Fusion of Foreground Object, Spatial and Frequency Domain Motion Information for Video Summarization

Md. Musfequs Salehin, Manoranjan Paul
School of Computing and Mathematics, Charles Sturt University
Bathurst, NSW-2795, Australia
{msalehin, mpaul}@csu.edu.au

Abstract. Surveillance video camera captures a large amount of continuous video stream every day. To analyze or investigate any significant events from the huge video data, it is laborious and boring job to identify these events. To solve this problem, a video summarization technique combining foreground objects as well as motion information in spatial and frequency domain is proposed in this paper. We extract foreground object using background modeling and motion information in spatial domain and frequency domain. Frame transition is applied for obtaining motion information in spatial domain. For acquiring motion information in frequency domain, *phase correlation* (PC) technique is applied. Later, foreground objects and motions in spatial and frequency domain are fused and key frames are extracted. Experimental results reveal that the proposed method performs better than the state-of-the-art method.

Keywords: background modeling; motion information; phase correlation; video summarization

1 Introduction

Every day an enormous amount of surveillance video is captured 24 hours throughout the whole world for providing security, monitoring, preventing crime, controlling traffic, and so on. In general, a number of surveillance video cameras are set up in a number of different places of a building, business area, or congested area. These cameras are connected to a monitoring cell for storing and investigating. To store this huge volume of video data requires tremendous memory space. In addition to this, to find out any important events from the stored video for investigating or performing analysis, operators need to access the stored videos. This process is very tedious, lengthy and expensive. To solve these problems, a method for generating the shorter version of original video containing important events is highly desirable for memory management and information retrieval.

Video summarization (VS) is the technique to select the most informative frames so that it can contain all the necessary events and reject unnecessary contents to make the summarized video as concise as possible. Therefore, a good video summarization method is one that has several important properties. First, it must have the capability to include all significant incidents within the original video. Second, it should be able to generate a smaller version of the provided long video. Third, it should not contain

repetative information. The main purpose of VS is to represent a long original video in a condensed version in such a way that a user can get the whole idea of the entire video in a constrained amount of time.

In a video, foreground objects generally contain more detail information [1]. Again, human usually concentrate more on the movements of objects [2]. Consequently, objects as well as their motion are important features for a video. Motivated by these findings, a video summarization scheme is proposed in this paper based on objects and their motion in a video. To include foreground object information, Gaussian mixture-based parametric *background modeling* (BGM) [3] has been applied. To acquire the complete information of object motion in a video, object motion is extracted not only in spatial domain but also in frequency domain. To obtain motion information in spatial domain, consecutive frame difference is applied. For achieving object motion in frequency domain, phase correlation technique [4] is applied. To the best of our knowledge, phase correlation is not applied for video summarization methods. Therefore, the main contribution of this paper is to apply phase correlation in video summarization method. The computational time of phase correlation is very low and rich motion information is obtained by phase correlation technique [4].

The structure of the remaining paper is as follows. Section 2 reviews related research. The proposed method is described in section 3. Experimental results as well as detail discussions are provided in section 4. Finally, a concluding remarks and future direction are drawn in section 5.

2 Related Research

In the literature, different approaches have been proposed for summarizing various types of videos. For egocentric video summarization, region saliency is predicted in [5] using a regression model and storyboard are generated based on region importance score. In the method proposed in [6], story driven egocentric video is summarized by discovering the most influential objects within a video. Gaze tracking information is applied in [7] for summarization. In case of user generated video summarization, adaptive submodular maximization function is applied in [8]. A collaborative sparse coding model is utilized in [9] for generating summary of the same type of videos. Web images are used in [10] to enhance the process of summarizing the user generated video. To summarize movie, aural, visual and textual are merged in [11]. Role community network is applied in [12]. Film comic is generated using eye tracking data in [13]. In addition to these, techniques proposed in [14][15][16][17] for wireless capsule endoscopic video summarization.

