Step3: Signature explicitly m, δ is calculated using the results of the first two steps as well as the extended Euclidean algorithm and modular inversion algorithm by $m=(x\gamma+k\lambda+t\delta)\bmod(p-1)$. It should be avoided to take the same random number and simple functional relationship existing between random numbers at the course of obtaining a number of signatures.

Step4: Discarded the random number k and t, then the required public key p, β and α are obtained. The private key is x. The signature of plain text m is (γ,λ,δ).

Step5: (γ,λ,δ) is sent to the corresponding customers by system. The customers use the following equation to verify the correctness of plaintext m digital signatures. If equal, the signature is correct. Otherwise, the signature is incorrect or transmission errors. The equation as follows:

$$\alpha^m=\beta^\gamma\gamma^\lambda\lambda^\delta \bmod p$$

Algorithm flow chart is shown in Fig. 1.

In the above-mentioned improved ElGamal digital signature algorithm, the same message m corresponded to the different digital signature (γ,λ,δ) for the different random number k, t. And they can be all verified through the validation algorithm, which characterizes with uncertainty of signature and improved security.

When signers take λ as a signature, they need to finish one more computing power each time, which increases the amount of the signer's operations. When taken λ as a public key, the signer calculates a value of λ, which could be used as many times as the value of β. So the signer's computation amount is almost the same as that of the original algorithm. But for authenticator, each authentication has one more computing power, but increased with only about 0.5 time computation. As for the computer which can easily verify the ElGamal-type digital signature, verifying the signature of the improved algorithm does not consume more time.

λ can be used as public key if it is taken into account that the amount of the signer's operation is not increased. But whether this is safe, it is still needed to be determined through the analysis of safety. If the public key λ as a random number will lead to insecurity, λ should be taken as a part of the signature. This would increase the operation of the signer. But as long as the signature is not carried out in large amount, this computation can still be accepted.

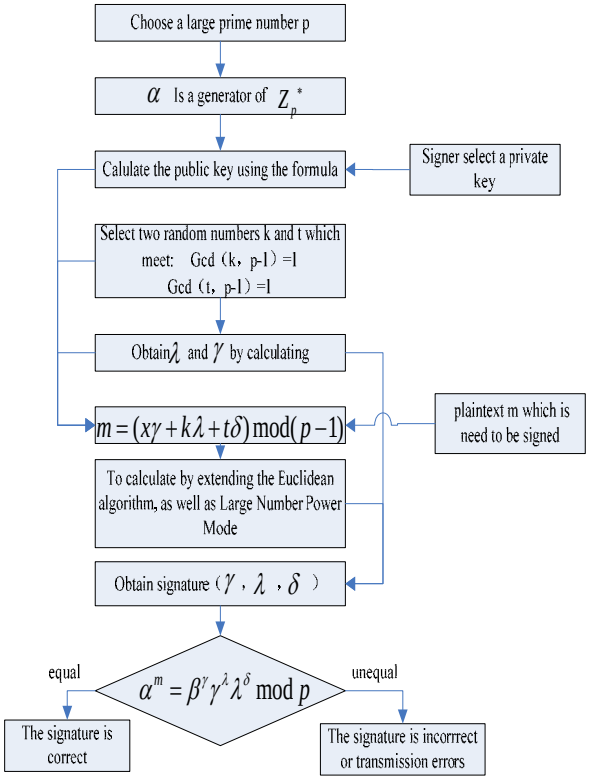


Fig. 1 Algorithm flow chart

V. ANALYSIS OF THE IMPROVED ALGORITHM

A. Security Analysis

ElGamal digital signature scheme is based on the difficult problems of seeking discrete logarithm in prime domain. At present, there are several types of hacks, which include a direct order to discrete logarithm algorithms, the use of special mathematical structure and parameters and so on [18]. But for the ElGamal digital signature scheme the hack comes from two aspects: on one hand, the key x is needed to be restored; on the other, the signature can be forged without restoring the key x. The improved program is analyzed from the perspectives of following five kinds of hacks.

1) A direct hack on the private key

If the ElGamal digital signature scheme with a added random number is used to sign , at the initial stage the hacker can compute the solution set of the private key x as described above, and determine the private key, but the two random numbers k and t by encrypting equation cannot be calculated in the next step. The encrypting equation as follows:

$$m=(x\gamma+k\lambda+t\delta)\bmod(p-1)$$

Similarly, the hacker will not be able to use (16) and (17) to verify the correctness of the private key x which has been known. Therefore, in improved ElGamal digital signature scheme, a hacker cannot attack the following formula:

$$m = (x\gamma + k\lambda + t\delta) \bmod (p-1) \tag{20}$$

Then the hacker will be able to think by attacking style (16) and (17) to derive the solution set of the two random numbers k and t, respectively. A signature is still needed to determine k, t values. And then it is substituted into (20) to verify.

In the course of the hack, the hackers apparently need to solve logarithm for three times, then test the solution sets obtained from solving logarithm, and test of each solution need to go through an inverse element and exponentiation. This process is clearly more complex than the traditional method mentioned above. This also brings about more difficulties for the hackers.

In addition, as long as the value of private key x that selected by the signer is large enough, we can ensure that a hacker in the foreseeable period of time can not figure out the value of the private key x. While this has increased the amount of the signer's operations, as for the public key β, it can be reused by calculating only one time. While the private key x in each signature involves only one multiplication operation, the increase in computational complexity by increasing x is not significant.

It can be said that such attacks on the security of ElGamal-type digital signatures are not a threat before the appearance of efficient algorithm for solving the discrete logarithm.

2) Arbitrary forged signature attack

The main attack will be focused on the digital signature verification equation in this way. However, algorithm verification equation is as follow after adding a random number:

$$\alpha^m = \beta^\gamma \gamma^\lambda \lambda^\delta \bmod p \tag{21}$$

δ is calculated is as follow:

$$\lambda^\delta = \alpha^m \beta^{-\gamma} \gamma^{-\lambda} \bmod p \tag{22}$$

Compared (21) with (22), the calculation of the latter is clearly more difficult because the latter one has one more inversion than the former.

And even hacked successfully, the value m of the selected files or messages m is the known quantity during the calculation process, so the calculation results can only be used for the signature of this document or message. The value of their attacks is clearly lowered.

3) Substitution attack according to known signatures

As for the improved algorithms, compared with many traditional digital signature algorithms, a hacker can still easily access signer's signature (γ, δ) of the message or document m, and then forged a number of legitimate digital signatures [19].

Since the improved algorithm introduces more parameters, the possibility that we want to adapt file m' can be reduced naturally. And forged methods need to be re-examined. Even if the method of forging signature can be found, the attack imposed on it is more complex than that on the ElGamal digital signature algorithm. Valid signature that is obtained in this way is still corresponding with value of m'. It is still a more daunting task to find the file or message to satisfy m'.

Methods of attack with an added signature data are more difficult to be constructed. Its computation is much more even if it succeeds. But the real difficulties for the hacker are still a valid signature that was forged at the same time, as well as produced the corresponding the value of m'. The attacker can successfully make the value of m' to the actual application. Hacker attacking in this way is only in the believers of his computing power consumption.

4) The assault in homomorphism

There are two cases:

a) Using the same random number k

Using the same random number k, same γ in two signatures from (04) can be introduced. A hacker could use these two signatures which are (γ_1, δ_1) and (γ_2, δ_2), files or messages named m1 and m2, and public key named γ. We can get group equation through (18).

$$\begin{cases} m_1 = x\gamma + k\lambda + t\delta_1 \\ m_2 = x\gamma + k\lambda + t\delta_2 \end{cases} \tag{23}$$

According to this group equation, the hacker can easily work out the value of t: $t = (m_2 - m_1)(\delta_2 - \delta_1)^{-1}$.

Taken λ as a public key, then the corresponding t value is equivalent to a private key, that is, the hacker has obtained the value of a private key. But substituted the calculated t value into (23), the value of another private key cannot be immediately calculated. Get a binary linear equation: $m_1 = x\gamma + k\lambda + \delta(m_2 - m_1)(\delta_2 - \delta_1)^{-1}$.

According to this equation we can get a solution set, and we can derive the value of x from verifying it by putting it into the signature equation one by one.

It has to go through a power calculation to verify the value of each x on the solution. But compared to the previous attack methods, it can be done at least in a limited period of time. That is, after an increase of a random number t and the corresponding x as the public key, when using the equal value of k to sign, the attacker can not obtain the value of the private key x through a simple calculation while the hacker can still attack successfully in a limited period of time. Therefore, it should be avoided to use the equal k-value to sign for the improved algorithm.

As for four numbers unknown, clearly, it is too difficult to solve with two equations. A hacker can get one ternary linear equation through this equation group and thus obtain a solution set, then test each solution set. Hackers have two ways to seek ternary linear equation. First, eliminating k-value, and second, eliminating x-value, then the solution set was about random numbers. Further, they need to solve the private key after the successful test. Eliminating k value, you can directly test the value of the obtained private key x.

In view of this shortcomings of the random number k, ElGamal-type digital signature algorithm will be required by using different values of k for each time. As long as the signers do this strictly, a hacker cannot take advantage of this simple and effective method to attack successfully [20].

b) To use the relevant random numbers k1, k2, k3. .

Suppose three random numbers k1, K2, and k3 as follows:

$$k_3 = k_1 + k_2 \tag{24}$$

From the signature equation (04) we obtain the relationship between λ:

$$\gamma_3 = \alpha^{k_3} = \alpha^{k_1+k_2} = \gamma_1\gamma_2 \tag{25}$$

The type (24) and type (25) are substituted into (18) obtained equations:

$$\begin{cases} m_1 = x\gamma_1 + k_1\lambda + t\delta_1 \\ m_2 = x\gamma_2 + k_2\lambda + t\delta_2 \\ m_2 = x\gamma_1\gamma_2 + (k_1 + k_2)\lambda + t\delta_3 \end{cases} \tag{26}$$

A hacker can obtain a binary linear equation from equations, and then get a solution set about the private key x and t. Then we verify the value of the solution set substituted into the signature equation one by one to find out the key pairs used by signer. This approach has one more power operation in the validation, but it can also be completed within a limited time .Moreover, for the same random number, using the relevant random number is more likely. A hacker would need to work out six unknowns from the three equations, and receive a quaternary linear equation. The complexity of its solution set and computing of the solution set of validation both have increased a lot.

Compared with the ElGamal digital signature algorithm, the improved algorithm is obviously much safer. Even in the state of insecurity, a hacker still needs to go through a great amount of computing to hack successfully. Signature can figure out the time limit, and change the private key and public key timely. As for the more complex link among the random number k, their attacks will be more difficult. When the degree of difficulty is so large that the hacker cannot figure out the value of the private key x within a limited period of time, algorithm can be considered safe.

Similar to ElGamal digital signature algorithm, the improved algorithm can still be used by a hacker when the random number meets the requirement of the more complex relationship. The process is more complicated, but it can still be completed within a limited time and calculations to obtain the private key. Signers also need to avoid using the same or related the value of a random number k to sign.

Time complexity

Digital signature scheme based on ElGamal involes a large integer arithmetic problem in the process of signature and verification. It is especially large integer modular arithmetic, including modular power, modular inversion and so on. These operations greatly affect its speed, especially in the signature process for each user In addition to select x as his private key, also need to select secretly random number k, and for safety reasons, k can not be reused, so users need to choose k continuously. And the inverse model of k needed to be calculated, while modular inversion of the large number is a very time-consuming thing. So in order to enhance the signature verification speed, on one hand, to improve the algorithm itself and improve processing speed; on the other, to improve digital signature scheme under the premise of not affecting the safety of signature.

The time complexity of ElGamal digital signature algorithm is (lnmln2n+ln2m+ n^2) [21]. The improved scheme proposed in this paper compared with the traditional ElGamal algorithm increases one more a random number t. with a corresponding increase in calculation. All of the calculations about the random number k in the traditional algorithms must be repeated in the calculation of t. The random number of modular inversion, modular computing power should be calculated more than once. For the signature equation of (19) and the verification equation of (20) calculation is more complicated. The amount of the calculation also increase compared to the former. The obvious efficiency compared to traditional algorithms has decreased, but its security has improved greatly, which is acceptable in real life.

For ElGamal algorithm's time complexity, algorithm of traditional computing a large integer modular power multiplication is a binary to index of m, it is said that m is expressed as a binary form. The time complexity is O ($\ln^2$m); Then a series of iterative calculations are done, which are from the peak of m's binary, if meets the bit of 1,it will multiply the results of the previous step iteration with base number. In view of safety, ElGamal choose the value of m that is relatively large, so speed of this method is slower, the time complexity is O(lnmln^{2}n). So, the time complexity of calculating a^m mod n is O(lnmln^{2}n)+O ($\ln^2$m) = O(lnmln^{2}n + $\ln^2$m).

Similarly, the function complexity of great common divisor of is O(n). The time complexity of extended Euclidean algorithm function is O(n), and the time complexity of generating the main signature function is about O (lnmln^{2}n + $\ln^2$m).

In summary, time complexity of ElGamal algorithm is about: (ln$m$ln^2n+ln^2m+ n^2)。

After algorithm being improved, the time complexity has also been changed, but there is no fundamental change. There is one more process of seeking the equation in signing step, the equation is $\lambda = \alpha^t \bmod p$ .When it seeks signatures, the calculation formula in comparison to traditional algorithms one more power of large numbers should be calculated. But these should not cause too much increases of the time complexity. At least, the order of magnitude of change is also not available, so time complexity of the improved algorithm is as follows: O(lnmln2n+ln2m+ n^2).

VI. CONCLUSION

This paper systematically analyzes the security of ElGamal digital signature algorithm under the four attacks scheme. Analysis show that there are two attacks against random number, thereby indirectly access the value of private key. And these are the two of the most likely schemes to succeed. In order to enhance the security of algorithm in random, we proposed two improved ideas: (1) Enhanced security of random numbers, making it difficult for the success of the random number of hacks; (2) Establish more complex link between the random number and the private key, so it is difficult for a hacker to use random number to attack the private key indirectly. Based on ideas that established the more complex link between the random number and the private key, we proposed to add the signature equation of the same form of an equation with the improvement of the program, thereby increasing a random number and a signature data.

The improved algorithm of security is enhanced, so that while a hacker needs greater computing capacity, the amount of signature and verification operations also will be increased. On the whole, the hacker's computational complexity is increased significantly. There are still some restrictions in the use of random numbers when the users sign. For example, we cannot make the two random numbers which were all the same and for the signature of two or more times. But generally speaking, the use of random numbers is restricted more lax than the restriction in the ElGamal digital signature algorithm. These are still to be continued to be improve. The analysis of the improved algorithm modeled on the ElGamal digital signature algorithm analysis is carried out by comparison. The new algorithm has its own characteristics. Especially, after the increase in random numbers, there may be an attacked method, which in the ElGamal-type digital signature algorithm it has not had. This is also the need for further study.

In this paper, security and efficiency analysis showed that the improved ElGamal algorithm in these two areas had significant increase or improvement, making the application wider in the production and life. But I still have to find a fact that, though in the vast majority of cases we can prevent the hacker from a variety of attacks, there are two real problem closely related to its vitality: (1) the current mathematical community for the discrete logarithm problem is still difficult for an effective solution[23]; (2) the signer must be very careful for the choice of random numbers. If there is a little vulnerability in these two areas, the ElGamal digital signature algorithm in relation should fade into history.

In conclusion, in terms of the improved algorithm in terms of security has been greatly improved, which makes its scope of application even greater. The impact due to the increase of computation in the signature and verification operations will be weakened with the enhancement of the computing power of the processor.

REFERENCES

[1] M. Bellare and P. Rogaway, “The Exact Security of Digital Signatures –Howto Sign with RSA and Rabin,” Proc. of Eurocrypt’96, Springer-Verlag, LNCS,pp.399–416, 1996. 378-379

[2] ELGAMAL T. A public key cryptosystem and a signature scheme based on discrete logarithms[J]. IEEE Trans Inform Theory.1985,31(4): 469-472.

[3] D. H¨ uhnlein, J. Merkle: An ecient NICE-Schnorr-type signature scheme, Procee-dings of PKC 2000, LNCS 1751, Springer, 2000, pp. 14-27.

[4] K. Nyberg and R.A. Rueppel, “A New Signature Scheme Based on the DSA Giving Message Recovery,” Proc. of the First ACM Conference on Computer and Communications Security, 1993. 378

[5] CHEN Hui-yan, LB Shu-wang, LIU Zhen-hua . Identity Based Signature Scheme with Partial Message Recovery [J]. ChineseJournal of Computers, 2006, 29 (9) : 1622-1627 .

[6] Chen Hai-peng, SHEN Xuan2jing, LIU Jin chan, et al . The Digital Signature Scheme All owingMessage Recovery Based on ElGamal Scheme [J]. Journal of J ilin University: I nfor mati on Science Editi on, 2008, 26 (5) : 531- 536 .

[7] CHEN Zhi-ming . An Improved Encryption Algorithm on ELGamal Algorithm [J]. Computer App licati ons and Sost ware,2005, 22 (2) : 822 85 .

[8] Jia-lun TSAI, Tzong-chen WU, Kuo-yu TSAI.A novel multisignature scheme for a special verifier group against clerk and rogue-key attacks[J]. Journal of Zhejiang University-SCIENCE C(Computers & Electronics), 2010,11(4): 290-295.

[9] HU Jian-jun, WANG Wei, PEI Dong-lin. Double-way Authentication Scheme Based on ELGamal Digital Signature[J]. Computer Engineering, 2010, 36(6): 173- 174.

[10] LI Wei-ke, LI Fang-wei.User authentication scheme based on the ElGamal signature for mobile communication system[J]. Journal on Communications, 2005, 26(11): 138-140.

[11] DONG Qing-Kuan, NIU Zhi-Hua, XIAO Guo-zhen. Research on the Freeness of Subliminal Channels in EIGamal-Type Signatures[J]. Chinese Journal of Computers, 2004, 27(6): 845-848.

[12] CAO Zhen-fu, LI Ji-guo. A Threshold Key Escrow Scheme Based on EiGamal Public Key Cryp to system[J]. Chinese Journal of Computers, 2002, 25(4): 346-348.

[13] Duanmu Qing-feng, ZHANG Xiong-wei, WANG Yan-bo, LI Bing-bing, LEI Feng-yu. ElGamal like Public key Cryptosystem and Digital Signature Scheme Based on 5F2L Sequence[J]. Computer Science,2010,37(5): 68-71.

[14] WANG Li, XING Wei, XU Guang-zhong. ElGamal public-key cryptosystem based on integral quaternions[J]. Journal of Computer Applications， 2008, 28(5): 1156-1157.

[15] D. Chaum, C. Cr´ epeau, and I. Damg˚ ard, “Multiparty unconditionally secure protocols,” STOC ’88.

[16] Yiannis Tsiounis, Moti Yung. On the Security of ELGamal Based Encryption[J]. Computer Science,1998.Vol.1431: 117-134

[17] C.P. Schnorr and M. Jakobsson : Security of Discrete Log Cryptosystems in the Random Oracle and Generic Model. TR report University Frankfurt and Bell Laboratories 1999.

[18] Claude Castelluccia ,Nitesh Saxena ,Jeong Hyun Yi. Self-configurable key pre-distribution in mobile Ad Hoc Networks[J] .Lecture Notes in Computer Science. 2005, 1083 - 1095.

[19] WANG Hua-qun, XU Ming-hai, GUO Xian-jiu. Cryptanalysis and improvement of several certificateless digital signature schemes[J]. Journal on Communications, 2008, 28(5): 88-92.
[20] DONG Qing-Kuan, NIU Zhi-Hua, XIAO Guo-zhen. Research on the Freeness of Subliminal Channels in EIGamal-Type Signatures[J]. Chinese Journal of Computers, 2004, 27(6): 845-848.
[21] Rafael C, RicardoI D. Two notes on the security of certificateless signature[A]. Provsec 2007[C]. Springer-Verlag, 2007. 85-102.
[22] You Lin, Sang Yong-xuan. Effective generalized equations of secure hyperelliptic curve digital signature algorithms[J]. The Journal of China Universities of Posts and Telecommunications, 2010,17(2): 100-108.
[23] C.P. Schnorr and M. Jakobsson : Security of Discrete Log Cryptosystems in the Random Oracle and Generic Model. TR report University Frankfurt and Bell Laboratories 1999.

Xiaofei Li, female, was born in Yantai County, Shandong Province, February, 1985. She received bachelor degree in 2009 from Changchun University and Technology.

Now she is a master candidate in the college of computer science and technology, Jilin University. Her research interests are digital image processing and pattern recognition.

Xuanjing Shen, male, was born in Helong County, Jilin Province, December, 1958. He received bachelor degree in 1982, master degree in 1984, and PhD degree in 1990 all from Harbin Institute of Technology respectively.

He is a professor and PhD supervisor currently in the college of computer science and technology, Jilin University. His research interests are multimedia technology, computer image processing, intelligent measurement system, optical- electronic hybrid system, and etc.

Haipeng Chen, male, was born in Cao County, Shandong, June, 1978. He received bachelor degree in 2003 and master degree in 2006 both from Jilin University.

Now he is a lecturer and a Ph.D candidate in the college of computer science and technology, Jilin University. His research interests are computer network security, digital image processing and pattern recognition.

Dr. Chen is is membership of China Computer Federation (E20-00 15167M).

Blind Channel Estimation with Lower Complexity Algorithm for OFDM System

Wensheng Zhu, Youming Li, Yanjuan Lu, Ming Jin
Institute of Communication Technology, Ningbo University
Ningbo, China
Email: liyouming@nbu.edu.cn

***Abstract*—Fast and reliable channel estimation is very important for wireless communication transmission. This paper addresses blind channel estimation for OFDM systems. A novel One-Cycle blind channel estimation algorithm based on cyclostationarity properties of OFDM signal is proposed in this paper, the new algorithm is formed from using the two related z-transform values of delay parameter. Furthermore, a lower complexity algorithm based on partial spectrum information can further reduce computational complexity. Computer simulation results verify that the performance of the proposed One-Cycle algorithm is superior to that of the Two-Cycle algorithm. Although the lower complexity One-Cycle algorithm has some performance loss in high SNR region, its computational complexity is dramatically reduced**

***Index Terms*—OFDM, Cyclostationarity, Blind channel estimation**

I. INTRODUCTION

The growing demand for wireless communications requires reliable and high-rate data transmission over the wireless channel. How to estimate the channel quickly and accurately is one of the key technologies to improve the reliability of information transmission. Generally, there are two kinds of channel estimation method, pilot based channel estimation [1-3] and blind channel estimation [4-9].

Blind estimation methods can be divided into high-order statistical method and second-order statistical method. The former needs a large amount of data samples, has a high complexity, and is difficult to be implemented in practical systems. Gardner first discussed the signal cyclostationarity characteristics and application in communication systems [5]. As the output second-order statistics contains the phase information of the channel due to the cyclostationarity, most non-minimum phase channels can be identified from the second-order statistics of the cyclostationary output sequence [6]. Since then, many elegant solutions such as subspace method in [8] and deterministic approach in [13] have been proposed for blind channel identification and equalization. In [9, 11], Giannakis etc use a precoder to induce cyclostationarity at the transmitter that guarantees blind identifiability of channels with minimal degradation of information rate. Heath and Giannakis [10] proposed a subspace method from using the cyclic correlation of the channel output to blindly estimate the channel in OFDM systems. This method is called Two-Cycle algorithm. In this algorithm, the spectrum information of two different non-zero cycle frequencies is used in the Two-Cycle algorithm for OFDM system, leading to lower utilization of spectrum resources and thus making degradation of the system estimation performance.

In this paper, we use the cyclostationarity induced by the cyclic prefix in the OFDM system for the Inter-Signal Interference (ISI) to develop a One-Cycle blind channel estimation algorithm for identifying the channel in OFDM systems. Different from [10], the new method is formed from analyzing the z-transform of delay variable. This algorithm is called One-Cycle algorithm due to only one cycle frequency of the spectrum information used. Furthermore, a lower complexity algorithm based on partial spectrum information can further reduce the computational complexity.

This paper is organized as follows. An overview of cyclostationarity frequency properties for OFDM system is provided in section II. The related blind channel estimation algorithm is outlined In Section III. The new proposed with lower complexity algorithms are developed in Section IV. Some computer simulation results are given in Section V. Section VI concludes this paper.

II. SYSTEM MODEL AND CYCLOSTATIONARY PROPERTIES

The block diagram for the transmitter of a typical QPSK-OFDM system is shown in Figure.1. At the OFDM transmitter, data are firstly converted from serial to parallel form. Each parallel data transmission is modulated by different carrier frequencies using QPSK or QAM scheme The OFDM modulator takes the M-point IFFT of a block input symbols. A sequence of $L < M$ symbols named cyclic prefix (see Figure.2) is appended to the beginning of each block in order to avoid inter-block interference (IBI). At the receiver, the data are retrieved by a FFT and, then, demapped with

The research is supported by China National Science Fund (60772126), Zhejiang Science and Technology Department Foundation (2009C34004), Zhejiang Province National Science Fund (Y1091155)and International co-operation project(2009DFA12120)
Manuscript received June 27, 2010.
Contacted author: Li Youming, Institute of Communication Technology, Ningbo University, Ningbo 315211, China.

doi:10.4304/jnw.6.5.783-790

corresponding scheme to obtain the estimated data as in Figure.3.

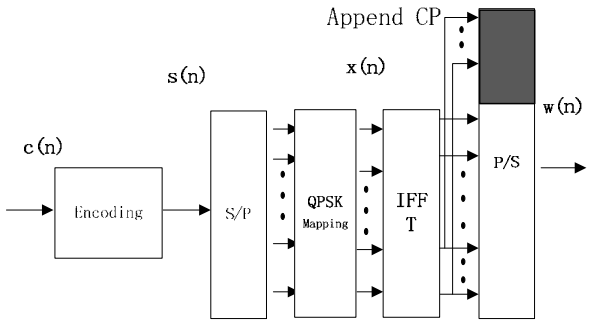


Figure 1. .The QPSK-OFDM transmitter

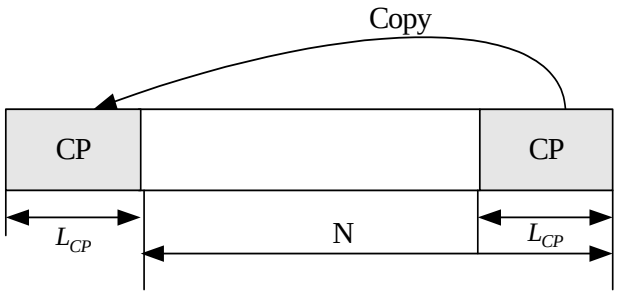


Figure 2. The OFDM frame structure

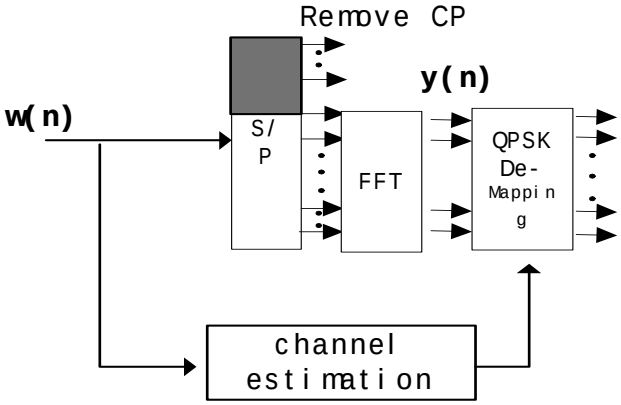


Figure 3. The QPSK-OFDM receiver

Without timing and frequency offset, the nth sample of the *m*th OFDM symbol can be described by

$$x_m(n)=x(nP+m)=\sum_{p=0}^{M-1}s_p(n)\exp(j\frac{2\pi}{M}p(m-L)) \quad (m=0,..,P-1) \quad (1)$$

where $s_m(p)$ is the *p*th subcarrier in the *m*th symbol at the input of the OFDM modulator and the $\exp(j2\pi p(-L)/M)$ accounting for the cyclic prefix, which is a repetition of the last L time domain symbols as in Figure.3.

Cyclic extensions are implemented with the cyclic repetition of the part of the IFFT output, as shown in Figure.4.Therefore periodicity is induced in the second-order statistics of the transmitted signals.

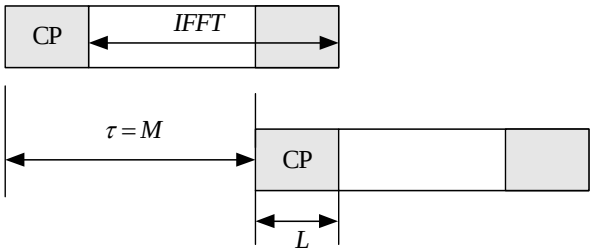


Figure 4. Autocorrelation of a OFDM block at a delay equivalent to one IFFT block Cyclic extensions

With the unit variance, independently and identically distributed (i.i.d) input stream, the transmitted signal autocorrelation appears at delays $\tau = \pm M$ as a pulse of length L and period P with the following form

$$R_x(n,\tau)=E\{x(n)x(n+\tau)^H\}$$
$$=\sigma_s^2[\delta(\tau)+\delta(\tau-M)\sum_{\tau=0}^{L-1}\delta(n-\tau)+\delta(\tau+M)\sum_{\tau=M}^{P-1}\delta(n-\tau)] \quad (2)$$

where X is the transmitted signal and superscript H stands for conjugate transpose. An example given in Figure.5 for M=12 and L=4 shows periodic pulse of length 4 at $\tau = \pm 12$.

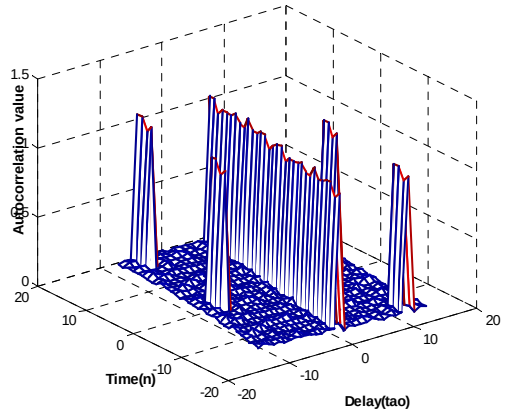


Figure 5. Example illustrating periodicity in transmitted signals autocorrelation

According to (2), the autocorrelation related to time n of the transmitted signal is a periodic function with period P. Therefore, the time-varying autocorrelation admits a Fourier series representation known as cyclic correlation. Each Fourier coefficient is defined at the k'th cyclic frequency as:

$$R_x(k,\tau)=\frac{1}{P}\sum_{n=0}^{P-1}R_x(n,\tau)e^{-j\frac{2\pi k}{P}n} \quad (3)$$

Substituting (2) into (3), the cyclic correlation $R_x(k,\tau)$ at a fixed cycle k, $0 \le k < P-1$, is given by

$$R_x(k,\tau)=\frac{\sigma_s^2 M}{P}\{\delta(\tau)\delta(k)+\delta(\tau+M)E_1(k)+\delta(\tau-M)E_2(k)\} \quad (4)$$

where

$$E_1(k)=e^{-j\frac{\pi k}{P}(L-1)}\frac{\sin(\pi kL/P)}{\sin(\pi k/P)} \quad (5)$$

$$E_2(k)=E_1(k)e^{-j\frac{2\pi Mk}{P}} \quad (6)$$

The cyclic correlation function of the transmitted signal obtained for (2) has the following properties

(a) an impulse is located at $\tau=0$ and $k=0$

(b) It has a sinc function at $\tau=\pm M$, and also is zero value located at value k with integer period P/L.

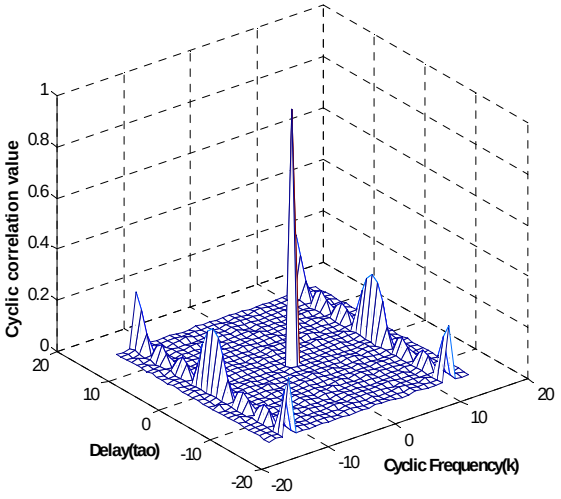


Figure 6. Transmitted signal cyclic correlation for the example in Figure.4

Assuming that the wireless channel is equivalent to the FIR filter of L_h+1 order, the receive signal is $y(n)=\sum_{l=0}^{Lh}h(l)x(n-l)+v(n)$, where $v(n)$ is stationary white Gaussian noise.

