World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:10, No:9, 2016

Surveillance Video Summarization Based on Histogram Differencing and Sum Conditional Variance

Nada Jasim Habeeb, Rana Saad Mohammed, Muntaha Khudair Abbass

Abstract**—**For more efficient and fast video summarization, this paper presents a surveillance video summarization method. The presented method works to improve video summarization technique. This method depends on temporal differencing to extract most important data from large video stream. This method uses histogram differencing and Sum Conditional Variance which is robust against to illumination variations in order to extract motion objects. The experimental results showed that the presented method gives better output compared with temporal differencing based summarization techniques.

Keywords**—**Temporal differencing, video summarization, histogram differencing, sum conditional variance.

I. Introduction

FOR security integration, video surveillance plays an increasingly important role for improving security. With the large amounts of video data being generated, large storage is needed. Traditional video surveillance technology based on hard disk or DVD disks suffers from many challenges. The limited throughput of these devices leads a limited storage problem for storing videos in the future. Without dropping frames, the servers lack the ability to serve simultaneous reads and writes. The surveillance video contains rich information; it should be analyzed in order to get the useful information. It is very important to summarize this type of video for facilitation of many image or video processing operations. Video summarization is a process which converts large videos to shorter videos containing useful information. Two basic types of video summarization are dynamic video summary and static video summary. Dynamic video is a shorter video of the original one containing the important information. The static video summary is a set of key-frames of a video or the salient frames [1]–[5] Key-frames extraction based approaches need to detect shot boundaries before summarization process. In surveillance videos, it is difficult to determine shot boundaries; therefore, Key-frames extraction is not suitable to be used for surveillance video summarization. In addition, surveillance video is continually generating video data over time, so surveillance video is considered as a dynamic video [6].

To generate a good video summary requires a good understanding of the structure of the original video and semantic content and summary video frames must be concise visually [7], [8].

Many approaches have been proposed for reducing the size of surveillance video containing the most important data. The best and most recently description of review to work of video summarization approaches can be found in [9]. Yang and Wei proposed a method for video summarization based on genetic algorithm. This method can be defined as flows. First, frames differences are computed. Second, the least similar frames from original video by using the technique of color histograms are selected. Finally, optimized frames by Fitness function which is defined for genetic algorithm are searched [10]. Lei et al. presented a video summary extraction approach. First, the original video is segmented into sub shots based on semantic structure by using distance measures, and then key-frames are extracted in each sub shot by Singular value decomposition [11]. Song et al. proposed surveillance video summarization approach based on event detection. First, the trajectories of motion objects are computed and then event detection is determined. Finally, video summary is generated with the most interested event and small number of frames [12]. Wang and Kato proposed a method which produces learned distance metric which can measure the similarity between videos. The metric is fused with supervised classification and unsupervised clustering in order to summarize the original video depending on events [13]. Li et al. proposed an approach for reducing the collisions in video sequence. First active objects during time are shifted to compact the original video. Second, when collisions occur, the objects sizes are reduced. The geometric object centroids are not chanced [14]. Ren et al. proposed a method for video summarization which is implemented in compressed-domain. Firstly, human event detection is performed. Fuzzy decision is used to classify frames into fuzzy domain. Secondly, Haar-like features are used for human objects detection. Finally, the event levels for each frame are determined and frames with same category are grouped to form the final video summary [15]. Astelo and Guillermo proposed a method for video summarization by using a color based feature extraction and clustering techniques. In this method, there is no need to shot extraction [16]. Lee et al. presented a method for video summarization with four stages. First, face detection is

Nada Jasim Habeeb is with Middle Technical University, Technical College of Management, Baghdad, Iraq (corresponding author; phone: +964 7902227895, e-mail: nadaj2013@gmail.com.).

Rana Saad Mohammed is with Computer Science Dept., Education College-Al-Mustansiriyah University, Baghdad, Iraq (e-mail: Ranasaad2014@gmail.com).

