International Journal of Engineering & Technology, 7 (4.39) (2018) 128-132

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



A Novel Approach for Active Event Based Video Summarization Using Foreground Analysis

Satyabrata Maity[1]*, Atanu Maji[2], Krishanu Maity[3], Sourav Biswas[4] , Jogendra Garain[5]

[1,4]Department of Computer Science & Engineering , Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, India.
[2] Centre for Applied Mathematics & Computing, Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, India,
[3] Department of Computer Science & Information Technology, Siksha 'O' Anusandhan Deemed to be University , Bhubaneswar, India.
[5]Department of Computer Science & Engineering, Siksha 'O' Anusandhan Deemed to be University, Bhubneswar, India.
**Corresponding author E-mail: satyabratamaity@soa.ac.in*

Abstract

Rapid growth of no-informative-videos is one of the major concerns of video analytics in recent time. The field like outdoor and indoor surveillance, home, office and shopping mall monitoring produces gigantic volume of no-informative-videos. A novel active event based video summarization is proposed in this research work to make the video analytics more applicable in those fields. Use of adaptive techniques for noise reduction, background modeling, foreground extraction and analysis make the proposed approach more robust towards active event based summarization and indexing. The results on publicly available datasets and a comparative study based on the objectives of the proposed approach with the same of related tesearch works justify the effectiveness of the proposed approach.

Keywords: *Adaptive thresholding; Background Modeling; Chronological ordering; Event detection; Foreground extraction; Spatiotemporal Redundancy.*

1. Introduction

The use of video based technology has been growing exponentially since last two decades due to the cost effective availability of video capturing and storing devices. Hence, video analysis becomes one of the major areas of research for preparing the roadway to make this prolific technology more efficient and convenient to use. Many researchers agree that the impact of the visual information is the key point towards the success of this technology. The fields like outdoor and indoor surveillance, home, office and shopping mall monitoring produce gigantic volume of *no-informative-video*. The video, which does not produce any information, is termed as *no-informative-video*. If there is no activity in front of active camera, this type of *no-informative-videos* is generated. This is the main motivation behind the work.

The videos are basically collection of interrelated image frames, which represent a sequence of events, which occur in chronological order either scripted or random. The videos, which follow some script to perform, are the scripted videos like movie, news etc. On the other hand, the videos like real world, sports, and surveillance videos are random, which never follow any script. Irrespective of types, videos are gigantic in volume containing higher amount of redundant information. Video technology makes us enable to visualize any past events, which were captured. The quality depends on the number of frames captured in a unit amount of time and the number of pixels are used to construct a frame. These two facts lead to temporal and spatial redundancy of information in a video. The temporal redundancy increases the number of frames to enrich the visual smoothness of experiencing any event and their transitions. The spatial redundancy increases the clarity of objects by increasing the number of pixel to represent a frame. On the other hand, this redundancy in video makes the world difficult for analysing the same. Finding the actual information from a video is like netting in a water-bank to take out the fishes. Increasing the redundancy is like an increment of the water in that bank.

Actual information extraction is required for automatic video analysis; otherwise one should watch the full video carefully. The human accuracy is inversely proportional to the time line, as the time passes away, the accuracy decreases. It would be more effective if we have a summary of the whole video with the chronological ordering of events, then one can easily scroll the actual happenings in a quick time like an out sketch. The main idea behind the proposed approach is described as a framework in Figure. 1. There would be some correlation among the consecutive frames, which represent same event, and it varies for the different events. This phenomenon assists us to find the event boundary. Figure. 1 shows an event detector, which takes the raw frames as input and produces the events as a result. On the other hand, each event is represented with the key frames. The similar group of frames are represented with same colour in Figure. 1.

