TRECVid 2012 Semantic Video Concept Detection by NTT-MD-DUT

Yongqing Sun, Kyoko Sudo, Yukinobu Taniguchi
NTT Media Intelligence Laboratories, Japan
1-1 Hikarinooka Yokosuka Kanagawa, 239-0847 Japan
Haojie Li, Lei Yi, Yue Guan
School of Software, Dalian University of Technology, China
yongqing.sun@lab.ntt.co.jp

ABSTRACT

In this paper, we describe the TRECVid 2012 video concept detection system first developed at the NTT Media Intelligence Laboratories in collaboration with Dalian University of Technology. For this year's task, we adopted a subspace partition based scheme for classifier learning, with emphasis on the reduction of classifier complexity, aiming at improving the training efficiency and boosting the classifier performance. As the video corpus used for TRECVid evaluation is ever increasing, two practical issues are becoming more and more challenging for building concept detection systems. The first one is the time-consuming training and testing procedures, which have taken up most of the evaluation activities, preventing the design and testing of novel algorithms. The second and the more important issue is that when using whole data for classifier training, the derived separating hyperplanes would be rather complex and thus degrade the classification performance. To address these issues, we propose to adopt the "divide-and-conquer" strategy for concept detector construction as follows. We first partition the whole training feature space into multiple sub-space with a scalable clustering method, and then build sub-classifiers on these sub-spaces separately for each concept. The decision of a testing sample is the fusion of the results a few fired sub-classifiers. Experimental results demonstrate the efficiency and effectiveness of our proposed approach.

Keywords
Concept Detection, Video Retrieval, Subspace Partitioning

1. Introduction

From the launching of the TRECVid Semantic Indexing (SIN, formerly called feature extraction and later high-level feature extraction) task at 2002, the benchmark dataset for evaluating the effectiveness of detection methods for semantic concepts has been increasing continuously. Tab. 1 shows the growth of SIN evaluation dataset for the recent five years.

Tab.1. The scale of datasets used for TRECVid semantic indexing or high-level feature extraction task

Year	Dataset length (hours)	Master shots
TV2008	~200	72028
TV2009	~380	133412
TV2010	~400	266473
TV2011	~600	403800
TV2012	~800	631646

It is noted that TRECVid also encourages the participants to include more training samples from other sources. With the rapid development of social multimedia on the Internet, researchers are getting interested in harvesting training samples from web data[1]. Thus for today's concept detectors construction, we have large number of data available.

For large-scale video database, the training and testing are really time-consuming, which have taken up most of the TRECVid evaluation activities, preventing the design and testing of novel algorithms.

* TRECVID'12, November 26 – 28, 2012, Gaithersburg, Maryland USA.

Moreover, due to the well-know semantic gap existed between low-level features and high-level concepts [2], the distribution of the entire training low-level features of unique concept is diversified and complex, thus it is difficult to learn a single classifier on the space. Motivated by the above issues, we propose to adopt the "divide-and-conquer" strategy, namely, we adopt an ensemble learning framework to improve the training efficiency and effectiveness of concept detectors.

In the rest of the paper, we first review the related work on concept detection in Section 2. Then, we introduce and describe the framework of our system in Section 3. In Section 4 we give the details of our subspace partition approach. The experimental analysis will be given in Section 5. Finally, we conclude our work in Section 6.

2. Related work

Concept detection, or referred as high-level feature extraction by TRECVid before 2010, plays fundamental role in multimedia information indexing, retrieval and managements [3][4], and in various related multimedia applications, such as video question answering [6], video summarization [7], etc., thus has attracted intensive research efforts recently. The goal of concept detection is to build mapping from low-level features to high-level concepts with machine learning techniques [4]. The main modules of state-of-the-art concept detection systems include feature extraction and fusion, and classifier training.