However, the importance of surveillance video for industrial application is very higher than other types of videos (e.g., egocentric, user generated, movie, etc). To summarize surveillance video, object centered technique is applied in [18]. Dynamic VideoBook is proposed in [19] for representing the surveillance video in a hierarchical order. Learned distance metric is introduced in [20] for summarizing nursery school surveillance video. In [21], salient motion information is applied. Maximum a posteriori probability (MAP) is used in [22] for synopsis generation. Recently, a method is proposed in [1] for a multi-view surveillance video summarization. Firstly, a single view summarization is generated in this approach for each sensor independently.

For this purpose, MPEG-7 color layout descriptor is applied to each video frame and an online-*Gaussian mixture model* (GMM) is used for clustering. The key frames are selected based on the parameters of cluster. As the decision of selecting or neglecting a frame is performed based on the continuously updates of these clustering parameters, a video segment is extracted instead of key frames. Lastly, multi-view summarization is produced by applying distributed view selection method using the video segments extracted for each sensor in the previous step.

To the best of our knowledge, phase correlation technique has not been applied for video summarization. In this proposed approach, phase correlation technique is applied to incorporate motion information in frequency domain and fused with moving foreground object and spatial motion information.

3 The Proposed Method

The proposed scheme is based on area of moving foreground objects and their motion information in spatial and frequency domain. The main steps of the proposed method are (1) moving foreground object extraction (2) motion information calculation in spatial domain, (3) motion estimation in frequency domain, (4) fusion of foreground object area and spatial as well as frequency motion information, and (5) video summary generation. The flow chart of the proposed method is shown in Fig. 1. The detail of each step is explained in the subsequent sub-section.

Fig. 1. Framework of the proposed method

3.1 Foreground Object Extraction

In the proposed method, Gaussian mixture-based parametric BGM [3] is applied. In this parametric BGM, each pixel is modeled by the K Gaussian distributions ($K=3$) and each Gaussian model represents either static background or dynamic foreground object during time frame. For instance, suppose a pixel intensity x_t at time t is modeled by k th Gaussian with recent value γ_k^t , mean μ_k^t , standard deviation σ_k^t and weight ω_k^t such that $\sum \omega_k^t = 1$. The learning parameter α is used to update parameter values, such as mean, standard deviation, etc. At the beginning, the system contains empty set of Gaussian models. After observing the first pixel ($t=1$), a new Gaussian model ($K=1$) is generated with $\gamma_k^t = \mu_k^t = x_t$, standard deviation $\sigma_k^t = 30$ and weight $\omega_k^t = 0.001$. Then for each new observation of pixel intensity X_t of the same location at t , it tries to find a matched model from the existing models such that $|x_t - \mu_k| \leq 2.5\sigma_k$.

To obtain gray scale background frame, background modelling [3] is applied after converting each color-frame into gray scale image. Then, A color video frame at time t is converted into gray image $I(t)$ and subtracted from the corresponding gray background frame $B(t)$ obtained by the background modelling. A pixel is considered foreground pixel and set the value to one, if the pixel intensity difference between $I(t)$ and $B(t)$ is greater than or equal to a certain threshold ($Thr1$). If the pixel intensity does not satisfy this condition, it is regarded as a background pixel and set to zero. In this way, a foreground pixel $G_{i,j}(t)$ is obtained as follows

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

where (i, j) is the pixel position. The value of $Thr1$ is set to 20 to avoid subtle changes between background and foreground. This is a common practice to set the threshold value to 20 to identify object from the background as mentioned in [24]. After that, the total number of non-zero pixels in $G_{i,j}(t)$ is used as area of foreground object feature $F(t)$ which is obtained by the following equation

$$G(t) = \sum_{i=1}^r \sum_{j=1}^c G_{i,j}(t) \quad (2)$$

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

According to the psychological theories of human attention, motion information is more significant than the static attention clues [2]. Therefore, motion information is included in the proposed method in addition to the foreground object.

