An Efficient Method for Video Summarization using Moving Object Information

Md. Musfequs Salehin, Manoranjan Paul

*School of Computing and Mathematics, Charles Sturt University
Bathurst, NSW-2795, Australia*

msalehin@csu.edu.au, mpaul@csu.edu.

Abstract— Video surveillance system captures continuous video for the purpose of security, monitoring, investigating and so on. It requires huge memory space to store as well as enormous time to retrieve important information manually from this high volume of videos. In this paper, we propose a novel video summarization scheme using moving object information by considering area of moving objects extracted from dynamic background modeling and frame-to-frame object motion. Through this scheme we rank all frames according to the importance of being key frame by combining moving object features through a fusion method so that users can select desired length of videos for summary. The experimental results show that the proposed method provides better video summary compared to the state-of-the-art method using a publicly available benchmark BL-7F video surveillance dataset.

Keywords—Background modelling; frame difference, video summarization

I. INTRODUCTION

Due to the advancement and availability of video technology, an enormous amount of video data is captured and/or stored every day for various purposes. Surveillance system has one of the major contributors to generate huge volume of video data. The surveillance video is used for the purposes of security, monitoring the environment, controlling traffic system, patient monitoring, human behaviour analysis, and so on. To store this huge amount of video data, it requires a tremendous memory space. In addition to this, to find out any important contents manually from this video is very time consuming and tedious work. To solve these problems, a method to manage the video with the most informative parts is highly expected.

Video summarization (VS) is one of approaches for extracting informative parts. In video summarization, a long video is presented in a short form keeping important contents and neglecting the unnecessary events so that a viewer can comprehend about the whole video within short time. Video summary can be represented mainly by a set of key representing the whole content of the original video using a small set of selected frames. Some researchers also define video summary as a video skimming, which presents a short video highlight of the original video. Key frames based methods are lack of motion information and enjoyment whereas video skimming based approaches contain more information and enjoyment but represent a small portion of video segments. Although video skimming provides smooth transition of the most important content of a video, however, it

might ignore some other important content of the video compared to key frame-based approach.

In general important information of a video is associated with the moving object i.e., foreground objects and their interaction in different frames. Thus, an ideal video summarization technique should capture those frames that have foreground objects and relative motion. When we need to select a number of frames that can represent the whole videos or highlight the content of the videos, we need to rank frames based on the foreground amount and motion as we assume that more foreground and more motion can represent more important information of the video. Literature also supports our hypothesis, for example, Ou *et al.* [1] concluded that foreground objects contain more detail information in video stream. Moreover, motion is another stimuli which attracts human visual attention significantly [2].

In this paper, an efficient method is proposed to prepare a summary of a surveillance video incorporating two important properties of a video, the area of the moving foreground regions and the frame-to-frame object motion regions. Foreground regions are obtained by finding the difference between the current frame and the equivalent dynamic background at that point. A *dynamic background modelling* (BGM) based on Gaussian mixture proposed in [3] is applied to generate dynamic background. This method creates the *most common frame in a scene* (McFIS) and used as a background frame. The generated McFIS has the adaptive capability. Therefore, it can provide a stable background frame even when there is a subtle change in illumination and background. To illustrate this property, when a foreground object remains static for longer time, this method starts converting the foreground object as background after a certain time later. During this time, the area of foreground object reduces gradually and finally becomes background. Therefore, key frame extracted based on only the area of foreground objects, may take some non-informative frames as key frames. To solve this problem, frame-to-frame difference is applied for including motion information. These two (area of moving foreground object and motion) important information of a video are linearly fused and generated a sorted list. Then the summary of a video is generated based on the threshold provided by the user.

The structure of the remaining paper is as follows. The related research on video summarization is described in Section II. The detail discussion of the proposed scheme is presented in Section III. Experimental results and discussion

are provided in Section IV. Lastly, a conclusion of this paper is discussed in Section V.