Various low-level visual features including global and local features have been proposed and tested for concept detection and other visual classification tasks. Color histogram, color moments, color correlogram, edge histogram, co-occurrence texture and wavelet texture are the most widely used global features in TRECVid evaluations [11][5]. Meanwhile, SIFT descriptors [11] extracted from keypoints have been the standard local features, showing impressive performances in concept detection [16]. While those above features are demonstrated complementary in representing video frames [5], lots of approaches have been proposed to fuse them to boost performance, which can be categorized into early-fusion and late-fusion methods. Early-fusion scheme concatenates multiple feature vectors into a larger single vector and feeds to classifiers while late-fusion method trains multiple classifiers using separate features and then fuses the results of classifiers. It has been shown by TRECVid participants [5] that early fusion achieves better performance than late-fusion. However, training on concatenated larger vector will greatly increase the computational cost than on multiple shorter vectors, making it impractical for most participants to design and testing algorithms with tight submission schedule of TRECVid.

As for classifier used for concept detection, a variety of classification techniques such as support vector machine (SVM) [11][5], Gaussian mixture model [8], etc, have been studied. The success of these classifiers largely depends on the separability of the underlying data structure. The scale of TRECVid training samples for single concept ranges from thousands to dozens of thousands, showing highly complex separating hyperplanes in feature space, presents grand challenges to the design of classification algorithms on such large scale dataset. Thus, it is highly desirable to partition the whole dataset into smaller collections to make the data structure more separable to improve classification performance.

To address the above issues, in this work, we propose to adopt a scalable clustering method, called Clara, to partition the training samples. Clara (Clustering for Large Applications) [10] is a partitioning algorithm which can deal with large data sets with a sampling scheme. After partition, we train set of classifiers on the subspaces. In testing, a few of classifiers are fired and fused to give the classification result of given testing keyframe.

3. System framework

The framework of the proposed concept detection system is illustrated in Fig.1. As is shown in the figure, the whole system consists of two stages: training stage and testing stage. We describe the main steps for each stage as follows.

3.1 Main steps for training stage

(1) **Feature Extraction.** Similar to our work of last year [11], we use two types of visual features extracted from keyframes as the signature of video shots: global and local features. The global features include color moments and edge histogram [12]. The widely used bag of visual words (BoW) representation is adopted as local feature, which is based on a visual vocabulary of visual words clustered by a set of SIFT features [13]. To enhance the discriminative power of the global features, we adopt the grid or patch based

manners. For color moment, we partition a keyframe into 5×5 grid and extract the first three moments of the three channels of LAB color space. Thus the final color moment is a 225 dimensional vector. To extract edge histogram, we separate the keyframe into 5 patches with 4 corner patches and a center overlapping patch [14], and each patch is represented as a 64 dimensional vector with 8 edge direction bins and 8 edge magnitude bins based on the Sobel filter, forming a final 320 dimensional vector for a keyframe. The local features are extracted and represented as follows. Difference of Gaussian (DoG) is used to detect salient keypoints and Scale Invariant Feature Transform (SIFT) descriptors [15] are generated as features of these keypoints. Then we sampled 1M keypoints from the whole data set and cluster them into 500 clusters, constructing a vocabulary of 500 visual words. Given a keyframe, the soft-weighting scheme [16] is used to quantize the SIFT descriptors, generating the 500 dimensional BoW representation.

After extraction of the four global features and the local feature, we normalized them and adopted the early-fusion strategy to form the final representation of a keyframe by concatenating these features as early fusion has been shown better performance than late fusion approaches [5].