Any frame in video is formed with two types of information like foreground and background. On the other hand, a video content is divided into two key areas, specifically static area and active/dynamic area. The information contains in the video is defined by the dynamic areas; otherwise it is idle or static. Hence, to analyse the video information, it is required to extract the dynamic or active areas from the video. The main objective of the proposed approach is to extract active events and index them in the form of a summary in chronological order, which can be used for further analysis. The active events refer to those events, where some movement in foreground is taken place, i.e. one or more moving objects can be present in foreground. Moving object extraction techniques are used to detect the activeness of a video.

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

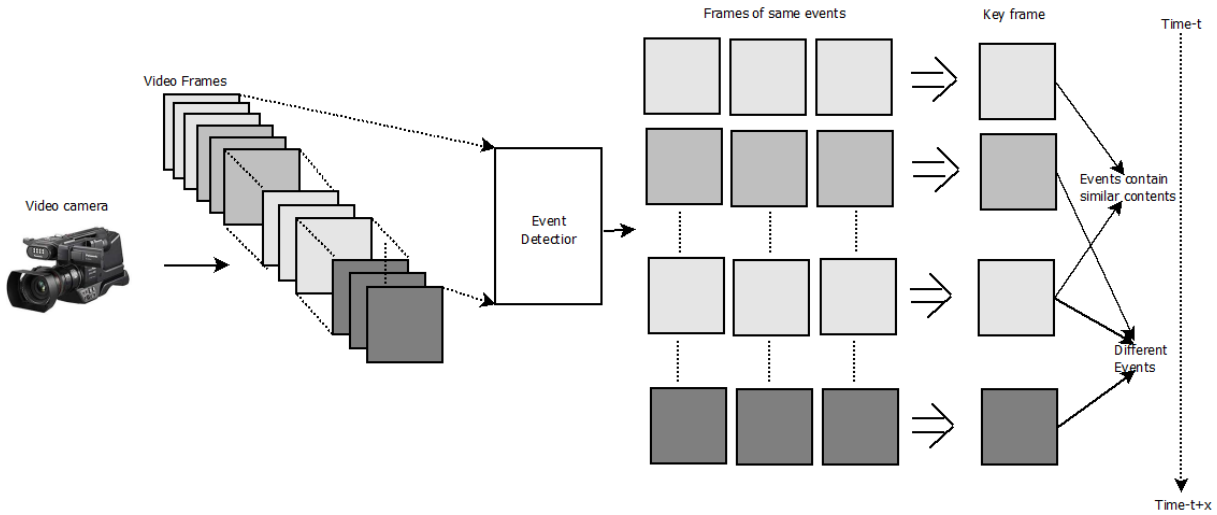


Figure 1: Framework of the proposed approach

The novelty of this work are stated below:

Intelligent summary generation : the summary is generated in the active areas, which can provide the brief information of whole video in quick time. It can be well suited in the applications like surveillance in office floor or shopping mall, elderly of kids monitoring, outdoor surveillance, virtual class room etc. Most of the times of the day, videos are found idle in such cases; human monitoring would be very difficult for those things. But, human can monitor after getting the alarm from the system when it tracks some activities. Thus, the system like this can assist to make the whole procedure more efficient.

Noise smoothing to get more effective results: The system uses noise reduction techniques to reduce the environmental effect like change in illumination of light, bad quality of acquisition devices etc. The illumination normalization technique reduces the effect of lighting condition. On the other hand, anisotropy based smoothing supresses the local variability and emphasizes the edge information.

Adaptive thresholding technique: All the threshold use in proposed approach are adaptive, which increases the accuracy.

False foreground removal and event detection: Actual fore-grounds are informative, but false foregrounds are affecting the originality of information. Hence, the removal of false foreground escalates the probability of effective results. De- spite, event detection based on background information and foreground changing ratio provide efficiency in summary generation.

2. Related Work

This section includes some of the milestones of related research work to justify the usability of the current work. *Damnjanovic et al.* [1] proposed a summarization technique for surveillance video based on important event clustering. The important events are defined with the amount of energy transfer in consecutive frames. Moreover, they have created two types of summaries namely static summary based on key frames, and dynamic summary based on short video segments.

