A NOVEL EVENT-ORIENTED SEGMENT-OF-INTEREST DISCOVERY METHOD FOR SURVEILLANCE VIDEO

Peng Cui[†] *Li-Feng Sun* *Zhi Wang* *Shi-Qiang Yang*

Department of Computer Science, Tsinghua University, Beijing, 100084, China

ABSTRACT

During recent years, the quick development of computer techniques has witnessed the ever-increasing surveillance video data, which essentially pose great challenge on the data storage, management, analysis and even retrieval. Considering that most of the high volume of data is with no interest, we mainly investigate the problem of effectively and efficiently discovering segments-of-interest (SoI) in this paper. To do so, we propose a novel event-oriented SoI discovery method in two steps: first, we represent an event by modeling pixels' change in temporal-spatial space, aiming to unify both the inter-frames and frames-background changes; second, with the benefit of unsupervised learning, the prototype-event models could be learned from these detected events and in turn exploited to measure the interest factor of each prototype-event. The experiment results demonstrate that the proposed method precisely discriminate different events and effectively discover SoIs.

1. INTRODUCTION

The digital era is facing the fact that, millions of surveillance cameras generate video data every minute, requiring high storage capability, as well as posing great difficulty on the retrieval of segments of interest (SoI) from the tremendous video database. Actually, the universal recognition - a large proportion of those surveillance contents are semantic-repeated and with no interest - indicates that it is both necessary and interesting to critically discover the helpful contents in surveillance video data, which substantially cuts down the cost for storage, and further facilitates the retrieval.

With the purpose of SoI discovery, a rich literature of existing works discussed various methods, which could be classified into two categories: The first category advocated highlights detection [1] in entertainment video, such as sports, news and movies, mainly by exploiting either the so-called editing cues information, e.g. anchor person and play-back, or multimodality features extracted from the video and audio streams, or both. While, the other category [2] utilized the abnormal event detection method, with aid of supervised learning on segmented objects and their motion trajectory. However, those two traditional methods suffer from the poor practicality in surveillance video and large set of prior knowledge. Recently, the pixel-wise methods [3][4]for event modeling offered a promising alternative in this area. The key idea of this technique mainly explored the model of pixels' change and further detected the motion according to the comparison between the pixels and background. Although it was evidenced that those methods performed quite well in terms of motion description and human behavior analysis, only the information of pixel-changes' recency and sustainment was far from sufficient to model complex events. Actually, how to effectively and efficiently discover SoIs in surveillance video is still a challenge issue due to the lack of good event representation mechanism and measurement metric of video-segments' interests.

In this paper, we are motivated by the intuition that the more frequently an event happens, the more common and uninterested to people, and vise versa, to fight the battle of SoI discovery at the other front, that is, evaluating the saliency value of a given event and deciding the corresponding SoI with those saliency values of its contained events. More detailedly, we firstly propose a novel pixel-wised event representation mechanism in next section, in which both inter frames and frame-background pixels' changes, as well as the change caused by very slow motions are recorded. Then in Section 3, we build prototype-event models by unsupervised learning so as to eliminate the visual multiformity of events. After that, the saliency values of prototype-events can be measured according to their happening frequencies, and finally the segments containing at least one salient prototype-event would be recognized as one SoI. To verify the performance of the proposed method, we have evaluated it in PETS2004 dataset in Section 4 and SoIs are effectively discovered without the need of prior knowledge and pre-training. Our original contributions come in two folds: on one hand, the proposed event representation mechanism outperforms other pixel-wise approach in terms of more complete change information; on the other hand, we are the first, at least in our research scope, to leverage the occurrence frequency of events as saliency values, which are proved to be effective.

*This work is supported by National Natural Science Foundation of China, No. 60573167; National High-Tech Research and Development Plan of China, No. 2006AA01Z118; and National Basic Research Program of China, No. 2006CB303103.

[†]Corresponding author: cuip05@mails.tsinghua.edu.cn

2. EVENT REPRESENTATION

In this section, we mainly discuss the event representation mechanism with a novel pixel-wise method as the basis of SoI discovery. As mentioned, pixel-wise method is prior-knowledge-free, but it's sensitive to noises from sensors and lamination spectrality. With the aim of improving the robustness of the proposed system, we downsample 8*8 pixels to a pixel block, which is termed as cells for the ease of understanding. In next two subsections, we firstly propose the single cell presentation to model changes of each cell, and then extend to cell-regions which substantially and semantically represent the events.