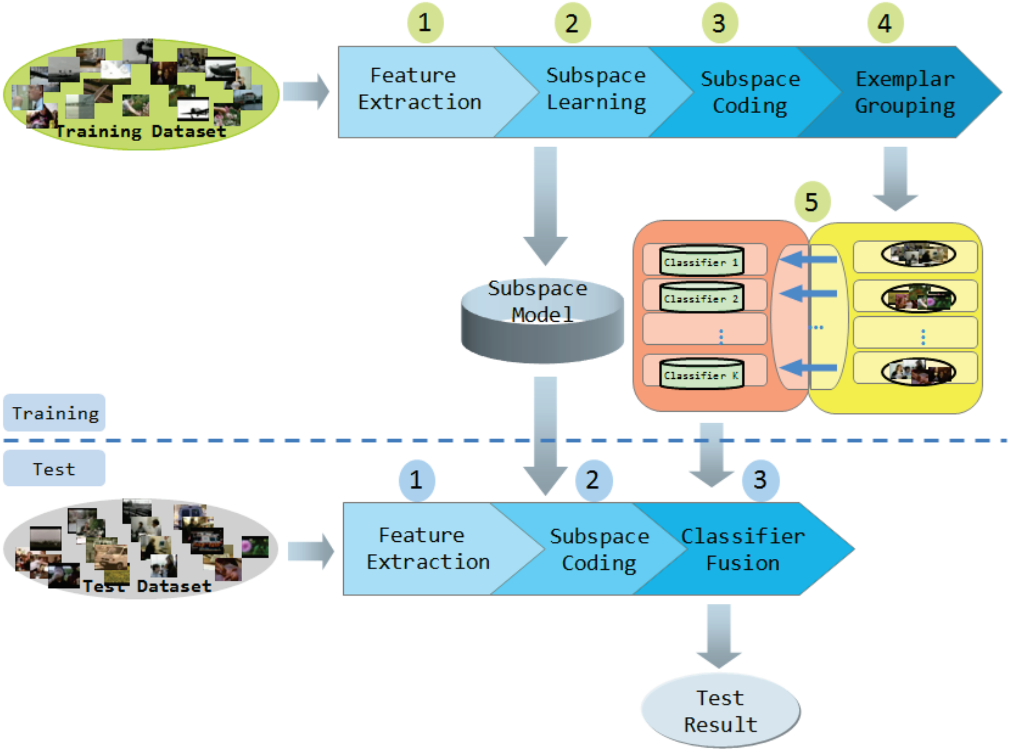

Fig.1 The framework of our video concept detection system

(2) **Subsapce Learning**. In this step we partition the whole training data into subspaces and learn the subspace model. Here, we adopt the Clara clustering method to cluster the training feature space into K subspaces. The generated K medoids' feature vectors from Clara form the subspace model. In our work, we experimentally set K to 100 and the details of subspace learning using Clara will be addressed in Section 4.

(3) **Subspace Coding**. For each training sample, we can calculate the similarities between it and the K medoids based on the cosine distance. Such similarity reflects the degree of membership of the sample to the corresponding subspace. The higher is the similarity to a certain medoid, the more likely the sample can be grouped into the subspace corresponding to that modoid. Thus, we can weight a sample's membership to a certain subspace using the calculated similarity. Then the resulted N×K weight matrix (N is the size of training sample) forms the outputs of subspace coding.

(4) **Exemplar Grouping**. With the weight matrix obtained by subspace coding, we assign a sample

(with id i) into multiple most adjacent subspaces by selecting the top L ($L<<K$) coefficients in the row i of the weight matrix. The motivations of this kind of soft-partition, *i.e.*, assigning a sample into L subspaces instead of a single subspace, are two-fold. The first one is to re-use the labeled samples of concept, since there are insufficient annotation samples for many concepts in the TRECVid evaluation. The second one is to improve classification performance by collaborating with classifier fusion module in the testing stage, working in the similar way of soft-weighting scheme used in SIFT quantization which has been shown better performance than hard-weighting. After assign all the training samples, we have the partitioned K subspaces ready for classifiers learning.