Rameswar et al. [2] proposed a video summarization methods for multi-view video. Their aim to fully exploit the intra-view and inter-view correlations in multi-view videos. They fulfil this criteria by introducing a unsupervised framework. The objective of this framework to capture the multi-view correlations and to select representatives shots for the summary. *Ali javed et al.* [3] modelled a theft prevention alarm based on video summarization algorithm. They proposed a theft prevention algorithm to model the theft prevention alarm. Their proposed model is able to captures each frame from the video and processes the frame. The model only retains the frame of interest otherwise it discards the frame. Video summarization model by selecting key-based frame has been proposed by *Antti et al.[4]*. The key-frame has been selected by two existing effective approach: visual content and motion content.

Authors in [5] have proposed a feature based approach for summarizing a video to extract interesting information. They have included the features like representativeness, uniformity, static attention, temporal attention etc. *Authors in [6]* have proposed a video summarization technique to make a synopsis of event of interest. They have introduced the cumulative moving average (CMA) and the preceding segment average (PSA) statistical metric as features to make their technique more effective towards gradual and sudden changes of moving objects.

3. Proposed Approach

The proposed approach is broadly divided into two sub steps, namely event detection and event based summarization. Once the first event is detected, these two steps can run in parallel.

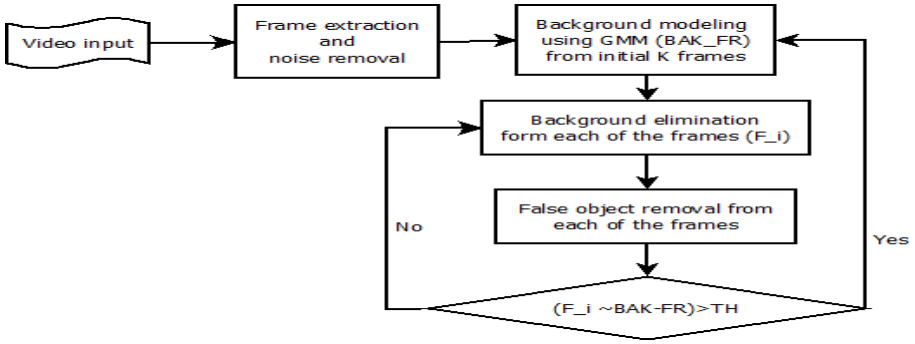


Figure 2: Work flow diagram of event detection

Event detection includes, video frame extraction, enhancement of image frame for better processing, background modelling in initial frames, background elimination and foreground detection, and difference comparison and event boundary detection. The event based summarization includes key frame extraction, event feature extraction and event based summarization and chronological ordering. All the steps are discussed in the following subsections.
The frame work of event detection is shown in Figure 2. Video input can be taken either in offline or online. After that, noise removal techniques are used in each of the frames to achieve more effective results. Background modelling using Gaussian Mixture Model (GMM) is applied on initial frames to estimate the background information of that event. Foreground extraction is done by eliminating the back- ground information from each of the frames. When the difference between the current frame and the background frame exceeds certain threshold and that continues to some extends, that point is considered as event boundary. The similar approaches are followed in the next sequence of frames starting with background modelling, and the detected event is analysed for key frame extraction and event based summarization.

3.1. Video Frame Extraction

The acquisition step of the proposed approach provides the input to the proposed system. In this step, the frames are extracted for use in the next sequence steps.

3.2. Enhancement of Image Frames for Better Processing

The moving regions are nothing but the difference between any two pair of consecutive frames. In other words, these moving regions are movement of foreground object in a particular event. Since, these differences can be affected by the noise or other external conditions present at the time of video acquisition. The contribution of this step is to reduce the amount of noise present in video. Two types of noise reduction techniques are used in the proposed approach as discussed below.

