Image Segmentation and Geometric Feature Based Approach for Fast Video Summarization of Surveillance Videos

Raju Dhanakshirur Rohan(✉), Zeba ara Patel, Smita C. Yadavannavar,
C. Sujata, and Uma Mudengudi

KLE Technological University, Hubli, India
rdshirur.cstaff@iitd.ac.in, rohanrd28296@gmail.com

Abstract. In this paper, we propose a geometric feature and frame segmentation based approach for video summarization. Video summarization aims to generate a summarized video with all the salient activities of the input video. We propose to retain the salient frames towards generation of video summary. We detect saliency in foreground and background of the image separately. We propose to model the image as MRF (Markov Random Field) and use MAP (Maximum a-posteriori) as final solution to segment the image into foreground and background. The salient frame is defined by the variation in feature descriptors using the geometric features. We propose to combine the probabilities of foreground and background segments being salient using DSCR (Dempster Shafer Combination Rule). We consider the summarized video as a combination of salient frames for a user defined time. We demonstrate the results using several videos in BL-7F dataset and compare the same with state of art techniques using retention ratio and condensation ratio as quality parameters.

Keywords: Video summarization · Graph cut · Geometric features · Dempster Shafer Combination Rule (DSCR)

1 Introduction

In this paper, we propose a feature based approach for video summarization. Video summarization aims to generate a summarized video with all the salient activities of the input video. We propose to retain the salient frames towards generation of video summary. Due to huge content available in the internet in form of videos, searching the most appropriate and effective information is time consuming for the user. Video summarization is the method to generate a short video containing the most effective frames of the available video. Video summarization finds its applications in video surveillance systems [3,24,26] in which computer vision algorithms, such as tracking, behavior analysis, and object segmentation, are integrated in cameras and/or servers. It also finds its applications in movie trailer generation, sport summary generation etc.

C. Arora and K. Mitra (Eds.): WCVA 2018, CCIS 1019, pp. 79–88, 2019.
https://doi.org/10.1007/978-981-15-1387-9_7

Many researchers have worked on video summarization. Objects and people within a video play vital role for video summarization [18]. This is because, we generally represent the events in a video by people/objects and their activities. Moreover, people/objects in the video have the high-level of the semantic perception. Also, along with this, humans usually are more attentive towards the moving objects in a video [6,8]. However, researchers consider the problem of extracting the moving objects from a video that has changes in illumination, high noise, bad contrast and multimodal environment as a challenging problem [2,22]. However, in the videos with low contrast, the edges of objects are given higher prominence [22]. Also, this method is further sensitive to the variation in the shape and position of the object.

To resolve these problems, we can apply the theory of edge-segments (i.e. groups of connected sequential edge pixels) [8]. But, [4,6] claims that the edge-segments based methods fail when the video has shape matching errors or local shape distortion. The *state of the art* methods for object detection use the ellipses or circles to represent curve fragments [6]. Even then, the problem persists if the video has low illumination [6]. Also, we observe that in real world, an object can take up any shapes other than circular, elliptical, parabolic, or hyperbolic curve. Thus, the object detection methods that approximate the shapes to the primitive structures fail in such circumstances. However, we can easily fit a conic part for simple objects.

In [10], authors use a set of similar objects to build a model for summarization. Authors in [16] present a part-based object movement framework. Authors in [14] apply object bank and object-like windows to extract the objects and then they perform story based video summarization. Authors in [5] propose a complementary background model. Pixel-based motion energy and edge features are combined in [23] for summarization. Authors in [12] propose a background subtraction method to detect foreground objects for video summarization. Authors in [13] modify the previous idea for Aggregated Channel Features (ACF) detection and a background subtraction technique for object detection.

In [21], authors propose a video summarization technique by merging three multi-modal human visual sensitive features, namely, motion information, foreground objects, and visual saliency.

Authors in [15] propose a min-cut based approach for generating storyboard. Also authors in [15] modify the previous idea and propose a Bayesian foraging technique for objects and their activities detection to summarize a video. The grid background model is applied in [7]. Authors of [17], deploy a key-point matching technique for video segmentation. Authors in [8] apply Spatio-temporal slices to select the states of the object motion.