3.2 Motion Information Calculation in Spatial Domain

Human usually concentrate more on the movements of objects in a video [2]. In order to obtain object motion information in spatial domain, frame-to-frame difference is applied. Consider two consecutive frames $I(t-1)$ and $I(t)$ at time $t-1$ and t in video. To find out spatial motion information, the color difference in red, green, and blue channel between these frames is calculated. If the differences at each pixel in three different channels are greater than or equal to a threshold, this pixel is considered as motion pixel and set to value one. Otherwise, it is sure that this pixel does not contain any

motion information. Therefore, the spatial motion information $S_{i,j}(t)$ in pixel (i,j) at time t can be obtained by the following equation

$$S_{i,j}(t) = \begin{cases} 1, & \text{if } |I_{i,j}^r(t) - I_{i,j}^r(t-1)| \geq T2 \\ & \text{and } |I_{i,j}^g(t) - I_{i,j}^g(t-1)| \geq T2 \\ & \text{and } |I_{i,j}^b(t) - I_{i,j}^b(t-1)| \geq T2 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where $I_{i,j}^r$, $I_{i,j}^g$, and $I_{i,j}^b$ represent red, green and blue color at (i,j) channels respectively. This is a common practice to set 20 as a threshold value to obtain information between two consecutive frames [24]. Therefore, the value of $T2$ is set to 20. The spatial motion information $S(t)$ is obtained at time t by summing all values in $S_{i,j}(t)$ as follows

$$S(t) = \sum_{i=1}^r \sum_{j=1}^c S_{i,j}(t) \quad (4)$$

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

However, motion extracted in spatial domain is not sensitive to diffuse phenomena [25]. For example, in case of global illumination changes, it does not work well. Additionally, spatial motion estimation is prone to local inaccuracies and small motion discontinuities [26].

3.3 Motion Estimation in Frequency Domain

To overcome the problem of spatial motion calculation, motion information is calculated in frequency domain. Motion estimated in frequency domain has some advantages over spatial domain [25]. It is efficient for global changes of illumination and robust to motion estimation near object boundaries. To obtain motion information in frequency domain, each frame is divided into a number of blocks of $\times 16$ pixels size. Then, phase correlation technique [4] is applied between the current block and reference block. The phase correlation peak (β) (i.e., the magnitude of the motion accuracy) extracted from phase correlation method is used as motion indicator for that block. The phase difference ($\emptyset$) is calculated between the current block and its co-located reference block after applying *Fast Fourier transformation* (FFT) on each block. The inverse FFT is performed on the calculated phase difference and finally two dimensional (2-D) motion vector (dx, dy) is obtained [27]. This 2-D motion vector contains peaks (β) where there are shift between the current and reference blocks.

$$\emptyset = \text{fftshift}|\text{ifft}(e^{j(\angle F_r - \angle F_c)})| \quad (5)$$

where F_r and F_c represent FFF of current and reference block respectively.

$$(dx, dy) = \max(\emptyset) - b/2 - 1 \quad (6)$$

$$\beta = \emptyset(dx + b/2 + 1, dy + b/2 + 1) \quad (7)$$

where b is block size. For example, b will be 16 if 16×16 is used.

If the value of β of a block greater than a threshold ($T3$), it is considered that this block contains sufficient motion information. In this method, the value of $T3$ is set to 0.6. All the values greater than $T3$ are summed to obtain motion information $F(t)$ in frequency domain.

$$F(t) = \sum_{n=1}^N \sum_{m=1}^M F_{n,m}(t) \quad (8)$$

where N and M represent row/ b and column/ b of F respectively.

The motion information obtained in frequency domain applying phase correlation technique at different blocks of frame no 3869 of bl-14 video is shown in Fig. 2. No motion is represented in block (4,4) with only a single highest pick (Fig. 2(b)). Single (block (10,14)) and complex motions (block (5,8)) are showed in Fig 2(c) and (d) respectively.