At the receiver side, after through a filter with a linear FIR channel, the autocorrelation of the received signal can be expressed as:

$$R_y(n,\tau)=\sum_{l=0}^{Lh}\sum_{q=-\infty}^{+\infty}h(l)h^*(l+\tau-q)R_x(n-l,q)+R_v(\tau) \quad (7)$$

where v represents the AWGN noise with autocorrelation $R_v(\tau)$. Linear times-invariant filtering is known to conserve cyclostationarity property. This is verified in Figure.7. In this example, the received signal is passed through h = [1 -0.8 + 0.2j 0.6 - 0.3j 0.8 - 0.5j 0.6 - 0.4j]. Cyclic correlation of the received signal is zero located at integer of P/L radio.

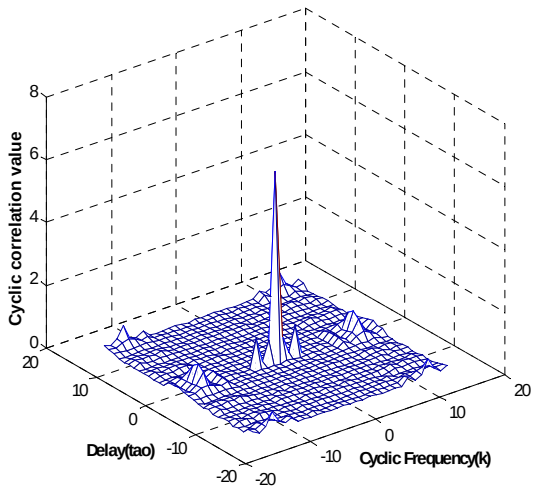


Figure 7. Received signal cyclic correlation for the example in Figure.4.Zeros are located at $k=\pm4,\pm8,\pm12,....$

The cyclic correlation of the receive signal in (7) is shown in (8)

$$\begin{aligned}R_y(k,\tau)&=\frac{1}{P}\sum_{n=0}^{P-1}R_y(n,\tau)e^{-j\frac{2\pi k}{P}n}\\&=\frac{1}{P}\sum_{n=0}^{P-1}[\sum_{l=0}^{Lh}\sum_{q=-\infty}^{+\infty}h(l)h^*(l+\tau-q)R_x(n-l,q)+R_v(\tau)]e^{-j\frac{2\pi k}{P}n}\\&=\sum_{l=0}^{Lh}\sum_{q=-\infty}^{+\infty}h(l)h^*(l+\tau-q)[\frac{1}{P}\sum_{n=0}^{P-1}R_x(n-l,q)]+R_v(\tau)\delta(k)\\&=\sum_{l=0}^{Lh}\sum_{q=-\infty}^{+\infty}h(l)h^*(l+\tau-q)R_x(k,\tau)e^{-j\frac{2\pi kl}{P}}+R_v(\tau)\delta(k)\end{aligned} \quad (8)$$

As the cyclic correlation of the noise is $R_v(k,\tau)=\sigma_v^2\delta(\tau)\delta(k)$, which is zeros for nonzero cycles k. Therefore, we will assume that $k\neq0$ to avoid stationary noise.

From (4), a conclusion can be drawn that the energy of transmitted signal autocorrelation is mainly distributed in $[-M,M]$, while that of receive signal in $[-(M+L_h),(M+L_h)]$ .In this case, the cycle correlation of receive signal is

$$R_y(k,\tau)=\sum_{l=0}^{L_h}\sum_{q=-M}^{M}h(l)h^*(l+\tau-q)R_x(k,\tau)e^{-j\frac{2\pi kl}{P}}+R_v(\tau)\delta(k) \quad (9)$$

The z-transform of the cyclic correlation with respect to τ defines the cyclic spectrum, the receive signal of which can be described by

$$S_y(k,z)=\sum_{\tau=-(M+L_h)}^{M+L_h} R_y(k,\tau)z^{-\tau} \tag{10}$$

Substituting (9) into (10), the cyclic spectrum for a special cycle $k \neq 0$ is

$$\begin{aligned} S_y(k,z) &= \sum_{\tau=-(M+Lh)}^{M+Lh}\sum_{l=0}^{Lh}\sum_{q=-M}^{+M} h(l)h^*(l+\tau-q)R_x(k,q)e^{-j\frac{2\pi kl}{P}}z^{-\tau} \\ &= \sum_{l=0}^{Lh} h(l)e^{-j\frac{2\pi kl}{P}} \sum_{q=-M}^{+M} R_x(k,q) \sum_{\tau=-(M+Lh)}^{M+Lh} h^*(l+\tau-q)z^{-\tau} \\ &\underline{\underline{l+\tau-q=\tau_1}} \sum_{l=0}^{Lh} h(l)e^{-j\frac{2\pi kl}{P}} \sum_{q=-M}^{+M} R_x(k,q)\Big[\sum_{\tau=-(M+Lh)}^{M+Lh} h(\tau_1)(z^*)^{l-q-\tau_1}\Big]^* \\ &= \sum_{l=0}^{Lh} h(l)e^{-j\frac{2\pi kl}{P}} z^l \sum_{q=-M}^{+M} R_x(k,q)z^{-q}\Big[\sum_{\tau=-(M+Lh)}^{M+Lh} h(\tau_1)(z^*)^{-\tau_1}\Big]^* \end{aligned} \tag{11}$$

and hence

$$S_y(k,z) = H(e^{j\frac{2\pi kl}{P}}z^{-1})S_x(k,z)H^*(z^*) \tag{12}$$

where (11) was obtained by changing the summation and denoting with $H(z)=\sum_{\tau=-(M+Lh)}^{M+Lh} h(\tau)(z)^{-\tau}$ the channel's Z transform and $S_x(k,z)$ is the cyclic correlation of transmitted signal.

III. THE EXISTING TWO-CYCLIC ALGORITHM

If two different non-zero cycle frequencies k1 and k2 available are taken into (12), we can get the linear equations as follows

$$\begin{cases} S_y(k_1,z)=H(e^{j2\pi k_1 l/P}z^{-1})S_x(k_1,z)H^*(z^*) \\ S_y(k_2,z)=H(e^{j2\pi k_2 l/P}z^{-1})S_x(k_2,z)H^*(z^*) \end{cases} \tag{13}$$

and hence

$$\frac{S_y(k_1,z)}{S_y(k_2,z)} = \frac{H(e^{j\frac{2\pi k_1 l}{P}}z^{-1})S_x(k_1,z)}{H(e^{j\frac{2\pi k_2 l}{P}}z^{-1})S_x(k_2,z)} \tag{14}$$

which is equivalent to

$$\begin{aligned} &S_y(k_1,z)H(e^{j\frac{2\pi k_2 l}{P}}z^{-1})S_x(k_2,z) \\ &= S_y(k_2,z)H(e^{j\frac{2\pi k_1 l}{P}}z^{-1})S_x(k_1,z) \end{aligned} \tag{15}$$

$H^*(z^*)$ can be canceled by the combination of the two above linear equations and (15) can be rewritten in the polynomial form as

$$\begin{aligned} &\sum_{\tau=-(M+L_h)}^{M+L_h} R_y(k_1,\tau)z^{-\tau} \times \sum_{q=-M}^{+M} R_x(k_2,q)z^{-q} \times \sum_{l=0}^{L_h} h(l)e^{-j\frac{2\pi k_2}{P}}z^l \\ &= \sum_{\tau=-(M+L_h)}^{M+L_h} R_y(k_2,\tau)z^{-\tau} \times \sum_{q=-M}^{+M} R_x(k_1,q)z^{-q} \times \sum_{l=0}^{L_h} h(l)e^{-j\frac{2\pi k_1}{P}}z^l \end{aligned} \tag{16}$$

In order to estimate the channel, some notations are used. Let $\mathrm{T}_y^{k_1}$ denote the $(4M+3L_h+1)\times(2M+L_h+1)$ Toeplitz matrix with first column $[R_y(k_1,M+L_h) \ \cdots \ R_y(k_1,-M-L_h) \ 0 \ \cdots \ 0]$ and first row $[R_y(k_1,M+L_h) \ 0 \ \cdots \ 0]$; $\mathrm{T}_y^{k_2}$ the $(4M+3L_h+1)\times(2M+L_h+1)$ Toeplitz matrix with first column $[R_y(k_2,M+L_h) \ \cdots \ R_y(k_2,-M-L_h) \ 0 \ \cdots \ 0]$ and first row $[R_y(k_2,M+L_h) \ 0 \ \cdots \ 0]$. Similarly, let $\mathrm{T}_x^{k_1}$ denote the $(4M+3L_h+1)\times(2M+L_h+1)$ Toeplitz matrix with first column $[R_x(k_1,M+L_h) \ \cdots R_x(k_1,-M-L_h) \ 0 \ \cdots 0]$ and first row $[R_x(k_1,M+L_h) \ 0 \ \cdots \ 0]$; $\mathrm{T}_x^{k_1}$ denote the $(4M+3L_h+1)\times(2M+L_h+1)$ Toeplitz matrix with first column $[R_x(k_2,M+L_h) \ \cdots \ R_x(k_2,-M-L_h) \ 0 \ \cdots \ 0]$ and first row $[R_x(k_2,M+L_h) \ 0 \ \cdots \ 0]$. With as transpose, let $h=[h(0) \ \ldots \ h(L_h)]^T$, $D_{k_1}=diag([1,e^{-j\frac{2\pi k_1\times 1}{P}},\ldots,e^{-j\frac{2\pi k_1\times L_h}{P}}])$ and $D_{k_2}=diag([1,e^{-j\frac{2\pi k_2\times 1}{P}},\ldots,e^{-j\frac{2\pi k_2\times L_h}{P}}])$.

The coefficients in both sides of (15) satisfy the following equations

$$\mathrm{T}_y^{k_1}\times\mathrm{T}_x^{k_2}\times D_{k_2}\times h = \mathrm{T}_y^{k_2}\times\mathrm{T}_x^{k_1}\times D_{k_1}\times h \tag{17}$$

which is equivalent to

$$(\mathrm{T}_y^{k_1}\times\mathrm{T}_x^{k_2}\times D_{k_2} - \mathrm{T}_y^{k_2}\times\mathrm{T}_x^{k_1}\times D_{k_1})\times h = 0 \tag{18}$$

The channel coefficient vector ***h*** can be uniquely recovered from (18), This forms the Two-Cyclic algorithm.

IV. THE CYCLOSTATIONARITY-BASED ONE-CYCLIC ALGORITHM AND LOWER COMPLEXITY FOR OFDM

Different from the Two-Cyclic algorithm, this paper addresses blind channel identification relying on the One-Cyclic algorithm through selecting the special variable z for the OFDM system.

A. The One-Cyclic algorithm

Firstly, taking variable z respectively as $e^{j2\pi k/P}z^{-1}$ and $e^{j4\pi k/P}z^{*}$ in (12), we generate the following two equations:

$$\begin{cases} S_y(k,e^{j2\pi k/P}z^{-1})=H(z)S_x(k,e^{j2\pi k/P}z^{-1})H^{*}(e^{-j2\pi k/P}(z^{-1})^{*}) \\ S_y^{*}(k,e^{j4\pi k/P}z^{*})=H^{*}(e^{-j4\pi k/P}(z^{-1})^{*})S_x^{*}(k,e^{j4\pi k/P}z^{*})H(e^{-j4\pi k/P}z) \end{cases} \quad (19)$$

Then canceling $H^{*}(e^{-j\frac{2\pi k}{P}}(z^{-1})^{*})$ in (19) form the polynomial as

$$\begin{aligned} &\sum_{\tau=-(M+L_h)}^{M+L_h} R_y^{*}(k,\tau)e^{j4\pi k\tau/P}z^{-\tau}\sum_{q=-M}^{+M}R_x(k,q)e^{-j2\pi kq/P}z^{-q}\sum_{l=0}^{L_h}h(l)z^{-l} \\ &=\sum_{\tau=-(M+L_h)}^{M+L_h} R_y(k,\tau)e^{-j2\pi k\tau/P}z^{\tau}\sum_{q=-M}^{+M}R_x^{*}(k,q)e^{j4\pi kq/P}z^{-q}\sum_{l=0}^{L_h}h(l)e^{j4\pi kl/P}z^{-l} \end{aligned} \quad (20)$$

Finally, four Toeplitz matrix named T_y^1, T_y^2, T_x^1, T_x^2 cab be constructed, where T_y^1 and T_y^2 are $(4M+3L_h+1)\times(2M+L_h+1)$ -dimensional Toeplitz matrix formed respectively from $R_y^{*}(k,\tau)e^{j4\pi k\tau/P}$, $R_y(k,\tau)e^{-j2\pi k\tau/P}$, and T_x^1 and T_x^2 are $(2M+L_h+1)\times(L_h+1)$ -dimensional Toeplitz matrix formed respectively in the same way from $R_x(k,q)e^{-j2\pi kq/P}$, $R_x^{*}(k,q)e^{j4\pi kq/P}$.

According to the Toeplitz matrix multiplication polynomial norms [12], we can rewrite (20) in the following matrix format

$$(\mathrm{T}_y^1\mathrm{T}_x^1-\mathrm{T}_y^2\mathrm{T}_x^2D_k)h=Th=0 \quad (21)$$

where $D_k=diag([1,e^{j4\pi k1/P},e^{j4\pi k2/P},...,e^{j4\pi kL_h/P}])$, the impulse response is $h=[h(0),h(1),.....,h(L_h)]^T$. The information of channel can be achieved by solving the linear equations (21).

B. The lower complexity One-Cyclic algorithm

Both the Two-Cyclic algorithm and the One-Cyclic algorithm have the shortcoming of large amount of computing. By analyzing the distribution of energy spectrum function, the lower complexity One-Cyclic algorithm based on partial spectrum information can further reduce the computational complexity.

Equation (4) tells us that the autocorrelation function has a non-zero value only in the M-point during $\tau>0$, this reveals that the main energy distribution of the autocorrelation function for the received signal is in $[M-L_h,M+L_h]$. Based on this result, (20) can be approximately simplified as

$$\begin{aligned} &\sum_{\tau=-L_h}^{L_h} R_y^{*}(k,M+\tau)e^{j\frac{4\pi k\tau}{P}}z^{2M-\tau}R_x(k,M)\sum_{l=0}^{L_h}h(l)z^{-l} \\ &=\sum_{\tau=-L_h}^{L_h} R_y(k,M+\tau)e^{j\frac{2\pi k\tau}{P}}z^{-\tau}R_x^{*}(k,M)\sum_{l=0}^{L_h}h(l)e^{j\frac{4\pi kl}{P}}z^{-l} \end{aligned} \quad (22)$$

Only two Toeplitz matrices named T_1, T_2 are needed and constructed respectively from $R_y^{*}(k,\tau)e^{j4\pi k\tau/P}$, $R_y(k,\tau)e^{j2\pi k\tau/P}$ with lower dimension as $(3L_h+1)\times(L_h+1)$.

With the T_1, T_2, we can rewrite (22) in the following matrix format

$$(R_x(k,M)T_1-R_x^{*}(k,M)T_2D_k)h=Th=0 \quad (23)$$

In consideration of the system performance, (23) can be expressed as

$$R_x(k,M)T_1-R_x^{*}(k,M)T_2D_k)+(R_x(k,-M)T_1-R_x^{*}(k,-M)T_2D_k)]h=0 \quad (24)$$

Finally, the linear equations can be solved by the least-square method.

Compared with the above two kinds of algorithms, the lower complexity One-Cyclic algorithm not only reduce the number of Toeplitz matrix from 4 to 2, but the dimension of the matrices is also from $(4M+3L_h+1)\times(2M+L_h+1)$ to $(3L_h+1)\times(L_h+1)$. Table 1 shows the comparison of the computational complexity among the Two-Cycle algorithm, One-Cyclic algorithm, and the lower complexity One-Cyclic algorithm.\

In generally, OFDM signal length M is much larger than the channel delay, so the lower complexity algorithm can greatly reduces the computational complexity of the original algorithm, while performance is not significantly reduced, with high practical value.

TABLE 1. THE COMPARISON OF COMPUTATIONAL COMPLEXITY FOR THE THREE METHODS

	Matrix dimension	Matrix number
The Two-Cycle algorithm	$(4M+3L_h+1)\times(2M+L_h+1)$	4
The One-Cyclic algorithm	$(4M+3L_h+1)\times(2M+L_h+1)$	4
The lower complexity One-Cyclic algorithm	$(3L_h+1)\times(L_h+1)$	2

V. SIMULATION RESULTS

In this section, the performances of the Two-Cycle algorithm, One-Cyclic algorithm, and the lower complexity One-Cyclic algorithm are compared. In order to measure the performance of the three algorithms, we define the mean-square error (MSE) as

$$\frac{1}{\| h \|^2}\sqrt{\frac{1}{N}\sum_{i=1}^{N}\| \hat{h}_i - h \|^2}$$

which is averaged over N Monte Carlos to evaluate the estimated channel error. The channel impulse response is h = [1 -0.8 + 0.2j 0.6 - 0.3j 0.8 - 0.5j 0.6 - 0.4j]. We used M=32, L=8. QPSK is used for modulation. All the results are averaged over 1000 independent runs.

In Figure 8 and Figure 9, with P=M+L=32+8=40, we respectively show the real and image part of the average estimated channel for SNR=15dB versus the number of blocks where '+','△' and '□', respectively represent the Two-Cycle algorithm, One-Cyclic algorithm, and the lower complexity One-Cyclic algorithm. The coefficients of the channel estimated respectively from the three algorithms are provided in Table 2.

TABLE 2 THE ESTIMATED CHANNEL COEFFICIENTS BY THE THREE METHODS

Two-Cycle algorithm	One-Cyclic algorithm	lower One-Cyclic
1.145 - 0.0041i	0.9788 + 0.023i	0.6046 - 0.4343i
-0.834 + 0.246i	-0.787 + 0.2038i	-0.8425 + 0.195i
0.4766 - 0.1635i	0.6341 - 0.2891i	0.6262 - 0.2970i
0.8335 - 0.6296i	0.7653 - 0.4951i	0.7097 - 0.4734i
0.5998 - 0.4507i	0.6046 - 0.4343i	0.5620 - 0.4269i

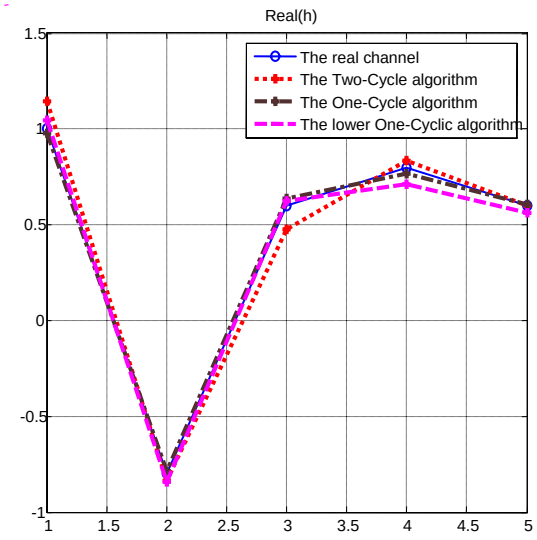


Figure 8. The real part of the channel estimates for P=40,M=32,N=1000 and SNR=20 by the three algorithm

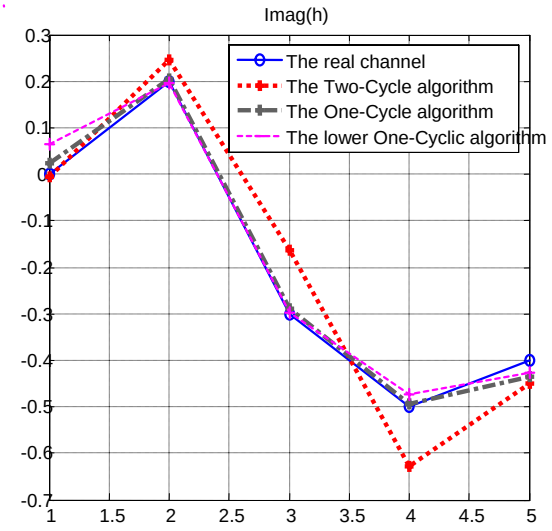


Figure 9. The image part of the channel estimates for P=40,M=32,N=1000 and SNR=20 by the three algorithm

Figure 10 shows the absolute value of autocorrelation function for the received signal versus the time delay variable when n = 5 and SNR = 15dB. From the figure we know that the main energy of autocorrelation function is concentrated in $[M - L_h, M + L_h]$ = [28, 36] interval. The lower complexity One-Cyclic algorithm is formed based on this property.

The symbol BER versus SNR is displayed in Figure 11, which demonstrates that the performance of the One-Cyclic algorithm is the best among the three algorithms, and is almost the same as the actual channel in the low SNR region, the BER for both the lower complexity algorithm and the One-Cyclic algorithm is less than that of the Two-Cyclic algorithm.

The curves of channel mean-square error (MSE) from three algorithms are plotted in Figure 12. It is seen that the Two-Cyclic algorithm has the largest MSE, matched well with the BER analysis in Figure 11. It's very interesting that the lower complexity One-Cyclic algorithm has the lowest MSE in lower SNR, and the One-Cyclic algorithm has the best performance with the high SNR. Further analysis will be discussed in another paper.

VI. CONCLUSION

A blind channel estimation algorithm and its lower complexity algorithm based on cyclostationarity properties induced by the cyclic prefix in OFDM system are proposed in this paper. In the new algorithms, the channel can be identified uniquely through analyzing the spectrum information of send and receive signal at the single frequency. The algorithms have better performance than the previous Two-Cyclic algorithm. Furthermore, a lower complexity algorithm is also developed which greatly reduce the computational complexity.

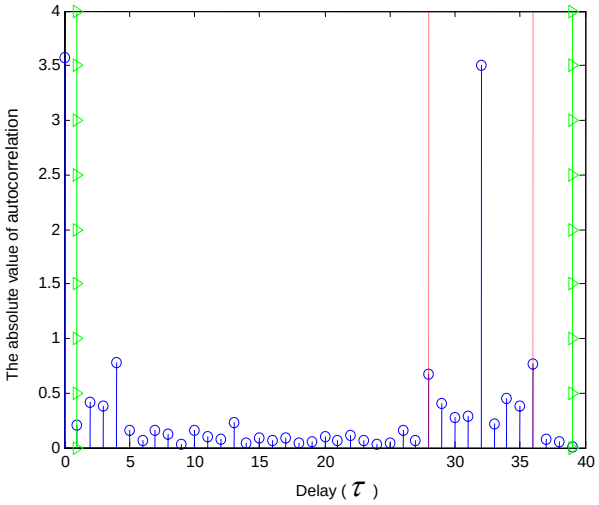


Figure.10.The received signal autocorrelation function in n = 5, SNR = 15dB.

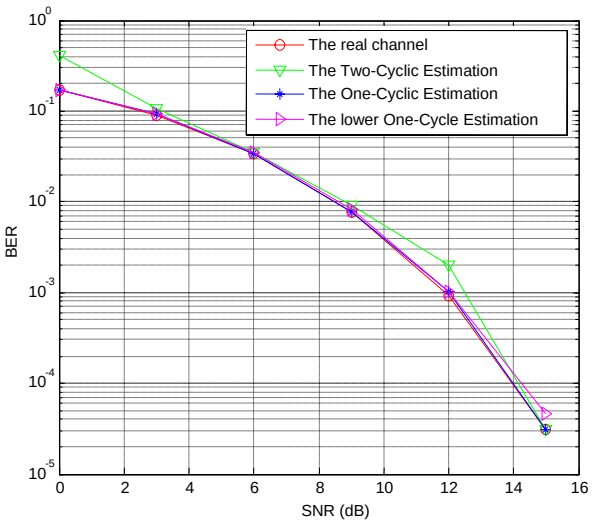


Figure.11. Symbol BER versus SNR for M=32, L=4, 1000M data

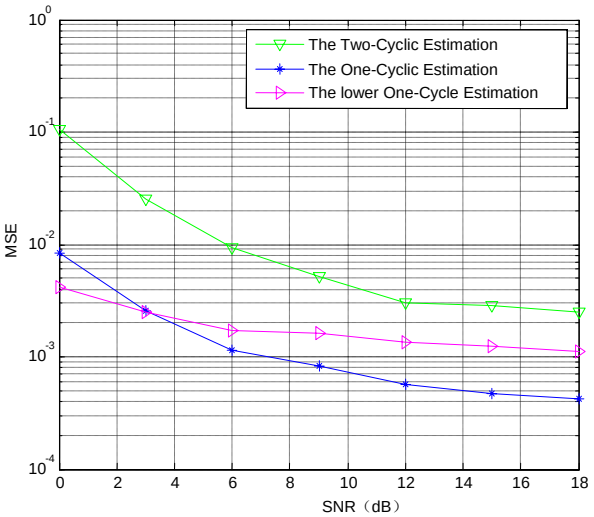


Figure.12. The channel MSE versus SNR for M=32, L=4, 1000M data

REFERENCES

[1] Dong X D, Lu W Sh, Soong A C K. Linear Interpolation in Pilot Symbol Assited Channel Estimation for OFDM [J]. IEEE Trans. on wireless communications, 2007, 6(5).

[2] Tang Z J, Cannizzaro R C, Leus G, Banelli P. Pilot-Assited Time-Varying Channel Estimation for OFDM Systems[J], IEEE Signal Processing, 2007, 5(7): 2226-2238.

[3] Kim J, Park J, Hong D. Performance analysis of channel estimation in OFDM systems[J], IEEE Signal Processing, 2005, 12(1): 60-62

[4] Bolcskel H. Blind estimation of symbol timing and carrier frequency offset in wireless OFDM systems. IEEE Trans. on Communications, 2001, 49(6): 988-999

[5] Gardner W A. Exploitation of spectral redundancy in cyclostationarity signals. IEEE Signal Processing Magazine. 1991, 39(14):14-36.

[6] Tong L, Xu G, Kailath T. A new approach to blind identification and equalization of multipath channel. In Proc. 25th Asilomar Conf. Signals, Syst., Comput., Pacific Grove, CA, Nov, 1991

[7] Amayrch A, Masson A L, Helard J. Blind Crosstalk Channel Identification in DMT-Based DSL Systems. IEEE “GLOBECOM”, 2008: 1-5

[8] Giannakis G B, Hua Y, Stoica P, Tong L. Signal Processing Advances in Wireless and Mobil Communication Volume I: Trends in Channel Identification and Equalization. Englewood Cliffs, NJ: Prentice-Hall, 2001.

[9] Chevreuil A, Serpedin E, Loubaton P, Giannakis G B. Blind channel identification and equalization using periodic modulation precoders： Performance analysis. IEEE Trans. Signal Processing. 2000, 48(6): 1570-1586.

[10] Heath R W, Jr and Giannakis G B. Exploiting Input Cyclostationarity for Blind Channel Identification in OFDM Systems. IEEE Trans. on Signal Processing, 1999, 47(3): 848–856

[11] Serpedin E, Giannakis G B. Blind Channel Identification and Equalization with Modulation-Induced Cyclostationarity. IEEE Trans. on Signal Processing, 1998, 46(7): 1930-1943

[12] Zhang Wei. Fast Polynomial Multiplocation and Toeplitz Matrix, JOURNAL OF BEIHUA UNIVERSITY. 2004,5(4):298-300

[13] Xu G, Liu H, Tong L and Kailath T. A least-square approach to blind channel identification. IEEE Signal Processing vol. 43, pp. 2982-2993, Dec. 1995.

Wensheng Zhu was born in Weifang, Shandong, China, on January 13, 1987. He received the B. S. degree in Electronic Information Science and Technology from Shandong University at Weihai, Weihai, China in 2008. He is currently pursuing his M.S. degree in Wireless Communication from Ningbo University Ningbo, China. His current research interests are in MIMO-OFDM.

Youming Li received the B. S. degree in Computational Mathematics from Lan Zhou University, Lan Zhou, China in 1985, the M. S. degree in Computationtional Mathematics from Xi'an Jiaotong University in 1988 and the Ph. D. degree in Electrical Engineering from Xidian University in 1995. From 1988 to 1998, he worked in the Department of Applied Mathematics, Xidian University where he was an Associate Professor. From 1999 to 2000, he was a Research Fellow in the School of EEE, Nanyang Technological University. From 2001 to 2003, he joined DSO National Laboratories. From Jan. 2004 to Dec. 2004, he was a Research Fellow in School of Engineering, Bar-Ilan University. Since 2005, he has joined Institute of Communication Technology, Ningbo University where he is now a Professor. His Research interests are in the areas of statistical signal processing and its application in wireless and wireline communications; MIMO-OFDM; Cognitive radio.

Neural Network with Momentum for Dynamic Source Separation and its Convergence Analysis

Hui Li
Institute of Communication Engineering, PLA University of Science and Technology, Nanjing, China
leehoo86@163.com

Yue-hong Shen and Kun Xu
Institute of Communication Engineering, PLA University of Science and Technology, Nanjing, China
chunfeng22259@126.com
xukunown@tom.com

***Abstract*—This paper addresses the problem of blind source separation (BSS) of n independent sources from their m linear mixtures in the over-determined cases ($m > n$) with unknown and dynamically changing number of sources. The system architecture including an on-line source number estimator and an auto-adjust separation mechanism is considered based on the feed-forward neural network (FNN). To speed up and stabilize the iteration procedure, we propose to modify the FNN by adding a momentum term, and convergence analysis for the new algorithm is also presented, provided that the learning rate is set as a constant and the momentum factor an adaptive variable. Computer simulation results confirm that our approach is feasible for dynamic BSS cases and has satisfied convergence speed and steady-state error performance. Moreover, the proposed algorithm can ensure the separation of weak or badly scaled signals.**

***Index Terms*—blind source separation (BSS), dynamic source number, feed-forward neural network (FNN), momentum, convergence**

I. INTRODUCTION

In the past two decades, BSS [1] has been studied extensively and found applications in many practical fields including wireless communication [2,3], speech and image processing, radar enhancement, biomedical signal processing [4,5]. The instantaneous mixing model solves the problem of extracting independent but unobserved source signals $\boldsymbol{s}$ from their observed mixtures $\boldsymbol{x}$ without available mixing coefficients as indicated in Eq.(1), where $\boldsymbol{x}(t) = [x_1(t), \cdots, x_m(t)]^T$ is the mixture vector of dimension m, $\boldsymbol{s}(t) = [s_1(t), \cdots, s_n(t)]^T$ is the unknown zero-mean source vector of dimension n and $\mathbf{A}$ is the $m \times n$ unknown nonsingular mixing matrix.

$$\boldsymbol{x}(t) = \mathbf{A} \bullet \boldsymbol{s}(t) \tag{1}$$

The separation problem is formulated to recover the waveforms of each source from output $\boldsymbol{y}$ through an un-mixing matrix $\mathbf{W}$ as follows:

$$\boldsymbol{y}(t) = \mathbf{W} \bullet \boldsymbol{x}(t) \tag{2}$$

Since the pioneering work by Jutten and Herant [6,7], many methods for BSS have been proposed, such as the independent component analysis (ICA) methods [1,8], the information-theoretic method [9], and the nonlinear principal component analysis (PCA) method [10]. Most of them assume that the number of sources is known as a priori, and typically, it should be equal to the number of sensors. In practice, however, such an assumption does not always hold, and the source number is unknown and even dynamically changing (take the random number of users within a cell in mobile communication for example). When there are more sensors than sources ($m > n$), the BSS problem is referred to as over-determined case. There are a variety of articles dealing with over-determined BSS with number of sources already known [11-14], but only a few researchers concerning about the unknown source number case [15-19]. Cichocki [15] firstly validated by extensive experiments that the natural gradient can be used directly to learn an $n \times n$ un-mixing matrix and showed that at convergence there are n independent components and $m - n$ remaining components rescaled copy of some independent components. A possible way consists a pre-whitening layer to estimate the source number and reduce the data dimension from m to n, but its separation results suffer from ill-conditioned mixing matrix or badly scaled source (i.e., some source signals are weak in comparison to others). Ye [16] has proved theoretically that the minimum mutual information criterion can still work as an effective contrast function for the over-determined BSS and proposed a generalized natural gradient algorithm in the unknown source number case. Unfortunately, its separating performance is incommensurate since the adjusting speed cannot always be applied to every case. Sun [17-19] has proposed to utilize an adaptive neural network with self-organized structure. The concept involves a system capable of instantly identifying the number of sources and adjusting the size of neural network accordingly. Nevertheless, the ANA algorithm in [19] enhances convergence stability at the expense of slower convergence speed.

doi:10.4304/jnw.6.5.791-798

The neural network algorithm which imitating the bio-nervous system is a well-known learning rule in many applications. Various BSS algorithms based on neural network have been proposed [6,15-24]. Some corresponding unsupervised learning rules are based on the observation that introducing a nonlinear function, allowing performing BSS if the distributions of the source signals meet certain conditions. Cichocki [20] utilized the feed-forward neural network (FNN) with the gradient method and proposed a unsupervised, self-normalizing, adaptive learning algorithm for roust blind identification or separation. Surprisingly it is proved to be equivalent to the subsequent natural gradient algorithm [25] latterly, the extension of the Infomax algorithm [9]. The convergence properties for two-layer FNN are discussed in [26,27]. To speed up and stabilize the training iteration procedure, a momentum term is often added to the increment formula for the weights, and it becomes one of the most popular modified methods [28,29]. However, its learning performance depends heavily upon the selection of the values of the step size and the momentum factor.