Authors in [9] propose a learning based approach for video summarization. They describe the Objects in a video by Histogram of Optical Flow Orientations and then apply a SVM based classifier. Authors in [19] propose unsupervised framework via joint embedding and sparse representative selection for video summarization. The objective function is two-stream in nature. The first objective is to capture multi-view correlations using an embedding, that assists in extracting a diverse set of representatives and the second is to use $L-2$ norm to model the sparsity while selecting representative shots for the summary. Authors in [28] uses RNN to exploit the temporal relationship between frames for saliency detection. Authors in [20] makes use of fully connected neural network for video summarization. However all these techniques need high computational capability which makes it highly impossible for low-cost real time implementation.

Authors in [27] apply a modularity cut algorithm to track objects for summary generation. Gaussian Mixture model based approach is employed in [4]. The key frames are selected based on the parameters of cluster. Authors in [4,6], use geometric primitives (such as lines, arcs) for distinguishable descriptors than edge-pixels or edge-segments.

These primitives are independent of the size of the object, and also they are efficient for matching and comparisons. They are also invariant to scale and viewpoint changes. Thus, these geometric primitives represent objects with complex shapes and structures effectively. Also, they are useful in cognitive system [11].

In this paper, we propose to fuse the techniques of foreground/background segmentation and the use of geometric features for saliency detection in order to achieve video summarization. Towards this, we make the following contributions:

- We propose to detect the saliency of a frame by detecting the saliency of its foreground and background separately and then combine the probabilities of foreground and background being salient to check the saliency of a frame.
  - We propose to model the image as an MRF and use MAP using graph-cut as final solution for foreground and background segmentation.
  - We propose to combine the probabilities of foreground and background being salient using the Dempster Shafer Combination rule (DSCR).
- We propose to use the changes in the variant of the geometric features (such as lines, arcs) to decide the saliency of a frame. For efficient extraction of geometric primitives,
  - We propose to extract the PCA features to detect the principle components of foreground and background frames.
  - We convert the image from RGB to YCbCr and compute PCA on Y channel of the frame to retain the chromic information.
- We demonstrate the results using the BL-7F dataset and compare the results using the state-of-the-art techniques with the help of the quantitative parameters such as condensation ratio and retention ratio.

2 Proposed Framework

We demonstrate the proposed framework in Fig. 1. We propose to detect the saliency of a frame by detecting the saliency of its foreground and background separately. We propose to detect the changes in the PCA and Geometric Primitives such as lines and contours by computing difference in standard deviation of the segments and comparing the difference with a heuristically set threshold.

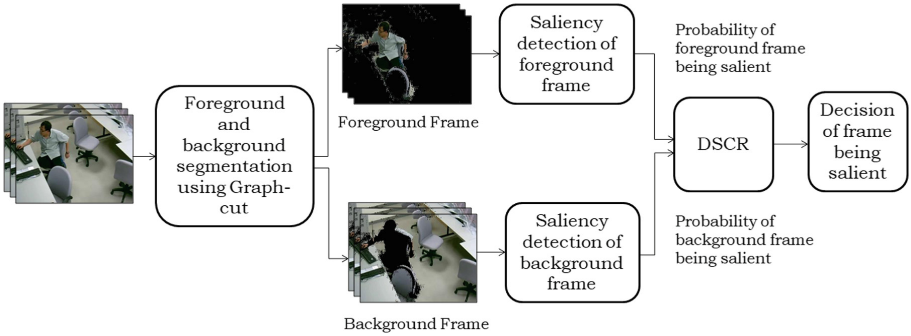
Foreground and background segmentation using Graph-cut
Foreground Frame
Background Frame
Saliency detection of foreground frame
Saliency detection of background frame
Probability of foreground frame being salient
Probability of background frame being salient
DSCR
Decision of frame being salient

Fig. 1. Proposed framework

The threshold for background is kept much lower as compared to that of foreground with an intuition that any small motion in background is much significant as compared to small motion in foreground. We find separate probabilities for foreground and background segments being salient. We combine the two probabilities using DSCR to obtain joint probability. We decide if the given frame is salient based on the decision boundary set upon the joint probability.