3.2.1 Illumination Rormalization

The lighting condition can vary over the time and can provide the different intensity values for same object in different time of the day. The normalization of illumination condition can reduce the effect of lighting situation on same object. The normalization is done using Eq. 1

$$P_{out} = (P_{in} - ll) \times \frac{(ul - ll)}{(hv - lv)} + ll \tag{1}$$

3.2.2. Anisotropy Based Smoothing

Traditional smoothing techniques use to smooth all regions of an image, which can smooth the vital inter regional or edge areas. The anisotropy based smoothing is use to smooth the intra-regional areas, whereas it cannot do anything in the intra-regional of edge areas. Sometimes, it is called edge stopping smoothing. The anisotropic filter, proposed by *Perona et al.* [7], provides the facility of smoothing of the low-frequency regions and emphasizes the edge regions using Eq. 2.

$$\frac{\partial I(x,y,t)}{\partial x} = div(\left[g(\left\|\nabla \perp\right\|)\nabla \perp\right] + a \tag{2}$$

Where, $\|\nabla \perp\|$ is the gradient magnitude, and $g(\|\nabla \perp\|)$ is an "edge-stopping" function. This function is chosen to satisfy $g(x) \rightarrow 0$ when $x \rightarrow \infty$ that the diffusion is "stopped" across edges.

3.3. Background Modelling in Initial Frames of an Event

The background modelling is one of the crucial steps for getting the efficient results of the proposed technique. The background refers to the extreme amount of redundant information throughout the frame sequence representing the event. Modelling and eliminating background remove the redundancy from frame sequence and help to extract the actual information. The proposed approach has used *Gaussian mixture model (GMM)* [8] to model the background.
To frame the randomness of the occurrence of background and foreground information in the form of pixels, they model the changing values of particular pixel over a period of time as a mixture of Gaussians. The background colour is determined based on the variance and the persistence of each of the mixture of Gaussians. The Gaussian distribution helps to determine the background and foreground. If any pixel value is not in the distribution of background, is treated as the foreground pixel. This technique was introduced in 1999 by *Stuffer et. al.*, and till date it is one of the most robust technique for modelling the background because of its adaptability over lighting changes, ambient motion of scene or background elements, slow-moving objects etc.
To model the background using *GMM*, the time series of a pixel of a particular location is required. The time series values are extracted using EQ. 3, where *i* is varying from *1* to *t* and giving the time series value of the position *(x0, y0)* for t different frames. This mechanism is applied in every pixel position to extract the corresponding time series value.

$$\{X_1, X_2, ..., X_{t-1}, X_t\} = IMG(x_0, y_0, i) : 1 \le i \le t \tag{3}$$

Let, the recent history of each pixel, (x1, x2,..., xt), which are considered model over *K* Gaussian, the probability of observing current pixel value is done using EQ. 4.

$$P(x_t) = \sum_{i=1}^{k} \omega_{i,t} \times \eta(x_t, \mu_{i,t}, W_{i.t}) \tag{4}$$

All the pixels are modelled using the same technique to provide the estimated background.

3.4. Background Elimination and Foreground Extraction

This step includes three sub-steps like background elimination, false foreground removal, and foreground labelling.

3.4.1. Background Elimination

Initial foreground FRGl1 is computed by taking the difference between current frame FRt and the background frame BAK with respect to an environmental resolution threshold THv as shown in EQ. 5, where, $1 \le i \le m$, and $1 \le j \le n$. THv is used to resolute the negligible change caused by environment. The value THv is the average of the deviation of histogram of initial k divided by 256 as given in EQ. 6. This threshold is adaptive with respect to the variability of the initial frames of the video since the histogram provides the overall distribution of intensities of a frame, and the threshold is recomputed with the change of each event.