In this paper, we concentrate on the over-determined BSS with unknown and dynamic source number. Based on the existing adaptive two-stage neural network architecture [19], we make contributions including the proposition of a modified FNN-based separation algorithm for better convergence performance by adding an momentum term instead of the adaptive learning rate, and the presentation of its theoretic convergence analysis when the learning rate is set to be a constant while the momentum factor an adaptive variable.

The remainder of this paper is organized as follows. Section II describes the framework of the adaptive FNN-based system model, the source number estimation based on singular value decomposition (SVD) and the self-organized criterion. Section III presents the modified algorithm with a momentum term. Simulation results and performance comparisons are given in Section IV. Some brief conclusions are drawn in Section V.

II. The Adaptive FNN-based System Model

To solve the over-determined BSS with unknown and dynamic number of source, it is necessary to estimate the source number instantly at the very beginning. Then the dimension of received data vector should be reduced to be the same as the estimated number, so that signal components of high correlation are abandoned. And the size of neural network is also trimmed accordingly to generate or remove nodes. Hence the over-determined BSS turns into our familiar determined BSS, which has been explored widely and deeply. At last the new FNN-based algorithm with momentum we proposed can update the un-mixing matrix and achieve rapid convergence within several iterations.

The framework diagram of the adaptive FNN-based system model is shown in Figure.1.

A. SVD-based source number estimation

Many methods for estimating signal number have been researched, such as AIC [30], MDL [31], GDE [32], et.al. However, above methods require storage of quantitative samples and may fail to realize real-time estimation. The SVD-based method, valid for small sample detection, can yield reliable estimation results in both noise-absence and non-uniform noise environment. It can be implemented by three steps:

Step1: Calculate the covariance matrix of observed signals,

$$\mathbf{C} = E\left\{\boldsymbol{x}(t)\boldsymbol{x}^H(t)\right\} \tag{3}$$

where E means the expectation value.

Step2: Perform SVD on $\mathbf{C}$,

$$\mathbf{C} = \mathbf{U}diag\left\{\sigma_1^2, \sigma_2^2, \cdots, \sigma_m^2\right\}\mathbf{V}^H \tag{4}$$

where $\mathbf{U}$ and $\mathbf{V}$ are both unitary matrixes, and $\sigma_1, \sigma_2, \cdots, \sigma_m$ are the so-called singular values of $\mathbf{C}$.

Step3: Under the presupposition of noise-absence in this paper, if $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\hat{n}}$, $\sigma_{\hat{n}+1} \approx \ldots \approx \sigma_m \approx 0$, the source number can be determined to be equal to $\hat{n}$.

To satisfy the requirement of on-line estimation, a sliding window length $wind = 35$ is defined. At instant sampling time t, $\hat{\boldsymbol{x}} = \left\{\boldsymbol{x}(t-(wind-1)), \ldots, \boldsymbol{x}(t)\right\}$ is used to generate the covariance matrix $\mathbf{C}$ as indicated in Eq.(3). More details can be referred to [33].

According to the estimation result $\hat{n}(t)$, received signals components of high correlation should be abandoned, so the observed vector will be cut to:

$$\tilde{\boldsymbol{x}}(t) = \left[x_1(t), x_2(t), \cdots, x_{\hat{n}(t)}(t)\right]^T \tag{5}$$

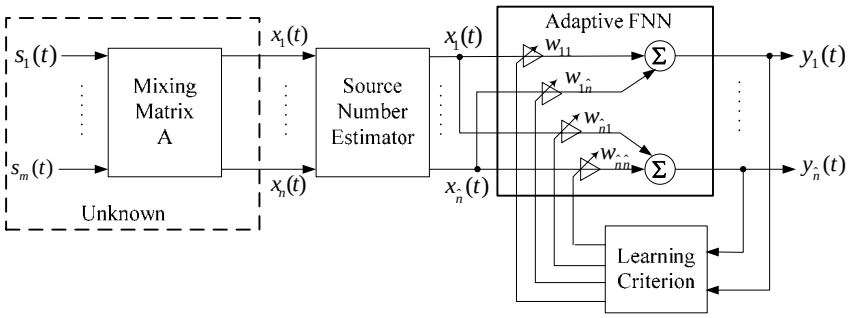


Figure 1. The framework diagram of the adaptive FNN-based system model

B. Self-organized criterion

As we find from Figure.1, the input layer and output layer of the FNN contain the same number of nodes. a pair of nodes from both layers will be generated or removed the time $\hat{n}(t)$ changes. Therefore the self-organized criterion [19] adjusts the dimension of un-mixing matrix $\mathbf{W}$ respectively according to the following three circumstances:

If $\hat{n}(t+1) > \hat{n}(t)$, which implies the number is increasing, then the dimension will be enlarged:

$$w_{ij}(t+1) = \begin{cases} w_{ij}(t), & i \le \hat{n}(t+1),\ j \le \hat{n}(t+1) \\ r, & i > \hat{n}(t+1),\ j > \hat{n}(t+1) \text{ and } i = j \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

Where w_{ij} is the i –th row j –th column element of $\mathbf{W}$, and r is a random value between $[-1,1]$.

If $\hat{n}(t+1) < \hat{n}(t)$, which implies the number is decreasing, then the dimension will be cut off:

$$w_{ij}(t+1) = w_{ij}(t),\ i \le \hat{n}(t+1),\ j \le \hat{n}(t+1) \quad (7)$$

If $\hat{n}(t+1) = \hat{n}(t)$, since the number hasn't been changed, then the dimension should keep the same:

$$\mathbf{W}(t+1) = \mathbf{W}(t) \quad (8)$$

III. The Modified Algorithm with Momentum

A. The classical neural network algorithm

Generally, the elements of un-mixing matrix $\left[w_{ij} \right]$ are deemed as the weights of neural network and updated by the gradient decent method, with the objective function a nonlinear correlation measurement.

Herault and Jutten [6] are the first to propose the following learning criterion:

$$\mathbf{W}(t+1) = \mathbf{W}(t) + \mu(t)\frac{d\mathbf{W}(t)}{dt} \quad (9)$$

$$\frac{d\mathbf{W}(t)}{dt} = f[\mathbf{y}(t)]g^T[\mathbf{y}(t)] \quad (10)$$

where $\mu(t) > 0$ denotes the learning rate, $f(\cdot)$ and $g(\cdot)$ are nonlinear different odd activation functions. In this paper, $f(y) = y^2 sign(y)$ and $g(y) = \tanh(10y)$ are adopted.

Later, Cichocki [20] considered an more efficient and robust learning criterion based on FNN:

$$\frac{d\mathbf{W}(t)}{dt} = \left[\mathbf{\Lambda} - f[\mathbf{y}(t)]g^T[\mathbf{y}(t)] \right] \mathbf{W}(t) \quad (11)$$

Where $\mathbf{\Lambda}$ is a diagonal matrix with its amplitude scaling factors $\lambda_i > 0\ \forall i$, with $\mathbf{W}(0) \neq 0$ and $\det(\mathbf{W}(0)) \neq 0$ (typically $\mathbf{\Lambda} = \mathbf{I}$, $\mathbf{W}(0) = \mathbf{I}$).

B. The proposed learning algorithm by adding momentum

We know easily from Eq.(9) that the smaller $\mu(t)$ can smooth the trajectory in the weight space at the cost of slower learning. On the contrary, if we make $\mu(t)$ large value to speed up, the resulted large changes of the weights from one iteration to the next may lead to unstable learning performance.

To overcome such contradiction, we based on the FNN-based Cichocki algorithm [20] indicated in Eq.(11) unchanged, include a momentum term in the weights update, and construct a new learning algorithm blow:

$$\mathbf{W}(t+1) = \mathbf{W}(t) + \mu(t)\frac{d\mathbf{W}(t)}{dt} + \eta(t)(\mathbf{W}(t) - \mathbf{W}(t-1)) \quad (12)$$

where the component $\eta(t)(\mathbf{W}(t) - \mathbf{W}(t-1))$ is the so-called momentum term. Obviously from Eq.(12), the convergence speed and steady-state performance of the proposed algorithm are determined by the learning rate $\mu(t)$ and the momentum factor $\eta(t)$ ($|\eta(t)| < 1$) . We choose to set μ a constant parameter and $\eta(t)$ as [34]

$$\eta(t) = \begin{cases} \gamma \left\| \frac{d\mathbf{W}(t)}{dt} \right\| \Big/ \left\| \mathbf{W}(t) - \mathbf{W}(t-1) \right\|, & if\ \mathbf{W}(t) \neq \mathbf{W}(t-1) \\ 0, & else \end{cases} \quad (13)$$

where $\gamma > 0$ is a constant parameter and $\| \cdot \|$ is the Euclidean norm.

Intuitively, if the previous weight change is large, then adding a fraction of this amount to the current weight update will accelerate the descent process towards the convergence stationary point $\mathbf{W}^*$. At convergence, assuming that μ and η are chosen sufficiently small, we have

$$E\left[\mathbf{W}(t+1)\right] = E\left[\mathbf{W}(t)\right] = E\left[\mathbf{W}(t-1)\right] = \mathbf{W}^* \quad (14)$$

As a result, (13) equals to the condition:

$$\frac{d\mathbf{W}^*(t)}{dt} = 0 \quad (15)$$

For clarity of presentation the theoretical convergence analysis of the proposed learning algorithm by the method in [34] is provided in the Appendix.

Combined above analysis, we decide to fix $\mu(t) = 0.005$ in the first learning phase (i.e. the "search" phase) and decrease it to zero exponentially in the second learning phase (i.e. the "converge" phase).

C. Summary

The procedure of our algorithm is illustrated as follows.

Beginning $\hat{n}(0) = 1$, $\mathbf{W}(0) = \mathbf{I}_{1\times 1}$, $\mathbf{\Lambda}(0) = \mathbf{I}_{1\times 1}$, $flag = 0$;

For $t = 1, 2, 3, \cdots$

Estimate the instant source number $\hat{n}(t)$ by (3) (4);

$\mathbf{\Lambda}(t) = \mathbf{I}_{\hat{n}(t)\times\hat{n}(t)}$;

if $\hat{n}(t) > \hat{n}(t-1)$,

enlarge $\mathbf{W}(t)$ by (6); $\mu(t) = 0.005$; $flag = 0$

else if $\hat{n}(t) < \hat{n}(t-1)$,

shorten $\mathbf{W}(t)$ by (7); $\mu(t)=0.005$; $flag=0$

else $\hat{n}(t)=\hat{n}(t-1)$;

keep $\mathbf{W}(t)$ unchanged by (8);

$flag = flag+1$;

if $flag=2000$;

$t_0=t$;

else if $flag>2000$;

$\mu(t)=0.005\exp\left[-2(t-t_0)\right]$;

end;

end;

Calculate the separated signals $\boldsymbol{y}(t)$ by (5) (2);

Update $\mathbf{W}(t)$ by (11) (12).

End.

VI. Simulation Results

The momentum algorithm proposed in this paper has been extensively simulated on computer. Simulation results fully confirm its validity for separation of unknown and dynamic source number and its considerable performance enhancement compared with the ANA algorithms [19]. Moreover, it is robust and efficient for weak signals which are badly scaled.

In our experiments, the sampling rate is 1 kHz and the sensors number is 8. The zero-mean mutually independent source signals are given as [19], and a piece of waveforms are displayed in Figure.2.

$$s(t)=\begin{bmatrix} s_1(t)=\sin(500t+5\cos(60t)) \\ s_2(t)=\sin(800t) \\ s_3(t)=\sin(450t)\sin(40t) \\ s_4(t)=\sin(90t) \\ s_5(t)=sign(\cos(2\pi\times 155t)) \\ s_6(t)=\text{Noise distributed uniformly in [-1,1]} \end{bmatrix} \quad (16)$$

The elements of mixing matrix $\mathbf{A}$ are randomly selected from [0,1] by computer itself, so the requirement of full column rank can be satisfied generally.

The cross-talking error [35], i.e. the performance index (PI) is utilized to evaluate separation performance:

$$PI=\frac{1}{m}\sum_{p=1}^{m}\left(\sum_{q=1}^{n}\frac{|c_{pq}|}{\max\limits_{k}|c_{pk}|}-1\right)+\frac{1}{n}\sum_{q=1}^{n}\left(\sum_{p=1}^{n}\frac{|c_{pq}|}{\max\limits_{k}|c_{kq}|}-1\right) \quad (17)$$

Where $\{c_{pq}\}=\mathbf{WA}$, multiplication of the mixing matrix and un-mixing matrix, is the so-called global matrix. The smaller PI is, the better separation performance is.

The simulation experiments include three different cases, static source number case, dynamic source number case and weak source signals case respectively. All simulation results will be displayed by averaging 80 independent runs. What should be noticed is that when we refer to $n=k(k=1,\ldots,6)$, it means that the received mixtures involve the source signals $s_i(t), i=1,2,\ldots,k$.

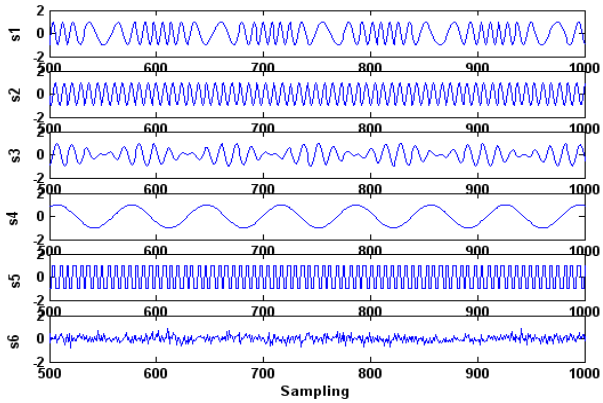


Figure 2. The waveforms of source signals

A. Static case

In this static case we set $n=4$ for sake of generality and keep it unchanged during the iteration of 10000 sampling time. Waveforms of the last 500 samples of separated signal are plotted in Figure.3. Windows without waveforms mean that no output exists. This denotation also applies to the rest of the paper. Figure.4 presents the average PI curves of the three algorithms (i.e, our proposed algorithm, the ANA algorithm [19] and the Cichocki algorithm [20]) in 80 independent runs.

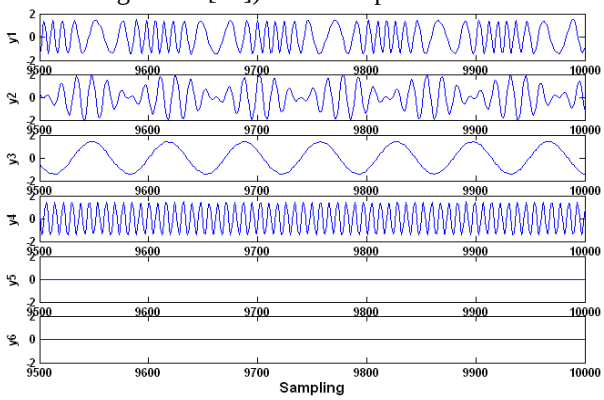


Figure 3. The waveforms of separated signals in $n=4$ case

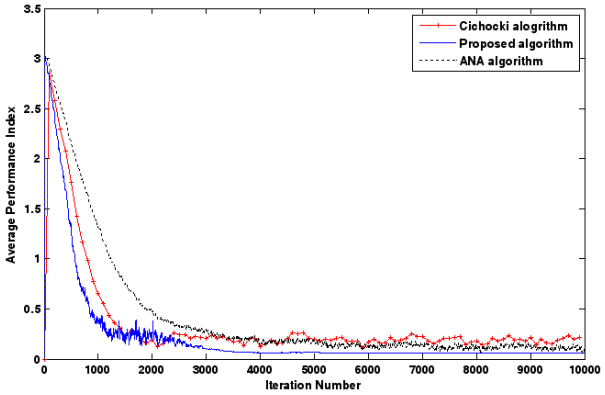


Figure 4. The average performance index comparison in $n=4$ case

B. Dynamic cases

We assume two dynamic cases each having three states switching every 6000 sample time.

(1) $n = 4,3,6$: the number of source decreases, then increases.

$$n = \begin{cases} 4, & \text{sample} \le 6000 \\ 3, & 6000 < \text{sample} \le 12000 \\ 6, & 12000 < \text{sample} \le 18000 \end{cases} \quad (18)$$

(2) $n = 3,6,2$: the number of source increases, then decreases.

$$n = \begin{cases} 3, & \text{sample} \le 6000 \\ 6, & 6000 < \text{sample} \le 12000 \\ 2, & 12000 < \text{sample} \le 18000 \end{cases} \quad (19)$$

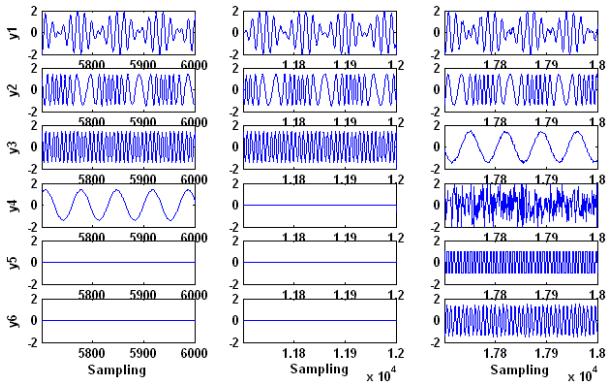


Figure 5. The waveforms of separated signals in $n = 4,3,6$ case

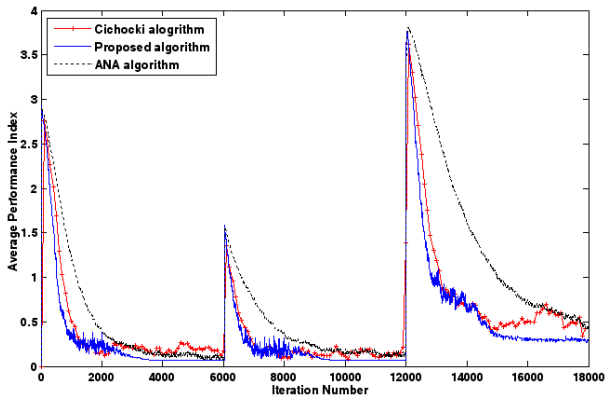


Figure 6. The average performance indexe comparison in $n = 4,3,6$ case

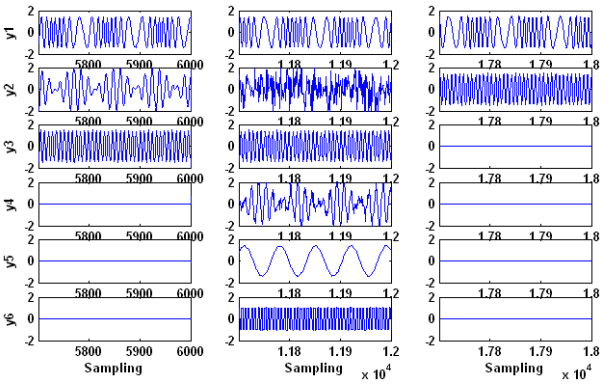


Figure 7. The waveforms of separated signals in $n = 3,6,2$ case

Presented in Figure.5 and Figure.7 are the waveforms of the last 300 separated samples in each state of the two dynamic cases. The average PI curves of the three different algorithms are compared in Figure.6 and Figure.8.

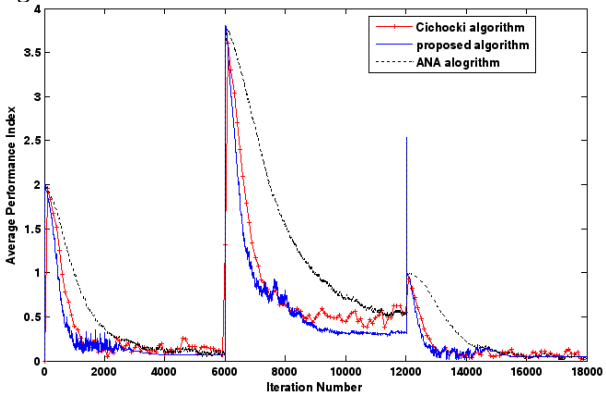


Figure 8. The average performance index comparison in $n = 3,6,2$ case

C. Discussion

From above simulation results in Figure 3, Figure 5 and Figure 7, it can be easily confirmed that the on-line SVD-based source number estimator is effective, and the proposed algorithm can separate the mixtures successfully no matter the number of sources changes or not. The reason for the separated signals' difference from the original source signals in the order is due to the inherent permutation indeterminacy of BSS.

From above simulation results in Figure 4, Figure 6 and Figure 8, we can find that the ANA algorithm has smooth trajectory at the cost of slow learning, and the Cichocki algorithm has rapid convergence speed but unstable learning procedure. Our proposed algorithm with momentum term addition is able to converge much faster than the ANA algorithm in the first search phase, and keep much more stable than the Cichocki algorithm in the second converge phase.

What should be noted is that the value of PI becomes intensively high for all the three algorithms when the number of source changes suddenly, no matter it increases or decreases, which may imply possible undesired divergence. And the mixing matrix $\mathbf{A}$ may not always maintain full column in practical environment, which may lead to separation failure.

D. Separation of weak source signals

In this experiment, three very badly scaled and weak source signals we used:

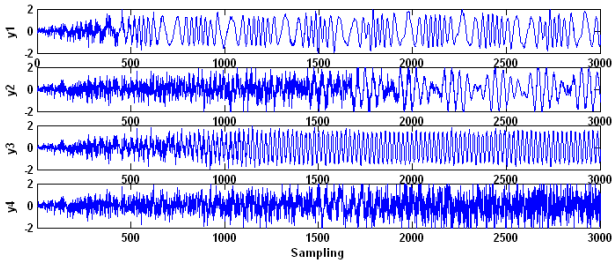
$$s_1(t) = 10^{-1}\sin(500t + 5\cos(60t))$$

$$s_2(t) = 10^{-2}\sin(800t)$$

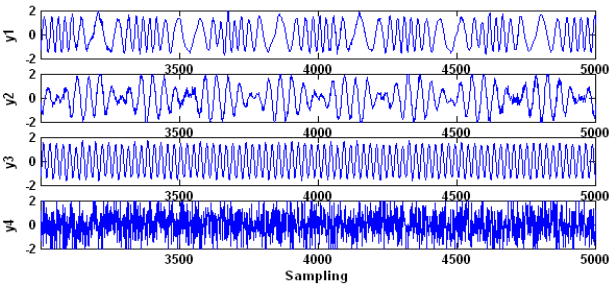
$$s_3(t) = 10^{-4}\sin(450t)\sin(40t)$$

are mixed together with a uniformly distributed noise $s_6(t)$ with amplitude 1. The mixing matrix $\mathbf{A}$ is chosen randomly as

$$\mathbf{A}_{8\times 4} = \begin{bmatrix} 0.5171 & 0.6044 & 0.4748 & -0.6571 \\ -0.2385 & -0.0580 & 0.7325 & -0.7387 \\ -0.3378 & -0.5945 & 0.9819 & -0.5624 \\ 0.0082 & 0.1592 & 0.0079 & -0.7890 \\ 0.1291 & 0.3330 & 0.2582 & -0.7171 \\ 0.5344 & 0.3535 & 0.5852 & -0.0861 \\ 0.5597 & 0.8850 & -0.1027 & 0.5763 \\ -0.0318 & 0.5403 & 0.0487 & -0.4379 \end{bmatrix} \quad (20)$$



(a)



(b)

Figure 9. (a) The waveforms of the separated signals for weak sources during the learning phase; (b) Continuation of (a).

As Figure 9 shows, after a period of learning time our proposed algorithm can ensure the successful separation of weak source signals which are badly scaled and mixed with considerable strong noise.

V. Conclusion

The paper resolves an advanced BSS issue with unknown and dynamic source number, which is much more likely encountered in practical environments. Under the system model presented, the procedure contains two stages: first is the SVD-based source number estimation and second is the adaptive FNN-based learning algorithm we proposed. The learning iteration can be divided into two phases: one is the rapid "search phase" due to the addition of the momentum term, and the other is the stable "converge phase" because of the learning rate decreased exponentially to zero. Compared with the ANA algorithm and the Cichocki algorithm, performances in both the convergence speed and the steady-state error have been enhanced obviously. Besides, the detailed deduction about its convergence analysis is also given. Moreover, the proposed algorithm can satisfy the requirement for on-line BSS, without storage of input data. And it is also efficient and robust when the sources are badly scaled weak signals. How to deal with BSS under noisy background and how to handle the possible resulted source number overestimation will be included in our next research.

Appendix

In this Appendix, we consider the convergence property of the proposed algorithm with momentum term.

Since the source signals $\boldsymbol{s}(t) = [s_1(t), \cdots, s_n(t)]^T$ are stationary zero-mean mutually decorrelated, i.e., the correlation matrix is a diagonal matrix,

$$\mathbf{R}_{ss} \triangleq E\left[\boldsymbol{s}(t)\boldsymbol{s}^T(t)\right] = \mathbf{D}_s = \mathbf{D}_s^{1/2}\mathbf{D}_s^{1/2} \quad (A\text{-}1)$$

then the un-mixing $\mathbf{W} \in \mathbb{R}^{n\times m}$ should ensure that the output signals $\boldsymbol{y}(t)$ are also mutually decorrelated, i.e.,

$$\begin{aligned} \mathbf{R}_{yy} &\triangleq E\left[\boldsymbol{y}(t)\boldsymbol{y}^T(t)\right] = E\left[\mathbf{W}\boldsymbol{x}(t)\boldsymbol{x}^T(t)\mathbf{W}^T\right] \\ &= \mathbf{W}\mathbf{R}_{ss}\mathbf{W}^T = \boldsymbol{\Lambda} \end{aligned} \quad (A\text{-}2)$$

where $\boldsymbol{\Lambda} = diag[\lambda_1, \lambda_2, \cdots, \lambda_n]$ is the diagonal matrix with positive entries (typically $\boldsymbol{\Lambda} = \mathbf{I}$).

Hence the cost (objective) function for developing an adaptive learning algorithm can be formulated as:

$$E_c = \left\|\mathbf{R}_{yy} - \boldsymbol{\Lambda}\right\|_F \quad (A\text{-}3)$$

or

$$E_c = \sum_{i=1}^{n}(r_{ii} - \lambda_i)^2 + \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n}\left|r_{ij}\right|^2 \quad (A\text{-}4)$$

where $\|\cdot\|_F$ means the Frobenius norm, r_{ij} are the elements of $\mathbf{R}_{yy}$, i.e.,

$$r_{ij} = E\left[y_i(t)y_j(t)\right] \quad (A\text{-}5)$$

$$y_i(t) = \sum_{j=1}^{m} w_{ij}(t)x_j(t) \quad (i = 1, 2, \cdots, n) \quad (A\text{-}6)$$

If we let $\boldsymbol{w} \in \mathbb{R}^{m\cdot n}$ to be

$$\boldsymbol{w} = vec(\mathbf{W}) = (w_{11}, \cdots, w_{1m}, \cdots, w_{n1}, \cdots, w_{nm}) \quad (A\text{-}7)$$

where *vec* denotes the vectorization operation of a matrix, then the cost function E_c can be rewritten as

$$E_c = g(\boldsymbol{w}) \quad (A\text{-}8)$$

where $\boldsymbol{w}$ plays as the variable and $g(\cdot)$ is a smooth activation function.

Therefore, the goal of the learning algorithm is to seek for $\boldsymbol{w}^* \in \mathbb{R}^{m\cdot n}$ such that

$$E_c(\boldsymbol{w}^*) = \min_{\boldsymbol{w}\in\mathbb{R}^{m\cdot n}} E_c(\boldsymbol{w}) \quad (A\text{-}9)$$

Applying the standard gradient descent approach to solve the minimization problem and we have

$$\frac{dw_{ij}}{dt} = -\mu\frac{\partial E_c}{\partial w_{ij}} \quad (A\text{-}10)$$

or

$$\frac{d\boldsymbol{w}}{dt} = -\mu g'(\boldsymbol{w}) \quad (A\text{-}11)$$

where μ is the learning rate mentioned before in Section III.

For convenience, we let

$$\Delta\boldsymbol{w}^{k+1} = \boldsymbol{w}(k+1) - \boldsymbol{w}(k),\ \boldsymbol{v}^k = g'(\boldsymbol{w}^k),$$

$$\eta_k = \begin{cases} \gamma \left\| \boldsymbol{v}^k \right\| / \left\| \Delta \boldsymbol{w}^k \right\|, & if\ \Delta \boldsymbol{w}^k \neq 0 \\ 0, & else \end{cases}$$

and rewrite Eq.(12) into

$$\Delta \boldsymbol{w}^{k+1} = \eta_k \Delta \boldsymbol{w}^k - \mu \boldsymbol{v}^k \tag{A-13}$$

Similarly as in [34], two assumptions shall be made:

A1): $|g(\boldsymbol{w})|$, $|g'(\boldsymbol{w})|$ and $|g''(\boldsymbol{w})|$ are uniformly bounded for $\boldsymbol{w} \in \mathbb{R}^{m \cdot n}$;

A2): $g(\cdot)$ is uniformly convex, i.e., there exits a constant $\beta_0 > 0$ such that

$$\beta_0 \left\| \boldsymbol{u} - \boldsymbol{v} \right\|^2 \leq (g(\boldsymbol{u}) - g(\boldsymbol{v})) \cdot (\boldsymbol{u} - \boldsymbol{v}), \quad \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{m \cdot n} \tag{A-14}$$

It is well known that (A-14) is equivalent to

$$\beta_1 \left\| \boldsymbol{z} \right\|^2 \leq \boldsymbol{z}^T H(\boldsymbol{w}) \boldsymbol{z}, \quad \boldsymbol{z}, \boldsymbol{w} \in \mathbb{R}^{m \cdot n} \tag{A-15}$$

where $\beta_1 > 0$ is a constant and and $H(\boldsymbol{w})$ is the Hessian matrix. From A1) we know that $H(\boldsymbol{w})$ is bounded, so there exits a constant $\beta_2 > 0$ such that $\left\| H(\boldsymbol{w}) \right\| \leq \beta_2$, thus

$$\beta_1 \left\| \boldsymbol{z} \right\|^2 \leq \boldsymbol{z}^T H(\boldsymbol{w}) \boldsymbol{z} \leq \beta_2 \left\| \boldsymbol{z} \right\|^2 \tag{A-16}$$

Theorem 1. If A1) and (A-17) are satisfied, then

$$E_c(\boldsymbol{w}^{k+1}) \leq E_c(\boldsymbol{w}^k)$$
$$\lim_{k \to \infty} E_c(\boldsymbol{w}^k) = E_c^{\ *} > 0$$
$$\lim_{k \to \infty} \left\| g'(\boldsymbol{w}^k) \right\| = 0$$

where

$$\begin{cases} \mu < \dfrac{1}{C_1} \\ 0 < \gamma < \dfrac{-1 - 2C_1\mu + \sqrt{1 + 8C_1\mu}}{2C_1} \end{cases} \tag{A-17}$$

Proof: Applying Taylor's formula, we expand $g(\boldsymbol{w}^{k+1})$ at $\boldsymbol{w}^k$:

$$g(\boldsymbol{w}^{k+1}) = g(\boldsymbol{w}^k) + g'(\boldsymbol{w}^k) \Delta \boldsymbol{w}^{k+1} + \frac{1}{2} g''(\boldsymbol{\theta}^k) (\Delta \boldsymbol{w}^{k+1})^2 \tag{A-18}$$

where $\boldsymbol{\theta}^k$ lies between $\boldsymbol{w}^{k+1}$ and $\boldsymbol{w}^k$. Incorporating (A-13) into (A-18), we get

$$\begin{aligned} E_c(\boldsymbol{w}^{k+1}) &= E_c(\boldsymbol{w}^k) + \boldsymbol{v}^k (\eta_k \Delta \boldsymbol{w}^k - \mu \boldsymbol{v}^k) + \frac{1}{2} g''(\boldsymbol{\theta}^k) (\eta_k \Delta \boldsymbol{w}^k - \mu \boldsymbol{v}^k)^2 \\ &= E_c(\boldsymbol{w}^k) - \mu \left\| \boldsymbol{v}^k \right\|^2 + \eta_k \boldsymbol{v}^k \cdot \Delta \boldsymbol{w}^k + \frac{1}{2} g''(\boldsymbol{\theta}^k) (\eta_k \Delta \boldsymbol{w}^k - \mu \boldsymbol{v}^k)^2 \end{aligned}$$

Noting that

$$\left| \eta_k \boldsymbol{v}^k \cdot \Delta \boldsymbol{w}^k \right| \leq |\eta_k| \left\| \boldsymbol{v}^k \right\| \left\| \Delta \boldsymbol{w}^k \right\| = \gamma \left\| \boldsymbol{v}^k \right\|^2$$

$$\left| \frac{1}{2} g''(\boldsymbol{\theta}^k) (\eta_k \Delta \boldsymbol{w}^k - \mu \boldsymbol{v}^k)^2 \right| \leq \frac{1}{2} \left| g''(\boldsymbol{\theta}^k) \right| (\gamma \left\| \boldsymbol{v}^k \right\| + \mu \left\| \boldsymbol{v}^k \right\|)^2 \leq C_1 (\gamma + \mu)^2 \left\| \boldsymbol{v}^k \right\|^2$$

where C_1 is a positive constant.