2.1 Foreground and Background Segmantation

We propose to separate the foreground of the scene from the background using Energy Minimization via Graphcut. We model every frame as MRF (Markov Random Field) and use MAP (Maxima A Posteriori) estimate as the final solution. In this framework, we use the grid graph containing image pixels for MRF. Here, we try to find the labelling for the pixels in the image f with minimum energy.

$$E(f) = Esmooth(f) + Edata(f)$$

Where $Edata(f)$ is defined by,

$$Edata(f) = \sum_{p \in P} Dp(fp)$$

Here $Esmooth(f)$ measures the extent to which f is not piecewise smooth, whereas the $Edata(f)$ measures the total disagreement between f and the observed data. Researchers have proposed many different energy functions. The form of $Esmooth(f)$ is typically,

$$Esmooth(f) = \sum_{p,q \in N} u\{p,q\}.T(f_p \neq f_q)$$

here, T is indicator function. It will output 1 if the input condition is true. We use Potts Model in which, discontinuities between any pair of labels are penalized equally. This is, in some sense, the simplest discontinuity preserving model.

We then obtain the two segments of the image, one corresponding to foreground and the other corresponding to background. The foreground and background segmentation for two datasets is shown in Fig. 2.

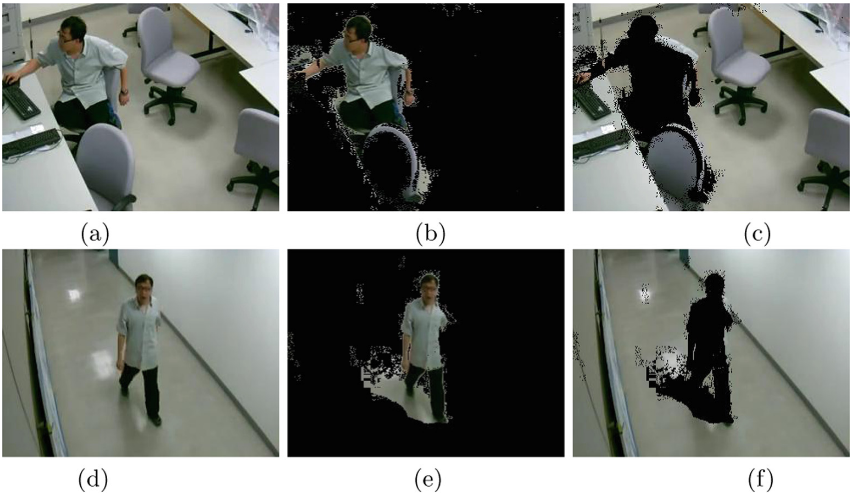

Fig. 2. Segmentation of image into foreground and background frames: (a), (d) are original images. (b), (e) are the corresponding foreground frames. (c), (f) are the corresponding background frames

2.2 Saliency Detection of Foreground and Background Frames

We demonstrate the saliency detection block in Fig. 3. The input for the saliency detection is the segmented frame (Foreground or background). We propose to use the changes in the variant of geometric primitives to decide the saliency of a frame. We extract the variant of geometric features, named the frame feature descriptors (FFD). The process of FFD extraction is demonstrated in Fig. 4. We then find the standard deviation between the extracted feature vectors of the consecutive frames. The probability of frame being salient is decided by the extent with which the obtained standard deviation is greater than a heuristically set threshold.

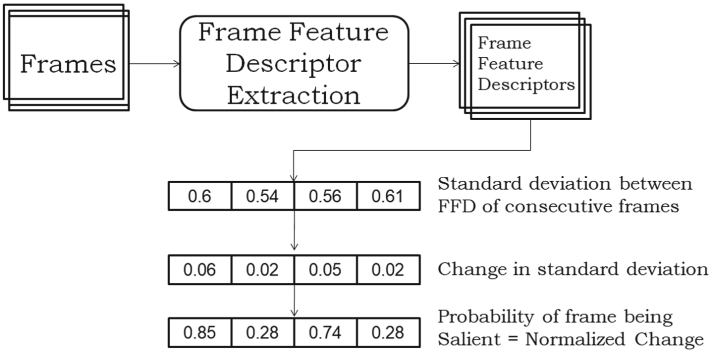
Frames
Frame Feature Descriptor Extraction
Frame Feature Descriptors
0.6
0.54
0.56
0.61
Standard deviation between FFD of consecutive frames
0.06
0.02
0.05
0.02
Change in standard deviation
0.85
0.28
0.74
0.28
Probability of frame being Salient = Normalized Change

Fig. 3. Saliency detection of foreground and background frames

2.3 Extraction of Frame Feature Descriptors (FFD)

The process of FFD extraction is demonstrated in Fig. 4. We convert the RGB frames of the video to YCbCr to retain the colour information. We apply PCA on 'Y' channel of the image to get PCA transformed 'Y' channel. We convert the output to RGB to obtain the images with enhanced principal components. We extract geometric features from images with enhanced principal components.