$$FRG_{l1} = FR_i(i, j) - BAK(i, j) \ge TH_v \tag{5}$$

$$TH_v = \frac{\sum_{t=1}^{k-1}\sum_{i=1}^{256}\left|Hist_t(i) - Hist_{t+1}(i)\right|}{k \times 256} \quad (6)$$

3.4.2. False Foreground Removal

Foregrounds of frames are sometimes mix up with some false objects, which are caused due to some unwanted noise. Most of the times the size of the false objects are respectively small with compare to actual objects. Hence, the size of objects below a certain threshold, TH_{fr} , are cut out as false objects. This threshold can be set with some known predefined value, which represent the minimum size of the object. Here, TH_{fr} is taken as the 1% of the frame size. Thus, the frame after foreground object removal would be FRG_{l2} as computed using EQ. 7.

$$FRG_{l2} = \{CC \mid \forall CC > TH_{fr}\} \quad (7)$$

3.5. Difference Comparison and Activity Boundary Detection

An array articulated as PER keeps the track of the size of foregrounds in consecutive frame sequence to maintain the persistence, as shown in EQ. 8, where S_i is the size of all foreground objects of frame (i). A change over 30% with compare to the previous frame is marked, and if that change continues in next k/2 number of frames, then the point is considered as the changing point or boundary of an event.

The consistency of the change in k consecutive frames ensures the change in scene. k is generally taken as the frame rate of the video. Hence, k and changing limit are adaptive with respect to frame rate and frame size as given in EQ. 9. The false change can occur due to the change in environmental condition like sudden change in illumination of light, the unwanted movement of acquisition devices, presence of noise etc.

$$PER_i = s_i \quad (8)$$

$$CHG_{pt} = i,\ IF\ \frac{\left|PER_i - PER_{i+1}\right|}{PER_i} \geq 0.3 \quad (9)$$

3.6. Key Frame Extraction and Event Summarization

The term key frame suggests that the frame(s) contain key information of that group. Key frame extraction from each of the group is essential for pursuing the summary of the events. Since, the background remains unchanged throughout the events; the foreground information among the frames in same event can be changed during the event. Hence, the frame(s) contains optimal information with respect to the size of foreground are considered as the key frames. This value is estimated from the array PER using EQ. 10. KF_k is the key frame of k^{th} event. i is ranging from j to $j+t_k$, where j is the starting frame index of the k^{th} event, which comprises t_k number of frames.

$$KF_k = MAX(PER(i)), \cdots where\ j \leq i \leq j + t_k \quad (10)$$

3.7. Similar Activity Grouping and Chronological Ordering

The summary of a video includes the key frame(s), which will be supported with some closely related frames to maintain the visual persistence, from all active events. Activity grouping is done as soon as one key frame is extracted from a new event. Suppose, the first event is extracted with the key frame and all details, but there is no previous events. Hence, it will be taken as first group. When the second event is detected, it is compared with the first group. If these two are similar, the 2nd event will be put in the 1st group, a new group will be created otherwise. In case of i^{th} event, if there are *k, (k ≤ i)* number of event groups previously, then the new event is compare with the all *k* groups with a prescribed threshold. The comparative results may create a new group, or it can be put in some previously defined group.

4. Results & Analysis

The proposed technique is applied on various available videos to verify the performance of the same. The datasets include several environmental conditions like indoor and outdoor videos, human centric and non-human centric videos, noisy conditions, background object motion etc. We have set some testing points while verifying the proposed approach, like actual foreground detection noise removal, background modelling, and true foreground estimation, key frame and event detection.