Fig. 2. An example of motion generated in each block of frame no 3869 of bl-14 video; (a) frame difference between 3868 and 3869 (multiplied by 6 for better visualization), phase correlation pick with no motion, complex motion and single motion are represented in (b), (c) and (d) respectively.

In contrast, frequency based motion estimation approach lacks of localization problem [25]. Therefore, motions obtained in both spatial and frequency domains are combined with moving foreground objects for generating video summary.

3.4 Fusion of Foreground and Motion Information

In order to select more accurate frame sequences, both area of foreground object and motion information are combined. In this approach, a weighted linear fusion is applied to combine the features for ranking each frame according to their representativeness in a video. Before applying fusion method, each feature is converted into z-score normalization using the following equation

$$Z(t) = (X(t) - \mu) / \sigma \quad (9)$$

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 this scheme, z-score

normalization is the preferred method because it produces meaningful information about each data point, and provides better results in the presence of outliers than min-max based normalization [28]. The weighted linear fusion is obtained as follows

$$A(t) = \phi_1 * Z_G(t) + \phi_2 * Z_S(t) + \phi_3 * Z_F(t) \quad (10)$$

where $A(t)$ is fusion value; $Z_G(t)$, $Z_S(t)$ and $Z_F(t)$ are z-score normalization of foreground feature ($G(t)$), spatial motion feature ($S(t)$), and motion information in frequency domain ($F(t)$) respectively at time t . Empirically, it is evaluated that if the values of weights ϕ_1 , ϕ_2 , and ϕ_3 are set to 15, 60, and 25 respectively, it provides better results for all videos in BL-7F dataset. The rationality to provide higher weight to motion feature compared to the foreground area is that according to the psychological theories of human attention, motion information is more significant than the static attention clues [2]. After that, $A(1, \dots, T)$ is sorted based on descending order where T is total number of frames in a video.

3.5 Video Summary Generation

The proposed method summarizes a video based on the threshold (Thr_{kf}) provided by a user. Thr_{kf} number of frames are selected from sorted list of $A(1, \dots, T)$ and a video summary is generated from the selected frames by maintaining their chronological sequence. Although a set of key frames are selected, the resulting video summary is skimming based. Because the area of foreground and frame-to-frame motion information change gradually over time and the proposed method maintains smooth transition among different frame sequences.

4 Results and Discussion

The proposed method is evaluated by the publicly available BL-7F dataset [1]. In this dataset, 19 surveillance videos are provided. The duration of each video is 7 minutes 10 seconds and contains 12,900 frames. This dataset also provides a complete list of selected key frames as a ground truth for each video in BL-7F dataset. The foreground object, spatial and frequency domain motion information are extracted by the proposed method are shown in Fig 3(c), (f), and (i) respectively.

(a)

(b)

(c)

Fig. 3. An example of foreground and motion information extracted by the proposed method; (a) and (b) are frame no 740 and 741 of bl-0 video, (c) is the background image of (b), (d) is the foreground image of (b) , (e) is the object motion between two frames and (f) is the motion obtained by phase correlation technique on frame no 741.

In Fig. 4, a number of ground truth frames of bl-11 video of BL-7F dataset [1] and the results obtained by GMM based method as well as the proposed method are shown. Although there exist significant contents in frame number 9963 and 12523, GMM based method fails to select these frames. In contrast, the proposed method is capable to select these frames successfully. The main reason of this success is that the proposed method combines area of foreground objects as well as frequency and spatial motion information. Foreground objects usually contain very informative contents of video frames [1] and object motion is visually more attractive for human being [2].