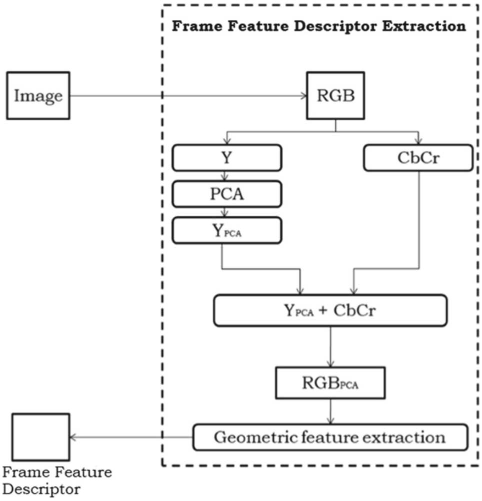
Frame Feature Descriptor Extraction
Image
RGB
Y
CbCr
PCA
YPCA
YPCA + CbCr
RGBPCA
Geometric feature extraction
Frame Feature Descriptor

Fig. 4. Extraction of frame feature descriptors (FFD)

We extract the objects present in the salient frames as geometric features. We apply Canny edge detection to find the edges of the objects. Using these edges we find the geometric features like line segments and contours. Contours represent the largest area in the frames. Hence we find the difference in the frames by monitoring the change in the position of the line segments present in the area of the contours.

2.4 Joint Probabability Using DSCR

We combine the two probabilities using Dempster Shafer Combination Rule (DSCR) to obtain the joint probability. We decide if the given frame is salient based on the decision boundary set upon the joint probability. Let P_1 and P_2 be the probabilities to be combined. DSCR combines two hypothesis consisting of three parameters, mass of belief, mass of disbelief and mass of uncertainty rather than two probabilities. We construct hypothesis, hyp_1 and $hyp2$ as a set of mass of belief $(m(b))$, disbelief $(m(d))$ and ambiguity $(m(u))$ respectively. We set mass of belief $(m_1(b))$ for hyp_1 as P_1 and mass of belief $(m_2(b))$ for hyp_2 be P_2. We assume mass of disbelief $(m_1(d))$ for hyp_1 and hyp_2 to be 0 and mass of ambiguity ($m_1(u)$ and $m_2(u)$) for hyp_1 and hyp_2 as $1-P_1$ and $1-P_2$ respectively. We combine hyp_1 and hyp_2 using combination table as shown in Table 1.

Table 1. Combination table

$\cap$	m_1^{belief}	$m_1^{disbelief}$	$m_1^{ambiguity}$
m_2^{belief}	ψ_1	$\emptyset$	ψ_1
$m_2^{disbelief}$	$\emptyset$	ψ_2	ψ_1
$m_2^{ambiguity}$	ψ_2	ψ_2	Ω

In the combination table, the product of mass of belief of one hypothesis and mass of disbelief of other hypothesis gives rise to conflict and is represented by $\emptyset$. The product of mass of belief and mass of belief or the product of mass of belief and mass of uncertainty represents joint belief and is represented by ψ_1. Similarly ψ_2 represents the joint disbelief.

The Combined belief of two evidences is considered as Joint probabilities and is given by:

$$JointProbability = \frac{\sum \psi_1}{1 - \sum \emptyset}$$

We decide if the given frame is salient based on the decision boundary set upon the joint probability. The advantage of using DSCR for combining the two probabilities is that it emphasis of the fact that if P_1 is the probability of frame being salient, then $1-P_1$ need not be the probability of frame being non-salient. It can be uncertainty as well.

3 Results and Discussions

We evaluate our approach using BL-7F dataset. In this dataset, 19 surveillance videos are taken from fixed surveillance cameras located in the seventh floor of the BarryLam Building in the National Taiwan University. Each video consists of 12,900 frames with a duration of 7 min and 10 s. We compare our results using Retention ratio and Condensation ratio as evaluation metrics.