II. RELATED WORK

There are many approaches has been proposed in the literature for generating a condensed version of the original long video. These methods can be classified mainly into two types- key frame extraction and dynamic video skimming generation. In key frame based VS, the original video is presented in a small set of salient frames. On the other hand, dynamic video skimming is a collection of sub-shots selected from the long original video sequence. Beside these two types of VS techniques, a summarized video can also be represented by storyboard and video synopsis. In a storyboard based VS, key frames are organized in a special way so that it provides higher level of view of a video [4][5]. In video synopsis based method [6], multiple objects as well as their activities are superimposed simultaneously.

To extract key frames from a video, various techniques have been proposed in the literature. In [7], a set of key frames is extracted from wireless capsule endoscopy using image moments, curvature, and multi-scale contrast. Heterogeneity image patch and min-max algorithm is applied in [8] for summarizing consumer video. Lecture videos are summarized in [9] using projection screen detection, and a set of key slide frames selection. A probabilistic summarization algorithm is introduced in [10] for summarizing multiple mobile videos. In [11] Bag-of-Importance model is proposed for key frame extraction. Visual attention features and contextual attention information are applied in [12] for generating key frames from sport videos.

In video skimming based approaches, a long original video is generated by a collect of short informative video clips. These video segments are selected in an optimal way so that the generated video contains all the important events in a video. To produce the skimming version of a video, temporal segmentation algorithm and category-specific supervised importance-scoring algorithm is applied in [13]. Online application with user interaction is proposed for video summary in [14]. Super-frame segmentation is performed in [15] using low, mid, and high level features for skimming generation. A sub-shot segmentation method is introduced in [16] for summarizing egocentric video.

In [1], a multi-view video summarization method is proposed for summarizing surveillance videos captured by multiple cameras. This method firstly generates a summary for each individual camera in the intra-view stage. For this purpose, MPEG-7 color layout descriptor is applied on each frame of a specific camera and generate cluster using online Gaussian mixture model (GMM). After this, key frames are selected based on clustering parameters- weights and variance of cluster. As the values of these parameters updates gradually, it extracts video segments from the captured video rather than key frames. Secondly, a shorter summary is generated in inter-view stage from the video segments extracted from each camera in the previous step. To generate the multi-view video summarization, a distributed view selection method is applied in the second stage.

In general, the most informative contents in a video are the area of foreground objects and their motion information between two consecutive frames. The existing techniques mentioned in this section did not apply the amount of moving objects with respect to the static background and frame-to-frame motion information. In the proposed method, we consider the amount of foreground in each frame against the background of the video and motion between consecutive frames as two features. Then we formulate a scoring criterion by combining these two features for ranking each frame in terms of their representativeness to the video so that a user can select necessary frames for summarization. The proposed method is compared with the GMM based technique proposed in [1]. Although both the methods apply GMM in learning phase, there is a major difference between these methods. The proposed method apply GMM on the intensity of each pixel whereas GMM based method apply this at frame level after applying the MPEG-7 color layout descriptor. The extensive experimental results reveal that the proposed method performs better than GMM based method [1].

III. PROPOSED METHOD

The introduced method has four main steps. They are as follows- (i) Foreground object extraction, (ii) Motion information estimation, (iii) Fusion of features, and (iv) Key frame selection.

The main steps of the proposed method are shown in Fig. 1. The details of each section is presented in subsequent sub-sections.

Fig. 1. Framework of the proposed method

A. Foreground Object Extraction

Foreground objects usually contain more detail information in a video stream [1]. To find out the objects in a single frame, any BGM technique can be used. In the literature, there exist two kinds of BGM such as parametric and non-parametric [17]. The parametric BGM techniques requires some prior information related to the content of the videos, however, they can provide an instant background frame from the beginning so that any online object detection can be possible using these techniques. Whereas, the non-parametric BGM techniques collect video-content information from the learning process, and require a number of video-frames for initial learning. Thus, sometimes they are not

