Surveillance Video Summarization Based on Object Motion Pattern Analysis

[1]Zaur Fataliyev, [2]Yadigar Imamverdiyev, [3]Rasim Alguliyev, [4]Orkhan Karimzada, [5]Hanseok KO
[1,2,3]ANAS Institute of Information Technology, Baku, Azerbaijan
[1,5]School of Electrical Engineering, Korea University, Seoul, Korea
[4]Department of Electrical Engineering, KAIST, Daejeon, Korea
[1]zaur@korea.ac.kr, [2]yadigar@lan.ab.az, [3]director@iit.ab.az, [4]orxan_karimzade@kaist.ac.kr, [5]hsko@korea.ac.kr

***Abstract*— This paper proposes a novel fusion method for summarization of surveillance videos based on motion pattern analysis of each detected object. The method allows to summarize long videos into a single index frame. Generated index frame is shortest possible summary of given video which also serves as a bookmark for a long video. Experimental results demonstrate its effectiveness.**

Keywords - pattern analysis, video summarization, Image Generation

I. INTRODUCTION

As inexpensive surveillance cameras become ubiquitous and find ways into many corners of applications. Such cameras record videos all day long and one organization may have dozens of surveillance cameras. Due to distraction, human supervision of dozens of cameras may lead to omitting important events and completely ignoring of such events. It becomes a daunting task to review and extract useful information from a vast amount of the video footages being generated. An intelligent summary of these videos by an algorithm may be a helpful alternative for quickly reviewing and archiving them. Research in video summarization has had several impressive outcomes [8, 9], and commercial companies has been built using those outcomes [8]. Without any doubt, more sophisticated methods different approaches are needed to meet requirements in this area. For such a purpose, it is proposed here to develop an algorithm for generating an Index Frame (IF) that captures key events in surveillance video by extracting Key Positions (KP) of pertinent objects in a scene with its timestamp. KPs is part of motion that is most important part of object motion such as falling person, fighting scene, turning point and others. Object may have one or more of such KPs. In this paper, multiple KP is extracted and for minimum summary, only one KP per object is used to generate IF. Such IF allows us to quickly review long video in a single frame and it can be used as a bookmark of important events in original videos.

Among Key Frame Extraction (KFE) algorithms, methods based on global motion energy have been shown to be effective and computationally efficient [2], [5]. However, KFE algorithms proposed by [1], [2], are not well suited for this task due to their inability of analyzing individual objects in a scene. This issue may be alleviated by methods proposed by [3] by considering objects extracted from the background and generating an IF. Since the method included all objects from the background subtraction and no further analysis was conducted for individual objects, its summary doesn't depict any particularly important information of object activities.

To extract more pertinent, concise, and useful summary, the proposed algorithm analyzes motion patterns of an individual spotted object in a scene. Major features of object motions i.e. linear acceleration, angular motion and perceived object size is used for decision-making. Local decisions based on its linear acceleration, angular motion and perceived size are fused to extract one or multiple KPs of pertinent objects. Those Key Positions are placed in IF with its timestamp that gives overview of long video in a single frame.

II. PROPOSED METHOD

As shown in Fig. 1, the proposed system is comprised of 3 steps. This paper mainly focuses on KP extraction and IF generation. For object extraction, adaptive background subtraction [4] with 3 Gaussian mixtures has been used. Morphological opening and closing has been used for noise reduction [7].

A. Key Position Extraction

Proposed method extracts 3 features from segmented blobs: linear acceleration, angular motion and perceived object size. It's followed by making local decisions based on each feature. Local decisions are fused for making final global decision.

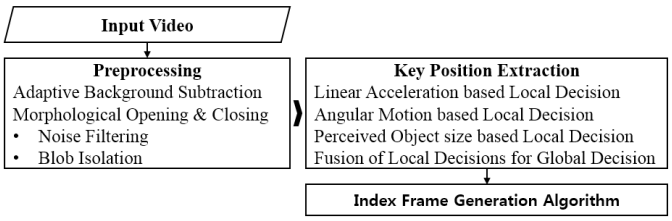

Figure 1: Proposed Index Image Generation System

1) Linear Acceleration based Local Decision

Acceleration characterizes object's kinetic energy and alteration in acceleration generally characterizes change in kinetic energy, thus embodies the key actions of object movement. Linear acceleration has been converted to probability of being KP at the point of acceleration alteration under a hypothesis that the ith object in kth frame is KP with

probability of $P_i(k_a)$. Due to segmentation algorithm and noise, extracted feature values may fluctuate. Therefore, moving average filter has been applied for n frames ($n = 5$ or 1/6 seconds) prior to making local decisions based on acceleration, motion angle as well as perceived object size. Both, acceleration ($a_i(t) > 0$) and deceleration ($a_i(t) < 0$) represent imperative actions of object at $(x_i(t), y_i(t))$, as first corresponds to increase in kinetic energy level and second corresponds to decrease in kinetic energy level, respectively. Thus, both are equally important features for detection of KP.