2.1. Single Cell Presentation

As the first step to present cell's change, we use following equation to assign gray-scale values to each cell:

$$V_t^c(i,j) = \sum_{m=(i-1)S}^{i*S} \sum_{n=(j-1)S}^{j*S} V_t^p(m,n)/s^2, \qquad (1)$$

where $V_t^p(m,n)$ is the pixel value of (m,n) at time point t, and S is the edge length of a cell.

If expanding the values of a certain cell along temporal axis, a Cell Curve could be achieved, which reflects all information of a dynamic cell. The more precisely to describe the Cell Curves, the more detailedly event can be modeled, and in turn the more computation is needed and the feature is more vulnerable. In the scenario of practical video surveillance, 'what is happening' is more vital than 'what is present' [5], so we focus on modeling cells' change to depict happening events.

Given a Cell Curve, there are two representative change-relevant characteristics: change frequency and change continuance. In the following parts, two features are exploited to respectively describe these two characteristics.

2.1.1. Cell Change Frequency

Cell Change Frequency (CCF) is an important feature to describe the severeness of object motions which is the causation of cell's changes. However, unstable luminance can also induce cells' tiny change, so we predefine a threshold for filtering:

$$CVC_{t,t-1}(m,n) = \begin{cases} 1 \; if \; (V_t^c(m,n) - V_{t-1}^c(m,n) > \alpha) \\ 0 \; otherwise \end{cases} \qquad (2)$$

where CVC represents cell's valid change.

Then, given a time duration δ, the CCF is computed as the number of CVC happening in δ:

$$CCF_\delta(i,j) = \sum_{\delta} CVC_{t,t-1}(i,j) \qquad (3)$$

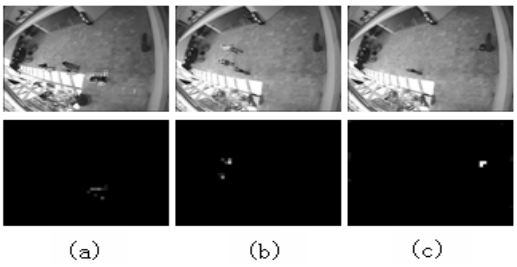

Fig. 1. Illustrated frames of CCF and CCA. The top ones are the original frames and cells marked by white squares indicate $CVCs$ there. The bottom ones are CCF and CCA images, and the brighter cells suggests changes with higher frequency or accumulation. In (a), a man is walking in a normal speed, in (b) two old man walking very slowly, and in (c) there is a left bag.

According to (3), we can estimate cell changes' frequency in most cases, as Fig. 1(a) shows. The only exception happens when an object moves very slow, in which no CVC can be detected although cells indeed change, since it only catches the changes between two consecutive frames, which, in this case, would not induce enough change as compared to the threshold. Without lowing down the threshold, a properly larger temporal distance of the two compared frames, which is determined by frame sampling period, will help to increase the change scale of cells.

So we propose a method to adaptively adjust the frame sampling period (SP) according to the motion condition. For a certain cell, under the current SP, if CVC isn't detected on one sampling time point, then the causing motion should be slower than expected, and the SP is increased, which results in higher change-scale but lower CCF; and vice versa. The upper and lower limit are predefined as 4 and 1 in our case. SP is adjusted at each sampling point, and the sampled frames of a certain cell would be:

$SF_{i,j} = \{C_0^{i,j}, C_{SP_0}^{i,j}, C_{SP_0+SP_1}^{i,j}, ...\}$

Then, given a time duration δ on original frame sequence, the CCF is computed on above SF:

$$CCF_{\delta'}(i,j) = \sum_{\delta'} CVC_{t',t'-1}(i,j), \qquad (4)$$

where

$$\delta' = \sum_{t-\delta}^{t} (SF(t) = 1). \qquad (5)$$

As shown in Fig. 1(b), this method successfully alleviates the slow motion problem. And CCF can rightly estimate the changing frequency of each cell no matter what condition is the causing motion.

However, CCF alone is somewhat ambiguous in the scenario that after the value of a cell (occupied by the left-bag in Fig. 1(c)) jumps to a new value, it is the same in CCF aspect

whether the cell turn back to the original value or maintain the new one, which motivates us to introduce a new feature below.