Table 2. Comparison of condensation ratio (in percentage) of the proposed method with the different state-of the art techniques [1,18,25] for different surveillance videos. Here RR = retention ratio and is seen to be 1 for the results, unless mentioned.

Video	Duration of given video (min:sec)	Duration of summarized video (min:sec)	Valdes et al. IAMIS 2008	Almedia et al. ISM 2010	S. Ou et al. JSTSP 2015	Proposed framework
bl-0	07:10	00:03	49.53	51.60	93.02	99.29
bl-1	07:10	00:08	36.27	91.6	83.02	97.97
bl-2	07:10	00:09	61.8	50	75.34	97.92
bl-3	07:10	00:02	56.744	98.83	96.27	99.45
bl-4	07:10	00:13	64.41	88.37	80.69	97.03
bl-5	07:10	00:04	36.27	90.46	85.58	99.05
bl-6	07:10	00:05	22.32	100 (RR = 0)	95.35	98.8
bl-7	07:10	00:05	30.93	95.34	88.37	98.8
bl-8	07:10	00:01	22.32	99.3	98.37	99.74
bl-9	07:10	00:09	17.9	95.58	90.93	98.01
bl-10	07:10	00:08	93.48	93.48	74.19	99
bl-11	07:10	00:07	68.6	62.09	73.95	98.31
bl-12	07:10	00:04	48.37	50	69.06	96.42
bl-14	07:10	00:14	63.72	94.88	83.25	96.62
bl-15	07:10	00:07	94.65	89.53	84.18	98.31
bl-16	07:10	00:26	89.53	89.53	76.15	93.85
bl-17	07:10	00:28	61.16	51.16	77.67	93.35
bl-18	07:10	00:03	61.62	95.11	85.16	99.24

Retention ratio is the ratio of number of objects in the summarized video to the number of objects in the original video.

$$RR = \frac{number\ of\ objects\ in\ summarized\ video}{number\ of\ objects\ in\ input\ video}$$

Condensation ratio is the ratio of length of summarized video to length of the input video.

$$CR = (1 - \frac{length\ of\ summarized\ video}{length\ of\ input\ video}) * 100$$

We find that the proposed method gives better results as compared to results obtained from the other state-of-the-art techniques. Retention ratio for the proposed method is unity for all videos and Condensation ratios are also very high compared to the existing methods. The comparison of the condensation ratio (in percentage) of the proposed method with the different state-of the art techniques [1,18,25] for different surveillance videos is demonstrated in Table 2.

4 Conclusions

In this paper, we have proposed a geometric feature and frame segmentation based approach for video summarization. We detected saliency in foreground and background of the image separately. We proposed to model the image as MRF (Markov Random Field) and use MAP (Maximum a-posteriori) as final solution to segment the image into foreground and background. The salient frame was effectively defined by the variation in feature descriptors using variant of geometric features. We proposed to combine the probabilities of foreground and background segments being salient using DSCR (Dempster Shafer Combination Rule). We modelled the summarized video as a combination of salient frames for a user defined time. We have demonstrated the results using several videos in BL-7F dataset and compared the same with state of art techniques using retention ratio and condensation ratio as quality parameters to prove the superiority of the proposed method over the other algorithms.