(5) **Classifier Training**. The positive and negative training samples of a concept can be grouped into multiple different subspaces. Thus for a subspace and a concept, there may exist labeled positives and negatives. We build classifiers for a concept on every subspace which has a certain amount of training samples (i.e., the number of training samples for the concept on the subspace should exceed some threshold). So we have at most K sub-classifiers for one concept, and totally $K\times M$ sub-classifiers (M is the number of concepts) for all the concepts, which construct the classifier pool of our system. We use Support Vector Machine (SVM), specifically LibSVM [9] to build up the concept sub-classifiers. In our experiments, we used the RBF kernel, and 3-fold cross validation to select the two optimal key parameters: cost parameter C and Gaussian kernel width g.

3.2 Main steps for testing stage

Given a testing keyframe and a concept, we conduct the following steps to calculate the detection score of the concept on the keyframe.

(1) **Feature Extraction.** The same features to the training stage are extracted to represent the keyframe.

(2) **Subspace Coding**. The membership weights that measure the degrees of the keyframe belonging to the L subspaces are calculated as described in the training stage.

(3) **Classifier Fusion**. The membership weights of the keyframe are ranked and the top L weights are selected. Then the most neighboring subspaces are determined, and the trained classifiers on such subspaces for the given concept are triggered. Suppose L' ($L'<L$) classifiers are triggered and their classification scores are denoted as score_i (i=1~L'). The corresponding weights of these classifiers are normalized and denotes as w_i (i=1~L'), then the final classification score for the keyframe over the given concepts is the weighted fusion of the L' results.

We conduct classifier fusion for all the testing keyframes on a given concept and then rank the fusion scores. The top 2000 results are used for the performance evaluation.

4. Subspace learning with Clara

The highlight of our concept detection system is to partition the entire training data into multiple subspaces and then train sub-classifiers on these subspaces to improve training efficiency and effectiveness. The subspace partition is actually a process of clustering thus we can adopt clustering algorithms to do this. However, the widely used clustering methods like K-Means or K-Medoids are not suitable to perform the task, because they conduct clustering or select representative medoids on the whole dataset, which is quite time-consuming and prevents them to be scaled up to large data set like TRECVid evaluation benchmark. Here we adopt Clara [10], a sampling-based clustering approach to deal with large data sets. Clara tries to find the best K medoids from fixed size sampled data subsets rather than the whole dataset, making the overall computation time and storage requirements linear to the total number of data rather than quadratic, thus is more suitable for TRECVid large corpus.

The process of Clara is listed as the following steps.

(1) For I = 1 to V (the number of sampling), iterate steps (2) to (4).

(2) Draws a subset (40+2K samples) of the dataset and applies PAM (Partitioning Around Medoids) on the subset to find the best K medoids.

(3) Calculate the distances for each sample in the whole dataset to the K medoids, and decide the most similar medoid for each sample.

(4) Calculate the total clustering cost in the last step based on the average distance between samples and their most similar medoids, and compare it with the current minimal cost. If it is smaller than the later, then the selected best K

medoids will be kept as the current best *K* medoids.

From the above steps we can seen that the complexity of each iteration is $O(K(40+2K)^2 + K(N-K))$ where N is the size of the dataset.

5. Experiments

We have submitted 4 runs totally. The description and MAP of each run are shown in the following Tab. 2, where *L* is the number of partitions that one sample can be assigned to. From the table we can see that the classification performance is influenced by the value of *L*, and the larger is *L*, the higher InfMAP is achieved. This validates the effectiveness of soft-partition scheme in **Exemplar Grouping** and **Classifier Fusion.** It should be also noted that the subspace learning with Clara is fast on the huge training dataset, taking about 9 hours on our PC (V=10). After partitioning the training space into subspaces, since the size of subspace is much smaller than that of the whole space and the subspace is of high separability, the training of sub-classifiers is very efficient and it took about 85 minutes on average to train all the sub-classifiers for one concept (for *L*=2).