2.1.2. Cell Change Accumulation

The essential purpose of Cell Change Accumulation (CCA) is to record the continuance of a cell's change. Actually, CCF can reflect CCA to some extend, with the notion that larger CCF corresponds to shorter continuance of each change in average aspect. But it cannot instead CCA in long term change (defined as a change sustained for a pre-defined period of time) as mentioned above. Targeting a high efficiency, only CCA of long-term change is calculated as:

$$CCA_{\delta_n}(i,j) = \begin{cases} \delta_n & if\ CCF_{\delta_n} = 0,\ CCF_{\delta_{n-1}} \neq 0\ and \\ & V_t^c(i,j) \neq V_b^c(i,j) \\ 0 & otherwise \end{cases} \tag{6}$$

where $V_b^c(i,j)$ is the background value (by dynamic background modeling method [6]) of cell (i,j).

The result of exemplary CCA computation is illustrated in Fig. 1(c), in which the left-bag (a typical kind of long-term change) is successfully detected.

As CCA is modeled with the condition of $CCF_{\delta_n} = 0$, given a certain cell, the CCA and CCF features are incompatible. So here we use Cell Change (CC) to uniformly represent the two features:

$$CC_\delta(i,j) = CCF_\delta(i,j) + CCA_\delta(i,j). \tag{7}$$

CC not only records cell's change between frame and background by CCA, but also describe changing process more detailedly by inter-frames changes, which is modeled by CCF under whatever condition of causing motions. So CC is a more complete cell change presentation feature compared to other pixel-wised method.

2.2. Cell Region Presentation

Each single cell's changing process in a period of time is described by CC and corresponds to a cell in Cell Change Image (CCI), with a gray-scale value to depict its changing characteristics. Actually, events cannot be displayed by only one cell's change, but are co-formed by a group of cells. The salient feature of these cells is that their changes are both coherent in spatial space and temporal space to reflect the progress of objects' motions.

In our case, given a temporal window of 50 frames, cells' changes happening within the temporal window are estimated as temporal coherence, and spatial coherence is modeled by Connected Component Method. Then, individual cells are merged into cell regions by temporal-spatial relations.

To this end, all cells' changes happening in a temporal window are compressed into a CCI, and the cell regions in CCI, which reflect the the cell groups' changing characteristics in temporal window, would be corresponding to actual events, as schematically shown in Fig 2.

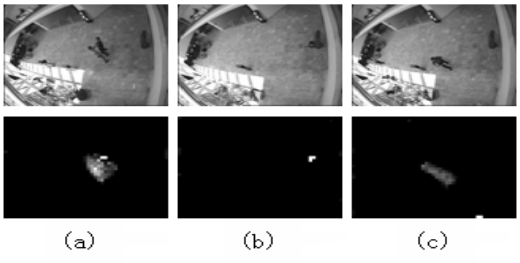

Fig. 2. CCIs of different events. The top images are original frames of different scenario and the bottom ones are respective C-CCIs. (a) is a event of two-man-fighting, (b) is a scenario of left-bag, and in (c), a man is walking.

3. SOI DISCOVERY

As known, the occurrence frequency of an event is necessary for discovering relevant SoIs. After event representation, each event happened in a temporal window has been transformed into a cell region in a CCI. But cell regions corresponding to the same type of event are not exactly the same in visual cues due to the structure alteration of events Therefore, in this section, we propose a prototype-event model to incorporate the variances of multiformed cell regions, and accordingly calculate the saliency value of each prototype-event with its occurrence frequency, through which the SoIs are determined.

With the reasonable assumption that in CCI, cell regions with similar visual features corresponds to the same type of events, we applied unsupervised learning on cell regions, with the final cluster centroid as prototype-event's features, the cluster radius as the variance of this type of event, and the cluster size as its happening frequency.

There are two main challenges in the modeling process: 1) what visual features should be selected to represent a cell region? 2) given a set of cell regions, how to determine the cluster number?

Given a CCI, we can extract cell regions' sizes, shapes, locations, orientations and luminance distributions, among which the luminance distributions is the most stable and intrinsic feature as it possesses immutability in distance, size and rotation. Therefore, as to the former challenge, cell region is represented by its luminance distribution, which is described by region luminance histogram (RLH) in our case. However, the feature dimensions of RLH are not equally discriminative, for efficiency and filtering noises induced by non-discriminative features, we use Principle Component Analysis (PCA) to project the RLH onto a lower dimensional subspace.