Muntaha Khudair Abbass is with Middle Technical University, Technical College of Management, Baghdad, Iraq (e-mail: mukeab2014@yahoo.com).

achieved using "Adaboost" algorithm. Decomposition of the face regions into basis and corresponding coefficients using Non-negative Matrix Factorization method is performed in the second stage. In the third stage, Support Vector Machine (SVM) classification is applied on the coefficients. Finally, frames are extracted containing the target person from the previous stage [17]. Luo presented a video summarization method using internet of things (IoT) information. Video summary frames by using IoT information to build a background model are selected. Key frames are extracted by computing differences between video frames and the background and then extract the important information. The final video summary is generated by clustering the extracted key frames with the most useful information [18].

The previous video summarization methods used similarity metrics without taking into consideration the illumination variances in the color intensities posed to the motion object during time. Due to these variations in the color, there is no guarantee that the color will be the same object in all frames in video sequence.

For more efficiency of time and space, in this paper, the development of summarization algorithm is presented using the combination of histogram differences and Sum Conditional Variance between two consecutive frames in video sequence. The goal is to improve video summarization of surveillance video considering illumination changes, and make it faster and accurate for many operations such as data mining, computer vision, image processing, video browsing, etc. Here we used Sum Conditional Variance as temporal differencing metric to detect motion objects and to solve the problem of the previous temporal differencing metrics.

The paper is organized as follows: In Section II, the background theory that are related to this work are described. Section III presents the method of this work to solve the current problem in video summarization techniques, while in Section IV, the performed experiments with comparative analysis between this approach and the existing methods are described. Finally, conclusions are given in Section V.

II. Background Theory

A. Temporal Differencing

Pixel differencing and histogram differencing are temporal differencing techniques. Pixel differencing is a process which can be used to detect motion objects by computing pixel-by-pixel difference between two or three consecutive frames ($f_i(x,y)$ and $f_{i-1}(x,y)$) in a video sequence. Pixel differencing method is used to extract motion objects. It can be defined by [19]:

$$(|f_i(x,y) - f_{i-1}(x,y)|) > Thr \quad (1)$$

where Thr is a specific threshold.

Histogram differencing is used to compare between the illuminations values of two or more consecutive frames. It is more efficient compared with the pixel differencing because it has less sensitive to motion and low computational complexity [20].

B. Sum Conditional Variance

Sum conditional variance (SCV) is a powerful similarity measure that is robust against global illumination variations. In the SCV function (see (2) and (3)), for every image at pixel x (the pixel coordinates), the value of s is found, where s is the parameter of transformation function $w(x, s)$ that minimize the SCV between the reference image Ir and the current image I [21], [22]:

$$SCV(s) = \sum_x (I(w(x,s)) - \widehat{Ir}_{(i,j)}(x))^2 \quad (2)$$

$$\widehat{Ir}(x) = E\left(I(w(x,s))|Ir_{(i,j)}(x)\right) \quad (3)$$

$E(.)$ is the expectation operation, (i,j) indicates the row and column of reference image. The w is one of transformation functions such as similarity, an affine, or holography transformation [23]. At each time, the reference image is adapted according to the illumination conditions of the current image I and it is replaced by the expected image computed by the expectation operator (3). The SCV was originally proposed for registration of multi-modal images. Then, the SCV was developed for visual tracking. It is invariant to illumination changes and has low computation cost.

C. Evaluation Metrics

The metrics for evaluation of the presented method compared with the existing methods can be summarized in Table I [24].

TABLE I
The Performance Evaluation Metrics for the Summarization Methods

Metric name	Format
Data compression ratio (DCR)	$\text{Compressed Ratio} = 1 - \frac{\text{Uncompressed size}}{\text{Ccompressed size}}$
Space saving	$\text{Space saving} = 1 - \frac{\text{Compressed size}}{\text{Uncompressed size}}$
Condensed Radio (CR)	$\text{CR} = \left[1 - \left(\frac{\text{number of output frames}}{\text{number ofinput frames}}\right)\right] \times 100$

III. The Proposed Method

Given a video sequence Vid contains n frames $(f_1, f_2, \dots, f_n)$. Let f_i and f_{i+1} be 2-D (n × m) matrices represent the current frame and the next frame in the video sequence. The histograms of the two consecutive frames are computed and the difference between them is calculated. Then, SCV similarity function between the consecutive frames is also computed. At each time, SCV adapts the next frame f_{i+1} to the illumination conditions or lighting changes of the current image f_i . Here, we used a dynamic threshold instead of using a global threshold for change detection between the two consecutive frames. In general, the dynamic threshold is defined as [25]:

$$Thr = \bar{x} + \alpha\sigma_x \quad (4)$$

where $\bar{x}$ and σ_x are the mean and the standard deviation for whole the sequence, α is a fitting parameter.