Tab. 2. Description and InfMAP of our 4 SIN runs

Submitted run	InfMAP	*L* (the number for soft-partition)
L_A_NTT_DUT_1_1	0.228	8
L_A_NTT_DUT_2_2	0.219	6
L_A_NTT_DUT_3_3	0.209	4
L_A_NTT_DUT_4_4	0.203	2

6. Conclusions and Future Works

To summarize, the proposed subspace partition based classifier learning scheme is efficient and effective, providing a practical solution for large scale concept training and detection. In the future, we will investigate more advances subspace partition approach, and meanwhile, more powerful feature representation like dense SIFT will be studied in the proposed framework.

REFERENCES

[1]Yongqing Sun, Kyoko Sudo, Yukinobu Taniguchi and Masashi Morimoto, Sampling of Web Images with Dictionary Coherence for Cross-domain Concept Detection, in MMM, 2013

[2]Jinhui Tang, Shuicheng Yan, Richang Hong, Guo-Jun Qi, Tat-Seng Chua: Inferring semantic concepts from community-contributed images and noisy tags. ACM Multimedia 2009: 223-232

[3]Jianping Fan, Hangzai Luo, Ahmed K. Elmagarmid: Concept-oriented indexing of video databases: toward semantic sensitive retrieval and browsing. IEEE Transactions on Image Processing 13(7): 974-992 (2004

[4]Cees G. M. Snoek, Marcel Worring, Concept-Based Video Retrieval, Foundations and Trends in Information Retrieval archive, pp.215~322, 2008

[5]Y. Dong, S. Gao, J. Zhang, K. Tao, The France Telecom Orange Labs (Beijing) Video Semantic Indexing Systems, in Proc. TRECVID 2011 Workshop

[6]Guangda Li, Haojie Li, Zhaoyan Ming, Richang Hong, Sheng Tang, Tat-Seng Chua: Question Answering over Community-Contributed Web Videos. IEEE MultiMedia 17(4): 46-57 (2010)

[7]Meng Wang, Richang Hong, et al.: Event Driven Web Video Summarization by Tag Localization and Key-Shot Identification. IEEE Transactions on Multimedia 14(4): 975-985 (2012)

[8] A. Amir, M. Berg, and S.-F. Chang et al., “IBM research trecvid-2003 video retrieval system,” in Proc. TREC Video Retrieval Evaluation Workshop, 2003

[9] Chih C. Chang and Chih J. Lin. LIBSVM: a library for support vector machines, Online Available: http://www.csie.ntu.edu.tw/ cjlin/libsvm

[10] Kaufman, L. and Rousseeuw, P.J., Finding Groups in Data—An Introduction to Cluster Analysis. Wiley, 1990

[11]Y. Sun, G. Irie, T. Satou, A. Kojima, K. Sudo, M. Morimoto, A. Kimura, TRECVID 2011 Semantic Indexing Task By NTT-SL-ZJU, in Proc. TRECVID 2011 Workshop

[12] Jinhui Tang, Haojie Li, Guo-Jun Qi, Tat-Seng Chua: Integrated graph-based semi-supervised multiple/single instance learning framework for image annotation. ACM Multimedia 2008: 631-634

[13]Zhiping Luo, Haojie Li, Jinhui Tang, Richang Hong, Tat-Seng Chua: ViewFocus: explore places of interests on Google maps using photos with view direction filtering. ACM Multimedia 2009: 963-964

[14]M. Campbell, A. Haubold, IBM Research TRECVID-2007 Video Retrieval System. In NIST TRECVID Video Retrieval Workshop. 2007

[15]Haojie Li, Xiaohui Wang, Jinhui Tang, Chunxia Zhao, Combining global and local matching of multiple features for precise item image retrieval, Multimedia Systems, 2012

[16]Y.-G. Jiang, J. Yang, C.-W. Ngo, and A. G. Hauptmann, “Representations of keypoint-based semantic concept detection: A comprehensive study,” IEEE Transactions on Multimedia, vol. 12, pp. 42–53, 2010