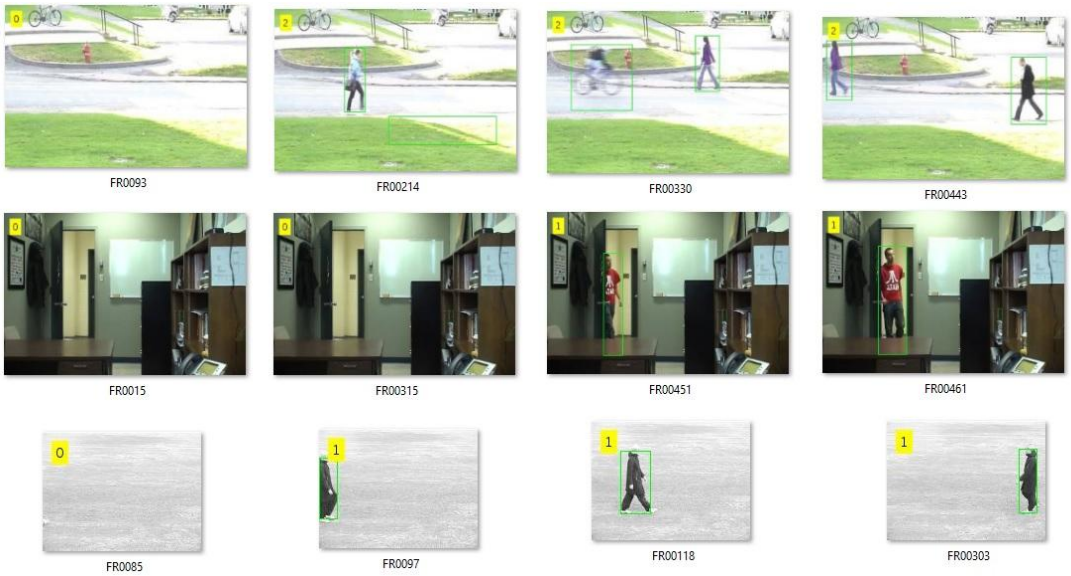


Figure 3: Key frames and labelled foreground in different shots of three different videos. The number below of each frame is the frame number of corresponding video. The highlighted value in left upper corner of each frame shows the number of foregrounds in the video. The bounding box(es) in each frame shows the labelled foreground(s).

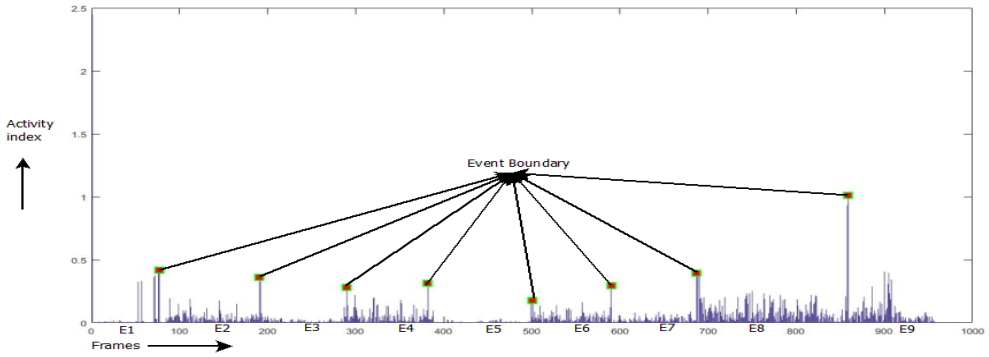


Figure 4: The activity index profile of a video. The highlighted points define the event boundary.

Two figures visibly describe our results. Key frames of some videos from different dataset are shown in Figure. 3. The key frames are labelled with number of foregrounds and the bounding box(es) shows the foregrounds. The actual extraction of foregrounds confirms the effectiveness of the sub-steps used by the proposed approach like noise removal, background estimation, and false foreground removall. In the figure, FR00214 in the first row, the extracted bounding box result shows one extra foreground, this is due to the shade of the actual foreground. On the other hand, Figure. 4 depict a graph, where frames are represented along X-axis and the Y-axis represents the activity index of the corresponding frames. The more y value of a frame represents the more activeness. The difference between activities in consecutive frames defines the event boundary, if the change is beyond some levels. The highlighted dots represent the event boundary.