Frame No	Ground Truth	GMM	Proposed
248			
8083			
9963		Not Selected	
12523		Not Selected	
12888			

Fig. 4. Evaluation of key frames extraction of bl-11 video of BL-7F dataset; first, second, third, and forth columns indicate frame no, 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, to estimate the quality of the summarized video, precision, recall, and F1-measure are calculated using the following equations

$$Precision = T_p / (T_p + F_p) \quad (11)$$

$$Recall = T_p / (T_p + F_n) \quad (12)$$

$$F_1 - measure = 2 \times \frac{Recall \times Precision}{Recall + Precision} \quad (13)$$

Where T_p , F_p , and F_n indicates true positive, false positive, and false negative respectively. A high value for both precision and recall indicate that F_1 -measure is also very high. Thus, a method with high F_1 -measure value confirms better summarization technique.

The proposed approach is compared with the single-view video summarization results provided by GMM based method [1]. 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, there are two key differences between GMM based method [1] and Gaussian mixture-based parametric BGM [3]. Firstly, GMM based method works at frame level whereas Gaussian mixture-based BGM works in pixel level. Secondly, GMM based method utilizes color descriptor as feature while Gaussian mixture-based parametric BGM uses pixel intensity. Another reason to compare with GMM based method is that both techniques use BL-7F standard dataset to verify the performance.

TABLE I						
<i>Precision, recall, and F_1-measure of GMM [1] and the proposed method</i>						
Video	GMM [1]			Proposed Method		
	Precision	Recall	F_1-measure	Precision	Recall	F_1-measure
bl-0	70.30	91.83	0.80	87.97	90.79	0.89
bl-1	28.90	78.99	0.42	62.26	72.79	0.67
bl-2	77.80	64.21	0.70	72.89	83.09	0.78
bl-3	27.40	98.53	0.43	71.89	96.38	0.82
bl-4	62.40	88.11	0.73	74.16	93.15	0.83
bl-5	26.10	97.62	0.41	68.98	89.63	0.78
bl-6	2.10	100.00	0.04	77.78	100	0.88
bl-7	63.90	80.88	0.71	80.45	100	0.89
bl-8	83.60	100.00	0.91	88.94	100	0.94
bl-9	97.50	60.22	0.75	99.06	82.25	0.90
bl-10	61.40	78.60	0.69	86.61	99.85	0.93
bl-11	98.30	62.71	0.77	95.47	81.91	0.88
bl-12	88.30	97.97	0.93	83.45	91.48	0.87
bl-13	0.00	0.00	0.00	100	100	1.00
bl-14	76.90	97.67	0.86	84.29	95.6	0.90
bl-15	92.40	96.51	0.94	98.6	99.09	0.99
bl-16	55.60	96.98	0.71	91.15	96.04	0.94
bl-17	88.10	81.92	0.85	87.28	94.74	0.91
bl-18	21.30	89.20	0.34	81.88	98.72	0.90
Mean	59.1	82.2	0.63	83.85	92.92	0.88
STD	32.01	23.81	0.28	10.69	7.84	0.08

In this proposed method, the summarization threshold Thr_{kf} is set to the total number of key frames suggested by the ground truth for each video. It is evaluated that the introduced scheme generates more accurate results if $Thr_{kf} \times 1.35$ frames are selected from the ranked sorted list of $A(1, \dots, T)$ where T is total number of frames in a video. The results of precision, recall, and F_1 -measure of GMM based method and the proposed method are shown in Table I.

It is observed from Table I that the mean values of the precision, recall, and F_1 measure of GMM based method are 59.1, 82.2, and 0.63 respectively. On the other hand, the proposed method shows enhanced performance compared to GMM based method with mean precision (83.6), recall (94.2), and F_1 -measure (0.88). In addition to this, *standard deviations* (STDs) of precision, recall, and F_1 -measure of the proposed method are much lower compared to the GMM based method (see Table I). This indicates that the proposed method not only performs in higher accuracy but also variance of the performance is more consistent in the different videos compared to the GMM based method [1].