References

1. Almeida, J., Torres, R.D.S., Leite, N.J.: Rapid video summarization on compressed video. In: 2010 IEEE International Symposium on Multimedia, pp. 113–120, December 2010
2. Bagheri, S., Zheng, J.Y.: Temporal mapping of surveillance video. In: 2014 22nd International Conference on Pattern Recognition, pp. 4128–4133, August 2014
3. Chan, W.K., Chang, J.Y., Chen, T.W., Tseng, Y.H., Chien, S.Y.: Efficient content analysis engine for visual surveillance network. IEEE Trans. Circuits Syst. Video Technol. **19**(5), 693–703 (2009)
4. Chang, W., Lee, S.Y.: Description of shape patterns using circular arcs for object detection. IET Comput. Vis. **7**(2), 90–104 (2013)
5. Chen, S.C., et al.: Target-driven video summarization in a camera network. In: 2013 IEEE International Conference on Image Processing, pp. 3577–3581, September 2013
6. Chia, A.Y.S., Rajan, D., Leung, M.K., Rahardja, S.: Object recognition by discriminative combinations of line segments, ellipses, and appearance features. IEEE Trans. Pattern Anal. Mach. Intell. **34**(9), 1758–1772 (2012)
7. Cui, Y., Liu, W., Dong, S.: A time-slice optimization based weak feature association algorithm for video condensation. Multimedia Tools Appl. **75**, 17515–17530 (2016)
8. Kovesi, P.D.: MATLAB and Octave functions for computer vision and image processing, January 2000
9. Fan, C.T., Wang, Y.K., Huang, C.R.: Heterogeneous information fusion and visualization for a large-scale intelligent video surveillance system. IEEE Trans. Syst. Man Cybern. Syst. **47**(4), 593–604 (2017)
10. Fei, M., Jiang, W., Mao, W.: Memorable and rich video summarization. J. Vis. Commun. Image Represent. **42**(C), 207–217 (2017)
11. Hu, R.X., Jia, W., Ling, H., Zhao, Y., Gui, J.: Angular pattern and binary angular pattern for shape retrieval. IEEE Trans. Image Process. **23**(3), 1118–1127 (2014)
12. Huang, C.R., Chung, P.C.J., Yang, D.K., Chen, H.C., Huang, G.J.: Maximum a posteriori probability estimation for online surveillance video synopsis. IEEE Trans. Circuits Syst. Video Technol. **24**(8), 1417–1429 (2014)

13. Li, X., Wang, Z., Lu, X.: Surveillance video synopsis via scaling down objects. IEEE Trans. Image Process. **25**(2), 740–755 (2016)
14. Lu, Z., Grauman, K.: Story-driven summarization for egocentric video. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2714–2721, June 2013
15. Napoletano, P., Boccignone, G., Tisato, F.: Attentive monitoring of multiple video streams driven by a Bayesian foraging strategy. IEEE Trans. Image Process. **24**(11), 3266–3281 (2015)
16. Nie, Y., Sun, H., Li, P., Xiao, C., Ma, K.L.: Object movements synopsis viapart assembling and stitching. IEEE Trans. Visual Comput. Graph. **20**(9), 1303–1315 (2014)
17. Otani, M., Nakashima, Y., Sato, T., Yokoya, N.: Textual description-based video summarization for video blogs. In: 2015 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6, June 2015
18. Ou, S.H., Lee, C.H., Somayazulu, V.S., Chen, Y.K., Chien, S.Y.: On-line multiview video summarization for wireless video sensor network. IEEE J. Sel. Top. Signal Process. **9**(1), 165–179 (2015)
19. Panda, R., Roy-Chowdhury, A.K.: Multi-view surveillance video summarization via joint embedding and sparse optimization. IEEE Trans. Multimedia **19**(9), 2010–2021 (2017)
20. Rochan, M., Ye, L., Wang, Y.: Video summarization using fully convolutional sequence networks. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11216, pp. 358–374. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01258-8_22
21. Salehin, M.M., Paul, M.: Adaptive fusion of human visual sensitive features for surveillance video summarization. J. Opt. Soc. Am. A: Opt. Image Sci. Vis. **34**(5), 814–826 (2017)
22. Salehin, M., Zheng, L., Gao, J.: Conics detection method based on Pascal's theorem. In: Proceedings of the 10th International Conference on Computer Vision Theory and Applications - Volume 1: VISAPP, (VISIGRAPP 2015), pp. 491–497. INSTICC, SciTePress (2015)
23. Shih, H.C.: A novel attention-based key-frame determination method. IEEE Trans. Broadcast. **59**(3), 556–562 (2013)
24. Taj, M., Cavallaro, A.: Distributed and decentralized multicamera tracking. IEEE Signal Process. Mag. **28**(3), 46–58 (2011)
25. Valdés, V., Martínez, J.M.: On-line video summarization based on signature-based junk and redundancy filtering. In: 2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services, pp. 88–91, May 2008
26. Valera, M., Velastin, S.A.: Intelligent distributed surveillance systems: a review. IEE Proc. - Vis. Image Signal Process. **152**(2), 192–204 (2005)
27. Zhang, S., Roy-Chowdhury, A.K.: Video summarization through change detection in a non-overlapping camera network. In: 2015 IEEE International Conference on Image Processing (ICIP), pp. 3832–3836, September 2015
28. Zhao, B., Li, X., Lu, X.: Hierarchical recurrent neural network for video summarization. In: ACM Multimedia (2017)