suitable in the online object detection especially at the initial stage. In the proposed technique we use Gaussian mixture-based parametric BGM [18]. In this technique, each pixel is modeled individually using a number of Gaussian models and each model is represented by a Gaussian distribution using *mean* and *standard deviation* of the pixel intensities experienced over the time through different frames. The basic assumption of the background modelling is to consider a pixel position (i.e., a physical position of a pixel in the temporal frames of a video) as a part of a background if the pixel intensity at that position does not change too much in a number of successive frames. A position may experience a background and a number of foregrounds for a given time. To accommodate different phenomena due to the changing of background and foreground, multiple Gaussian models are applied for a pixel position where each model represents each experience (i.e., backgrounds or foregrounds). The detail explanation of BGM is described in [18].

In the introduced scheme, a video frame at time t is converted into gray-scale image $I(t)$ first. Then background modelling [18] is applied on it to obtain its corresponding background frame $B(t)$. Later a background-foreground separating threshold ($Thr1$) is applied between $I(t)$ and $B(t)$ to obtain the foreground image $O(t)$. If the pixel intensity difference between $I(t)$ and $B(t)$ is greater than $Thr1$, this pixel belongs to foreground object and set the pixel value to one. Otherwise, this pixel belongs to background and set the pixel value to zero. In this way, a foreground pixel $O_{p,q}(t)$ is obtained as follows

$$O_{p,q}(t) = \begin{cases} 1, & \text{if } |I_{p,q}(t) - B_{p,q}(t)| \geq Thr1 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where (p, q) is the pixel position. The background-foreground threshold ($Thr1$) value is set to 20 as per suggested in [3]. This threshold value is robust against subtle changes between foreground and background. Then, the foreground object feature $F(t)$ is obtained counting all the foreground pixels in $O(t)$ using the following equation

$$F(t) = \sum_{p=1}^r \sum_{q=1}^c O_{p,q}(t) \quad (2)$$

where r and c represent the row and column of O respectively.

Fig. 2 shows a demonstration for key frame selection based on a number of features on frame no 1 to frame no 1500 of bl-1 video in BL-7F dataset. According to the video content, at frame 1001 a moving object has stopped moving. It is clearly visible from Fig. 2 that if a moving object becomes static for a while, background-modelling technique starts converting it as a part of background. This conversion process takes few times to complete. As a result, it may select some unnecessary frames as key frames during this conversion. Therefore, method proposed based on foreground object only does not guarantee better result. To avoid this situation and make the proposed method robust, motion information is applied as another feature in the proposed method. Dynamic

and static object can easily be identified by motion information of the object.

Fig. 2. Features applied in the proposed method on 1 to 1500 frames of bl-1 video. The blue, black and red curves represent ground truth i.e., key frame, foreground and motion features respectively.

B. Motion Information Extraction

As discussed in the previous section, foreground object alone is not an efficient feature to extract key frames. To solve this problem, motion information of an object is also included in the introduced method. The motivation regarding this is that object motion in a video is very attractive to human being [2]. To include motion information, frame-to-frame difference is calculated. Consider, $C(t)$ and $C(t-1)$ are two consecutive frames at t and $t-1$ respectively. To obtain motion information, $C(t-1)$ is subtracted from $C(t)$ at red, green, and blue channels. A pixel is considered motion pixel if the pixel value of the subtracted frame is greater than or equal to a threshold ($Thr2$) in three different color channels. Otherwise, it is considered non-motion pixel. The values of motion and non-motion pixel are set to one and zero respectively. Thus, a pixel containing significant motion $M_{p,q}(t)$ is calculated using the following equation

$$M_{p,q}(t) = \begin{cases} 1, & \text{if } |C_{p,q}^r(t) - C_{p,q}^r(t-1)| \geq Thr2 \\ & \text{and } |C_{p,q}^g(t) - C_{p,q}^g(t-1)| \geq Thr2 \\ & \text{and } |C_{p,q}^b(t) - C_{p,q}^b(t-1)| \geq Thr2 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where (p, q) is pixel position and $C_{p,q}^r$, $C_{p,q}^g$, and $C_{p,q}^b$ represents red, green and blue color channels respectively. The value of $Thr2$ is also set to 20 [3]. The motion feature $G(t)$ is obtained by summing all non-zeros in $M_{p,q}(t)$ as follows

$$G(t) = \sum_{p=1}^r \sum_{q=1}^c M_{p,q}(t) \quad (4)$$

where r and c represent row and column of M respectively.