The graphical representations of F_1 -measure of GMM based method (F-GMM), the proposed method using only foreground objects (f-foreground), spatial motion (f-spatial), motion in frequency domain (f-frequency), and combining all these features (f-proposed) are shown in Fig. 5. After examining this graph, it is obvious that the proposed using only foreground objects shows less performance in bl-2, bl-4, and bl-12 videos compared to GMM based method. If only spatial motion is considered, it fails to outperform in bl-12 and bl-17 videos. Again, the proposed method applying only phase correlation technique shows poor result in bl-2, bl-12, and bl-17 videos. Therefore, the proposed approach combines all these features and performs superior to GMM based method in 18 videos. However, the proposed technique fails to perform better for bl-12 video compared to the GMM based method [1].

Fig. 5. F_1 -measure of the proposed and GMM based approach

After observing the key frames extracted by the proposed method for bl-12 video, the reasons of failure have been explored. In bl-12 video, there are some frames with significant object, and motion, however, they are not selected as ground truth frames by [1]. Again, although there is no foreground object and/or motion, in some frames, they are considered as ground truth. For example, frame no 4083, 4120, and 4563 contain sufficient amount of object, and motion as shown in first row of Fig. 6. In these frames, it is clearly visible that a person is working near the door. However, these frames are not selected as ground truth (key frames). On the other hand, there is no object, and significant motion exist in frame no 12615, 12675 and 12750 (see the

second row of Fig. 6). Nonetheless, they are selected as key frames (ground truth). There is no explanation found about this incident in [1].

Fig. 6. Sample frames of bl-12 are not selected as ground truth (first row) and considered as key frame (second row).

After evaluating the proposed method quantitatively and qualitatively, it is revealed that the proposed method based on foreground objects, and motion in spatial and frequency perform superior to the state-of-the-art GMM based method [1].

5 Conclusion

In this paper, a novel framework is proposed to summarize surveillance video combining foreground object along with motion information in spatial and frequency domain. According to [1], foreground objects usually contain details information of the video contents. Moreover, human being naturally gives more attention to object motion in a video [2]. Therefore, there two important properties of a video are included in this approach. To include motion information in frequency domain, phase correlation technique [4] is applied. To the best of our knowledge, phase correlation technique is applied for the first time for video summarization. Extensive experimental results reveal that the proposed method outperforms the state-of-the-art method.

References

1. Ou, S., LEE, C., Somayazulu, V., Chen, Y., Chien, S.: On-line Multi-view Video Summarization for Wireless Video Sensor Network. *IEEE J. Sel. Top. Signal Process.* 9, 165–179 (2015).
2. Gao, D., Mahadevan, V., Vasconcelos, N.: On the plausibility of the discriminant center-surround hypothesis for visual saliency. *J. Vis.* 8, 1–18 (2008).
3. Paul, M., Lin, W., Lau, C., Lee, B.: Explore and model better I-frames for video coding. *IEEE Trans. Circuits Syst. Video Technol.* 21, 1242–1254 (2011).
4. Paul, M., Lin, W., Lau, C.T., Lee, B.-S.: Direct intermode selection for H.264 video coding using phase correlation. *IEEE Trans. image Process.* 20, 461–73 (2011).