The active events are extracted on the basis of activities of the underlying frames of the event. From the graph, it can be concluded that events like E2, E4, E6, E8, and E9 are active events. On the other hand, E1, E3, E5, and E7 are static events.

Six different objective parameters like type of video summarization (VS), event detection or shot boundary detection (ED/SBD), adaptive thresholding (AT), key frame detection (KFD), active events (AE), similar activity grouping (SAG) are considered, which can influence the effectiveness of ultimate outcome. A comparative study on those parameters among the proposed technique and others in the related research work is tabulated in Table. 1. The proposed apporache outperforms the other in objectives considerations.

Table 1: A comparative study of proposed work with the same of the related research work.

Approach	ST	ED/ SBD	AT	KFD	AE	SAG
EDC,2008 [1]	Normal	SD		Yes	Yes	No
MVSV, 2017 [2]	Miltiview	SD	Yes	No	No	No
TBA, [3]	Surveillance	SD		Yes	Yes	No
RBA, 2016 [5]	Ranking based	SD		Yes		No
Proposed	Active event based	ED	Yes	Yes	Yes	Yes

5. Conclusions & Future Work

This paper has successfully described the event based video summarization technique based on foreground analysis. The image enhancement technique provides better inputs towards effective background modelling. On the other hand, the result describes the adaptive techniques for threshold selection for background elimination and false foreground reduction helps to extract the actual foreground. The active event detection can be used in the field of application like outdoor and indoor surveillance, home, office and shopping mall monitoring. The proposed system can produce the active events, which can be analysed for higher order decision making. In this way, a huge amount of human effort can be preserved. Moreover, the accuracy of those applications would be improved.

Although the proposed work provided effective results, yet some modification is required for using it in more purposeful way. The shadow detection is not considered in the proposed approach, which sometimes misguide the system. An effective shadow detection technique can make the system more accurate. The proposed approach does not use foreground object classification technique, but an effective classification technique of foreground objects can help to understand the activity of the scene. The foregrounds may include human, cars, pets etc. If we can understand what is there in the foreground, the event detection will be more purposeful as we can eliminate the event, which are not in our interest. The extension of this work will include those important features.

Acknowledgement

This is a text of acknowledgements. Do not forget people who have assisted you on your work. Do not exaggerate with thanks. If your work has been paid by a Grant, mention the Grant name and number here.

References

[1] U. Damnjanovic, V. Fernandez, E. Izquierdo, and J. M. Martinez. Event detection and clustering for surveillance video summarization. In *2008 Ninth International Workshop on Image Analysis for Multimedia Inter- active Services*, pages 63–66, May 2008.

[2] R. Panda and A. K. Roy Chowdhury. Multi-view surveillance video summarization via joint embedding and sparse optimization. *IEEE Transactions on Multimedia*, 19(9):2010–2021, Sept 2017.

[3] A. Javed and N. Sidra. Aj theft prevention alarm based video summa- rization algorithm. *International Journal of Information and Education Technology*, 2(1):23, 2012.

[4] J. Lankinen A. E. Ainasoja, A. Hietanen and J. Kämärä. Keyframe- based video summarization with human in the loop. In *VISIGRAPP (4: VISAPP)*, pages 287–296, 2018.

[5] R. M. Pai M. Srinivas, M. M. Manohara Pai. An improved algorithm for video summarization–a rank based approach. *Procedia Computer Science*, 89:812–819, 2016.

[6] Debi Prosad Dogra, Arif Ahmed, and Harish Bhaskar. Smart video summarization using mealy machine-based trajectory modelling for surveillance applications. *Multimedia Tools Appl.*, 75(11):6373–6401, June 2016.

[7] Malik J. Perona P. Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. Pattern Anal. Mach. Intell. 1990*, 12(7):629–639, 1990

[8] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, volume 2, pages 246–252 Vol. 2, June 1999.