Motion feature $G(t)$ is shown by red curve in Fig. 2. It is obvious that including motion information, unnecessary frames obtained from the foreground feature can be removed (see Fig. 2). Although motion is a possible feature, it alone is not sufficient. If significant motion information and very trivial foreground object are observed in a frame, it should give lower importance than frame with considerable amount of motion and foreground objects for the selection of being key frame. To exemplify this, Fig. 3 shows some frames where a small portion of human head is visible. Although there is significant motion information near human head, the area of foreground object is negligible. Therefore, they cannot

be key frames. Thus, these frames should be given lower preference in the key frame selection process.

Fig. 3. Sample frames taken from bl-3 video. Each row (Frame no 1730-1730 and frame no 4027- 4037) contains small object with adequate motion.

C. Fusion of Features

As neither foreground object area nor object motion information alone cannot provide better result, these features are combined by a weighted linear fusion method in the proposed approach. The main purpose of this fusion scheme is to provide an importance index to each frame based on its contents of foreground area and object motion information. Before combining these features, z-score normalization is applied on each feature by the following equation

$$Z(t) = (X(t) - \mu)/\sigma \quad (5)$$

where $X(t)$ is a feature value at time t , μ is the mean and σ is standard deviation of the feature values. Z-score, $Z(t)$ is a normalized form of $X(t)$. In the proposed method, z-score normalization is applied instead of min-max based normalization for two main reasons. First, it generates the meaningful information at each data point. Second, it provides good results in the presence of outliers [19]. The weighted linear fusion is obtained as follows

$$A(t) = \alpha * Z_G(t) + (100 - \alpha) * Z_F(t) \quad (6)$$

where α is a weight, $A(t)$, $Z_F(t)$ and $Z_G(t)$ are fusion value, z-score normalization of foreground feature ($F(t)$) and motion feature ($G(t)$) respectively at time t . Empirically it is evaluated that if the value of weight α is set to 90, it provides better results for all videos in BL-7F dataset. According to the psychological theories of human attention, human being naturally pay more attention to dynamic objects than static objects [2]. Therefore, higher weight is provided to object motion information than the area of foreground object. Next, the fusion values $A(1), A(2), A(3), \dots, A(T)$ are ranked based on descending order where T is the number of frames in a video.

D. Key Frame Selection

In the final step, key frames are selected from the sorted fusion values. The introduced approach offers the user to select the number of key frames (Thr_{kf}) in the summarized video. It generates summarized video based on the frame number (Thr_{kf}) selected by the user. From the sorted fusion values $A(1), A(2), A(3), \dots, A(T)$; Thr_{kf} number of frames are selected from the top. Finally, summarized video is produced

from these selected frames keeping their sequential order in the original video.

IV. RESULTS AND DISCUSSION

To evaluate the proposed method, 19 surveillance videos from publicly available BL-7F dataset [1] have been used. In this dataset, all the videos are taken from fixed surveillance cameras in the seventh floor of the BarryLam Building in National Taiwan University. Each video is 7 minutes 10 seconds long and contains 12,900 frames. A ground truth video comprises the selected key frames for each video in BL-7F dataset is also provided.

Fig. 4. An example of foreground and motion of objects where (a) and (d) are two consecutive (765 and 766) frames of bl-0 video respectively, (b) and (e) are background image of (a) and (d) respectively generated by the proposed algorithm, (c) is the foreground image of (d) and (f) is the object motion between two frames.