5. Lee, Y.J., Ghosh, J., Grauman, K.: Discovering Important People and Objects for Egocentric Video Summarization. *IEEE Conf. Comput. Vis. Pattern Recognit.* 1346–1353 (2012).
6. Lu, Z., Grauman, K.: Story-Driven Summarization for Egocentric Video. *IEEE Conf. Comput. Vis. Pattern Recognit.* 2714–2721 (2013).
7. Xu, J., Mukherjee, L., Li, Y., Warner, J., Rehg, J.M., Singh, V.: Gaze-enabled Egocentric Video Summarization via Constrained Submodular Maximization. *IEEE Conf. Comput. Vis. Pattern Recognit.* 2235–2244 (2015).
8. Gygli, M., Grabner, H., Gool, L. Van: Video Summarization by Learning Submodular Mixtures of Objectives. *IEEE Conf. Comput. Vis. Pattern Recognit.* 3090–3098 (2015).
9. Liu, Y., Liu, H., Sun, F.: Outlier-attenuating summarization for user-generated-video. *IEEE Int. Conf. Multimed. Expo.* 1 – 6 (2014).
10. Khosla, A., Hamid, R.: Large-scale video summarization using web-image priors. *IEEE Conf. Comput. Vis. Pattern Recognit.* 2698 – 2705 (2013).
11. Evangelopoulos, G.: Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Trans. Multimed.* 15, 1553–1568 (2013).
12. Tsai, C., Kang, L.: Scene-Based Movie Summarization Via Role-Community Networks. *IEEE Trans. Circuits Syst. Video Technol.* 23, 1927–1940 (2013).
13. Sawada, T., Toyoura, M., Mao, X.: Film Comic Generation with Eye Tracking. *Adv. Multimed. Model.* 467–478 (2013).
14. Schoeffmann, K., Del Fabro, M., Szkaliczki, T., Böszörményi, L., Keckstein, J.: Keyframe extraction in endoscopic video. *Multimed. Tools Appl.* (2014).
15. Spyrou, E., Diamantis, D., Iakovidis, D.K.: Panoramic Visual Summaries for Efficient Reading of Capsule Endoscopy Videos. 2013 8th Int. Work. Semant. Soc. Media Adapt. Pers. 41–46 (2013).
16. Mehmood, I., Sajjad, M., Baik, S.W.: Video summarization based tele-endoscopy: a service to efficiently manage visual data generated during wireless capsule endoscopy procedure. *J. Med. Syst.* 38, 109 (2014).
17. Ismail, M. Ben: Endoscopy video summarization based on unsupervised learning and feature discrimination. *Vis. Commun. Image Process.* 1–6 (2013).
18. Fu, W., Wang, J., Zhao, C., Lu, H., Ma, S.: Object-centered narratives for video surveillance. *IEEE Int. Conf. Image Process.* 29–32 (2012).
19. Sun, L., Ai, H., Lao, S.: The dynamic VideoBook: A hierarchical summarization for surveillance video. *IEEE Int. Conf. Image Process.* 3963–3966 (2013).
20. Wang, Y., Kato, J.: A distance metric learning based summarization system for nursery school surveillance video. *IEEE Int. Conf. Image Process.* 37–40 (2012).
21. Mehmood, I., Sajjad, M., Ejaz, W., Wook, S.: Saliency-directed prioritization of visual data in wireless surveillance networks. *Inf. Fusion.* 24, 16–30 (2015).
22. Huang, C., Chung, P.J.: Maximum a Posteriori Probability Estimation for Online Surveillance Video Synopsis. *IEEE Trans. Circuits Syst. Video Technol.* 24, 1417–1429 (2014).
23. Chakraborty, S., Paul, M.: A novel video coding scheme using a scene adaptive non-parametric background model. *IEEE 16th Int. Work. Multimed. Signal Process.* 1–6 (2014).
24. Haque, M., Murshed, M., Paul, M.: A hybrid object detection technique from dynamic background using Gaussian mixture models. *IEEE 10th Work. Multimed. Signal Process.* 915 – 920 (2008).
25. Ahuja, N., Briassouli, A.: Joint Spatial and Frequency Domain Motion Analysis. *Int. Conf. Autom. Face Gesture Recognit.* 203–208 (2006).
26. Briassouli, A., Ahuja, N.: Integration of Frequency and Space for Multiple Motion Estimation and Shape-Independent Object Segmentation. 657–669 (2008).
27. Paul, M., Frater, M.R., Arnold, J.F.: An Efficient Mode Selection Prior to the Actual Encoding for H . 264 / AVC Encoder. *IEEE Trans. Multimed.* 11, 581–588 (2009).
28. Han, J., Kamber, M., Pei, J.: Data Mining, Southeast Asia Edition: Concepts and Techniques. Morgan Kaufmann (2006).