Clips	SN	EN(GT)	EN	$Pre.$
Browse1	20	3	3	95%
Fight_OneManDown	19	4	4	100%
Fight_RunAway2	11	4	3	91%
LeftBag	28	4	5	93%
Rest_WiggleOnFloor	25	3	3	100%
Overall	188			96%

Table 1. Results of event presentation

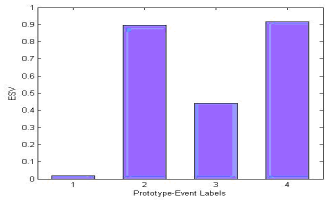

Fig. 3. Saliency values of prototype-events.

After acquiring RLHs, we wish to cluster them into clusters, using the well-known K-means. In the face of challenge 2, a method similar to [7] is adopted, in which the intra-cluster information is used to evaluate the tightness of clusters and inter-cluster information is used to characterize the overall quality of a clustering.

Hereto, several clusters form from the given set of cell regions, with each cluster corresponding to a prototype-event. So the occurrence frequency of a prototype-event could be calculated by its corresponding cluster size. As mentioned, salient events mostly mark up themselves with comparatively low occurrence frequency, so the Event Saliency Value (ESV) is computed as:

$$ESV_i = \exp(-nf_i / \sum_{j=1}^{n} f_j), \tag{8}$$

where f_i is the appearing frequency of i^{th} prototype-event and n is the total number of prototype-events.

To this end, with the proposed event representation and modeling method, we can discover salient events according to ESV, and segments consist of at least one salient event are recognized as SoIs. In next section, we will evaluate the proposed method with extensive experiments.

4. EXPERIMENT

We conduct experiments on PETS2004 dataset with resolution of 384*288 pixels. The data set consists of 28 video sequences with in total 26419 frames to describe 5 scenarios, that is, Walking, Browsing, Collapse, Leaving objects, Meeting and Fighting appended with ground truth. In order to testify the event presentation mechanism, we applied the proposed method on each single clip to differentiate different events, with Precision as criterion:

$$Pre. = \frac{numberofrightclusteredcellregions}{totalnumberofcellregions} \tag{9}$$

Some typical results and the overall precision are given in Table 1,where SN is the number of segments in each video clip; EN is the number of prototype-events in experiment and $EN(GT)$ is the Ground Truth of EN. Observe that, the proposed event representation mechanism can discriminate different events in high precision.

Then we evaluate the SoI discovery method by simulating a real surveillance scenario. We put PETS clips together as a data pool, and then make 10 copies of Walk to simulate the severe asymmetry of dominant events and interest events. The saliency value of each prototype-event is shown in Fig. 3.

Totally, 227 segments are clustered into 4 proto-type events. As observed, the prototype-event 1 corresponds to the dominant event type - walking, and the other three correspond respectively to Fighting, leftbag & people_lying, and people_resting & Browsing. In the data pool, segments only containing prototype-event 1 occupies more than 63%, and the figure should be much higher in reality. And the remained segments containing events with higher $ESV s$ are indeed interest to users apparently.

5. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a novel method for discovering segment-of-interest in surveillance video by firstly represent events in a pixel-wised method and then measure the saliency value of each event according to its occurrence frequency. The experiment results have demonstrated its effectiveness and precision. But the result is preliminary and the future work will focus on inducing context information and exploiting more cell region features to eliminate the representational ambiguity problems.

6. REFERENCES

[1] TS Chua, SY Neo, HK Goh, etc., TRECVID 2005 by NUS PRIS, TRECVID 2005 Workshop, 2005.

[2] Li, X., Porikli, F.M., A Hidden Markov Model Framework for Traffic Event Detection Using Video Features, ICIP, Vol. 5, pp. 2901-2904, 2004.

[3] A.Bobic, J.Davis, The recognition of human movement using temporal templates, IEEE Trans. PAMI, 23(3):257-267, 2001.

[4] T Xiang, and S Gong, Beyond Tracking: Modelling Activity and Understanding Behaviour, IJCV, Vol 67, pp. 21-51, 2006.

[5] AP Graves, S Gong, Spotting scene change for indexing surveillance video, BMVC, pp. 469-478, 2003.

[6] C Stauffer, WEL Grimson, Learning patterns of activity using real-time tracking, IEEE Trans. PAMI, Vol 22, pp. 747-757, 2000.

[7] JM Odobez, D. Gatica-Perez, M. Guillemot, Video Shot Clustering Using Spectral Methods, CBMI, 2003.