Let $\beta = \mu - \gamma - C_1(\gamma + \mu)^2$, then

$$E_c(\boldsymbol{w}^{k+1}) \leq E_c(\boldsymbol{w}^k) - \beta \left\| \boldsymbol{v}^k \right\|^2 \tag{A-19}$$

If μ and γ satisfy (A-17), $\beta > 0$ is ensured. As shown in (A-19), the sequence $\left\{ E_c(\boldsymbol{w}^k) \right\}$ is monotonically decreasing. Since is $E_c(\boldsymbol{w}^k)$ nonnegative, it must converge to some $E_c^{\ *} > 0$. Since $\sum_{k=1}^{\infty} \left\| \boldsymbol{v}^k \right\|^2 < \infty$, then $\lim_{k \to \infty} \left\| g'(\boldsymbol{w}^k) \right\| = \lim_{k \to \infty} \left\| \boldsymbol{v}^k \right\|^2 = 0$.

Theorem 2. If A1), A2) and (A-17) are satisfied, then there exist a unique $\boldsymbol{w}^* \in \mathbb{R}^{m \cdot n}$ such that

$$E_c(\boldsymbol{w}^*) = \inf_{\boldsymbol{w} \in \mathbb{R}^{m \cdot n}} E_c(\boldsymbol{w})$$
$$\lim_{k \to \infty} \left\| \boldsymbol{w}^k - \boldsymbol{w}^* \right\| = 0$$

Proof: Because E_c is a uniformly convex function, the existence of a unique minimum $\boldsymbol{w}^*$ follows from the inherent property of uniformly convex functions.

Noting that $g'(\boldsymbol{w}^*) = 0$, so

$$\begin{aligned} \left\| g'(\boldsymbol{w}^k) \right\| \cdot \left\| \boldsymbol{w}^k - \boldsymbol{w}^* \right\| &\geq g'(\boldsymbol{w}^k) \cdot (\boldsymbol{w}^k - \boldsymbol{w}^*) \\ &= (g'(\boldsymbol{w}^k) - g'(\boldsymbol{w}^*)) \cdot (\boldsymbol{w}^k - \boldsymbol{w}^*) \\ &\geq \beta_0 \left\| \boldsymbol{w}^k - \boldsymbol{w}^* \right\|^2 \end{aligned}$$

which indicates

$$\left\| \boldsymbol{w}^k - \boldsymbol{w}^* \right\| \leq \frac{1}{\beta_0} \left\| g'(\boldsymbol{w}^k) \right\| \to 0, \quad k \to \infty$$

REFERENCES

[1] P. Comon, "Independent Component Analysis: a new concept?" *Signal Proc.*, vol. 36, pp.287-314, 1994.

[2] T. Routtenberg and J. Tabrikian, "Blind MIMO-AR system identification and source separation with finite-alphabet," *IEEE Trans. on Signal Processing*, vol. 58, pp.990-1000, March 2010.

[3] R. Dubroca, C. D. Luigi, M. Castella, et.al, "A general algebraic algorithm for blind extraction of one source in a MIMO convolutive mixture," *IEEE Trans. on Signal Processing*, vol.58, pp.2484-2493, May 2010.

[4] E. Oja, J. Karhunen, A. Hyvarinen, et.al, "Neural independent component analysis – approaches and applications," *Brain-Like Computing and Intelligent Information Systems*, Springer, Singapor, pp.167-188, 1997.

[5] Z. Shi and C. Zhang, "Semi-blind source extraction for fetal electrocardiogram extraction by combining non-Gaussianity and time-correlation," *Neurocomputing*, vol.70, pp.1574-1581, 2007.

[6] J. Herault and C. Jutten, "Space or time adaptive signal processing by neural network models," in *Proc. AIP Conf.* Snowbird, UT, in *Neural Networks for Computing*, J. S. Denker (Ed.), New York: Amer, Inst. Phys., pp.206-211, 1986.

[7] C. Jutten and J. Herault, "Blind separation of sources, Part I: An adaptive algorithm based on neuromimetic architecture," *Signal Processing*, vol.24, pp.1-20, 1991.

[8] A. Hyvarinen, J. Karhunen and E. Oja, *Independent component analysis*, New York: Wiley, 2001.

[9] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Compuationt.*, vol.7, pp.1129-1159, 1995.
[10] J. Karhunen and J. Joutsensalo, "Representation and separation of signals using nonlinear PCA type learning," *Neural Network.*, vol. 7, pp.113-127, 1994.
[11] S. Amari, "Natural gradient for over-and under-complete bases in ICA," *Neural Computation*, vol.11, pp.1875-1883, 1999.
[12] J. F. Cardoso, "Blind signal separation: statistical principles," *Proc. IEEE*, vol.86, pp. 2009-2025, 1998.
[13] L. Q. Zhang and A. Cichocki, "Blind deconvolution of dynamical systems: a state-space approach," *Journal of Signal Processing*, vol.4, pp.111-130, 2000.
[14] L. Q. Zhang, A. Cichocki and S. Amari, "Natural gradient algorithm for blind separation of overdetermined mixtures with additive noise," *IEEE Signal Processing Letters*, vol.6, pp.293-295, 1999.
[15] A. Cichocki, J. Karhunen, W. Kasprzak, et.al, "Neural networks for blind separation with unknown number of sources," *Neruocomputing*, vol. 24, pp.55-93, 1999.
[16] J. M. Ye, X. L. Zhu and X. D. Zhang, "Adaptive blind source separation with an unknown number of sources," *Neural Computation*, vol.16, pp.1641-1460, 2004.
[17] T. Y. Sun, C. C. Liu, S. T. Hsieh., et.al, "Blind separation with unknown number of sources based on auto-trimmed neural netwoek," *Neurocomputing* vol.71, pp.2271-2280, 2008.
[18] C. C. Liu, T. Y. Sun, K. Y. Li, et.al, "A self-organized neural network for blind separation process with unobservable sources," *International Symposium on Intelligent Signal Processing and Communication Systems*, pp.177-180, 2005.
[19] T. Y. Sun, C. C. Liu, S. J. Tsai, et.al, "Blind source separation with dynamic source number using adaptive neural algorithm," *Expert System with Application*, vol.36, pp.8855-8861, 2009.
[20] A. Cichocki, R. Unbehauen, and E. Rummert, "Robust neural networks with on-line learning for blind identification and blind separation of sources," *IEEE Trans. on Circuits and Systems*, vol.43, pp.894-906, Nov. 1996.
[21] J. Cao, N. Murate, S. Amari, et.al, "A robust approach to independent component analysis of signals with high-level noise measurement," *IEEE Trans. on Neural Network*, vol.14, pp.631-645, 2003.
[22] Y. C. Zhang and S. A Kassam, "Robust rank-EASI algorithm for blind source separation," *IEEE Proc. Commun.*, vol.15, pp.15-19, 2004.
[23] A. Tonazzini, L. Bedini and E. Salerno, "A markov model for blind image separation by a mean-field EM algorithm," *IEEE Trans. on Image Process*, vol.15, pp.473-482, 2006.
[24] K. I. Diamantaras and T. Papadimitriou, "Applying PCA neural models for the blind separation of signals," *Neurocomputing* vol.73, pp.3-9, 2009.
[25] S. Amari, A. Cichocki and H. H. Yang, "A new learning algorithm for blind source separateion," In *Advances in Neural Information Processing Systems 8*, pp.757-763, MIT Press, 1996.
[26] Z. X. Li, W. Wu and Y. L. Tian, "Convergence of an online gradient method for feedforward neural networks with stochastic inputs," *Journal of Comput. Appl. Math.*, vol.163, pp.165-176, 2004.
[27] L. Behera, S. Kumar and A. Patnaik, "On adaptive learning rate that guarantees convergence in feedforward networks," *IEEE Trans. on Image*, vol.17, pp.1116-1125, Sep. 2006.
[28] J. J. Shynk and S. Roy, "Analysis of a perceptron learning algorithm with momentum updating," *International Conference on Acoustics, Speech and Signal Processing*, vol.3, pp.1377-1380, April 1990.
[29] C. C. Yu and B. D. Liu, "A backpropagation algorithm with adaptive learning rate and momentum coefficient," *Proceedings of the International Joint Conference on Neural Networks*, vol.2, pp.1218-1223, 2002.
[30] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. on Automatic Control*, 1974, vol.19, pp.716-723, 1974..
[31] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol.6, pp.461-464, 1978.
[32] H. T. Wu, J. F. Yang and F. K. Chen, "Source number estimators using transformed gerschgorin radii," *IEEE Trans. on Signal Processing* , vol.43, pp. 1325-1333, 1995.
[33] S. T. Lou and X. D. Zhang, "Fuzzy-based learning rate determination for blind source separation," *IEEE Trans. on Fuzzy Systems*, vol.11, pp.375-383, 2003.
[34] N. Zhang, W. Wu and G. F. Zheng, "Convergence of gradient method with momentum for two-layer feedforward neural networks," *IEEE Trans. on Neural Networks*, vol.17, pp.522-525, March 2006.

Hui Li received her Bachelor degree in Communication Engineering from PLA University of Science of Technology in 2007, Nanjing, China. She is currently pursing for her Ph.D. degree in the same university. Her reasearch interest includes: blind source separation and cooperative communication.

Yue-hong Shen received his Ph.D. degree in Communication Engineering from Nanjing University of Science & Technology in 1999. And now, he is professor and doctor advisor of wireless department at the Institute of Communication Engineering, PLA University of Science and Technology, China. His interests are in high-efficient modulation/demodulation and mobile communications.

Kun Xu received his Bachelor degree in Communication Engineering from PLA University of Science of Technology in Nanjing, China in 2007. He is now working for his Ph.D. degree in the same university. His research interest includes: cooperative communication, wireless resource management and multi-user information theory.

Traffic-Aware Multiple Regular Expression Matching Algorithm for Deep Packet Inspection

Kefu Xu, Jianlong Tan, Li Guo, Binxing Fang
National Engineering Laboratory for Information Security Technologies
Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China
xukefu@software.ict.ac.cn

***Abstract*—Deep packet inspection sometimes is called application level semantic detection, which is capable of examining the content of data packets in order to provide application-specific services and improve network security. Application traffic classification based on regular expressions is an essential step for deep packet inspection. However regular expression, especially multiple regular expression matching is known to require intensive system resources and is often a performance bottleneck. Currently, the DFAs of regular expression are constructed in the preprocessing stage and the context of network streams is excluded which leads to low throughputs.**

In this paper, we analyzed the application level protocols and found that the protocols are asymmetrically distributed and it is changing dynamically. From the protocol distribution characteristic, we proposed an adaptive multiple regular expression matching method for application traffic classification with deep packet inspection. The adaptive method, schedule the multiple DFAs through splay tree by matching probability other than linear scheduling in linked list, can adjust scheduling sequence according with the changing dynamic traffics. We evaluate the proposed method with the L7 rules; experiments proved that our method can improve the throughputs more than three times.

***Index Terms*—regular expressions, deep packet inspection, traffic adaptive, high-speed network**

I. INTRODUCTION

Deep packet inspection, as the process is called, arises as networks incorporate increasingly sophisticated services into their infrastructure. Such application-aware services use specific data found in packet payloads. The standard packet inspection process (a.k.a. shallow packet inspection) extracts basic protocol information such as IP addresses (source, destination) and other low-level connection states. This information typically resides in the packet header itself and consequently reveals the principal communication intent. Deep packet inspection, on the other hand, provides application awareness. This is achieved by analyzing the content in both the packet header and the payload over a series of packet transactions, consequently, provides the ability to analyze network usage and optimize network performance, thereby playing a crucial role in the equation between supply and demand faced by every network operator.

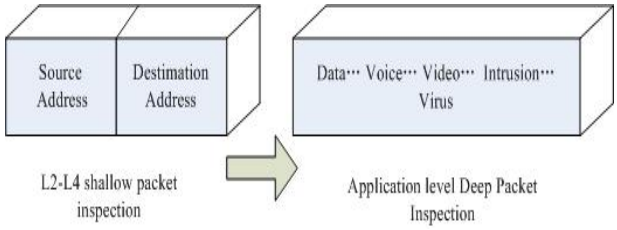


Figure 1. Deep packet inspection analyzes the application level content.

Deep packet inspection is sometimes called application level semantic detection, which associate packets into one data stream, maintain the state when searching for the signatures of the applications. It demands on-line speed to analyze, detect and reassembly the application level packet streams for high throughputs, especially for high-speed network.

Application traffic classification is an essential step for deep packet inspection process to forbid applications, bill on the content, detect intrusion or malicious attacks. Traditional approach to classify traffics of network flow is mainly port-based, which examines the TCP and TDP server port number in packet header, and maps the port to a higher layer application using IANA (Internet Assigned Number Authority) list of registered or well-known ports [1]. For example, if the port is 80 we can identify the flow as HTTP according to IANA. However, with the continuous development of the network and emergence of new network applications, port-based method becomes ineffective for lots of applications use random ports to transmit packets, even use well-known port to hide themselves. Some recent studies show that port-based approach can only identify 30%-70% today's Internet traffic [2].

Deep packet inspection focuses on the packet payload for application identification. It not only views the packet header, but also examines the payload deeply to determine the application level protocol. Deep packet

This work is supported by the National Basic Research Program of China (973) under grant No. 2007CB31110 National Natural Science Foundation of China under grant No. 61003295, xukefu@software.ict.ac.cn.

doi:10.4304/jnw.6.5.799-806

inspection matches packet payload with signatures of network applications, which are represented by explicit string or regular expression, in order to check whether or not a pattern appears in the packet payload. It has high identification accuracy and can map a flow to an application classes. This technique has been used in many traffic classification systems, such as L7-filter [3] (the Linux Application Protocol Classifier), Ethereal [4] (the world's most popular network protocol analyzer) and OpenDPI [5] (an open source version of Ipoque's commercial DPI engine, released in September 2009), all of them contain a large number of signatures which describe the characters of applications.

The biggest problem for deep packet inspection is that regular expression, especially multiple regular expression matching is known to require intensive system resources and is often a performance bottleneck. It needs a high storage and computational cost to match every packet that traverses a link, which makes it impossible to classify traffic at very high-speed links online. The DFAs of the multiple regular expressions are constructed and organized in the preprocessing stage and the information of network streams is excluded, which leads the poor performances in network environments that often lead to low throughputs. In this paper, we proposed an adaptive multiple regular expression matching method for application traffic classification. The adaptive method, schedule the multiple DFAs through splay tree by matching probability other than linear scheduling in linked list, can adjust scheduling sequence according to the change of dynamic traffics, which is suitable to identify applications in online speed. We make the following contributions:

- We analyzed the application level protocol distribution and summarize the characteristic of the distribution.
- From the characteristic, we propose a more adaptive method scheduling the multiple DFAs through splay tree by matching probability other than linear scheduling in linked list.
- We evaluate the proposed method with the Linux L7 system. Experiments proved that our adaptive method can improve the throughputs more than three times.

In the following section, we present some of the related work (Section II). In Section III we analyzed the asymmetry of the protocol and its distribution. Then the technique will be explained in detail and analyzed in Section IV. After that in Section VI, the technique's performance is evaluated and discussion about its applicability is shown. Finally, we end with final remarks, conclusion and future work in Section VII.

II. RELATED WORKS

Application traffic classification is a helpful technique for Internet service providers (ISPs) and enterprises. There are 3 parallel lines of research for traffic classification: port-based, payload-based and machine learning, each of them has its own strong points and shortcomings. The researchers of machine learning strongly disapprove of deep packet inspection, because privacy laws will make the payload inaccessible and it it need a high storage and computational cost.

Traditionally, the patterns of applications are represented by explicit strings. A. C. Yao analyzes the low bound of time complexity of string matching for a random string [6], and G. Navarro et al. extend the bound to exact and approximate multiple string matching [7]. With the development of network, there is an increasing requirement of high throughout and supporting large-scale patterns for string matching. Various new concepts and algorithms, either software-based or hardware based, have been proposed and implemented, such as ACBitmap [8], reconfigurable silicon hardware [9] and TCAM based method [10]. The bloom filters [11] have received much attention, and they are now being used in many systems, such as web caches, database systems etc.

Recently, regular expressions are replacing explicit strings as the choice of pattern describing language in packet payload scanning applications. For example, all protocol patterns in L7-filter [3] are expressed as regular expressions. More and more newer systems are replacing strings with regular expressions. The Snort NIDS [12] has evolved from no regular expressions in its rule set in April 2003 to 1131 (out of 4867 rules) using regular expressions as of February 2006. The widespread use is due to the expressive power, simplicity and flexibility for describing useful patterns.

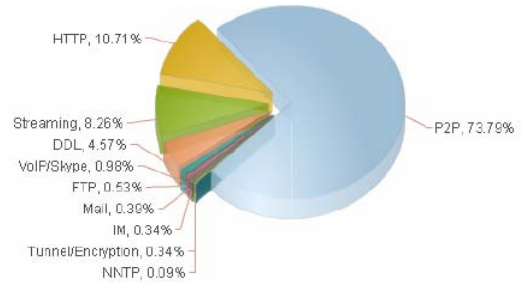
In regular expression matching algorithms, the regular expression is first parsed into an expression tree, which is transformed into a Nondeterministic Finite Automaton (NFA) in several possible ways, the most interesting in practice is the Thompson construction [13] and the Glushkov construction [14]. It is possible to search directly with the NFA, and there are various ways to do that, but the process is quite slow. The algorithm consists in keeping a list of active states and updating the list each time a new character is read. The search is normally worst-case time $O(m^{2n})$ (n is the length of packet payload), but it requires little memory. Another approach is to convert the NFA into a Deterministic Finite Automaton (DFA), which permits $O(n)$ search time by performing one direct transition per text character. On the other hand, the construction of such an automaton is worst-case time and space $O(2^m)$.

The traditional techniques to search for regular expressions can not adaptive to network traffic context. The DFAs are constructed in the preprocessing stage and the context of network streams is excluded. This leads that the performances for matching regular expressions vary in network environments and often leads to low throughputs. The network stream is naturally dynamic and protocol distribution is asymmetrical, which should be included into the design of regular expressions searching algorithms.

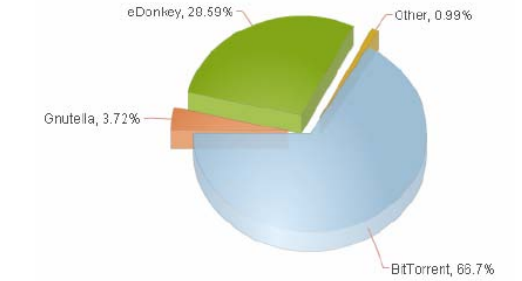
III. PROTOCOL DISTRIBUTRION

Previous studies of Internet trace have shown that a very small percentage of flows consume most of the

network bandwidth [15]. So far, there are already many researchers such as Karagiannis [16] investigated the application level protocol distribution of different network traffics. Ipoque Company analyzed the Internet traffics in five regions of the world between August and September 2007 [17]. Figure 2 shows the protocol type distribution based on the size of protocol traffic. In figure 2 HTTP do not include any audio or video streaming content embedded in Web pages like YouTube, which is counted separately. P2P is composed of BitTorrent and eDonkey etc., and the corresponding proportion is showed in figure 2(b).



（a）Protocol type distribution, Germany, 2007



（b）P2P protocol distribution by volume, Germany, 2007

Figure 2. Deep packet inspection analyzes the application level content.

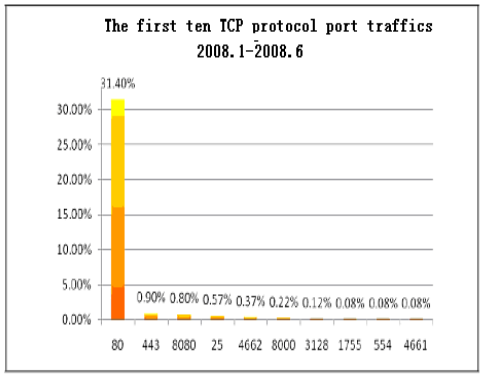


Figure 3. The first ten TCP protocol port traffics.

According to the CNCERT's (Chinese Network Center for Emergency and Reponses Technology) traffic statistics in 2008 [18], in TCP, there are four applications, Web exploring, P2P downloads, Emails and the chat tools

TABLE I. THE FIRST TEN TCP PROTOCOL PORT SERVICES

UDP ports	Traffic ranking	Percentage	Main services
15000	1	4.43%	P2P download
53	2	1.34%	DNS seivice
80	3	1.22%	Web pages
8000	4	1.20%	QQ comunication
29909	5	1.12%	Unknown port
7100	6	0.95%	Online games
1026	7	0.88%	MS Messenger
1027	8	0.75%	MS Messenger
6881	9	0.41%	P2P download
1434	10	0.38%	SQL Server Resolution

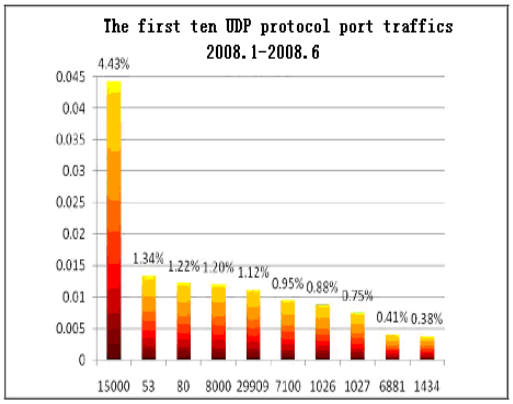


Figure 4. The first ten TCP protocol port traffics.

TABLE II. THE FIRST TEN UDP PROTOCOL PORT SERVICES

UDP ports	Traffic ranking	Percentage	Main services
15000	1	4.43%	P2P download
53	2	1.34%	DNS seivice
80	3	1.22%	Web pages
8000	4	1.20%	QQ comunication
29909	5	1.12%	Unknown port
7100	6	0.95%	Online games
1026	7	0.88%	MS Messenger
1027	8	0.75%	MS Messenger
6881	9	0.41%	P2P download
1434	10	0.38%	SQL Server Resolution

that consume the most traffic bandwidth. The first ten TCP protocol port traffics are as figure 3 and the first ten TCP protocol port service are as table I.

In UDP, the ports consume the most traffic bandwidth are the P2P downloads applications, such as eMule, BT. The DNS services also consume large bandwidth, 1.34% of the total. The first ten UDP protocol port traffics are as figure 4 and The first ten UDP protocol port services is as table II.

Many Internet and private traces shows that the frequency distribution for some of the traffic properties appears to be highly skewed [19]. Utilizing traffic characteristics to the optimization process was addressed by many researchers [20]. In [21] the work was extended

to multiple fields in firewall policies with the capability of parallel processing. The authors in [22] proposed a technique that is based on a specialized policy encoding (i.e., policy segments) in order to build Huffman trees that adapt to the traffic statistics. The technique can also be parallelized and its worst case could be bounded. An approach to find an optimal ordering of rules while maintaining policy semantics was addressed in [23]. The maintained form of the rules makes it a plausible preprocessing phase to any other technique. In [24], a hybrid approach between software and hardware was proposed; it also incorporates the traffic statistics to dynamically build new rules in the form of a cache. These new rules have better matching ratio than using original rules from the rule set. In [25], depth-constrained alphabetic trees are used to reduce lookup time of destination IP addresses of packets against entries in the routing table. The authors show that using statistical data structures can significantly improve the average-case lookup time. As the focus of the paper is on routing lookup, the scheme is limited on search trees of a single field with arbitrary statistics. In addition, the paper provides no further details on traffic statistics collection and dynamic update of the statistical tree.

IV. ADAPTIVE MULTIPLE REGULAR EXPRESSION MATCHING METHOD

A. Multiple Regular Expression

Definition 1 (Finite Automaton). Finite automaton is a finite set of states Q, among which one is initial (state I Q) and some are final or terminal (state set $F \subseteq Q$). Transitions between states are labeled by elements of Σ $\{\varepsilon\}$. These are formally defined by a transition function D, which associates to each state q Q a set $\{q_1, q_2, \ldots, q_k\}$ of states of Q for each α Σ $\{\varepsilon\}$. An automaton is then totally defined by $A = (Q, \Sigma, I, F, D)$.

In practice, there are two general types of automata, depending on the form of the transition function.

Definition 2 (Nondeterministic Finite Automaton). If the function D is such that there exists a state q associated by a given character α to more than one state, say $D(q, \alpha) = \{ q_1, q_2, \ldots, q_k \}$, $k > 1$, or there is some transition labeled by ε, then the automaton is called a nondeterministic finite automaton (NFA), and the transition function D is denoted by the set of triples $\Delta = \{(q, \alpha, q'), q$ Q, α Σ $\{\varepsilon\}, q'$ $D(q, \alpha)\}$.

Definition 3 (Deterministic Finite Automaton). Deterministic finite automaton D is denoted by a partial function $\delta: Q \times \Sigma \rightarrow Q$, such that if $D(q, \alpha) = \{q'\}$, then $\delta(q, \alpha) = q'$. We give examples of both types of automata in figure 5. In both, the state 0 is initial state and the double-circled states are terminal. The left automaton is nondeterministic since from the state 0 by T we reach 2 and 6. The right one is deterministic because for a fixed transition character all the states lead to at most one state.

Definition 4 (Pattern Recognized by Automaton). A pattern is recognized by the automaton $A = (Q, \Sigma, I, F, \Delta)$ or $A = (Q, \Sigma, I, F, \delta)$ if it labels a path from an initial to a final state. The language recognized by an automaton is

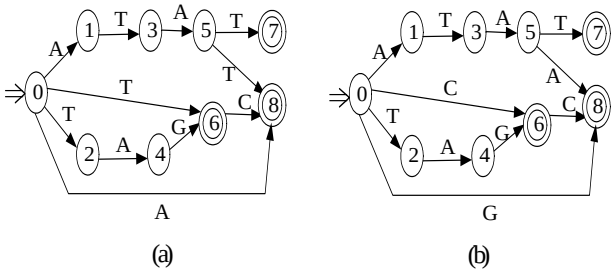


Figure 5. Two type of automatons.

the set of patterns it recognizes. For instance, the language recognized by the automaton in figure 5 (a) is

the set of patterns: *A* in state 8, *ATAT* in states 7 and 8, *T* in state 6, *TC* in state 8, *TAG* in state 6, and finally *TAGC* in state 8.

Application traffic classification by regular expressions can commonly divided into several steps. Firstly, the signatures of each application protocol are analyzed, extracted and written in the form of regular expressions. Then the regular expressions are compiled into NFAs and the NFAs are converted into DFAs. The DFAs require too much storage space, it is impossible to compile all the NFAs into one DFA. So NFAs are accordingly compiled into individual DFAs. The DFAs are organized into a linked list and when the packets arrive, DFAs in the linked list are selected one by one from left to right linearly to search for the signatures in the packets, as is showed in figure 6.

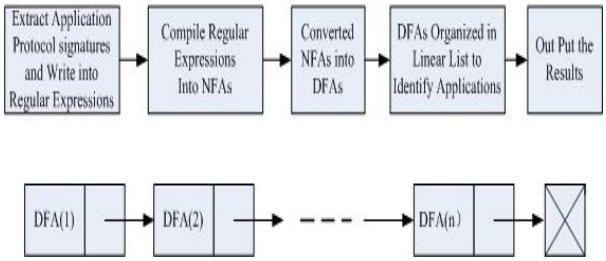


Figure 6. DFAs are organized into a linked list and scheduled linearly to search in the packets for the signature.

B The Problem of DFA Scheduling

As we analyzed in the section III, application level protocol distribution is asymmetric, DFAs in the linked list has different probability to matching successfully. Currently, DFAs are ordered in fixed sequence, and are scheduled in fixed linear sequence to identify the traffic; it will try too many times to make a successfully matching. We know that the regular expression algorithms matching are slow and need much computing resource; linear scheduling will consume much more time and slow down the throughputs. To reduce the regular expression matching time and improve the throughputs, a method is to adjust the schedule sequence to make the highest matching probability DFA be select to check the

traffic firstly. Another problem is that the network stream is naturally change dynamically, sometimes, some DFAs have the higher probability to be matched, in the other time, the probability changed and the schedule sequence should be changed accordingly.

In deep packet inspection applications, the DFAs scheduling problem can be described as follows:

- Is it returned when any rule is matched, or it needs to find all rules? If it is the former, the signatures searching will stop when any DFA is successfully matched; otherwise, all the DFAs need to be checked. Here we assume it is the former, as is used in most network security applications.
- Supposing there are k ($1 \leqslant i \leqslant k$) DFAs, and the i-th DFA has the probability p_i to be matched, then the probability of all the DFAs that not matched is $q = 1 - \sum_{1<i<k} p_i$. If the value of q is large, the time for the matching is nearly to the worst case, on the other words, it needs to traverse all the DFAs.
- If the p_i is changed dynamically, how to schedule the DFAs to adaptive to the flows?

We suppose that:

(a) DFAs are scheduled in linear sequence;
(b) It is returned when any rule is matched.

The DFAs scheduling can be illustrated as figure 7.

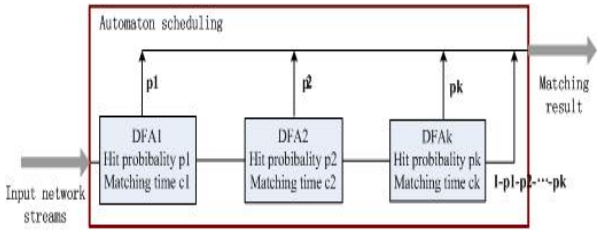


Figure 7. Problem of DFAs scheduling.

As is showed in figure 7, the mean time price for scanning is:

$$T = p_1 c_1 + p_2(c_1 + c_2) + \cdots + p_k(c_1 + c_2 + \cdots + c_k) + (1 - p_1 - p_2 - \cdots - p_k)(c_1 + c_2 + \cdots + c_k) \quad (1)$$

It can be proved that to minimize the mean time price, the optimal scheduling sequence must suffice the following conditions:

$$\frac{p_1}{c_1} > \frac{p_2}{c_2} > \cdots > \frac{p_k}{c_k} \quad (2)$$

If $c_1 = c_2 = \cdots = c_k = O(n)$, the mean time price is:

$$T = p_1 + 2p_2 + \cdots + kp_k + k(1 - p_1 - p_2 - \cdots - p_k) \quad (3)$$

The optimal scheduling sequence is:

$$p_1 > p_2 > \cdots > p_k. \quad (4)$$

C Adaptive DFA Scheduling Method Based on Improved Splay Tree

According to the dynamic and asymmetric distribution network traffic, the traditional DFA scheduling method should be adapted. Here we use an adaptive DFA scheduling method based on improved splay tree.

Assuming that there are a serial lookup operations in tree-based data structures, to minimize the times of the lookup operations, the most frequent items should be near to root of the tree. We need to adjust the structure of the tree every time after the lookup operations and try to move the item that currently matched near to the root. After some time, the high frequent items will centralized on the root of the tree.

Splay tree is a binary sort tree, and is constructed by Daniel Sleator and Robert Tarjan [27]. Splay trees are standard example of self-adjusting binary search trees. They have great advantages over explicitly balanced trees, as they automatically adapt to various non-uniform access patterns. Nodes of splay balanced trees have the relative values while DFAs are equality, so we improved the splay tree.