In Fig. 4, an illustration of foreground and motion information extracted by the proposed method is shown. Two consecutive (765 and 766) frames (Fig. 4(a) and (d)) of bl-0 video from BL-7F dataset are selected to demonstrate the foreground area and motion between them. The background images of 765 and 766 frames obtained by dynamic background modeling proposed by [3] are shown in Fig. 4(d) and 4(e). The foreground object of frame 766 after applying Eq. 1 is provided in Fig. 4(c) and the object motion between frames (765 and 766) using Eq. 3 is displayed in Fig. 4(f). It is demonstrated by Fig. 4 that the proposed technique can extract moving foreground region applying dynamic background modeling [3] and motion information from frame-to-frame difference for selecting important information.

In Fig. 5, some sample ground truth frames from bl-9 video of BL-7F dataset [1], the experimental results to select these key frames by the proposed approach and GMM based method are displayed. GMM based technique unable to select key frames no 10112 and 11255 even if these frames contain significant amount of foreground object and motion information. However, the introduced scheme is successful to identify these key frames as it considers two significant contents of video frames – foreground objects area and motion information between consecutive frames. Foreground objects contain more detail information in a video stream [1] and frame to frame object motion information is another stimuli which attracts human visual attention significantly [2].

Therefore, the proposed method provides superior performance to GMM based method.

Fig. 5. Comparison of key frames extraction of bl-9 video from BL-7F dataset. The left, middle, and right columns show ground truth, results obtained by GMM based method and the proposed method respectively.

In this experiment, an objective evaluation has been performed. In this regard, a set of quantitative metrics are measured to estimate the quality of the summary video. The following terms are used.

- True positive (T_p): A frame is in the set of ground truth and it is selected by the proposed method.
- False positive (F_p): A frame is not in the set of ground truth though it is selected by the proposed method.
- True negative (T_n): A frame is not in the set of ground truth and it is not selected by the proposed method.
- False negative (F_n): A frame is in the set of ground truth however it is not selected by the proposed method.

These terms are used to define the precision and recall.

$$Precision = T_p / (T_p + F_p) \quad (7)$$

$$Recall = T_p / (T_p + F_n) \quad (8)$$

The precision indicates that the number of frames obtained from the proposed method match with the ground truth frames and the recall is how many ground truth frames match with the frames obtained by the proposed method. However, precision or recall could not indicate the overall performance of a method individually; for example, the high value of precision

(or recall) does not guarantee a good measure. To overcome this, both precision and recall must be high. In order to get this, the precision and recall are combined in another metric known as F_1 -measure, which is defined as follows:

$$F_1 = 2 \times \frac{Recall \times Precision}{Recall + Precision} \quad (9)$$

If both precision and recall are high, F_1 -measure will also be high. Therefore, a video summarization method with higher F_1 -measure can guarantee better summarization.

TABLE I Precision, recall, and F_1 -measure of GMM [1] and proposed

Video	GMM [18]			Proposed		
	Precision	Recall	F_1 -measure	Precision	Recall	F_1 -measure
bl-0	70.3	91.8	0.80	79.86	95.68	0.87
bl-1	28.9	79.0	0.42	54.48	73.04	0.62
bl-2	77.8	64.2	0.70	70.96	86.38	0.78
bl-3	27.4	98.5	0.43	65.41	87.68	0.75
bl-4	62.4	88.1	0.73	70.55	94.99	0.81
bl-5	26.1	97.6	0.41	59.36	80.04	0.68
bl-6	2.1	100.0	0.04	77.78	100	0.88
bl-7	63.9	80.9	0.71	74.17	100	0.85
bl-8	83.6	100.0	0.91	88.16	100	0.94
bl-9	97.5	60.2	0.75	96.48	88.83	0.92
bl-10	61.4	78.6	0.69	83.26	99.85	0.91
bl-11	98.3	62.7	0.77	94.99	85.53	0.9
bl-12	88.3	98.2	0.93	82.69	94.03	0.88
bl-13	0	0	0	100	100	1.00
bl-14	76.9	97.7	0.86	76.83	99.42	0.87
bl-15	92.4	96.5	0.94	98.01	99.09	0.99
bl-16	55.6	97	0.71	78.84	97.38	0.87
bl-17	88.1	81.9	0.85	85.96	93.74	0.9
bl-18	21.3	89.2	0.34	71.7	96.57	0.82
Mean	59.1	82.2	0.63	79.4	93.3	0.85
STD	32.00	23.81	0.28	12.73	7.73	0.10