Algorithm (1) gives steps of this algorithm. The algorithm has two loops for reading the whole video sequence, one for generating the two dynamic thresholds, (Thr_1) is used for Histogram differences and (Thr_2) is for SCV differences. The next loop is used for generating video summary.

Algorithm (1): Single Video Summarization

Input: Vid is a video sequence, n is the number of frames, α is a constant.
Output: video summary
Steps:

1. Given video sequence (Vid) of frames $(f_1, f_2, \dots, f_n)$.
2. For *i*=1 to n
3. Read the successive two 2D frames $(f_i \; and \; f_{i+1})$ from Vid.
4. Compute Histogram difference (Hdiff) between the two frames using Euclidian distance:
$$h_1 = hist(f_i), h_2 = hist(f_{i+1}),$$
$$Hdiff_i = sqrt(sum((h_1 - h_2).\hat{}2))$$
5. Compute Sum of Conditional Variances (SCV) similarity measure between the two frames $(f_i \; and \; f_{I+1})$:
$$scv_i = SCV(f_i, f_{I+1})$$
6. End for
7. Compute the dynamic threshold (thr_1) for $Hdiff$:
$$thr_1 = Mean(Hdiff) + \alpha \times std(Hdiff),$$
8. Compute dynamic threshold (thr_2) for SCV:
$$thr_2 = Mean(scv) + \alpha \times std(scv)$$
9. For *i*=1 to n
10. Read the successive two frames $(f_i \; and \; f_{i+1}) \; from \; the \; Vid.$
11. Compute Histogram difference (Hdiff) between the two frames using Euclidian distance:
$$h_1 = hist(f_i), h_2 = hist(f_{i+1}),$$
$$Hdiff_i = sqrt(sum((h_1 - h_2).\hat{}2))$$
12. Compute Sum of Conditional Variances (SCV) similarity measure between the two frames $(f_i \; and \; f_{I+1})$:
$$scv_i = SCV(f_i, f_{I+1})$$
13. if ($Hdiff_i > thr_1$) and ($scv_i > thr_2$)
14. store the current frame f_i in the video summary
15. Otherwise, f_i is canceled
16. End if
17. End for

The proposed summarization method extracts meaningful information depending on motion object event detection. In general, the motion object detection is based on observing the change in the background of scene over time. For example, in Fig. 1, each point represents the variance of difference between each two consecutive frames. If the variance of differences have lowest values, it means that there is no moving object in the sequenced frames. The background will be presented in these frames and the spread of special-temporal vectors will be close to each other. In contrast, if there is a moving object in sequence of frames, the spatial-temporal will be spread fast in the space of their coordinates.

Fig. 1 shows an example of how to partition the video data stream and how to represent the motion information on a sequence of frames. After partitioning the input video into motion parts and no motion parts, frames containing motion objects are selected and stored based on the motion sequence frames (B) while no motion frames (A) are removed.

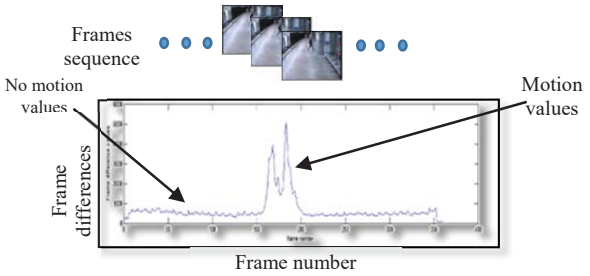


Fig. 1 Normal distribution of video pixels

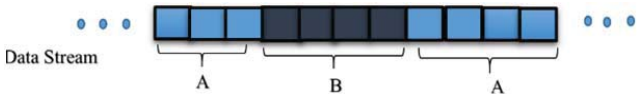


Fig. 2 General partitioning of data sequence

IV. Experiment Results

For testing our presented method, we used two samples from surveillance video streams one for indoor scene which is a metro scene and one for outdoor scene which is a road scene. From each one, we cut a sample with 1000 sequenced frames for a specific period of time. The proposed algorithm based on combination of sum of conditional variances (SCV) and histogram differencing has been implemented. The graph results of implementation of this algorithm are shown in Figs. 3 and 4 for the metro video and the road video, respectively.