At first, we construct a complete binary tree to stand for a splay tree, compile the rules into DFAs and insert every DFA as nodes into the tree, as is shown in figure 8. Then levelly traverse the tree and find the DFA that can identify the current application protocol. If any DFA is successfully matched, stop the traverse and adjust the tree according to adaptive policy by moving the node contain the DFA up nearly to the root of the tree.

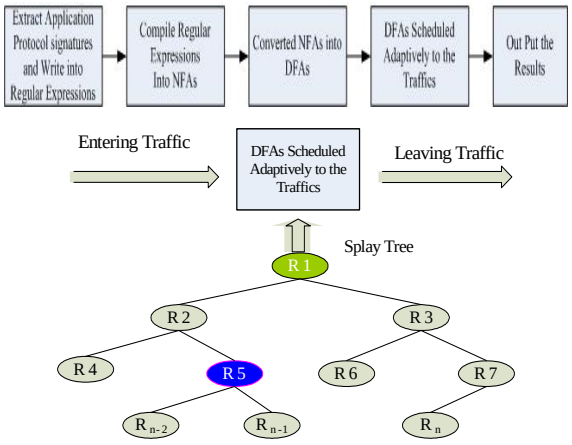


Figure 8. DFAs scheduled by splay tree.

After a period of time adjusting, the higher probability DFAs will centralized on the root, the others will sinking down to the bottom of the tree and the scheduling could also be adaptive to dynamic traffics. By statistics, the times for searching DFAs will reduce and the throughput will increase.

The new algorithm may be described as follows:

Procedure Adaptive-Multiple-RE(TCP-links, Splay-tree, DFA)

(1) Compile Regular Expressions Into NFAs;
(2) Converted NFAs into DFAs;
(3) Construct the Improved Splay-tree;

(4) Insert the DFAs into the Splay-tree;
(5) Searching // Check each TCP-links with the DFAs;
(6) Read the packets from the TCP-links;
(7) For (TCP-link0, TCP-link-Last) output protocol of search of TCP-links
(8) Do
(9) Level traverse the splay tree from the root;
(10) If any DFA matching successfully, report the DFA;
(11) Stop the traverse;
(12) Adjust the tree according to adaptive policy in advance;
(13) Go back to the root of the splay tree;
(14) End If;
(15) End Do;
(16) End For;
(17) End;

V. EVALUATION AND ANALYSIS

A Experiments Environment

We tested the two scheduling method and contrasted the scheduling number of times and throughputs between them. The testing machine is VMware Workstation 6.5-7.0 over windows XP OS, the CPU is Intel(R) Core(TM)2 Duo 2.40GHZ, 1.00GB of RAM. The traffic classification rules are extracted from Linux L7-filter. Three are total 125 regular expressions. We gather 350GB data set from the Chinese core network; there are total 615,924 TCP stream links in the data set.

B Scheduling Number of Times

We tested the scheduling number of times in linear scheduling and adaptive scheduling. At the start, the method of adaptive splay-tree needs a little time to adjust to the traffic. Figure 9 is the scheduling number of times for the first five hundreds TCP links of the linked list method. The DFAs organized in the linked list need to try average 36.2 times. The method base on splay tree adjusting speed is fast, after a little time for adjusting, the scheduling number of times reduced notability and the meaning scheduling times is 13.7 as is shown in figure 10.

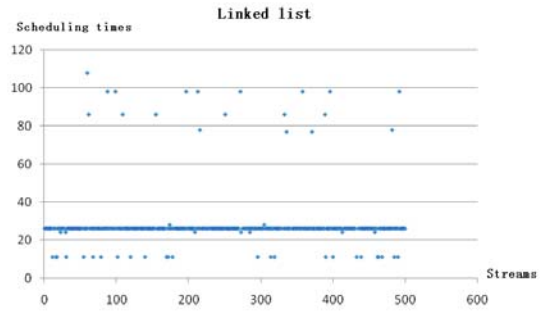


Figure 9. Scheduling times for the first five hundreds TCP links based on linked list

Figure 11 is the scheduling number of times contrast between the two methods. At the beginning, there are nearly the same scheduling times for the two methods, but after a little time adjusting, the scheduling times reduced rapidly for the adaptive method based on splay tree. The method based on linked list need more than two times scheduling numbers before a successfully matching. After a little more time, the times for adaptive scheduling will be even less.

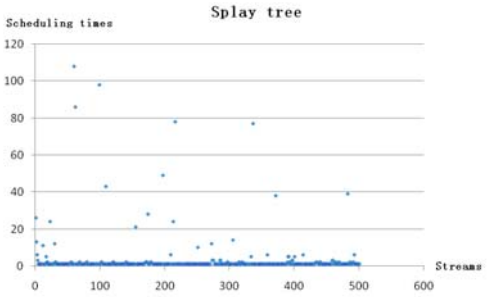


Figure 10. Scheduling times for the first five hundreds TCP links based on splay tree

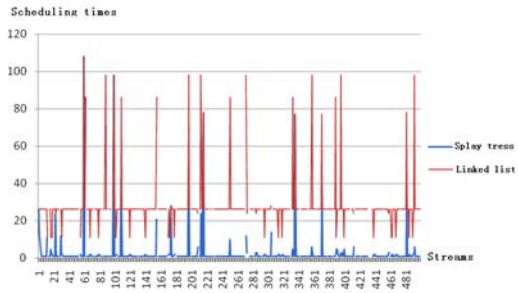


Figure 11. Scheduling times contrast between the two methods.

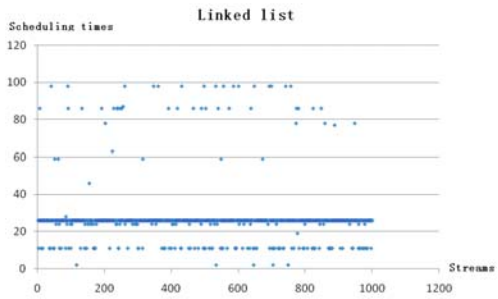


Figure 12. Scheduling times for one thousand stream links of the linked list method.

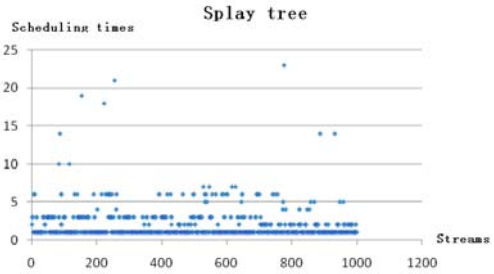


Figure 13. Scheduling times for one thousand stream links of the splay tree method

In figure 12, 13 and 14, we give out the scheduling times and the contrast for one thousand TCP links from 300,001th to 301000th. The adaptive scheduling times is less than 10 after adjusting and the scheduling number of

times is obviously less than the linked list, as is shown in figure 14.

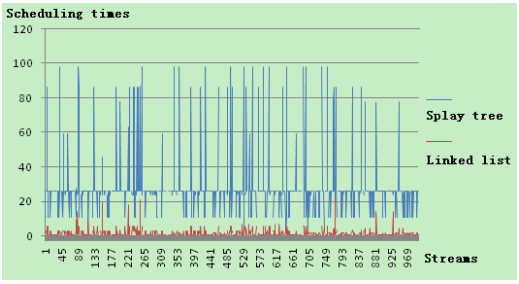


Figure 14. Scheduling times contrast between the two methods for one thousand stream links.

C *Throughputs Contrast*

There are total 615,924 TCP stream links in the data set. We observe the run time from the fixed linked list method and the traffic-adaptive method. Above 90% TCP stream links could be recognized. The throughputs could be raised more than 2.5 times by the traffic-adaptive method than the fixed linked list method as is shown in table III. Consider that the 2.5 times performance upgrade is of the whole system, there are many resource and time consuming operations include the data reading, TCP stream assembly, stream state maintain, stream searching. So only consider the scheduling algorithms, the actually performance optimization by traffic-adaptive scheduling algorithms should be about 3.5-4 times.

VI. CONCLUSIONS

In this paper, we analyzed the application level protocol distribution and proposed an adaptive multiple regular expression matching method for application traffic classification with deep packet inspection. The adaptive method, schedule the multiple DFAs through splay tree by matching probability substituted as linear scheduling in linked list, can adjust scheduling sequence according to the change of dynamic traffics. We evaluate the proposed and compare it with the L7 system and the actually performance improve by traffic-adaptive method is about 3.5-4 times. The Traffic adaptive is suitable to identify applications online.

TABLE III. TIME CONSUMING CONTRAST BETWEEN LINKED LIST AND SPLAY TREE

	Numbers of TCP links	Time for Linked list	Time for Splay tree	Recognition ratio
Test1	615924	450s	188s	89.1%
Test2	615424	421s	190s	89.1%
Test3	615924	474s	209s	89.1%
Mean result	615924	445s	195.6	89.1%

REFERENCES

[1] IANA: TCP and UDP port numbers". http:// www.iana.org/ assignm ents/port-numbers.
[2] A. Madhukar and C. Williamson. "A Longitudinal Study of P2P Traffic Classification". MASCOTS 2006.
[3] J. Levandoski, E. Sommer, and M. Strait." Application Layer Packet Classifier for Linux". http://l7-filter.sourceforge.net/.
[4] The world's most popular network protocol analyzer". http://www.ethereal.com/.
[5] Broadband World Forum Europe 2009. http://blog.ipoque.com/ 2009/09/dpi-the-end-of-the-internet/#more-377.
[6] A. C. Yao. "The complexity of pattern matching for a random string". SIAM Journal of Computing, 8(3):368-387, 1979.
[7] G. Navarro, and K. Fredriksson. "Average complexity of exact and approximate multiple string matching". Theoretical Computer Science, 321(2-3):283-290, 2004
[8] N. Tuck, T. Sherwood, B. Calder, and G. Varghese. "Deterministic memory-efficient string matching algorithms for intrusion detection". In Proceedings of the IEEE INFOCOM Conference, 2004, pp. 333-340.
[9] J. Moscola, J. Lockwood, R. P. Loui, and M. Pachos. "Implementation of a content-scanning module for an internet firewall". Proceedings of IEEE Symposium on Field-Programmable Custom Computing Machines (FCCM), Napa, CA, April 2003, pp.31-38.
[10] F. Yu, R.H. Katz, and T.V. Lakshman, Gigabit rate packet patternmatching using TCAM, Proceedings of the network protocols, 12th IEEE International Conference on (ICNP04), Oct. 2004, pp.174-183.
[11] S. Dharmapurikar, P. Krishnamurthy, T. Sproull, and J. Lockwood. "Deep packet inspection using parallel Bloom filters". IEEE Micro, Vol. 24, No. 1, 2004, pp. 52-61.
[12] SNORT Network Intrusion Detection System". http://www.snort. org.
[13] K. Thompson. "Programming Techniques: Regular expression search algorithm". Communications of the ACM, 11(6):419-422, 1968.
[14] V-M.Gluskov. "The abstract theory of automata". Russian Mathematical Surveys, 16:1-53, 1961.
[15] K. Lan and J. Heidemann. On the correlation of internet flow characteristics. Technical Report ISI-TR-574, USC/ISI, 2003.
[16] T. Karagiannis, A. Broido, N. Brownlee, K. Claffy, and M. Faloutsos. "Is P2P dying or just hiding?". In IEEE Globecom 2004, Dallas/Texas, USA, Nov, 2004.
[17] http://www.ipoque.com/resources/internet-studies/internet-study-2007
[18] CNCERT's (Chinese Network Center for Emergency and Reponses) traffic statistics in 2008.
[19] A. El-Atawy, T. Samak, E. Al-Shaer, and H. Li. On using online traffic statistical matching for optimizing packet filtering performance. In IEEE INFOCOM'07, May 2007.
[20] Adel El-Atawy, Ehab Al-Shaer, Tung Tran, Raouf Boutaba, "Adaptive Early Packet Filtering for Defending Firewalls against DoS Attacks", In the 28th Annual IEEE Conference on Computer Communications (INFOCOM'09), Rio De Janeiro, Brazil, April 2009.
[21] H. Hamed, A. El-Atawy, and E. Al-Shaer. Adaptive statistical optimization techniques for firewall packet filtering. In IEEE INFOCOM'06, April 2006.
[22] A. El-Atawy, T. Samak, E. Al-Shaer, and H. Li. On using online traffic statistical matching for optimizing packet

filtering performance. In IEEE INFOCOM ’ 07, May 2007.
[23] Hazem Hamed and Ehab Al-Shaer. Dynamic rule-ordering optimization for high-speed firewall filtering. In ASIACCS ’ 06: Proceedings of the 2006 ACM Symposium on Information, computer and communications security, pages 332–342, New York, NY, USA, 2006. ACM.
[24] Qunfeng Dong, Suman Banerjee, Jia Wang, and Dheeraj Agrawal. Wire speed packet classification without tcams: a few more registers (and a bit of logic) are enough. SIGMETRICS Perform. Eval. Rev., 35(1):253–264, 2007.
[25] H. Hamed and E. Al-Shaer. Adaptive statistical optimization techniques for firewall packet filtering. Technical Report TR-05-012, DePaul University, 2005.
[26] Daniel Dominic Sleator , Robert Endre Tarjan, Self-adjusting binary search trees, Journal of the ACM (JACM), v.32 n.3, p.652-686, July 1985.

Kefu Xu received the Ph.D. degree in computer science from The Center South University of Technology, Guangdong, China, in 2009.

He is now a researcher as postdoctoral at Institute of Computing Technology; Chinese Academy of Sciences. He has published over 10 research papers in journals and conferences in the field of computer networks. His research interests include network measurement and monitoring, network security. Particularly, his research interests include algorithm design, high-speed deep packet processing systems and converged networks.

Jianlong Tan received the Ph.D. degree in computer science from The Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2003.

He is an associate researcher and master supervisor at Institute of Computing Technology, Chinese Academy of Sciences. His main research interests include algorithm design, data stream management and information security. He has published over 30 research papers in journals and conferences in computer network.

Li Guo received the Master degree in computer science from The Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 1999.

She is now a senior research engineer and master supervisor at Institute of Computing Technology, Chinese Academy of Sciences. Her main research interests include information security, and data stream processing. She has published over 40 research papers in leading journals and conferences in computer network and security.

Accurate Detection of Peer-to-Peer Botnet using Multi-Stream Fused Scheme

Jian Kang and Yuan-Zhang Song
Department of Computer Science & Technology, Jilin University, Changchun, China
Email: {kj885788,tao8565}@gmail.com

Jun-Yao Zhang
Department of EECS, University of Central Florida, Orlando, FL USA
Email : zhangjunyao5518@gmail.com

***Abstract*—Nowadays decentralized botnets pose a great threat to Internet. They evolve new features such as P2P Command and Control(C&C), which makes traditional detection methods no longer effective for indicating the existence of the bots. In this paper, based on several of the new P2P botnet characteristic properties, we propose a novel real-time detecting model – MSFM (Multi-Stream Fused Model). MSFM considers multiple types of packets' unique characteristics and handle them with corresponding strategies. Extensive experiment results show that our model can accurately detect peer-to-peer botnet with relatively low false-positive and false-negative rates.**

***Index Terms*—decentralized botnet, Hurst, discrete Kalman filter, Multi-chart CUSUM, peer to peer**

I. INTRODUCTION

Botnets, mainly formed by compromised machines which are infected by various ways such as worms and Trojans, are engaged in malicious activities like DDOS, Spam and etc. It is recently considered as one of the most important threats to the Internet security. From SANS security report 2008[1], the Botnet is listed at 2th out of 10 crucial risks in the new millennium.

In recent years the newly-evolved P2P botnets are flourishing in the Internet. Partly because the "traditional" ones, such as IRC and HTTP botnets, are now easily detected or controlled by taking measurements on the Command and Control(C&C) servers, while in P2P network, every peer serves as both client and server, which made it influence little by single-point(server) failure. Also, the new P2P botnets are using new techniques like Rootkits, Fast-flux and etc. One example is Storm Botnet, appeared in early 2007[2] and quickly developed to the "biggest" botnet of the Internet world. It indicates a more sophisticated P2P approach by using an embedded decentralized architecture. As its threats to the Internet Security are increasing, searching for detecting and mitigating methods has become urgent. Therefore, in this paper, 1) we present a brief overview of Storm's mechanism; 2) we propose a novel real-time detecting model- MSFM (Multi-Stream Fused Model). In this model, we incorporate different detecting methods based on unique characteristics of network flow packets.

Firstly, MSFM lays emphasis on the UDP flow, which is most related with the botnet C&C (Command and Control). We use the Hurst Parameter to retrieve the whole condition, and detect the abnormality based on the UDP packets' self-similarity. Secondly, MSFM applies the discrete Kalman filter to find ICMP and SMTP packets' anomaly and Multi-chart CUSUM acts as the amplifier to make the abnormality clearer. Finally, we consider the impact on the botnet detection which web applications generate, especially the P2P applications, and use the properties of TCP flow to analyze that the abnormalities are caused by the botnets or the P2P applications. In addition, this paper uses the Kaufman algorithm to dynamically adjust the threshold to minimize the false positives and false negatives. After series of experiments, the results prove that the model can detect the Storm botnet in a relatively high precision with both low false-positive and false-negative rate.

II. RELATED WORK

Current research works on decentralized P2P botnets are still at the beginning stage:

Julian B. Gizzard et al. [3] study and analyze the storm's mechanism thoroughly, including the infection steps, communication ways and so forth, which is valuable for later studies.

Sandeep Sarat et al. [4] and Thorsten Holz et al. [5] analyze and monitor the Storm in a similar way. The former draws the conclusion: the distribution of peer IDs of Storm is irregular and there are many unreachable/private IPs. The later offers a novel botnet mitigation thought—infilter into the Overnet network as a peer and publish large number of keys to delay or disrupt the communication between bots.

Reference[6] present some of the behavior characteristics of the P2P botnets, and offer thoughts of controlling bots on hosts, such as using the System Service Table (SST) Hooking and etc.

doi:10.4304/jnw.6.5.807-814

Metthew STEGGINK et al. [7] analyze the botnets' net flow, found some unique characteristics of Storm when comparing them with other software's net flow.

In SRI technical report [8], Phillip Porras et al. present a penetrating analysis in Storm's logic, and provoke the Dialog-based Detecting Method—using Snort for discovering the dialogs and BotHunter for matching up later.

Carlton R.Davis et al. [9] use Sybil attacks to mitigate Storm botnet—infiltrate the botnet with large number of fake nodes (sybils) to reroute or disrupt "real" C&C traffic.

Brent ByungHoon Kang et al. [10] develop the Passive P2P Monitor (PPM) to enumerate the infected hosts in the Storm botnet regardless whether or not they are behind a firewall or NAT.

Reference [11] proposed a structured-P2P-based botnet detection strategy through the aggregation and stability analysis of network traffic, considering that the flows related to the structured-P2P-based bot exhibit stability on statistical meaning due to the impartial position in botnet and performing pre-programmed control activities automatically. And they develop a small flow-aggregation extraction subsystem to exclude a majority of flows unlikely for C&C ahead of stability detection.

Reference [12] presented a new general detection framework which currently focuses on P2P based and IRC based botnets. The framework is based on definition of botnets: a group of bots that perform similar communication and malicious activity patterns within the same botnet. And there is no need for prior knowledge of botnets such as botnet signature in the framework.

Above all, researches in decentralized botnets detecting are still in a beginning period. It is difficult for most of the studies to identify that whether the net flow characteristics and the host behaviors are led by botnets or normal activities.

III. BACKGROUND: P2P BOTNET AND STORM BOTNET

Recent years have witnessed an increasingly appearance of various P2P botnets-trying to "borrow" the P2P networks as their communication tools. Storm is a new P2P botnet based on Kademlia algorithm [13] and used the complete decentralized architecture for C&C. It is hard to detect partly because its communication channel is encrypted. More importantly, new techniques like fast-flux, rootkit and even anti-reverse engineering are used into this botnet.

The Storm infection and communication could be briefly divided into two different phases - the "Bot Initialization" and the "Secondary Injection".

In the initializing step, the worm's executable binary is downloaded by users from their email boxes. Then it makes several configurations to prepare for the "secondary injection" part, such as opening up several ports for later communication. The worm tries to contact the other peers from an encoded list, which is in the initial downloaded worm binaries. If successful, it joins into the Overnet/eDonkey net as well as steps into the second phase. In this phase, the bot uses hard coded keys to search on Overnet and download a value, which is an encrypted URL that points to the location of a secondary injection executables. The bot decrypts the value with its hard coded key and follows the executable directions such as upgrading, email spamming or others likewise.

Within the two phases, some unique network flow characteristics are worth noticing: 1) the number of UDP packets is also detected sharply increasing because Storm are using them for publishing itself in the Overnet, peer discovery, and other functions, as depicted in Fig. 3 Section V. 2) When initializing, the bot randomly sends requests to connect to the other peers, thus leads to an unusual amount of ICMP "Destination unreachable" packets, which are rarely seen in regular circumstances. 3) As in [7], the number of SMTP packets is in a rising trend when botnets are activated.

Besides, unique characteristics are detected in UDP packets [5]: 1) Most packets are fixed sized. 2) Same message type is found when compromised machine are connecting to other bots. 3) The storm used the same source port in controlling and communicating.

These characteristics are symbolization for the existence of bots. In this paper, we propose MSFM, which deal with different flows in different methods, to detect them.

IV. MSFM: MULTI-STREAM FUSED MODEL

MSFM is designed to provide real-time and accurate detection for indicating the existence of P2P botnets. The detecting procedure is in the following steps, as shown in Figure 1.

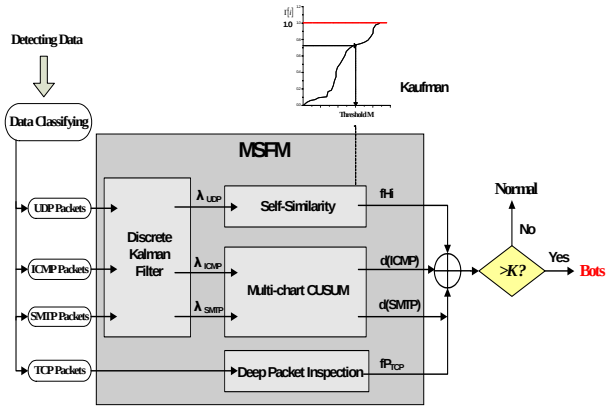


Figure 1. Process mapping of MSFM

To begin with, we detect the abnormality in packets (UDP, ICMP and SMTP) numbers using discrete Kalman filter. Then for UDP flow specifically, we analyze its self-similarity: lower self-similarity indicates the bot existence. For ICMP and SMTP flow, Multi-chart CUSUM is adopted as an amplifier to clarify the abnormalities. Meanwhile, considering the impact on the botnet detection which web applications generate, especially the P2P applications, we use the characteristics

of TCP flow to analyze that the abnormalities are caused by the botnets or the P2P applications.

A. UDP Flow

In the paper, we lay emphasis on the UDP flow, because it is most related with the C&C (Command and Control) of botnet. We use the Hurst Parameter to obtain the whole condition of it.

- Self-Similarity

In the subsection, we will use the self-similarity to analyze the whole condition of UDP flow.

In 1994, Leland et al. [14] demonstrated that Ethernet LAN traffic is statistically self-similar after analysis on the Bellcore, which is different from the previous Poisson processes (which is short-range dependence). Beran et al. declared that the Variable-Bit-Rate (VBR) video traffic is also self-similar [15, 16]. Furthermore, from the research about WAN [17], FASTPAC [18] and other network, we can see that the network traffic has an inherent feature: self-similarity. In this paper, the self-similarity property of UDP flow [19] is used, because the C&C of botnet is mainly about UDP flow, so it is able to reveal the properties of botnet clearer.

The self-similarity refers to a scale invariance property, which intuitively means that plots of traffic intensity at different time-scales look very similar.

Let *X(t)* (such as the number of UDP packets) be a stochastic process. If the following equation is always true for all a>0, then *X(t)* is self-similar.

$$x(t) = a^{-H} x(at) . \tag{1}$$

The "=" represents that they are equal in the statistical sense. That's to say, if the statistical properties of the process *X(t)* remain unchanged after it is compressed or extended in timescale, then *X(t)* has self-similarity. And *H* in (1) represents the degree of self-similarity, which is called the Hurst Parameter. A value of *H* closer to 1 means a larger degree of self-similarity or long range dependence (LRD).

As mentioned in Section III, Storm botnet causes abnormities in network flows: the number of UDP packets will increase because of communication between bots, which will weaken the self-similarity of UDP flow. Because the Hurst Parameter represents the degree of self-similarity, the change of Hurst Parameter will reveal the abnormities of UDP flow. Thus, we sampling the UDP flow, and calculate the Hurst Parameter of it. The value of Hurst Parameter closer to 0.5 means abnormities.

Finally, we define a function to decide the UDP abnormities:

$$fH_i = \begin{cases} 1, Hurst < M_{UDP} \\ 0, Hurst \geq M_{UDP} \end{cases} . \tag{2}$$

M_{UDP} is a threshold, and it adjusts dynamically as introduced in Section IV D. fH_i=1 means that the UDP flow is abnormal.

There are many method to calculate the Hurst Parameter, such as the rescaled range(R/S) method, variance-time plots(VTP) method, the Whittle's Estimator method and so on. Basis on the previous research [20, 21], we choose the R/S method, which is more stable and less affected by the parameters.

For a given time series $X = \{X_K, K = 1, 2, ..., L\}$, it is divided into *d* subsequences, and the length of every subsequence is *n=L/d*. For every subsequence *m=1,…,d*

1) Calculate the average (E_m) of every subsequence:

$$E_m = \frac{1}{n} \sum_{i=(m-1)*n+1}^{m*n} X_i . \tag{3}$$

2) Calculate the standard deviation(S_m) of every subsequence:

$$S_m = \sqrt{\frac{1}{n} \sum_{i=(m-1)*n+1}^{m*n} (X_i - E_m)^2} . \tag{4}$$

3) Calculate R_m of every subsequence:

$$R_m = \max_{1 \leq i \leq n} \{Y_{i,m}\} - \min_{1 \leq i \leq n} \{Y_{i,m}\} . \tag{5}$$

$$Y_{i,m} = \sum_{j=1}^{i} (Z_{j,m} - E_m) = \sum_{j=1}^{i} Z_{j,m} - i * E_m . \tag{6}$$

where $Z_{j,m}$ represents the value of j^{th} element in m^{th} subsequence.

4) Calculate the average of $\frac{R_m}{S_m} (m = 1, ..., d)$ of every subsequence:

$$\left(\frac{R}{S}\right)_n = \frac{1}{d} \sum_{i=1}^{d} \frac{R_m}{S_m} . \tag{7}$$

Accord to the research, the relationship between $\left(\frac{R}{S}\right)_n$ and the length of subsequence (*n*) when $n \to \infty$ can be expressed as:

$$\left(\frac{R}{S}\right)_n = Cn^H . \tag{8}$$

where *C* is a positive constant independent of *n*, and *H* is Hurst Parameter.

Take the logarithm on both sides of the equation (8):

$$\log\left(\frac{R}{S}\right)_n = \log C + H * \log n . \tag{9}$$

where $\log C$ is a positive constant independent of *n*.

A graph could be plotted taking the $\log\left(\frac{R}{S}\right)_n$ as longitudinal coordinates, $\log n$ as the abscissa. A line can be achieved after linear fitting, and the slope of this line is *H*.

To enhance the algorithm's real-time performance, we use a slide window to improve the R/S method, which is presented in reference [22].

B. ICMP and SMTP Flow

In the section, we will use the discrete Kalman filter to find ICMP and SMTP flow anomaly, and Multi-chart CUSUM acts as the amplifier to make the abnormality clearer.

- Discrete Kalman Filter

The discrete Kalman filter estimates a process by using a form of feedback control: the filter estimates the process state at some time and then obtains feedback in the form of (noisy) measurements. So, the equations for the Kalman filter fall into two groups: time update equations and measurement update equations. The time update equations are responsible for projecting forward (in time) the current state and error covariance estimates to obtain a priori estimate for the next time step. The measurement update equations are responsible for the feedback—for incorporating a new measurement into the priori estimate to obtain an improved posteriori estimate[23, 24]. The time update equations can be thought of as predictor equations, and the measurement update equations can be thought of as corrector equations .And the posteriori estimate is the result of the Kalman Filter.

We assume that the current time step is *k*.

1) Discrete Kalman filter time update equations

The priori estimate ($X_{k|k-1}$) is calculated given knowledge of the process prior to step *k*:

$$X_{k|k-1} = AX_{k-1|k-1} + BU_{k-1}. \quad (10)$$

where $X_{k-1|k-1}$ is the posteriori estimate at step *k-1*, U_{K-1} represents the control input, A and B are system parameters.

The priori estimate error covariance of $X_{k|k-1}$ ($P_{k|k-1}$) is:

$$P_{k|k-1} = AP_{k-1|k-1}A' + Q. \quad (11)$$

where $P_{k-1|k-1}$ is the posteriori estimate error covariance of $X_{k-1|k-1}$, A' represents the transposed matrix of *A*. *Q* is the process noise covariance.

So, the time update equations project the state and covariance estimates forward from time step *k-1* to step *k*.

2) Discrete Kalman filter measurement update equations

The posteriori estimate ($X_{k|k}$) is calculated given knowledge of the measurement Z_k and $X_{k|k-1}$:

$$X_{k|k} = X_{k|k-1} + Kg_k(Z_k - HX_{k|k-1}). \quad (12)$$

where *H* is the system parameter for measurement. And *Kg* is the Kalman Gain, which is calculated as following:

$$Kg_k = P_{k|k-1}H'/(HP_{k|k-1}H' + R). \quad (13)$$

where H' represents the transposed matrix of *H*. *R* is the measurement noise covariance.

The posteriori estimate error covariance of $X_{k|k}$ ($P_{k|k}$) is calculated as following:

$$P_{k|k} = (I - Kg_k H)P_{k|k-1}. \quad (14)$$

where *I* is an identity matrix.

After each time and measurement update pair, the process is repeated with the previous posteriori estimates used to project or predict the new priori estimates. That is, $X_{k-1|k-1}$ in (10) is substituted by $X_{k|k}$ in (12), and $P_{k-1|k-1}$ in (11) is substituted by $P_{k|k}$ in (14).

As mentioned in Section III, Storm botnet causes abnormities in network flows. The number of UDP, ICMP and SMTP packets are all increasing because of communication between bots, spamming or some other behaviors. Thus, we choose these 3 abnormal changes, turn them into proportion of net flow, and calculate the posteriori estimate of the proportions with the discrete Kalman filter (assuming that the current time step is *k*):

$$\lambda_i = X_{k|k}. \quad (15)$$

$X_{k|k}$ can depict network flow characteristics more precisely than the only measurement. Further more, the discrete Kalman filter is real-time. Hence, if λ_i increases, it implies that the network flow characteristic estimated by the discrete Kalman filter is abnormal. To make the increment clearer, we use Multi-chart CUSUM on the output of discrete Kalman filter (which will be introduced below).

- Multi-chart CUSUM

Internet traffic could be viewed as a complex random model: any abnormity in the traffic will bring obvious changes. However, since the observation series are blurred in the Internet security issues, it is hard to build up a specific model. For this reason, a nonparametric CUSUM (NP-CUSUM) that uses minimum a priori information is needed for abnormity detection. [25, 26] Moreover, more details in net flow changing are needed in the botnet detection, so multi-chart NP-CUSUM detection algorithm is used for multi-factor detection in the network. [27]

Now, we use the output of discrete Kalman filter as the input of Multi-chart CUSUM to amplify the fluctuation, and the following X_i can be substituted by λ_i in Section IV.B.

For random series $\{X_1,\ldots,X_n\}$, let H_k^i denotes the abnormity happens at time step k, and detected at channel *i*.($i \in \{1,\ldots,N\}$), H_∞ denotes no abnormity happens. Let $\sum_{s=k}^{n} g_{i,s}(X_i(1),\ldots X_i(s))$ stand for the score function that measures "likelihood" when H_k^i is true. So the detection statistics $S_n(i)$ is

$$S_n(i) = \left\{ \max_{1 \le k \le n} \sum_{s=k}^{n} g_{i,s}(X_i(1), ...X_i(s)) \right\}^+ . \quad (16)$$

where $i = 1, \ldots, N$ and $x^+ = \begin{cases} x, x>0 \\ 0, otherwise \end{cases}$ .Then each $g_{i,s}(X_i(n))$ could be changed into

$$g_{i,s}(X_i(n)) = X_i(n) - \mu_i - c_i . \quad (17)$$

where $\mu_i = E_\infty X_i(n)$ means the average data in normal state, and c_i is a turning positive constant for turning the $g_{i,s}(X_i(n))$ into a negative value so the positive ones could accumulate. So (16) could be represented in a recursive way

$$S_n(i) = \left\{ S_{n-1}(i) + X_i(n) - \mu_i - c_i \right\}^+ . \quad (18)$$

$$S_0(i) = 0$$

Finally, the definition of judging function is

$$d_M(S_n(i)) = \begin{cases} 1, S_n(i) > M \\ 0, S_n(i) \le M \end{cases} . \quad (19)$$

where M is a threshold, and it will be adjusted dynamically as introduced in Section IV D.