The proposed approach is compared with the single-view video summarization by GMM based method [1] as the proposed method also uses single-view video. The GMM method is the most relevant and the state-of-the-art method to compare with because both techniques use GMM for the training purpose; however, they used different features for training models. The GMM based method in [1] exploited color descriptor difference to train different models at frame level and determined key frame based on the weight and variance of the models in the frames of a video whereas the proposed method uses pixel intensity to train different GMMs at pixel level. Then the proposed method generates background frame from the model in each pixel individually, which has the highest weight and lowest variance. Both techniques use BL-7F standard dataset to verify the performance. The results of precision, recall, and F_1 -measure of GMM based method have been taken from the single-view summarization result of [1]. In this experiment, the total key frames number in the ground truth [1] for each video is used as the value of Thr_{kf} . It is observed that if 0.35 times of Thr_{kf} more frames are taken from the sorted list of fusion values $A(1), A(2), A(3), \dots, A(T)$; where T is the total frame number in a video, the proposed approach can provide better summarized video. The results obtained by GMM-based method and the proposed approach are shown in Table I. It is easily seen that mean of precision, recall, and F_1 -measure of the proposed method are higher than those of GMM-based method. For example, the mean of precision, recall, and F_1 -measure of the proposed method are 79.4, 93.3, and 0.85 respectively. On the other hand, the mean of precision, recall, and F_1 -measure of

GMM-based method are 59.1, 82.2, and 0.63 respectively. Moreover, standard deviations (STDs) of precision, recall, and F1-measure of the proposed method are much less than those of GMM-based method. Therefore, the proposed method provides not only better but also consistent results compared to GMM-based method [1].

The graphical representations of F_1 -measure of the proposed method and GMM based method are shown in Fig. 6. After examining all these graphs, it is obvious that the proposed approach performs superior to GMM based method in 17 videos. However, the proposed technique almost provides similar performance for the bl-14 video and fails to perform better for the bl-12 video compared to the GMM based method [1].

Fig. 6. F_1 -measure of GMM based method and the proposed method.

The reason of the failure to perform better in the video bl-12 is that there are a number of frames that contains informative contents such as object as well as motion, however, they are not considered as key frames in BL-7F dataset [1]. Three sample frames of such category are shown in Fig. 7. In frame 4255, 4341 and 4520, a person is moving near door. It is easily visible that there are significant content exists in these frames, however, for the unknown reason those frames are not selected in the ground truth. As the introduced method selects frames based on the area of foreground and object motion, it includes these frames as key frames. For this reason, our method fails to perform better than the method in [1] for the bl-12 video. After visual observation and objective evaluation based on a set of quantitative metrics, it is certain that the proposed method is more efficient and robust than GMM based method for summarizing surveillance video.

Fig. 7. Sample frames are not selected as key frames in bl-12 video

V. CONCLUSION

In this paper, an efficient method has been proposed combining area of foreground object obtained a Gaussian mixture-based dynamic background modelling [3] and frame-to-frame object motion. Foreground objects usually contain very rich information of the video. Again, human pay more attention to object motion. Extensive experiment has been performed using 19 videos of publicly available benchmark

BL-7F dataset. Experimental results have shown the superiority of the proposed method compared to the state-of-the-art method.