$$P_i(k_a) = \begin{cases} \left|\frac{a_i(t)}{\max(a_i)}\right|, if\ a_i(t) > 0 \\ \left|\frac{a_i(t)}{\min(a_i)}\right|, if\ a_i(t) < 0 \\ 0, \quad otherwise \end{cases} \text{ where } a_i(t) = \frac{d^2x_i(t)}{dt^2} + \frac{d^2y_i(t)}{dt^2} \quad (1)$$

2) Angular Motion based Local Decision

Second key action of object which characterizes its KP is represented by change in its motion direction. Change in motion direction characterizes features such as movement direction change, turning point, abnormality in motion, and others. Change of angle at $(x_i(t), y_i(t))$ relative to motion direction at $(x_i(t-1), y_i(t-1))$ is given by (2).

$$\Delta\varphi_i = \tan^{-1}\frac{dy_i(t)}{dx_i(t)} - \tan^{-1}\frac{dy_i(t-1)}{dx_i(t-1)} \quad (2)$$

Change in motion direction converted into probability by (3). Two cases, $\Delta\varphi_i > 0$ and $\Delta\varphi_i < 0$, are essential to describe motion to the left and motion to the right, respectively.

$$P_i(k_\varphi) = \begin{cases} \left|\frac{\Delta\varphi_i(t)}{\max(\Delta\varphi_i)}\right|, if\ \Delta\varphi_i(t) > 0 \\ \left|\frac{\Delta\varphi_i(t)}{\min(\Delta\varphi_i)}\right|, if\ \Delta\varphi_i(t) < 0 \\ 0, \ otherwise \end{cases} \quad (3)$$

3) Perceived Object Size based Local Decision

Assuming object of interest is within Field of View (FOV) at Z distance and perceived width of $w(t)$ and height of $h(t)$. Perspective projection of line L at distance of Z from the camera of f focal length on XY plane is mapped as $l = fL/Z$. That's to say, perceived object size is proportional to camera-to-object distance. Corresponding local decision is calculated as given in (4), where $\max(w_i)$ and $\max(h_i)$ is width and height of object respectively at closest distance to camera.

$$P_i(k_S) = \frac{1}{2}\left(\frac{w_i(t)}{\max(w_i)} + \frac{h_i(t)}{\max(h_i)}\right) \quad (4)$$

Height and width are weighed equally, which averages out possible error in perception that may occur due to detection algorithm

B. Fusion and Key Position Selection

Probability of particular point being KP is given by (1), (2) and (3), based on three different features. Overall probability of particular point of i^{th} object at time t being KP is $P_i(k_a, k_\varphi, k_S)$. Strictly speaking, local decisions are not independent, but dependence between local decisions don't imply particular advantage, since every local decision extracts feature-specific position. Thus, assuming all local decisions are independent, global decision is defined by (5).

$$P_i(k_a, k_\varphi, k_S) = P_i(k_a)P_i(k_\varphi)P_i(k_S) \quad (5)$$

Fig. 2 is a result of video in which frames between 100 and 150 correspond to an activity of a person turning back from walking. While probability decision based on the perceived object size doesn't change significantly, it decreases for linear acceleration and fluctuates for angular motion. The result shows that a position in frame 110 ($P = 0.5677$) corresponds to the best KP of the object in the scene. In case of an overlap with other object KP coordinates, alternative second or third best KPs such as frame 74 ($P = 0.4173$) or frame 128 ($P = 0.36988$) is selected to avoid the issue (Fig. 3).

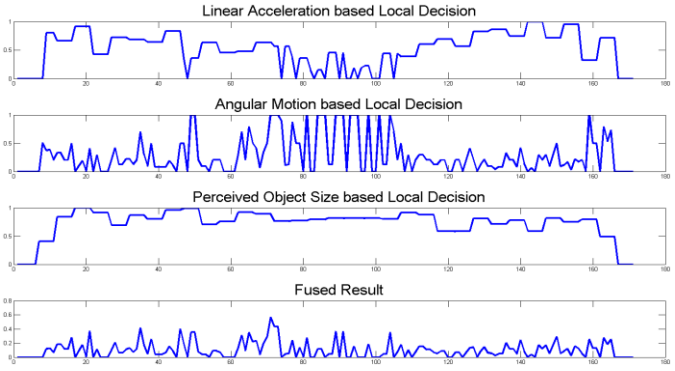

Figure 2: Sample Fused Result for one detected object

Fig. 3 shows two KP for same person. First (Frame #110) is turning point and waving of object, while second KP (Frame #74) is position of scene entrance and maximum acceleration. Other KPs are less important than those two KPs.

Figure 3: Frame number 110 and Frame number 74

In order to handle overlapping KPs, algorithm for IF generation has been developed. Alternative second or third KP with highest score is used to handle such issues. Flow of algorithm is given in Fig. 4.