C. TCP Flow

The new decentralized botnets are very similar to the traditional P2P applications, so the abnormalities mentioned above are also possibly caused by the traditional P2P applications in the network. Thus, some method is needed to analyze that the abnormalities are caused by the botnet or the P2P applications.

The normal P2P applications always use the UDP packets to transfer the control information, and use the long TCP packets (the size is always bigger than 1300 bytes) to transfer the data. As mentioned is Section III, the botnet only has the process of C&C with UDP packet. And the only data transmission of botnet is the "secondary injection" process, which is transferred by the HTTP protocol and the data volume is small.

So, we will statistics the proportions of the long TCP packets- P_{TCP}, the size of which is above 1300 bytes. The value of P_{TCP} smaller means the abnormalities mentioned above are more possibly caused by the botnet.

Let N_{TCP} represents the number of the long TCP packets; the value of N_{TCP} is 0 in the beginning.

Monitoring every TCP packet P_i:

1) If the destination port of P_i is well known ports, such as 80(HTTP), 21(FTP) and so on, then deal with the next TCP packet P_{i+1}. Otherwise, go to step 2);

2) Use the Deep Packet Inspection(DPI)[28] to deal with TCP packets. If P_i belongs to a known P2P application, such as BT， eMule, Skype and so on, then deal with the next TCP packet P_{i+1}. Otherwise, go to step 3);

3) If the size of P_i is bigger than 1300 bytes, then $N_{TCP} = N_{TCP} + 1$. Otherwise deal with the next TCP packet P_{i+1}.

At last, we define a decision function:

$$fP_{TCP} = \begin{cases} 1, P_{TCP} < M_{TCP} \\ 0, P_{TCP} \ge M_{TCP} \end{cases} . \quad (20)$$

$$P_{TCP} = N_{TCP} / N_{total} . \quad (21)$$

N_{total} is the total number of TCP packets. M_{TCP} is a threshold, and it adjusts dynamically as introduced in Section IV D. fP_{TCP}=1 means that the abnormalities mentioned above are more possibly caused by botnet.

D. Dynamic threshold adjusting

Enlightened by Load-Shedding method and using the Kaufman algorithm [29], we adjust the threshold dynamically to improve the detection precision.

Let the $\Gamma[i]$ denotes the mapping variable of the system effective payload and detection algorithm threshold in the (i+1)th time span. Defined $\Gamma[0] = 1$ and $\Gamma[i]$ values in $[\Gamma[\min], 1]$, where $\Gamma[\min]$ is a rather small but not 0 constant. This is because if $\Gamma[\min]$ is 0, no data flows are allowed to pass through. Hypothesize that right at the ith time over, the actual payload in the system is $\rho[i]$, and $\rho[target]$ is the maximum payload, so we get $\varphi[i] = \rho[target] / \rho[i]$. Thus, $\Gamma[i]$ could be presented in a recursive way

$$\Gamma[i] = \Gamma[i-1] * \varphi[i] . \quad (22)$$

And since $\Gamma[i] \in [\Gamma[\min], 1]$, we can get the final equation of $\Gamma[i]$, that is

$$\Gamma[i] = \max \left\{ \min \left\{ \Gamma[0] * \prod_{j=1}^{i} \varphi[i], 1 \right\}, \Gamma[\min] \right\} . \quad (23)$$

where $i = 1, \ldots, n$. In this way, threshold could be computed out by $\Gamma[i]$.

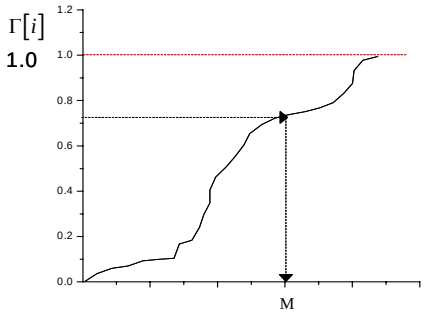


Figure 2. Mapping between threshold M and $\Gamma[i]$

E. Process of MSFM

The process of MSFM (shown in Fig. 1) is:

1) Retrieve data from monitoring device, and turn them into proportions—C_{UDP}, C_{ICMP}, C_{SMTP}. And deal with the TCP flow, we get fP_{TCP}.

2) Calculate the posteriori estimate of the proportions with the discrete Kalman filter—λ_{UDP}, λ_{ICMP}, λ_{SMTP}.

3) Deal with different packets in different strategies.

①Deal with the UDP packets

Calculate the Hurst Parameter of UDP packets, and finally get fH_i.

②Deal with the ICMP,SMTP packets

Input λ_{ICMP} and λ_{SMTP} into the Multi-chart CUSUM algorithm to amplify the fluctuation, and output $d(S_i(ICMP))$ and $d(S_i(SMTP))$.

4) Synthesize these outputs and make the decision. The decision method is

$$D_t = \eta_t * fH_i + \theta_t * d(S_n(ICMP)) + \tau_t * d(S_n(SMTP)) + \omega_t * fP_{TCP} \quad (24)$$

$$\eta_t + \theta_t + \tau_t + \omega_t = 1. \quad (25)$$

Where η_t, θ_t, τ_t and ω_t are the weight values generated by Exponential Weighted Moving Average (EWMA) algorithm. If $D>K$ (K is a constant decided by different network situation), it is judged abnormal, and consider that botnet exists, otherwise not.

V. PERFORMANCE EVALUATION

The environment is set up following paper [7]. It consists of a protected net. Several hosts are logging the traffic with the Wireshark. Some of the hosts are served as bots in storm botnet.

A. Netflow Comparison

In experiment A, we monitor the net flows under different situations. The sample packets are gathered every 10 seconds. After running normally for a while, Storm bots are injected into the network.

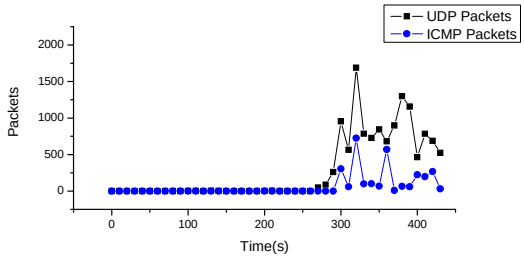


Figure 3. UDP and ICMP packets comparison

From the 280th seconds, as illustrated in Fig. 3, the number of UDP packets increases dramatically from 49 to 1689 for the reason that the bots are communicating using UDP packets. Meanwhile, the ICMP packets also showed a skyrocketing trend when bots were activated (from 100 to 700). Few SMTP packets were found in our experiment due to the time delay for spamming, so we don't consider them in the below experiment. Therefore, we can see that it is the botnet communication mechanism that decides the increasing of UDP and ICMP packet number, which has critical significance for measurements.

B. Self-Similarity Experiment

In experiment B, we will observe the change of the Hurst Parameter of UDP packets. In regular circumstances, the Hurst Parameter will keep a relatively stable value interval, which is from 0.65 to 0.85. It may fluctuate, but usually in a small range.

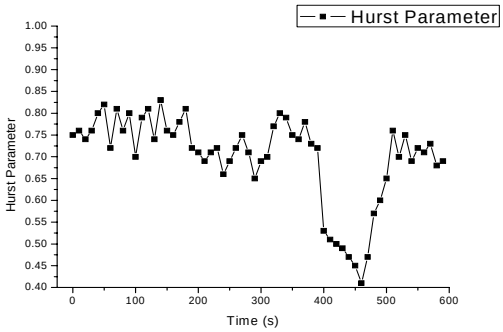


Figure 4. Hurst Parameter of UDP packets

From Fig.4, after the storm worms are injected to the network, the Hurst Parameter of UDP packets decreased sharply down to nearly 0.51 at t=420s and even to 0.41 at the lowest point when t=460s. And it fluctuates acutely because of the periodic behaviors in C&C. Hurst parameter can clearly indicate the UDP flow abnormal increasing caused by bots communication.

C. Multi-chart CUSUM Experiment

In experiment C, we use the output of ICMP and SMTP data packets in discrete Kalman filter as the input of multi-chart CUSUM to enlarge the difference and improve the measure precision. We take the ICMP data packet as our experimental example. From Fig. 3, at t=290s, ICMP packets began sharply increase, but in Fig. 5, at t=310s, Multi-chart CUSUM detected it with little delay. It proves that the model is sensitive enough to detect the activated bots in the real-time network.

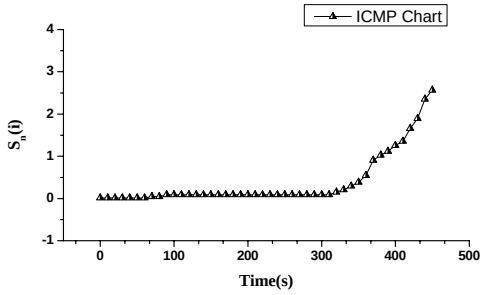


Figure 5. Output of ICMP-Chart CUSUM

D. False-positive and False-negative Comparison

Experiment D is mainly to compare the algorithm's false-positive and false-negative rate with previous works. We define the model from paper [30] as M-CUSUM, model from paper [31] as KCFM, and MSFM without TCP deep packet inspection as MSFM-T. We select five groups of data, which are using different combination of protocols and net flow rates, especially the variation of flow intensity in p2p applications. The first 3 samples are

in the environment without bots. The last 2 samples are collected when bots are injected.

TABLE I. FALSE-POSITIVE AND FALSE- NEGATIVE

	No.1	No.2	No.3	No.4	No.5
M-CUSUM	3	1	16	81	133(79)
KCFM	1	0	10	85	128(83)
MSFM-T	1	0	8	85	125(83)
MSFM	1	0	2	89	105(87)
Real	0	0	0	100	100

From Table 1, the detecting precision of our MSFM is desirable, as its false-positive and false-negative rate in sample 1 and 2 (normal office environment) are approximating to the real situation. However, in the third group of experiment, we add the P2P application such that the network background flow consists of a large amount of P2P protocols packets. In this case, all models show error reports. The M-CUSUM model only using CUSUM algorithm has a relatively high false-positive rate; the KCFM model which combines Kalman filter and CUSUM obtained a reduced false-positive rate, and the MSFM-T model has even lower ERR since the self-similarity model is added into the system. MSFM, equipped with TCP deep packet inflection functions, can differentiate normal P2P application data flow and the botnet C&C flow. Even under the extreme condition, it also performs well in a relatively low false-positive rate.

In the fourth experiment, we add bot attacking flow in normal network environment without P2P applications. All the four models have relatively low false-negative rate. In the fifth experiment, we add normal P2P flows based on the fourth experiment and all models shows the situation of false-negative and false-positive. For instance, the M-CUSUM detects 133 times of attack in total with 79 correct. KCFM and MSFM-T reduce the false-negative times from 133 to 128 with 83 correct and 125 with 83 correct, respectively. TCP deep packet inflection function in MSFM will further differentiates the normal network flow and botnets flow, thus MSFM can detect P2P botnets more accurately and effectively with reduced false-positive and false-negative rates.

VI. CONCLUSIONS

In this paper, we briefly described the infection and communication mechanism of new P2P botnet, and presented its unique characteristics. In order to accurately detect this botnet, we propose a novel real-time detecting model—MSFM (Multi-Stream Fused Model). We deal with different types of packets in different methods: detecting increasing number of UDP, ICMP and SMTP packets by discrete Kalman filter; amplifying the results by multi-chart CUSUM; detecting UDP packets using self-similarity; and improving detection precision using Kaufman algorithm to dynamically adjust the threshold. Furthermore, we considered the web applications-generated impacts on the botnet detection, especially the P2P applications, and differentiated normal flow with botnet communications. The results is good within detecting, false-positive and false-negative experiment, which prove that proposed model has its own advantages in accurately detecting this new botnets on experimented network platform. We are planning to improve the detecting precision in large-scale network environment, and try to mitigate its harm to the Internet.

ACKNOWLEDGMENT

This work was supported by Technology Development Plan of Jilin Province of China (No.20090110).

REFERENCES

[1] S. Northcutt, E. Skoudis, M. Sachs, J. Ullrich, T. Liston, E. Cole, E. Schultz, R. Dhamankar, A. Yoran, H. Schmidt, W. Pelgrin, and A. Paller, “Top Ten Cyber Security Menaces for 2008”, *SANS Institute*, SANS Press Room, 2008.

[2] J. Stewart, “Storm Worm DDOS Attack”, *SecureWorks*, Inc, Atlanta GA, 2007.

[3] J. Grizzard, V. Sharma, C. Nunnery, B. Kang and D. Dagon, “Peer-to-Peer Botnets: Overview and Case Study”, In *HotBots ’07 conference*, Usenix, 2007.

[4] S. Sarat and A. Terzis, “Measuring the Storm Worm Network”， *Technical Report 01-10-2007*, HiNRG Johns Hopkins University, 2007.

[5] T. Holz, M. Steiner, F. Dahl, E.W. Biersack and F. Freiling, “Measurements and Mitigation of Peer-to-Peer-based Botnets: A Case Study on Storm Worm”, *1st USENIX Workshop on Large-Scale Exploits and Emergent Threats*, Usenix, San Francisco, 2008.

[6] A. Nummipuro, “Detecting P2P-Controlled Bots on the Host”, *Seminar on Network Security*, Espoo, Helsinki, 2007.

[7] M. STEGGINK and I. IDZIEJCZAK, “Detection of peer-to-peer botnets”, University of Amsterdam, Netherlands, 2007

[8] P. Porras, H. Saidi and V. Yegneswaran, “A Multi-perspective Analysis of the Storm (Peacomm)Worm”, *Computer Science Laboratory*, SRI International , CA, 2007.

[9] C. R. Davis, J. M. Fernandez, S. Neville, and J. McHugh, “Sybil attacks as a mitigation strategy against the storm botnet”, *Proc. 3rd Int. Conf. on Malicious and Unwanted Software (Malware '08)*, Alexandria, VA (2008) pp. 32-40.

[10] B. Kang, E. Chan-Tin, C. Lee, J. Tyra, H. Kang, C. Nunnery,Z. Wadler, G. Sinclair, N. Hopper, D. Dagon and Y. Kim, “Towards complete node enumeration in a peer-to-peer botnet”, *ACM Symposium on Information, Computer & Communication Security (ASIACCS 2009)*, 2009.

[11] Zhitang Li, Binbin Wang, Dong Li, Hao Chen, Feng Liu, ZhengBin Hu, “The Aggregation and Stability Analysis of Network Traffic for Structured-P2P-based Botnet Detection”,*Journal of Networks*, Vol.5, No.5 ,2010,pp.517-526, May 2010.

[12] Hossein Rouhani Zeidanloo,Azizah Bt Abdul Manaf, “Botnet Detection by Monitoring Similar Communication Patterns”,*(IJCSIS) International Journal of Computer Science and Information Security*,Vol.7, No.3, 2010,pp.36-45.

[13] P. Maymounkov and D. Mazieres, “Kademlia: A peer-to-peer information system based on the XOR metric”, *1st International Workshop on Peer-to-Peer Systems (IPTPS'02)*, Springer, NY, 2002.

[14] Leland W E, Taqqu M S,Willinger W, et al. “On the self–similar nature of Ethernet traffic(extended version) ”.*IEEE/ACM Trans on Networking*, 1994,2(1) : 1- 15.
[15] Beran J, Sherman R, Traqqu M S, et al. “Long range dependence in variable bit rate video traffic”.*IEEE Trans on Communication*,1995, 43(2/3/4) : 1566- 1579.
[16] Garrett M W,Willinger W. Analysis, “modeling and generation of self-similar VBR video traffic”,*Proc ACM Sigcomm’94*, 1994:269-280.
[17] Paxson V, Floyd S. “Wide area traffic: the failure of poisson modeling”,*Proc ACM Sigcomm’94, 1994*: 257-268.
[18] Addie R. “Fractal traffic: measurements, modeling and performance evaluation”,*Proc of INFOCOM’95*, Boston, MA, 1995: 977- 984.
[19] KIM J S, KAHNG B, KIM D, et al. “Self-similarity in fractal and non-fractal networks”. *Journal of the Korean Physical Society*, 2008,52: 350-356.
[20] T Karagiannis,M Molle,M Faloutsos. “Understanding the limitations of estimation methods for long-range dependence”.*University of Califomia,Tech ReP*:TRUCR-CS-2006-10245,2006.
[21] T Karagiannis,M Molle,M Faloutsos. “Long-range dependence:Ten years of Internet traffic modeling”.*IEEE Intenet Computing*,2004,8(5):57-64.
[22] Hagiwara T, Doi H, Tode H, et al. “High-speed calculation method of the Hurst parameter based on real traffic”,*LCN 2000: Proceedings 25th Annual IEEE Conference on Local Computer Networks*,2000: 662- 669.
[23] R. E. KALMAN, “A New Approach to Linear Filtering and Prediction Problems”, *Transaction of the ASME—Journal of Basic Engineering*, pp. 35-45 (March 1960).
[24] G. Welch and G. Bishop, “An introduction to the Kalman filter,” Dept. Comp. Sci., Univ. North Carolina, Chapel Hill, TR95-041.
[25] A.G. Tartakovsky and V. Veeravalli, “Change-point detection in multichannel and distributed systems with applications”, *Applications of Sequential Methodologies, Marcel Dekker, Inc.*, pp. 339–370, New York, 2004.
[26] A.G. Tartakovsky, “Asymptotic properties of CUSUM and Shiryaev’s procedures for detecting a change in a nonhomogeneous Gaussian process”, *Mathematical Methods of Statistics*, No. 4, pp. 389–404, 1995.
[27] A.G. Tartakovsky, B. Rozovskii and K. Shah, “A Nonparametric Multichart CUSUM Test for Rapid Intrusion Detection”, *Proceedings of Joint Statistical Meetings, Minneapolis*, MN, 2005.
[28] Subhabrata S, Spatscheck O, W ang D. Accurate, “scalable in-network identification of p2p traffic using application signatures”, *Proceedings of the 13th International Conference on World Wide Web*. New York: ACM Press, 2004:512-521.
[29] S Kasera, J Pinheiro, C Loader, M Karaul, A Hari, T LaPorta. “Fast and robust signaling overload control”. In *Proceedings of Ninth International Conference on Network Protocols, Riverside, USA*: IEEE, 2001. pp.323-331.
[30] Jian Kang, Jun-Yao Zhang, “Detecting New P2P Botnet with Multi-chart CUSUM”, *NSWCTC '09*, pp. 688 – 691, April 2009.
[31] Kang Jian, Song Yuan-zhang, “Application KCFM to Detect New P2P Botnet Based on Multi-Observed Sequence”, *Geomatics and Information Science of Wuhan University*, vol.35(5), pp.520-523, May 2010.

Jian Kang (1975-), Ph.D in 2006 from JiLin University, with major in Computer Science and Technology. He is currently an Associate Professor in department of Computer Science and Technology in Jilin University, Changchun, China. His research interests include distributed computing system and network security.

Yuan-Zhang Song(1986-), is currently completing his Master’s degree in Computer Architecture in the department of Computer Science and Technology in Jilin University. His area of interest is mainly about network security.

Jun-Yao Zhang(1986-), is currently pursuing his Ph.d degree in Computer Science in the department of EECS in Univeristy of Central Florida. His research interests includes distributed computing system and cloud computing.

A Novel Cluster-head Selection Algorithm Based on Hybrid Genetic Optimization for Wireless Sensor Networks

Lejiang Guo, Qiang Li, Fangxin Chen
Air Force Radar Academy, Department of Early Warning Surveillance Intelligence, Wuhan, China
radar_boss@163.com

***Abstract*—Wireless Sensor Networks (WSN) represent a new dimension in the field of network research. The cluster algorithm can significantly reduce the energy consumption of wireless sensor networks and prolong the network lifetime. This paper uses neuron to describe the WSN node and constructs neural network model for WSN. The neural network model includes three aspects: WSN node neuron model, WSN node control model and WSN node connection model. Through learning the framework of cluster algorithm for wireless sensor networks, this paper presents a weighted average of cluster-head selection algorithm based on an improved Genetic Optimization which makes the node weights directly related to the decision-making predictions. The Algorithm consists of two stages: single-parent evolution and population evolution. The initial population is formed in the stage of single-parent evolution by using gene pool, then the algorithm continues to the next further evolution process, finally the best solution will be generated and saved in the population. The simulation results illustrate that the new algorithm has the high convergence speed and good global searching capacity. It is to effectively balance the network energy consumption, improve the network life-cycle, ensure the communication quality and provide a certain theoretical foundation for the applications of the neural networks.**

***Index Terms*—wireless sensor networks, energy efficiency, coverage, the routing protocol, the network lifetime**

I. INTRODUCTION

Wireless sensor networks (WSN) have an important practical value in the military, environmental monitoring, industrial control, intelligent home and urban transport, etc. In Wireless Sensor Networks, the efficient routing protocol plays a critical role for data packet transition. However, the traditional routing protocol is little regard for the energy consumption of the node. Because the energy of WSN node is limited, the maximum of the lifetime of WSN become an important goal to the designation of routing protocol [1]. Therefore, the sensor network routing protocols must consider not only the energy consumption of a small message transmission path, but also the consumption of energy balance in the whole network routing. On the other hand, due to the number of sensor network nodes are often large, the node can only get partial topology information, so the routing protocol also can choose the right path on the basis of the information of the part network.

In 1975, John Holland proposed a global optimization algorithm Genetic Algorithm (GA). In recent years, based on genetic algorithm in the WSN, the routing optimization research is also very active [2]. For path optimization, genetic algorithms have shown a tremendous advantage. Based on the model of the energy multi-path routing protocol , this paper present a new algorithm which abandons the randomness in the generation of initial population and replaces the gene fragments by gene pool. Through simulation, the novel routing protocols is effective and extend the network life time.

II. THE ISSUES OF SUB-CLUSTER STRUCTURE IN WSN

In sub-cluster structure of wireless sensor network, the network nodes are divided into several clusters. Each cluster usually consists of a cluster head node (CH) as well as several members of the node (MN) component. The MN communicates with the cluster head node; CH and CH constitute a high-level virtual backbone which is responsible for clusters of data fusion and data forwarding between clusters [2]. Because a larger energy consumption in the cluster head node, a cyclical way to select cluster head nodes in the network nodes is used for balancing energy consumption. From Fig.1, it shows the sub-cluster within the cluster structure and the data flow between clusters.

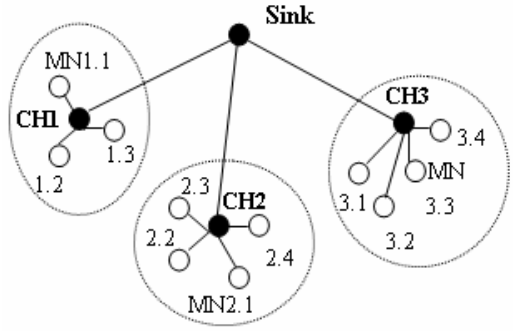


Figure 1. The Cluster Structure in WSN

WSN is designed to maximize the network lifetime as the ultimate goal, thus making each node as much as possible is extremely important to balance the energy

doi:10.4304/jnw.6.5.815-822

consumption. In the clustering algorithm, the energy consumption of cluster-head node is generally much higher than normal, it likely makes cluster-head node die prematurely because of energy depletion [3]. In order to avoid this situation, one way is to use cluster-head rotation mechanism. Each node of each cluster-head rotates from time to time. The remaining energy of each node is as close as possible. The cluster-head rotation mechanisms are often independent of the cluster algorithm, and the cluster algorithms complement each other. Common mechanism of cluster-head rotation is two kinds of passive and active. The former leads to a fixed time intervals which require a threshold when the monitored parameters exceed a certain threshold value, a common threshold are residual energy, node number and so on. Both the passive and active cluster-head rotation mechanism select the appropriate parameters of the algorithm, the final result will be a significant impact [4]. If the cluster-head rotation is too frequent, it will bring a lot of additional overhead and network disruption. If the cluster-head rotation frequency is too low, it may cause some nodes run out of energy prematurely. Therefore, only a reasonable compromise can achieve the most optimal network lifetime.

Wireless sensor networks usually use the energy principle of giving priority; it needs to consider the energy consumption of the node and energy balance of WSN. Hierarchical routing protocols have been proven to be effective in saving energy in Fig.1.All nodes in the network is divided into cluster head nodes and common nodes .Common node is responsible for data collection and send it to cluster head node, cluster head node within the cluster receive the data sent by ordinary nodes fusion and then transmit to the sink node, this algorithm is called clustering algorithms. Representative cluster algorithms are LEACH, PEGAGIS, and HEED [5].

In LEACH, the cluster-head node generation and distribution network, the sensor node is nothing to do with the uneven distribution of sensor nodes. The Cluster algorithm uses the multi-hop communications, close to the sink node of the cluster- head node forward a lot of data which led to the excessive consumption of energy and the nodes in its own is easy to lapse, thus known as network partitioning. Density gravity is using a cluster algorithm may result in cluster-head node, the node sparse are with little or no situation. Node traffic aggregation may arise in some aggregation node load imbalance, thereby resulting in the convergence node congestion, packet loss and buffer overflows [6]. Even in the nodes uniformly distributed and have the same flux density distribution in WSN, using the multi-path routing protocols transmit data along any of a multi-path distribution, Only in the main path to failure, the backup path can take transmission task which limits the potential of the backup path, frequent use of the primary path to transmit data, it will cause the nodes on the path prematurely because of excessive energy consumption death, so that the whole network is divided into isolated and disconnected parts[7]. It reduces the overall network lifetime. In addition to the central gateway nodes and network nodes in some key positions in the WSN, nodes will also form a network bottleneck. Different from ordinary node, gateway or network node, the central node, these nodes are likely to have a great load. Therefore, the load-balancing for improving the WSN can be obtained to expand throughput and enhance the nature of WSN is vital. Stochastic algorithm based on a certain degree of probability determine whether a cluster head node. In LEACH, the probability of the node becomes cluster-head only with the past several rounds of sub-nodes in the own state. The HEED algorithm relates the probability and residual energy, there are some algorithms which are taken into account various parameters of the node degree. Stochastic optimization algorithm for the degree of the cluster results is usually less deterministic algorithm, but the convergence speed, less overhead, especially suitable for large-scale networks.

III. THE WSN MODEL

A. WSN node neuron model

WSN node neuron model is shown in Fig.1.q is the node data fusion; $s_1, s_2, ..., s_n$ for the sensor nodes to collect information; $\omega_1, \omega_2, ..., \omega_n$ for the weight value; θ for the threshold. Relationship between input and output nodes in accordance with the following formula:

$$q(t) = f(\sum_{i=1}^{n} \omega_i \varsigma_i(t).\theta) \qquad (1)$$

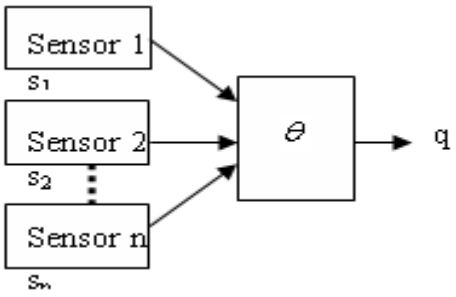


Figure 2. WSN neuron model

B.WSN node control model

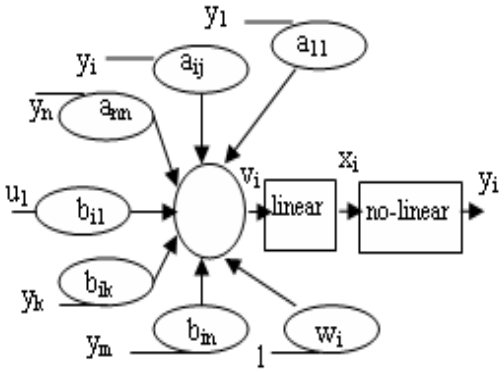


Figure 3. WSN neuron control model

Similar to the neural network control model, WSN node control model based on neural model is shown in Fig.2.

(1) Weighted adder

$$\upsilon_i(t)=\sum_{j=1}^{n}a_{ij}y_i(t)+\sum_{k=1}^{m}b_i k u_k(t)+\omega_i \qquad (2)$$

The input of the sensor nodes is expressed as W_i:

$$\mathrm{W_i}=\begin{bmatrix} y_1 \\ \vdots \\ y_n \\ u_1 \\ \vdots \\ u_m \\ 1 \end{bmatrix}, A_i=\begin{bmatrix} a_{i1} \\ a_{i2} \\ \vdots \\ a_{in} \end{bmatrix}, B_i=\begin{bmatrix} b_{i1} \\ b_{i2} \\ \vdots \\ b_{im} \end{bmatrix} \qquad (3)$$

Among the formula (3), A_i , B_i is he connection matrix.

$$\upsilon_i(t)=\mathrm{A_i y_i(t)+B_i u_k(t)}+\omega_i \qquad (4)$$

(2)Linear dynamic functions:

$$\mathrm{X(s)=H(s)V(s)} \qquad (5)$$

Where $\mathrm{X(s)}$, $\mathrm{V(s)}$, $\mathrm{H(s)}$ is the Laplace transform from $\chi_i(t),\nu_i(t),h(t)$. $h(t)$ is linear dynamic function of the impulse response.

(3)Static nonlinear function:

$$y_i=g(\chi_i) \qquad (6)$$

$$g(\chi)=\begin{cases}1, x>0\\ 0, x\le 0\end{cases}$$

$$g(\chi)=\frac{1}{1+e^{-x}}, g(\chi)=\arctan(\chi)$$

$$g(\chi)=e^{-x^2/o^2}$$

C.WSN node connection model

Suppose neurons are static, t *H(S) =1*.The neuron can be expressed as

$$\begin{cases} X(t)=A\gamma(t)+BU(t)+W \\ \gamma(t)=g(X(t)) \end{cases} \qquad (7)$$

X is N-dimensional vector. *g(.)* is a nonlinear function. A, B is the connection matrix. The WSN connectivity for three neural networks can be expressed as:

$$\begin{bmatrix} x^1(t) \\ x^2(t) \\ x^3(t) \end{bmatrix}=A\begin{bmatrix} y^1(t) \\ y^2(t) \\ y^3(t) \end{bmatrix}+B\begin{bmatrix} u^1(t) \\ u^2(t) \\ u^3(t) \end{bmatrix}+\begin{bmatrix} \omega^1(t) \\ \omega^2(t) \\ \omega^3(t) \end{bmatrix} \qquad (8)$$

The superscript expresses the level. First level is general node level.

$$\begin{cases} X^1(t)=B^1U^1(t)+W^1 \\ \gamma^1(t)=g(X^1(t)) \end{cases} \qquad (9)$$

Second level is the sink node layer.

$$\begin{cases} X^2(t)=A^2\gamma^1(t)+W^1 \\ \gamma^2(t)=g(X^2(t)) \end{cases} \qquad (10)$$

Third level is the user node layer.

$$\begin{cases} X^3(t)=A^3\gamma^2(t)+W^1 \\ \gamma^3(t)=g(X^3(t)) \end{cases} \qquad (11)$$

Sensor network node energy is extremely limited. In order to prolong the life of the network and the whole system, all the information processing strategies must reduce node energy consumption as far as possible. Hopfield neural network is an artificial neural network model to optimize computing, associative memory, pattern recognition and image restoration, etc. Hopfield energy function is a reflection of the overall state of multidimensional neuronal scalar function. It can form a simple circuit of artificial neural networks which adopts a parallel computing interconnected mechanism [8]. Hopfield is built on energy with the same Lyapunov function. Hopfield considers that the internal stored energy is gradually decreased with time increases in the system movement process. When the movement to equilibrium, the energy of the system runs out or becomes miniature [9]. The system will naturally balance and go to stabile state. Therefore, the system can solve the stability problem if we can find a complete description of the process of energy function. For continuous feedback network circuit, the state equations is

$$\begin{cases} C\frac{du_i}{dt}=-\frac{u_i}{R}+\sum_{j=1}^{r}w_{ij}v_j+I_i \\ v_i=f_i(u_i) \end{cases} \qquad (12)$$

When the system reaches steady output, Hopfield energy function is defined as:

$$E=-\frac{1}{2}\sum_{i=1}^{r}\sum_{j=1}^{r}w_{ij}v_iv_j-\sum_{i=1}^{r}v_iI_i+\sum_{i=1}^{r}\frac{1}{R}\int_0^{v_i}F^{-1}(\eta)d\eta \qquad (13)$$

D. The general design method of energy function

Suppose the optimization objective function is $f(u), u\in R^n$ is the state of artificial neural network which is also the objective function of the variable.