REFERENCES

- [1] S. Ou, C. LEE, V. Somayazulu, Y. Chen, and S. Chien, "On-line Multi-view Video Summarization for Wireless Video Sensor Network," *IEEE J. Sel. Top. Signal Process.*, vol. 9, no. 1, pp. 165–179, 2015.
- [2] D. Gao, V. Mahadevan, and N. Vasconcelos, "On the plausibility of the discriminant center-surround hypothesis for visual saliency," *J. Vis.*, vol. 8, no. 7, pp. 1–18, 2008.
- [3] M. Haque, M. Murshed, and M. Paul, "A hybrid object detection technique from dynamic background using Gaussian mixture models," *IEEE 10th Work. Multimed. Signal Process.*, pp. 915 – 920, 2008.
- [4] J. Kwon and K. M. Lee, "A Unified Framework for Event Summarization and Rare Event Detection," pp. 1266–1273, 2012.
- [5] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering Important People and Objects for Egocentric Video Summarization," *CVPR*, pp. 1346–1353, 2012.
- [6] X. Zhu, C. C. Loy, S. Gong, Q. Mary, and E. London, "Video Synopsis by Heterogeneous Multi-Source Correlation," *IEEE Int. Conf. Comput. Vis.*, pp. 81 – 88, 2013.
- [7] I. Mehmood, M. Sajjad, and S. W. Baik, "Video summarization based tele-endoscopy: a service to efficiently manage visual data generated during wireless capsule endoscopy procedure," *J. Med. Syst.*, vol. 38, no. 9, p. 109, Sep. 2014.
- [8] C. Dang and H. Radha, "Heterogeneity image patch index and its application to consumer video summarization," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2704–2718, 2014.
- [9] K. Li, J. Wang, H. Wang, and Q. Dai, "Structuring Lecture Videos by Automatic Projection Screen Localization and Analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1233–1246, 2015.
- [10] L. Zhang, Y. Xia, K. Mao, H. Ma, and Z. Shan, "An effective video summarization framework toward handheld devices," *IEEE Trans. Ind. Electron.*, vol. 62, no. 2, pp. 1309–1316, 2014.
- [11] S. Lu, Z. Wang, T. Mei, G. Guan, and D. Feng, "A bag-of-importance model with locality-constrained coding based feature learning for video summarization," *IEEE Trans. Multimed.*, vol. 16, no. 6, pp. 1497–1509, 2014.
- [12] H. Shih, "A novel attention-based key-frame determination method," *IEEE Trans. Broadcast.*, vol. 59, no. 3, pp. 556–562, 2013.
- [13] D. Potapov and M. Douze, "Category-specific video summarization," *Comput. Vision–ECCV ...*, 2014.
- [14] J. Almeida, N. J. Leite, and R. D. S. Torres, "VISON: Video Summarization for ONline applications," *Pattern Recognit. Lett.*, vol. 33, no. 4, pp. 397–409, Mar. 2012.
- [15] M. Gygli and H. Grabner, "Creating summaries from user videos," *Comput. Vision–ECCV ...*, 2014.
- [16] Z. Lu and K. Grauman, "Story-Driven Summarization for Egocentric Video," *2013 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2714–2721, Jun. 2013.
- [17] S. Chakraborty and M. Paul, "A novel video coding scheme using a scene adaptive non-parametric background model," *IEEE 16th Int. Work. Multimed. Signal Process.*, pp. 1–6, 2014.
- [18] M. Paul, W. Lin, C. Lau, and B. Lee, "Explore and model better I-frames for video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 9, pp. 1242–1254, 2011.
- [19] J. Han, M. Kamber, and J. Pei, *Data Mining, Southeast Asia Edition: Concepts and Techniques*. Morgan Kaufmann, 2006.
- [20] N. Ejaz, I. Mehmood, and S. Wook Baik, "Efficient visual attention based framework for extracting key frames from videos," *Signal Process. Image Commun.*, vol. 28, no. 1, pp. 34–44, Jan. 2013.
- [21] Y. Fu, Y. Jia, and Y. Kong, "Interactive phrases: Semantic descriptions for human interaction recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 9, pp. 1775–1788, 2014.