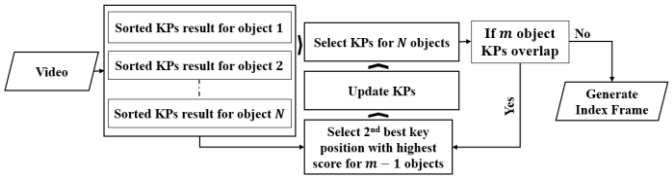

Figure 4: Algorithm for generating the Index Frame

Performance of the IF generation algorithm was evaluated using CAVIAR dataset [6] and sequence is depicted in Fig. 4. Additionally, two sample IFs are shown in Fig. 5. As it is apparent, the index image gives a concise overview of the entire video. As it can be verified from the CAVIAR dataset, the proposed algorithm correctly identifies the KPs. For instance, direction of the individual movement in frame (1325, 517, 250), action (110) and position of scene entrance (250, 1325, 517) are evident by looking at the index image. The IF can also serve as a bookmark for the original videos.

III. Experimental Results

Subjective measurement was conducted using the CAVIAR dataset. KPs such as fighting scenes, browsing, or turning point in walking activity is labeled as accurate, while minor deviation from such scenes is labeled as acceptable. KPs that does not depict key action is labeled as inaccurate. As shown in Table 1, the result demonstrates robust performance of cumulative 96% accuracy.

Figure 5: Examples of Generated Index Frames (2 videos)

Tested CAVIAR dataset compised of 6 different scenarios, i.e., walking, browsing, resting, fighting, leaving. Those senarios cover almost all possible events that may happen in area under surveillance. Moreover, this dataset is publicly avaiable in [6] which maybe used for verification.

TABLE 1. Subjective Evaluation Result

CAVIAR Dataset: Clips from IRNIA				
Scenario	Accurate	Acceptable	Inaccurate	Total
Walking	87.5%(7)	12.5%(1)	0%(0)	8
Browsing	80%(8)	10%(1)	10%(1)	10
Resting	80%(8)	20%(2)	0%(0)	10
Fighting	75%(9)	16.7%(2)	8.3%(1)	12
Leaving	90%(9)	10%(1)	0%(0)	10
Average	**82%(41)**	**14%(7)**	**4%(2)**	**50**

IV. Conclusion

Large amount of data is generated by widespread usage of surveillance cameras is difficult to analyze and search for particular events. Besides, such cameras are being monitored by humans, and due to distraction, such supervision may lead to omission of important events. In order to make analysis easier and reliable, it is proposed here to develop an algorithm for generating an Index Frame that captures key events in surveillance video by extracting Key Positions of pertinent objects in a scene with its timestamp. With that in mind, a novel method of video summarization was proposed to provide essential information about the video in a single frame, KPs of spotted objects were extracted and positioned in the corresponding IF with timestamp of a such occurrence. KPs were extracted by fusion of local decisions based on linear acceleration, angular motion and perceived object size of spotted objects. The results show robust performance, implying that the proposed method is suitable for

summarizing long video footages into several IFs for providing quick review and bookmark to original videos.

REFERENCE

[1] Tianming Liu, Hong-Jiang Zhang, and Feihu Qi, “A Novel Video Key-Frame-Extraction Algorithm Based on Perceived Motion Energy Model”, IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, no 10, pp. 1006 – 1013, Oct. 2003

[2] Weiming Hu, IEEE, Nianhua Xie, Li Li, Xianglin Zeng, and Stephen Maybank, “A Survey on Visual Content-Based Video Indexing and Retrieval”, IEEE Transactions on Systems, Man, and Cybernetics, vol. 41, no 6, pp. 797 – 819, Nov. 2011

[3] Chris Pal and Nebojsa Jolic “Interactive Montages of Sprites for Indexing and Summarizing Security Video” in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 20-25 June, 2005

[4] Stauffer, C. and Grimson, W.E.L, “Adaptive Background Mixture Models for Real-Time Tracking”, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 2246 – 252, 06 August 1999

[5] W. S. Chau, O. C. Au, and T. S. Chong, “Key frame selection by macroblock type and motion vector analysis”, IEEE International Conference on Multimedia and Expo, vol. 1, pp. 575–578, June 2004

[6] Robert Fisher (2011). CAVIAR Test Case Scenarios. Retrieved from http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/

[7] Robert M. Haralick, Stanley R. Sternberg, Xinhua Zhuang“Image Analysis Using Mathematical Morphology”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 9, no 4, pp. 532 – 550, July 1987

[8] Yael Pritch, Alex Rav-Acha, Shmuel Peleg, “Nonchronological Video Synopsis and Indexing”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no 11, pp. 1971 – 1984, Nov. 2008

[9] Yu-Fei Ma, Xian-Sheng Hua, Lie Lu, Hong-Jiang Zhang, “A Generic Framework of User Attention Model and Its Application in Video Summarization”, IEEE Transactions on Multimedia, vol. 7, no 5, pp. 907 – 919, Oct. 2005