Optimizing constrained conditions is *g(u)=0*. Optimization problem is to meet the constraint conditions and to minimum objective function. The equivalent minimum energy function E is expressed as:

$$E=f(u)+\sum\left|g_i(u)\right| \qquad (14)$$

$\left|g_i(u)\right|$ is penalty function. When $\left|g_i(u)\right|=0$ is not satisfied, the value $\sum\left|g_i(u)\right|$ is always greater than zero. According to Hopfield energy function and gradient descent, E is limited in the negative direction, $|E|<E_{\max}$, $\frac{dE}{dt}\le 0$ .System can always reach the final minimum

E and $\frac{dE}{dt}=0$ which is the stability point $\frac{du_i}{dt}=0$. When solving optimization problems, *E* is often a function of the status u , so $\frac{dE}{dt}\le 0$ .It is turned into the conditions on the state derivative

$$\frac{\partial E(u_i,v_i)}{\partial u_i}=-\frac{dE}{dt} \qquad (15).$$

$$\frac{du_i}{dt}=-\frac{\partial E}{\partial u_i} \ , \ \frac{dE}{dt}=\sum_i \frac{\partial E}{\partial u_i}.\frac{du_i}{dt}=-\sum_i (\frac{du_i}{dt})^2 \le 0$$

The gradient descent method can ensure that E is always down until the arrival of a local minimum [10]. With the energy function in solving optimization problems, the first question is to change the problem into the objective function and constraints, and then construct the energy function, calculate energy function of the parameters by using the conditional formula; at last the result is the artificial neural network connection weights.

IV. THE CLUSTER-HEAD SELECTION ALGORITHM BASED ON NEURAL NETWORK

BP artificial neural network imitate the human brain structure and function, the non-linear processing of information can increase as well as to deal effectively with ambiguous, incomplete. There is the contradiction between the understandings of complex environments to determine the problem. It can reflect the people's way of thinking. Networks have self-learning function. It can extract from the general principles contained in and also learn to deal with specific issues. After training, the neural network of free weights is that the knowledge acquired [11]. Neural network learns through the existing program and the evaluation of the results of the study, which access to the experiences, knowledge and a perspective on the importance of various objectives such as intuitive thinking. Once it is used for evaluation, the network can reproduce these experiences, knowledge and intuitive thinking. When evaluation, the network can reproduce the experience, knowledge and intuitive thinking on complex issues to make sound judgments; thus it not only embodies the people's subjective judgments, but also greatly reduces the accuracy of assessment method. The right the value of the distribution is more objective and accurate, and it uses the program proposed by the weighted average of BP neural network model to enable members of the weights which is assigned to their decision-making predictions. Predictions directly relate to the right to exclude the value of an incorrect evaluation of man-made factors.

The weight distribution of the members is objective while the weighted average calculation of the program will also be more accurate, and the members of the weights w_i also has a dynamic adjustment of the self-learning function, you can improve the accuracy and efficiency. Clustering decision-making is calculated by the base station. At the same time, each node in the algorithm has been advised before the implementation of their location. When the base station collects all the nodes in the network about residual energy and location information, the program weighted average of the BP neural network model is three layers which are shown in Fig.4.There include respectively the input layer, the hidden layer, and the output layer. Neurons connection forms a fully connected model. But within the various levels, there is no connection between neurons. Input vector *X* composes with the node residual energy, covering the number of nodes within a radius of the node to other nodes, the node position of four components, the composition of the hidden layer nodes in the network learning speed selection. Score value is converted to values between 0 and 1 .The output through a particular transformation can become a program of the weighted average [12]. Training samples result from the previous decision-making members of the same program rating value and the correct result value. After training, the connection between neurons is the corresponding weights into the weights, the knowledge and experience. Training program of the weighted average of neural network computing model can be used to calculate the weighted average. When a group of decision-making members of the rating the value continues, the output vector *Y* contains the node which can become cluster-head probability. Base stations in the network determine the number of cluster-head the number of *k*, and select the highest probability of *k* nodes to become cluster head round times.

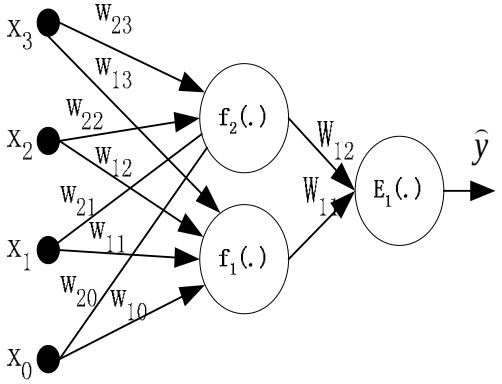


Figure 4. The Cluster-head Selection with neural network

It has three main features: (1) anonymous responses. Members circulate a Consultation on the table from their responses to be anonymous; (2) Iteration and controlled feedback. The method is a step by step approach which includes several iterations. Every iteration is called one round, and each round regards the views collected through the statistical processing back to the base members. Through this feedback, the opinions of the group will gradually concentrate; (3) Statistical group response. Received the final round of the views of members, it combines groups of views. After each round of consultations is required to collect the views of members of the group were statistically treated. In order to improve the algorithm accuracy, each member assigned a weight factor according to the issue of the

degree of familiarity with the area. The weight coefficients of its members are mainly based on the location of the members of the nearby nodes and residual energy. Neural network algorithm described as follows:

Input: Input vector $X = [x_0, x_1, \cdots, x_{n-1}]^T$, in *F1* there is a corresponding *N*-month treatment Unit

Output: Output vector(*Y*) is an *M*-dimensional real vector, $Y=[y_0, y_1, \cdots, y_{M-1}]^T$, in the *F2* there are *M* neurons, corresponding to a sample of model types to be identified .

(1) Initialization

$W_{ij}(0) = 1/(n+1)$;

/ * W_{ij} is the weight coefficient of *F1* to *F2* * /

$W_{ji}=1$;

/ * W_{ji} for *F1* to *F2*, weights * /

$0 \le \rho \le 1$;

/ * Threshold ρ * /

int *L=M*;

(2)After the calculation of *F1* to be the output vectors *S*.

(3)Match calculation (*K1*)

For *F2* each neuron *j* {

$$t_j = \sum_i w_{ij} . S_i ;$$

/ * t_j for *j* the activation value; S_i is the layer neurons * /

} end for

(4) Choose the best matching neuron

If Activate t_c=max (t_j) then output y_j;

$$y_j = f(t_j) = \begin{cases} 1 & t_j > \theta_j \\ 0 & t_j \le \theta_j \end{cases}$$

/* θj is neuron j which is a threshold* /

/* y_j is the output of neuron j * /

(5) Compare and test the value of vigilance

If $R = \|S\| / \|X\| > \rho$ then goto *K3*;

/* *S* for the *F1* layer output vectors* /

/* *X* as the input vector * /

(6)Optimal matching is invalid and their treatment (*K2*)

If (*L*=φ) then automatically increase the network model is to represent a class of new model categories;

/ * identify the layers of neurons similar to the rate achieved (*R*) are less than the threshold (ρ) * /

Else {*L=L*-1;

System Reset signal issued, and placing neuron is zero, do not allow their further participation in competition,

Goto *K*1 ;}

End if.

V. THE HYBRID GENETIC ALGORITHM

Nowadays, in the stage of the whole evolution the genes involved in genetic operator are mostly from the individual itself. The quality level of the individual determines the efficiency of the algorithm [13]. If the fitness of all individuals is poor, the algorithm performance will be affected. In order to overcome these weaknesses, this paper sorts n points, constructs n*n matrix of the gene pool, prepares for the genetic operators .This method greatly improves the efficiency of the algorithm [14].

A. Network Model

The network models as an undirected connection diagram *G(S, V, P)*, where *S* on behalf of the node of sink, V presents the sensors. Set the nodes numbered *1, 2, 3... n*, $P_K=V_1^K\ V_2^K\ V_3^K$ for a feasible path, the first *k* nodes of the path starting point is V_1^K, the aggregation node is V_n^K, then the first *k* paths) the total length can be expressed as:

$$f(P_k) = \sum_{i=1}^{n-1} PE(V_i^k, V_{i+1}^k) + PE(V_n^k, V_1^k)$$

PE (V_i^K, V_j^K) is the energy consumption between the nodes. In the algorithm, $f(P_K)$ is evaluated the individual's good or bad by using P_K.

B. Construction of Gene Pool

According to the cost between the nodes *PE(i,j)* ,*PE(i,j)* construct a *n*n* matrix *D={PE(i ,j)}*.

$$num[i, j] = \begin{cases} j+1 & j \ge i \\ j & j < i \end{cases}$$

i=1, 2, 3… n, j=1, 2, 3… n, then it defines a n*(n-1) matrix, num= {num [i, j]}.For each node *i*, according to the size of *PE (i, j)*, the num in the first line of the corresponding *i* elements in accordance with the order from small to large order, and before all the elements i come in, so will the expanded num which increased in first column to get a *n*n* of the square, gene pool is formed in this paper. Each row element in gene pool is the problem a feasible solution, the start node of the feasible solution is the line number of the gene pool. Solving the problem is divided into single-parent evolution and the evolution of population. By single-parent evolution, the initial population *P= {P1, P2...Pn}* is generated.

C. The Single-parent Evolution Algorithm

(1) Randomly generate one individual *V*, calculate fitness value *f (V)*;

Step1-1: Randomly generate gene length

$l \le len \le m$ a *xlen* ;

Step1-2: Randomly generating a gene location

$l \le pos \le n - len$;

Step1-3: The gene pool in the pos line of len elements replace the current total of the individual V in the genes, their location from the beginning of *pos*, get a new individual *Vnew*;

Step1-4: Calculation *f (Vnew)*;

Step1-5: If *f (Vnew)* < *f (V)* then *V: =Vnew*;

Step1-6: If (no end) then goto Step1-1;

(2)Randomly generate two integers

$l \le pos\,1, pos\,2 \le n$, $pos\,1 \ne pos\,2$

Step2-1: The gene segment between *pos1*, *pos2* of the individual *V* is reversed in order to be new individuals *Vnew*;

Step2-2: Calculation *f (Vnew)*;

Step2-3: If *f (Vnew)* < *f (V)* then *V: =Vnew*;

Step2-4: If (no end) then goto (2);

(3) Randomly generate two integers

$l \le pos\ 1, pos\ 2 \le n$, $pos\,1 \ne pos\,2$

Step3-1: The genome *Vpos2* of individual genes *V* will be inserted into the genome *Vpos1* to be new individual *Vnew*;

Step3-2: Calculation *f (Vnew)*;

Step3-3: If *f (Vnew)* < *f (V)* then *V: =Vnew*;

Step3-4: If (no end) then goto (3);

Step3-5: Output.

In the single-parent evolution algorithm, only a single individual is evaluated. The speed of evolution which produces a good general is very fast. At the same time, a global optimal path is generated.

D. The Population Evolution Algorithm

The population evolution algorithm in this paper, the algorithm presents only as an amendment to the role, the purpose is to improve the solution quality. After producing the initial population *P* with the single-parent evolution algorithm, select two individual *pr1, pr2* randomly to hybrid operation, if the new individual is better than the mother, the new individual replace the mother directly [15]. The evolution which repeats the hybrid process will not end until the best individual is generated. In another the mother *pr2* the gene position which is equal to *p1* and *p2* in the gene *pr1* is found. The same value of *pr2* between *pos1* and *pos2* from the gene *pr1* is deleted in its entirety, the genes of *p1* and *p2* will be moved to the adjacent location [16]. Then the genes between *pos1* and *pos2* will be cut and inserted *pr2* between the individual *p1* and *p2* according to the order of the position of *p1* and *p2*, as shown in Fig.5.

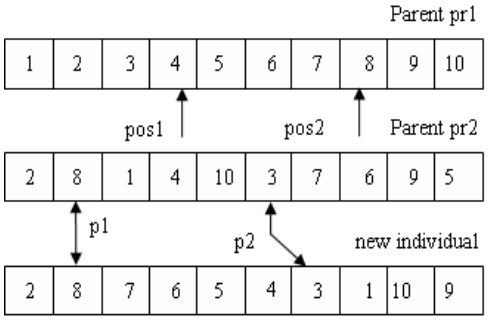


Figure 5. The procession of Hybrid gene operator

VI. SIMULATION

A. Simulation Environment

The experiment environment is based on a WSN cluster model. It uses the distributed data fusion. The energy consumption is mainly communication transmission. Suppose Hopfield neural network state vector $V=[v_1,v_2,\cdots,v_n]^T$ is the output vector $I=[I_1,I_2,\cdots,I_n]^T$ is the network input vector. As the time went on, the evolution of the solution in state space moves towards the direction of movement of energy E decreases. The final network output V is the network's stable equilibrium point and minimum points. The problem is mapped to the dynamic process of neural networks. Hopfield uses transposed matrix, *N * N* matrix represent the node which is accessed with the cluster-head node, the data is the integration of data is not sent back to the cluster-head node until traversing N nodes. The energy function is defined as follows

$$E = \frac{A}{2}\sum_{x=1}^{N}\left(\sum_{i=1}^{N}V_{xi}-1\right)^2+\frac{B}{2}\sum_{i=1}^{N}\left(\sum_{x=1}^{N}Vxi-1\right)^2$$
$$+\frac{D}{2}\sum_{x=1}^{N}\sum_{y=1}^{N}\sum_{i=1}^{N}V_{xi}d_{xy}V_{yi}$$

The dynamic equation of Hopfield network is

$$\frac{dU_{xi}}{dt}=-\frac{\partial E}{\partial V_{xi}}(x,i=1,2,\cdots N-1)=-A\left(\sum_{i=1}^{N}V_{xi}-1\right)$$
$$-B\left(\sum_{Y=1}^{N}V_{yi}-1\right)-D\sum_{y=1}^{N}d_{xy}V_{yi}$$

B. Experimental steps

The network solves this problem using an algorithm described as follows. Solving the local minimum and unstable problem, it should be chosen large enough coefficient of A, B, D to ensure the effectiveness of solution.

(1) Set its initial value and weight, t = 0, A = B = 1500, D = 1000, U_0= 0.02.

(2)Read $d_{xy}(x,y)$ distance among the cluster nodes *N*.

(3) Neural network input $U_{xi}(t)=U_0^{'}+\delta_{xi}$,

$U_0^{'}=\frac{1}{2}U_0\ln(N-1)$, *N* is the neuron number, δ_{xi} is the random value in (-1, +1).

(4) Use dynamic equation, Calculate $\frac{dU_{xi}}{dt}$.

(5) According to the first order Euler method

$U_{xi}(t+1)$, $U_{xi}(t+1)=U_{xi}(t)+\frac{dU_{xi}}{dt}\Delta T$.

(6) Use sigmoid function to calculate $V_{xi}(t)$

$$V_{xi}(t)=\frac{1}{2}\left[1+\tan\left[\frac{U_{xi}(T)}{U_0}\right]\right]$$

(7) According to (3), computational energy function *E*. Check the path of legality, judge whether the end of the number of iterations. If the termination is end, the program is over. Otherwise it returns to step (4).

(8) Output the optimal path, optimal energy function, path length, energy function changes with time.

Under the Matlab Environment, the simulation environment for wireless sensor networks is established. The number of sensor nodes is 210, communication radius is 20.The sensors randomly are distributed in the area of 100 × 100 area, the initial energy of sensor nodes are uniformly distributed between 200 J and 400J, data packet collected by the sensor nodes is 512byte. Each node sends a packet consumes 0.2J; each node receiving

a packet consumes 0.01J. The network life-cycle is end when the number of nodes in the network is below 85%.

C. Simulation Results

In this paper, the largest energy path routing mechanism (MCP) and the maximum energy path switching routing mechanism (MCP-PS) are also simulated [17]. Three kinds of protocol simulation results are shown in Fig.6, 7. In Fig.6, When the node density is small, the network work cycle of three kinds of algorithms is similar, however the number of nodes is increased, compared with MCP and MCP-PS, MCP-GEN algorithm has more work cycle; This algorithm is also satisfied of the requirements of real-time path, the network life cycle of MCP-GEN is longer than the other two algorithms.

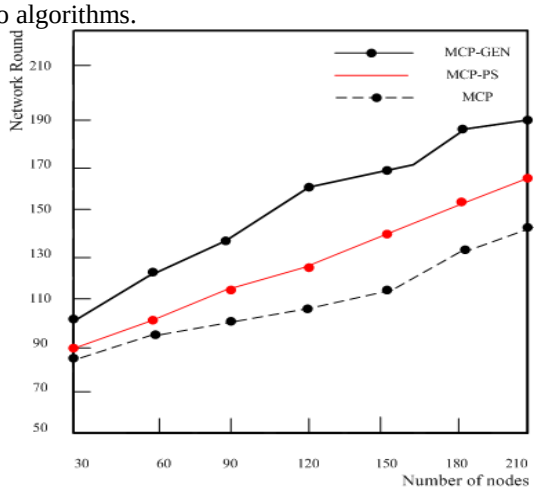


Figure 6. Network round with number of nodes

From Fig.7, it shows the packet loss rate of three kinds of protocol. When the node density is small, the packet loss rate of the MCP protocol is higher than the protocol of MPC-GEN or MCP-PS. When the node density is large, Three kinds of protocol packet loss rate is very similar and below 10%, while the range of the MCP-GEN slightly lower of than the other two kinds of routing protocol in the whole procession.

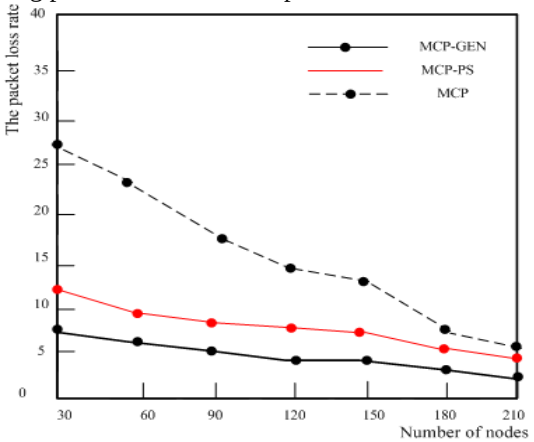


Figure 7. The packet loss rate with number of nodes

Simulation, the first step is to sow in the network within the region of 100 sensor nodes randomly and then proceed to the election of cluster-head node. The first cluster number of the network life-cycle has a certain influence in the area 200*100. From Fig.8, 9, the blue circle represents the survival of the first cluster node, while the colored stars point is the failure of cluster node. The nodes away from the gateway are close to death, but the majority is still far away from the gateway to survive. This is because far from the gateway node transmission distance less energy, but from the gateway nodes near large amount of data transmitted.

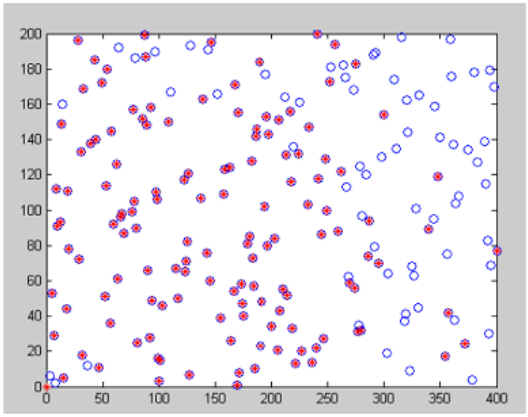


Figure 8. The Cluster-head Selection with LEACH

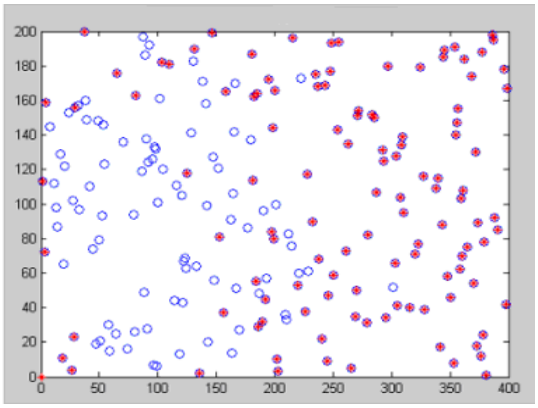


Figure 9. The Cluster-head Selection with GEN

The result can be concluded that: after the same number of rounds, the node selection algorithm based on GEN technology distribute less the death of node than the cluster head selection of LEACH. It can still monitor the entire network. Experimental results show that the number of trained data smaller than the original, the training time the network correspondingly is shortened. From above, the value of the allocation of their decision-making predictions is directly related to the allocation of weights in algorithm based on neural network. The weighted value calculations are more accurate, and it has a certain degree of intelligence.

VII. CONCLUSIONS

In this paper, neuron describes WSN nodes, the wireless sensor networks are expressed by a neural model. It introduces the design and realization methods of the neuronal model and neural network model. We build a gene pool for the design of genetic operators to avoid the algorithm into a part optimal solution which results in

premature convergence problem. At the same time, the algorithm balances energy consumption and extends the network life cycle; the network efficiency of WSN is improved. In future, we focus on how to build a high-quality gene pool.

ACKNOWLEDGMENT

This work is supported by two grants from the National Natural Science Foundation of China (No. 60773190, No. 60802002).

REFERENCES

[1] YOUNISM A, "An Energy-aware QoS Routing Protocol for Wireless Sensor Networks." Proceedings of the 23rd International Conference on Distributed Computing Systems Workshops, Los Alamitos, USA: IEEE Computer Society, 2003, pp.710-715.

[2] PerrigA,SzewczykR,WenV, "SPINS: security protocols for sensor networks", Proceedings of the 7th Annual International Conference on Mobile Computing and Networking ,USA: ACM, 2001,pp. 189~199.

[3] Heinemann W R, Chandrakasan A, Balakrishnan H, "An application-specific protocol architecture for wireless micro sensor networks", IEEE Transaction on Wireless Communications, 2002, 1(4): 660~670.

[4] B. Liu and D. Towsley, "A Study on the Coverage of Large-Scale Sensor Networks," Proc. First IEEE Int'l Conf. Mobile Ad-hoc and Sensor Systems (MASS '04), pp. 475-483, Oct. 2004.

[5] Zhu Xiaorong, Shen Lianfeng, "RBF-based cluster-head selection for wireless sensor networks", Journal of Southeast University (English Edition), 2006, 22 (4): 451-455.

[6] Ulakov A, Davcev D, "Data Mining in Wireless Sensor Networks Based on Artificial Neural Networks Algorithms," Proc of the 1stInternational Workshop on Data Mining in Sensor Networks of SDM,USA: SIAM Press, 2005,pp. 10-16.

[7] Y.Liu, H.Ngan and L.M.Ni, "Power-Aware Node Deployment in Wireless Sensor Networks", Int'l J. Distributed Sensor Networks, vol.3, pp. 225-241, April. 2007.

[8] Xiaorong Zhu, Lianfeng Shen, and Tak-Shing Peter Yum, "Hausdorff Clustering and Minimum Energy Routing for Wireless Sensor Networks", IEEE Transactions on Vehicular Technology. Vol58, pp. 990 – 997, February. 2009.

[9] L. Xing, A. Shrestha, "QoS reliability of hierarchical clustered wireless sensor networks", IEEE International Conference on Performance, Computing, and Communications, 2006, pp. 641-646.

[10] T. Zhong, S.Wang, S. Z. Xu,H. F. Yu, D. Xu, "Time Delay Based Clustering in Wireless Sensor Networks" ,Wireless Communications and Networking Conference,2007, pp.3956-3960.

[11] Lejiang Guo, Qiang Tang, "An Improved Routing Protocol in WSN with Hybrid Genetic Algorithm", The 2nd International Conference on Networks Security, Wireless Communications and Trusted Computing,2010, pp. 289-292.

[12] Sadjadi BA, Kiayias A, Mercado A, Yener B, "Robust key generation from signal envelopes in wireless networks.", Proceedings of the 14th ACM Conference on Computer and Communications Security, 2007,pp.401-410.

[13] Wang Dong, Zhang Jin-rong, Cao Chang-xiu, "The estimating calculation and distributing regularity of wireless sensors", Proceedings of the International Conference on Mechanical Transmissions. Chongqing, China: Science Press, 2006, pp. 532-1535.

[14] Lejiang Guo, Bingwen Wang, Weijiang Wang, "Research of energy-efficiency algorithm based on on-demand load balancing for wireless sensor networks",2009 International Conference on Test and Measurement, 2009, pp. 22-25.

[15] Akyildiz F, Su W, Sankara subramaniam Y, Cayirci E, "A surveyon sensor networks", IEEE Communications Magazine ,2002,pp.102–114.

[16] S. Lin, J. Zhang, G. Zhou, L. Gu, T. He, J.A. Stankovic, "ATPC: adaptive transmission power control for wireless sensor networks", Proceedings of the International Conference on Embedded Networked Sensor Systems, November 2006, pp. 223-236.

[17] Kannan AA, Mao G, Vucetic B, "Simulated annealing based wireless sensor network localization", Journal of Computers ,2006, 1(2): 15-22.

Lejiang Guo received the B.Eng. and M.Eng. degrees from the Department of automation science and technology, Xi'an University, Xi'an, China,in 1998 and 2005, respectively. He is currently working toward the Ph.D.degree at department of Control Science and Engineering, Huazhong University of Science and Technology, Wuhan, China. His research interests include optimization and control, with applications in communication systems, wireless systems and data networking, with a current focus on cross-layer design and optimization for energy-constrained wireless networks.

Qiang Li is a Senior IEEE member and a full professor at Air Force Radar Academy, China. He is leading a research group on networking and multimedia systems (LRSM Lab). He holds an engineer diploma in electronic from USTHB (1988), Algeria and a Master and Phd degrees in computer science respectively in 1989 and 1992 from Huazhong University of Science and Technology, China.

Fangxin Chen received his BE degree in computer engineering from Tsinghua University in 1982, ME degree in computer engineering from Tianjin University in 1991 and Ph.D. degree in computer science from Huazhong University of Science and Technology in 1999. Now, Dr.Chen is an assistant professor in Air Force Radar Academy. Dr.Chen has published more than 70 academic papers in the areas of wireless networks, optical networks, etc. His current research interests are wireless networks, optical networks, performance analysis of computer networks, etc. He currently serves as an associate editor for IEEE Communications Letters and an editor for IEEE Communications Surveys & Tutorials.

Security Issues and Solutions in 3G Core Network

Xuena Peng, Yingyou Wen and Hong Zhao
College of Information Science and Engineering, Northeastern Universiy
Shenyang, China
Email: {pengxn, wenyy, zhaoh}@neusoft.com

***Abstract*—Nowadays, the 3G network plays a very important role in mobile communication system. But the security concern of such network, especially the core network, is far from being satisfied. With the continuously development in the security enhancement in RAN, core network would become the future target of attackers. GPRS Tunnel Protocol (GTP), which is one of the key protocols in the core network, is quite vulnerable to attacks in the flat, full IP environment. Therefore solving such problem properly is very urgent and important for the operation of 3G network. In this paper, the security challenges in the 3G core network and the security issues in GTP are discussed, a defense solution for these security threats and an event-based description language are proposed. The experiment result shows the potential of our solution.**

***Index Terms*—3GPP; core network; security; GTP; event-based description language; Intrusion Prevention System**

I. INTRODUCTION

With the fast deployment of 3G system in the world, 3G-based applications will be a very prospective field in the next a few years. While more and more applications and services are provided by countless providers in the 3G network, the security of these applications as well as the fundamental 3G system will become one of the most important concerns of the public.

In fact, 3G system has its unique security problems that are not quite similar with the security problems of either the traditional GSM system or the IP network[1]. With the introduction of IP techniques to the mobile communication system, 3G system becomes more open, more flexible, and provides more interfaces to the other systems, but at the same time it is also more vulnerable. The vulnerabilities come from two sides. On one hand, the widely use of IP techniques breaks the technical barriers in the core network, so that the vulnerabilities hidden in the traditional system may be absolutely exposed in a full IP-based 3G system. On the other hand, the vulnerabilities of IP techniques may also be introduced into the 3G system, so that the security vulnerabilities of the system are further increased.

As a result, 3G security is one of the most important concerns as to 3GPP. TSG SA WG3 is a dedicated work group in researching the 3G security issues. TS21.133[1] and TS33.200[2] are some of their products. In TS21.133, the threats and requirements of 3G security are briefed, and in TS33.200, the security issues of both the access network and the core network are discussed. Although 3GPP have made many efforts in 3G security, generally speaking, most of current efforts are focused on the 3G radio access network (RAN), while the security issues in the core network (CN) have been underestimated. With the complexity and importance of the 3G core network, more efforts should be made to ensure its security.

There are three kinds of 3G core network techniques, namely TD-CDMA, WCDMA and CDMA2000. TD-CDMA and WCDMA are developed from GSM system, and they inherit GPRS (General Packet Radio Service) techniques in their core-network, which means, in these two 3G systems, GPRS network is responsible for connecting the internet with the system; while CDMA2000 gets rid of the GPRS transformation, and connect the internet directly with PSTN devices. To make it clear, in the context of this paper, we only focus on the security of the GPRS-based 3G core network, and the term core network is commonly used to infer GPRS core network if not explained specifically.

As an enrichment of the researches in this important area, our paper focuses on the security of the 3G core network, mainly discusses the security issues and solutions of GPRS Tunnel Protocol (GTP), which is the most important carrier protocol suite in 3G core network.

The rest of the paper is organized as follows. Section II discusses the security issues in the 3G core network and GTP in detail. Section III introduces a network event-based description language, which would be used as the fundamental technique in our proposed solution. Section IV proposes an Intrusion Prevention System for GTP as a solution to the former mentioned security issues. Section V provides the experiments and results. We give the conclusion in Section VI.

II. 3G CORE NETWORK SECURITY ISSUES

A. Introduction of GPRS Network and GTP

GPRS network techniques was first introduced to WCDMA in R99, developed in R4, R5, R6, R7/R8, and currently, they are still on the way of evolution. The GPRS core network is the kernel part of the GPRS system, and it provides support for WCDMA and TD-CDMA based 3G networks. In 3G networks, the GPRS core network provides mobility management, session management, and transport for Internet Protocol packet services. Besides, it also provides support for other

doi:10.4304/jnw.6.5.823-830

additional functions such as billing and lawful interception. Fig. 1 shows the general logical architecture of GPRS network. For simplicity, some of the secondary function nodes are ignored in this figure. In this architecture, the core network mainly refers to the sub network composed of SGSN (Serving GPRS Support Node) and GGSN (Gateway GPRS Support Node). Both SGSN and GGSN are known as GSN, which supports the use of GPRS in 3G core network. All GSNs should have a Gn interface and support the GPRS tunneling protocol. SGSN stands for Service GPRS Support Node, and it is responsible for the delivery of data packets from and to the mobile stations within its geographical service area. Its tasks include packet routing and transfer, mobility management (attach/detach and location management), logical link management, and authentication and charging functions. GGSN stands for Gate GPRS Supporting Node, and it is responsible for the inter-working between the GPRS network and external packet switched networks, like the Internet. GGSN can be viewed as a router to a sub-network, for it hides the GPRS infrastructure from the external network. Besides, it also plays an important role in mobility management and packet forwarding between the core network and external network.

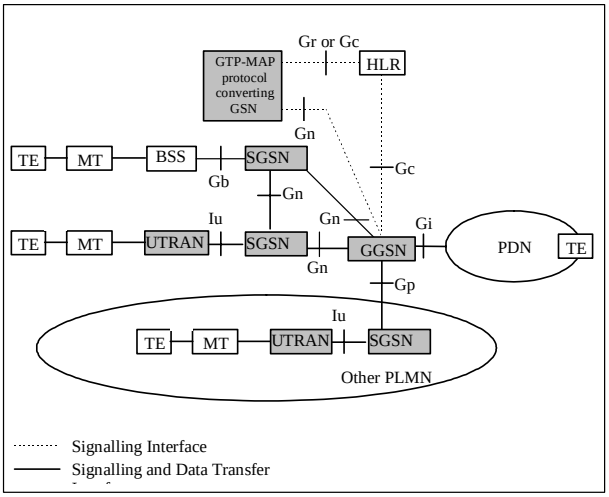


Figure 1. GPRS logical architecture with interface name denotations

GTP is the most important carrier protocol suite in 3G core network. It stands for a suite of IP-based communication protocols used to carry GPRS within GSM and most 3G networks. Primarily it is the protocol which allows end users of WCDMA or TD-CDMA network to move from place to place while continuing to connect to the Internet as if from one location at the GGSN. It does this by carrying the subscriber's data from the subscriber's current SGSN to the GGSN which is handling the subscriber's session.

GTP can be decomposed into separate protocols, GTP-C (GTP control plane), GTP-U (GTP user plane) and GTP' (GTP prime).GTP-C is used within the GPRS core network for signaling between GGSN and SGSN. This allows the SGSN to activate a session on a user's behalf (PDP context activation), to deactivate the same session, to adjust QoS parameters, or to update a session for a subscriber who has just arrived from another SGSN. GTP-U is used for carrying user data within the GPRS Core Network and between the RAN and the CN. GTP' uses the same message structure as GTP-C and GTP-U, but has an independent function. It can be used for transfer of charging data from GSNs to the charging function.

B. Security Challenges in 3G Core Network (CN)

Since 3G RAN has adopted CDWA technology, and is equipped with the security mechanisms, such as access control, encryption, authentication and etc., it is well protected to some degree. But with the continuously development on the security in RAN, core network would become the future target of attackers. Although 3GPP and 3GPP2 has proposed the security framework of the core network, but their proposals about the possible security threats in 3G core network are neither comprehensive nor sufficient.

Compared with RAN, the core network, especially the packet switch (PS) domain, has more interfaces connecting with the outside world, so it is possibly facing much more security challenges.

The attacks could be originated from other networks which are interfacing with the core network, such as internet and global roaming partners, or they could be comes from the inside of the core network, such as content provider who provides their services from the inside of the core network. The attacks could possibly be originated from any network interfaces. These attacks may target the overall performance of the network, resulting in system downtime; or they may target specific applications such as accounting and billing systems. In fact, one of the most severe security issues comes from GPRS Tunnel Protocol (GTP), which is used as the protocol connecting SGSN and GGSN. GTP is based on UDP, and it has no embedded security mechanisms. Besides, the different implementations of a variety of vendors further bring about a tremendous threat to the system. So security protection is needed at corresponding network interfaces, such as Gi, Gn, Ga, and etc.

The attacks could also originate from the end users of the 3G system. With the enhancement on mobility and processing ability of the terminals, the security threats of the core network are no longer limited to the interfaces to other networks, validated subscribers could also launch attacks to the core network, and it is much more difficult to defend. For example, a DDoS attack originated from the subscribers may disable the service capabilities in a sudden. So security protection is also needed at RAN-CN interface.

Beside, in order to comply with local laws and regulations, ISP may also need to be responsible for decreasing DDoS attack to the outside network, forbidding the end users from viewing pornographic sites or delivering reactionary remarks, thus these user behavior related security issues should also be handled by ISP. Actually, this is not direct security threat to the core network, but as the information about the user behavior, user identity association and etc can be easily obtained in the core network, it makes the core network to be the perfect place to handle such problem.

As identified above, there are many security challenges in the core network, but, until now, not many researches have been done in protecting it. A. Prasad discussed the infrastructure security of 3GPP UMTS network[3]. K. Boman proposes the MAPSec-based 7 signaling MAP protection, IPSec-based IMS SIP protection and WTLS-based application data protection schemes[4]. Unter Schafer focuses on the DOS and privacy issues in the IP-based core network[5]. Xiaoming Fu discusses the security implications of the session identifier[6]. C. K. Dimitriadis has proposed honeynet solution to secure 3G core network[7]. Fabio Ricciato discussed the unwanted traffic in 3G networks and their impacts for core network[8,9], and developed an extensive monitoring system for 3G core network[10].

All the above solutions only solve a small margin of the security challenges in 3G core network, there's still a long way to go in this research area.

C GTP Security Issues

Since protocols are one of the most important scheme that make system work as expected, it is considered a possible solution to protect the network through protecting the protocols. In our point of view, among all the protocols used in the core network, GTP (GPRS Tunnel Protocol) is considered to be the most important protocol, for almost all user operations, protocols and data are carried by GTP. GTP is an IP-based tunnel encapsulation protocol in the GPRS core network, and it is an important transforming scheme to connect the PLMN with other networks (including the Internet).

The design of GTP is lack of security concern, and there are no embedded security schemes in GTP. With the consideration of compatible with the 2G network, the protocol itself is also not perfect. Particularly, in the 3G network interconnection process, the messages delivered between SGSN and GGSN from different domains may incur attacks from the outside, so that GTP itself will be fully exposed to outside threats. Although GTP has been upgraded for 3 versions, more and more vulnerabilities have been exposed. Currently GTP has been considered as the most obvious security issues in 3G core network.

With the widely usage of GTP, attacks toward this protocol could come from different directions, such as the air interface, internet and other PLMN (Public Land Mobile Network), so the attacks could make very huge damage not only to the core network infrastructure, but also to the internet and mobile users. Typical attacks range from over billing attack to infrastructure attack. Basically, GTP security issues can be roughly classified as followings.

1) Protocol abnormal attack.

This kind of attack often generates abnormal or damaged PDU packet, or PDUs not comply with the protocol. In regular scenarios, there are no such packets, but for malicious intentions, attackers can generate such packets to make use of the vulnerabilities in protocol processing programs, so that they can degrade the performance of the system or gain illegal permissions. Some of such attacks are listed as follows.

- ✓ Invalid Reserved fields: There are many reserved "SPARE" fields in the header of GTPv0 and GTPv1. For GTPv0 these fields should be set and for GTPv1 these fields should be unset. So it would be abnormal, if the contents of such fields in the corresponding packet don't comply with such specifications. Depending on the nature of vulnerability within the device from different vendors, the attack ranges from Denial of Service to remote compromise.
- ✓ Invalid Reserved message type field: In both GTPv0 and GTPv1, message type field is one byte and allows 255 different message types. Message type values that are listed as reserved or for future use should not be used. If the header does not comply with the protocol specifications, it is an indication of an attack involving malformed or corrupt packets. Depending on the nature of vulnerability within the device from different vendors, the attack ranges from Denial of Service to remote compromise.
- ✓ GTP over GTP: As GTP is used to encapsulate packets originating from a mobile terminal, it is possible for a mobile terminal to create a GTP packet and forward it along to the SGSN. Upon receiving the GTP packet, the SGSN will encode it again, and forward it to the GGSN, through the relative PDP context. This embedded GTP packet may be potentially decoded via the GGSN and forwarded into the GGSN infrastructure, or decoded a second time, allowing an attacker to spoof GTP packets coming from a range of different answers. Another potential attack would be attackers sending recursive GTP packets, which is a GTP packet that contains X number of other GTP packets embedded within.

2) Infrastructure attack (GTP Deception).

The impacts of this kind of attack often include illegal access of the infrastructure devices, such as SGSN, GGSN, OAM system and mobile terminals. By modifying his own address, the attacker can connect with the core network, and encapsulate attack packets into GTP, thus attack any mobile targets or targets from other network.

The end user may also encapsulate the attack message into a GTP packet, and such attack message may be routed to any mobile targets in the network, and even the targets of the outside network, through a vulnerable GGSN (in which GTP packet maybe resolved more than once).

3) Resource consumption attack.

By now, this kind of attack can be launched from two positions, namely, mobile terminals and other in-network devices which can set up valid connections with the infrastructure. In the long term, this kind of attack may also be launched from the outside network. Depending on the nature of vulnerability within the GSN device from different vendors, the attackers may launch SYN-like attacks, initializing thousands of PDP contexts, so that it

will disable the GGSN from allocating new PDP context, resulting in denial of service.

Besides, terminals may mobile between PLMNs (Public Land Mobile Network). When terminal mobiles from one PLMN (for example, home ISP) to another PLMN (for example, roaming partner), the old SGSN should do TCP-like 3-phase handshakes with the new SGSN, and after the handshakes, new SGSN would take over the connection from the old one, and relay data for the terminal. If malicious or compromised SGSN keep invoking the handshakes while not finishing it, then the invoked SGSN would deny services to the other legal terminals.

All the above are the major security issues of GTP itself. As the only tunnel protocol connecting mobile terminals with the outside network, such as internet, GTP protocol suite also is responsible for relaying user application messages. Keeping the terminals from being attacked by malicious attackers from the internet should be considered in the GTP protection solutions. Besides, as mentioned in II.B, core network work is a perfect place to regulate user behaviors, so such requirement should also be considered in GTP protection context.

For protecting GTP from the above security issues, we consider GTP traffic analysis and filter as a possible solution. So in the next two sections, we will describe our solution in detail.

III. NETWOKR EVENT-BASED DESCRIPTION LANGUAGE

In order to analyze and filter the GTP stream more flexibly and more efficiently, we introduce an event-based description language. In this context, event refers to any possible activity which can be detected in the packet or packet stream, and attack refers to a single malicious event or a serial of malicious or non-malicious events formed in a certain logic pattern to implement a malicious aim. The language that we propose is a simple script language which can be used to define the logic of the application protocol as well as that of the attacks. When application level protocol analyzer is defined in our event language, the underlying event engine can analyze the corresponding traffic according to the definition of the protocol and generate the events for further inspection. Besides, it can also be used to maintain the protocol state and deal with the security events. When an attack signature and logic is defined in our language, the underlying event engine can analyze the incoming event generated by the protocol analyzer according to the attack definition, and take action when attacks are detected. By unifying the description of the protocol and attack in the same descriptive manner, the system's extensibility can be easily improved. By coupling protocol analysis and attack detection tightly in the runtime, the system efficiency can also be improved. Besides, by defining application protocol in this language, we simply need to focus on the protocol logic while let the event engine to deal with other less relevant processing details.

There are two categories of event in the context of our attack description language, namely atom event and abstract event. Atom event refers to the very specific event that is generated by the protocol analyzer module, such as "UDP event", while the abstract event is made of one or several events (atom event or abstract event), the purpose that we introduce abstract event is to indicate any serials of atom events with different intensions, so that we can use the most proper abstract event to describe the attack.

In our language, we use "**Atom**" and "**Event**" as the key words to describe atom event and abstract event respectively. They can be used in the following style:

```
Event <type> abstract_event;
Atom <type> atom_event;
```

In GTP protocol description, when describing an atom event, such as SIMPLE_IE, we can put it in the following pattern:

```
Event struct hdr_v0 SIMPLE_IE;
```

In which, *struct hdr_v0* is one of our predefined GTP_v0 data structure, and *SIMPLE_IE* is an atom event which refers to a GTP v0 packet with Information Element (IE).

When defining more abstract event, we need to use rules. Rules can be used to define the logic among events. A rule is composed of four parts: the left part, operator, the right part, and the action part. The following example shows what a rule looks like:

```
BAD_LONG_GTP_V0 : header_v0 ($1->gtp_hdr_v0_len >=160)
{
        GTP_v0_deny($0, "fault lenth header! Len = %d\n,
                $1->gtp_hdr_v0_len);
}
```

The left part of the rule is composed of an abstract event, which is viewed as the object to be defined in the rule. In this example, it is the abstract event *BAD_LONG_GTP_V0*. The right part of the rule is composed of one or more events (including both atom event and abstract event) with predicates, in this example, the event is an atom event *header_v0* with a predicate *$1->gtp_hdr_v0_len >=160*. The ":" in the middle of the first line is the operator of the rule, meaning the abstract event at the left part is a sum up of the events at the right part. So the whole line 1indicates that when a header_v0 event with it parameter gtp_hdr_v0_len no less than 160 is detected, then BAD_LONG_GTP_V0 event can be generated.

Lines 2-5 are the action part of the rule. *GTP_v0_deny* in line 3 is the rule action, which is enforced when the abstract in the left part is generated. In this example, when a BAD_LONG_GTP_V0 event is generated, the action *GTP_v0_deny* should be enforced, and then the system will simply deny and discard the corresponding abnormal packet.

IV. INTRUSION PROTECTION SYSTEM FOR GTP

A. System Architecture

In 3G core network environment, network traffic filtering and security event analysis must be fast and reliable responded. As a network protection system deployed inline, it must satisfy the critical real time requirement.

We have proposed a GTP IPS to protect GTP protocol against attacks as well as satisfying the real time requirement. The system is based on the event-based description language and event analysis engine introduced in section III. The architecture of the GTP IPS is shown in Fig. 2. In this system, we have applied hardware based data stream capture techniques to capture network packets in real time, while for the upper layer protocol event analysis and filtering, we also accurately control the response speed as well as its reliability.

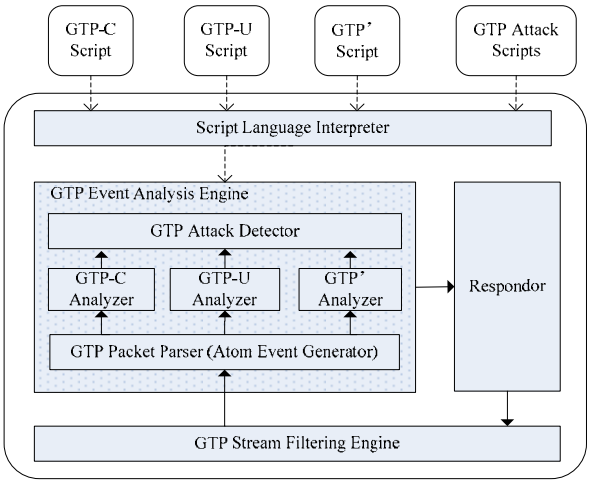


Figure 2. GTP IPS Architecture

There are mainly four modules in this architecture. One is for system initialization and the other three are for runtime analysis.

The *Language Interpreter* is mainly used to fulfill the initialization and configuration phases of the system. During system initialization, it translates the GTP protocol scripts and attack scripts into event engine based *GTP-X Analyzers* and *GTP Attack Detectors*, and these analyzers and detectors are to be used to analyze GTP protocols and attacks during runtime.

The runtime GTP analysis and filtering scheme is implemented by three modules:

- *GTP Stream Filtering Engine* is implemented by a set of parallel processing hardware, which can perform simple analysis and filtering of thousands of packets simultaneously. Parallel analysis and filtering can assure the secure packets to pass through the system as soon as possible, without affecting the overall performance of the system.
- *GTP Event Analysis Engine* is composed of three parts: *GTP Packet Parser* (also *Atom Event Generator* in the context of e*vent engine* mentioned in III.A), *GTP Protocol Analyzers* and *GTP Attack Detector*. The GTP Packet Parser parses the incoming GTP packets according to the protocol specification at the lexical level, distills the essential protocol elements, encapsulates these protocol elements into "Atom Events", and submits them to the corresponding protocol analyzers. The *GTP Protocol Analyzer* then matches the atom events to the predefined rules, which are predefined either by default or by the users, and generate various "Abstract Events" to facilitate further protocol analysis and attack detection. The system default rules are defined according to the various 3GPP specifications, such as R5, R6 or higher versions. The *GTP Attack Detector* finally analyzes the attributes of the relevant "Abstract Events" generated by *GTP Protocol Analyzer*, and examines whether the signatures of the attacks are existed.
- If attacks are detected, the *Responder* Module will be responsible for dealing with the reactions, such as logging, alerting, dropping the packets, and etc. If the relevant packets are to be dropped, the *GTP Stream Filtering Engine* is responsible for enforcing the corresponding action, as well as maintaining and updating the protocol state of the relevant GTP tunnel.

With the idea of application level purging, the protocol parser parses and checks the packets according to the 3GPP specifications, distills the protocol elements, encapsulates the atom events, and submits them to the corresponding event analysis engines. Whenever a new attack is appeared, it is unnecessary to manually modify the source code of the protocol analysis engine, but only to add a new piece of attack description script, and load it to the system. This scheme is easy to use, and really facilitates both the system producer and end user when new detection requirements are generated and updated. Developing and updating GTP IPS in such manner, especially in 3G core network environment, can be much simpler and quicker, and the protection for newly emerged attacks are much more in time and robust. Furthermore, the end user has more control on the reaction to the attacks. They can configure the reaction levels up to their needs, such as dropping the abnormal packets, disconnecting current connection and sending the administrator an alert. In this way, the end user can be actively involved in the real time protection of the GGSN and SGSN nodes.

B. Signalling Message Analysis and Detection for GTP-C

The protection for the signaling message of the GTP control plane is of the most importance for securing the GGSN and SGSN node in the core network. On one hand, some of the attacks could be detected during the analysis and detection process of GTP-C; on the other hand, by parsing the GTP-C messages, important information about user connections may be easily obtained, thus the GTP-U analyzer could analyze GTP-U messages and detect attacks according to such information.

The signaling messages of the GTP control plane are composed of path management messages, tunnel management messages, location management messages, and mobility management messages. In our GTP protection system, the detection for GTP-C oriented attacks is performed by the *GTP*-C specific *Event Analysis Engine*, which includes *GTP Packet Parser*, *GTP-C Analyzer* and *GTP Attack Detector*.

GTP Packet Parser can perform the message parsing for GTP-C protocol, and examine the protocol abnormity according to the protocol specification. It usually generates one header and several IE atom events for each GTP message, and submits them to the corresponding analyzer in the parsing sequence. The parser only parses and distills the type of each IE, such as TLV or TV, but do not parses the detail segments in the IE message. It is up to the *GTP-C Analyzer* and the *GTP Attack Detector* to decide whether to parse the details or not. After the packet parsing phase, the atom events are encapsulated and submitted to the upper modules to be further analyzed. The *GTP-C Analyzer* and the *GTP Attack Detector*, which are automatically generated by the protocol and attack detection scripts, are used to detect protocol abnormity and specific attacks. These two modules guarantee the validity and security of the GTP messages passing through the GTP IPS, and avoid possible attacks to the protocol and system vulnerabilities.

In order to guarantee the validity of the state of the GTP messages, PDP context information must be maintained and synchronized with the GGSN and SGSN. So in the whole analysis and detection process, a PDP context record is set up for each created and in use tunnel, and is updated whenever the signaling message passes the corresponding tunnel. The PDP context record is indexed in the data structure of binary tree, and it is indexed by each tunnel direction (up stream and down stream). In this way, the state-based protocol abnormal detection is implemented. Besides, this state-based mechanism can also be used to support the extended protection and system maintenance.

C. User Data Analysis and Detection for GTP-U

GTP user plane traffic monitoring is very important for mitigating ISP's operating risks. On one hand, it could protect the vital equipments in the core network from being attacked, for example, GTP-over-GTP attack. In such attack, mobile station creates GTP packet and forward it into the core network through SGSN, it exists in the GTP user plane traffic, so it could only be detected and prevented in the user plane. On the other hand, the ISP could also enforce their regulations towards mobile stations' behavior, for example, preventing mobile stations from accessing illegal or insecure sites.

The key in GTP user plane traffic analysis and detection is to distill and parse user data from GTP-U traffic, and according to the PDP context information distilled by GTP-U analyzer, analyze and detect whether the user behavior is secure or not.

The overall processing procedure for GTP-U traffic is quite similar to that of the GTP-C traffic, so we do not go into it here. What we need to emphasize is that GTP-U analyzer focus more on the user data analysis relayed by GTP-U traffic, but itself. Depending on the requirements, GTP-U analyzer could parse the user traffic from network layer to application layer, and store and maintain user traffic in a 2-dimension chain, thus the system could detect those events avoiding ISP's security regulations.

V. EXPERIMENTS AND RESULTS

Based on the architecture that is proposed in section IV, a GTP IPS prototype system is implemented. Currently, the prototype system supports deep analysis and filtering functions for both GTP-C and GTP-U, and GTP' protection will also be supported in the future.

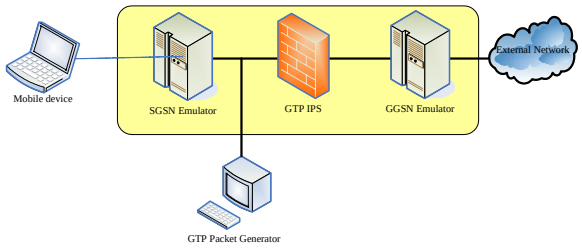


Figure 3. GTP IPS Experiment Environment

The evaluation of the system is a very important phase to validate its functionality. Currently, since it is unrealistic to test our prototype system in a live GPRS system, the best way to test it is to carry out the experiments under the lab environment. Fig. 3 shows the emulation environment for testing the system. We employ OpenGGSN emulator suite[11] to emulate the SGSN and GGSN nodes in the GPRS network.

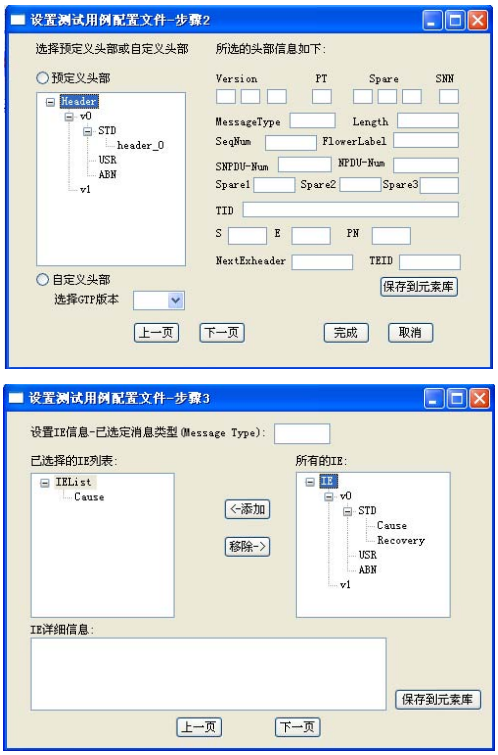
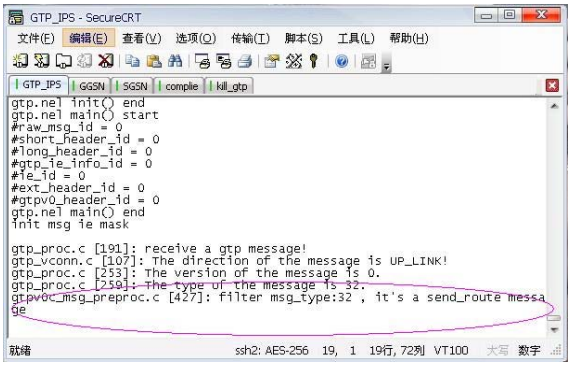


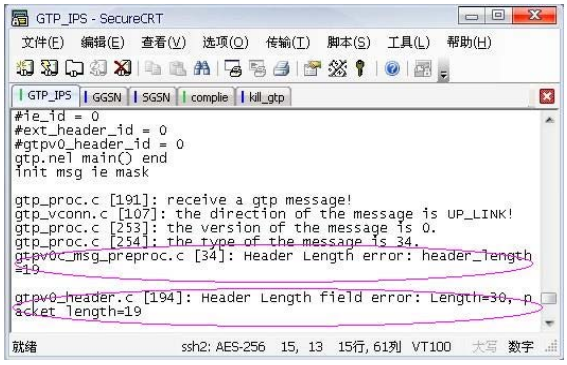
Figure 4. Dynamic configuration UI of GTP packet generator

In order to validate the effectiveness of the proposed filtering and protection methods, we implement a GTP packet generator, which is compatible to 3GPP R4/R5 specifications. The packet generator can generate various abnormal GTP packets with incorrect headers or IEs as well as attack packets. Fig. 4 shows the dynamic configuration UI of the packet generator. For testing the functionality of GTP IPS prototype system, we need to

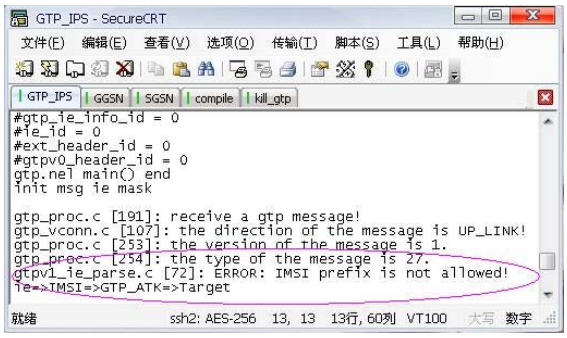
deploy this generator between the SGSN and the GGSN so that GTP messages can be intercepted and filtered.



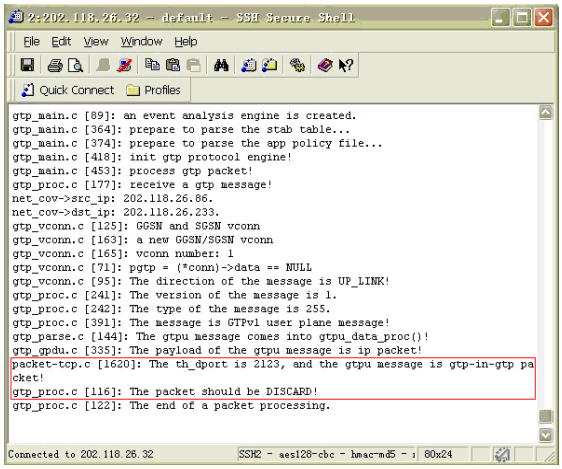
(a) Message type filtering



(b) Abnormal message detection



(c) IMSI prefix filtering



(d) GTP in GTP message detection

Figure 5. GTP IPS output in different experimental scenarios

A serial of experiments have been set up to evaluate the validity of the GTP IPS. For GTP-C protection functioning, the experimental scenarios include message filtering, abnormal message detection, access control, and traffic control; for GTP-U protection functioning, the experimental scenarios include content auditing and GTP-in-GTP attack detection. The experiment results show that the system can work correctly under all these scenarios. Fig. 5 shows some of the system output in different experimental scenarios.

VI. Conclusion

The security of core network is very important to the overall healthy of 3G network. Since GTP is the protocol to tunnel user data across the core network, its security detrimentally affects the operations of the whole 3G network. In this paper, we have discussed the security challenges in the 3G core network, and analyzed the security issues of GTP protocol suite in detail, then proposed an event-based GTP IPS as a countermeasure to protect this key protocol suite, and also evaluate the prototype system in various scenarios. The experiment results show that the system can work pretty well under lab environment. In the future, we will embody more functions to the system as well as improve the performance of the system. Besides, we will also try to evaluate the system in a more real environment.

Acknowledgment

This work is under the sponsorship of NSF of China, with the grant ID of 60602061and 60803131.

References

[1] 3GPP TS 21.133 V4.1.00:"Security Threats and Requirements". Available from: http://www.3gpp.org.

[2] 3GPP TS33.200 V5.1.0. "3G Security; Network Domain Security; MAP application layer security". Available from: http://www.3gpp.org

[3] Prasad, A., H. Wang, and P. Schoo, Infrastructure Security for Future Mobile Communication System, in WPMC 2003. 2003: Yokosuka, Japan.

[4] Boman, K., et al., UMTS security. IEEE Ectronics&Communication Engineering Journal, October, 2002.

[5] Shafer, G., Research Challenges in Security for Next Generation Mobile Networks, in Workshop on Pioneering Advanced Mobile Privacy and Security (PAMPAS). 2002: Royal Holloway University, Egham, Surrey, United Kingdom.

[6] Fu, X. and H. Tschofenig, Security Implications of the Session Identifier. 2005, Institute for Informatics, University of Goettingen, Germany.

[7] Dimitriadis, C.K., Improving Mobile Core Network Security with Honeynets. IEEE Security & Privacy, 2007. 5(4): p. 40-47.

[8] Ricciato, F., Svoboda P. and etc, On the Impact of Unwanted Traffic onto a 3G Network. Forschungszentrum Telekommunikation Wien. Ftw. Technical Report FTW-TR-2006-006.

[9] Ricciato, F. Unwanted Traffic in 3G Networks. ACM SIGCOMM Computer Communication Review, 2006. 36(2):p 53-56.
[10] METAWIN and DARWIN projects. http://www.ftw.at/ftw/research/projects.
[11] openggsn. Available from: http://sourceforge.net/projects/ggsn/

Xuena Peng was born in 1979, in Shenyang, China. She received her PH.D. degree in computer engineering and science from Northeastern University in China, in 2007. Her main research interest include: Intrusion detection, 3G network security, security alert analysis, P2P streaming, and sensor network.

Currently, she is working as a postdoc in Northeastern University, China. Her current research focuses on WSN Middleware and network security.

Yingyou Wen was born in Shenyang, China on March 12, 1974. He received the PH.D. degree in computer science from Northeastern University, China, 2005.his main research include: network security, wireless communication and sensor network.
He is an associate professor of Northeastern University, and vice president of Neusoft Research， China. His current research focuses on Middleware of WSN and network security.

Hong Zhao was born in Hejian, China in 1954. He received his PH.D. degree in computer science from Northeastern University, China, 1991. His main research include: computer network and information security.

He is a professor of Northeastern University, and super vice president of Neusoft Group, China. His current research focuses on network security and wireless communication.

Call for Papers and Special Issues

Aims and Scope.

Journal of Networks (JNW, ISSN 1796-2056) is a scholarly peer-reviewed international scientific journal published monthly, focusing on theories, methods, and applications in networks. It provide a high profile, leading edge forum for academic researchers, industrial professionals, engineers, consultants, managers, educators and policy makers working in the field to contribute and disseminate innovative new work on networks.

The Journal of Networks reflects the multidisciplinary nature of communications networks. It is committed to the timely publication of high-quality papers that advance the state-of-the-art and practical applications of communication networks. Both theoretical research contributions (presenting new techniques, concepts, or analyses) and applied contributions (reporting on experiences and experiments with actual systems) and tutorial expositions of permanent reference value are published. The topics covered by this journal include, but not limited to, the following topics:

- Network Technologies, Services and Applications, Network Operations and Management, Network Architecture and Design
- Next Generation Networks, Next Generation Mobile Networks
- Communication Protocols and Theory, Signal Processing for Communications, Formal Methods in Communication Protocols
- Multimedia Communications, Communications QoS
- Information, Communications and Network Security, Reliability and Performance Modeling
- Network Access, Error Recovery, Routing, Congestion, and Flow Control
- BAN, PAN, LAN, MAN, WAN, Internet, Network Interconnections, Broadband and Very High Rate Networks,
- Wireless Communications & Networking, Bluetooth, IrDA, RFID, WLAN, WMAX, 3G, Wireless Ad Hoc and Sensor Networks
- Data Networks and Telephone Networks, Optical Systems and Networks, Satellite and Space Communications

Special Issue Guidelines

Special issues feature specifically aimed and targeted topics of interest contributed by authors responding to a particular Call for Papers or by invitation, edited by guest editor(s). We encourage you to submit proposals for creating special issues in areas that are of interest to the Journal. Preference will be given to proposals that cover some unique aspect of the technology and ones that include subjects that are timely and useful to the readers of the Journal. A Special Issue is typically made of 10 to 15 papers, with each paper 8 to 12 pages of length.

The following information should be included as part of the proposal:

- Proposed title for the Special Issue
- Description of the topic area to be focused upon and justification
- Review process for the selection and rejection of papers.
- Name, contact, position, affiliation, and biography of the Guest Editor(s)
- List of potential reviewers
- Potential authors to the issue
- Tentative time-table for the call for papers and reviews

If a proposal is accepted, the guest editor will be responsible for:

- Preparing the “Call for Papers” to be included on the Journal’s Web site.
- Distribution of the Call for Papers broadly to various mailing lists and sites.
- Getting submissions, arranging review process, making decisions, and carrying out all correspondence with the authors. Authors should be informed the Instructions for Authors.
- Providing us the completed and approved final versions of the papers formatted in the Journal’s style, together with all authors’ contact information.
- Writing a one- or two-page introductory editorial to be published in the Special Issue.

Special Issue for a Conference/Workshop

A special issue for a Conference/Workshop is usually released in association with the committee members of the Conference/Workshop like general chairs and/or program chairs who are appointed as the Guest Editors of the Special Issue. Special Issue for a Conference/Workshop is typically made of 10 to 15 papers, with each paper 8 to 12 pages of length.

Guest Editors are involved in the following steps in guest-editing a Special Issue based on a Conference/Workshop:

- Selecting a Title for the Special Issue, e.g. “Special Issue: Selected Best Papers of XYZ Conference”.
- Sending us a formal “Letter of Intent” for the Special Issue.
- Creating a “Call for Papers” for the Special Issue, posting it on the conference web site, and publicizing it to the conference attendees. Information about the Journal and Academy Publisher can be included in the Call for Papers.
- Establishing criteria for paper selection/rejections. The papers can be nominated based on multiple criteria, e.g. rank in review process plus the evaluation from the Session Chairs and the feedback from the Conference attendees.
- Selecting and inviting submissions, arranging review process, making decisions, and carrying out all correspondence with the authors. Authors should be informed the Author Instructions. Usually, the Proceedings manuscripts should be expanded and enhanced.
- Providing us the completed and approved final versions of the papers formatted in the Journal’s style, together with all authors’ contact information.
- Writing a one- or two-page introductory editorial to be published in the Special Issue.

More information is available on the web site at http://www.academypublisher.com/jnw/.

(Contents Continued from Back Cover)