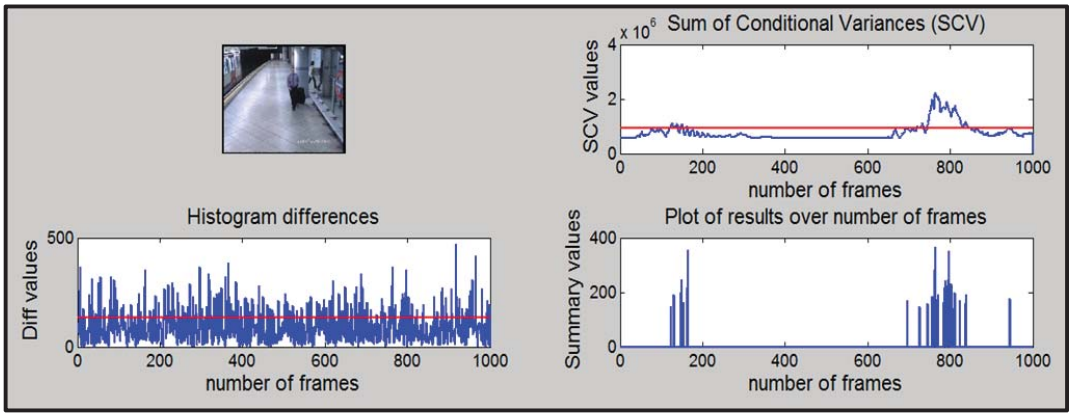
TABLE II
DCR, S_{SAVING} AND CR RESULTS COMPARISON BETWEEN THE TRADITIONAL METHODS AND THE PRESENTED METHOD FOR MERO VIDEO

Video/ methods		Frame-Diff	Histogram-Diff	SCV	Proposed method
Metro video sequence of 1000 frames	Video size before	15.5 MB	15.5 MB	15.5 MB	15.5 MB
	Video size after	4.21 MB	4.07 MB	2.70 MB	644 KB
	Number of frame before	1000	1000	1000	1000
	Number of frame after	260	259	167	36
	DCR	-2.6817	-2.8084	-4.7407	-23.6462
	S_{saving}	07284	0.7374	0.8258	0.9594
	CR	74%	74%	83%	96%

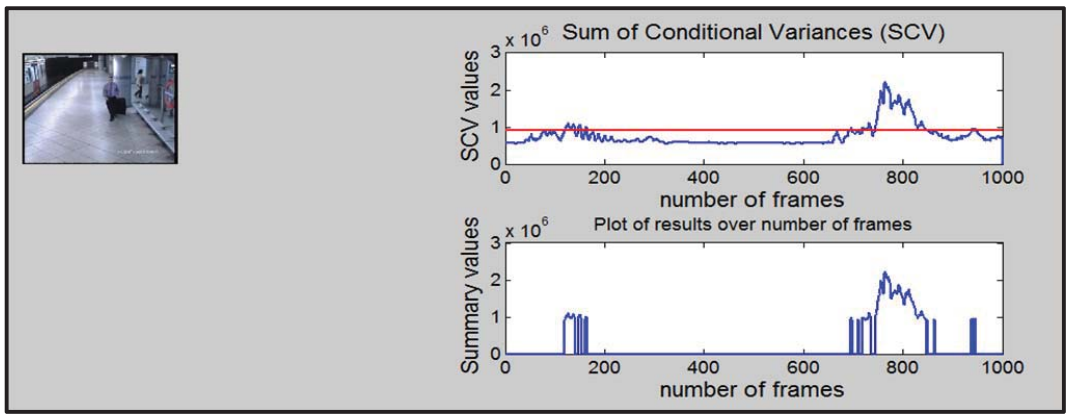
TABLE III
DCR, S_{Saving} AND CR RESULTS COMPARISON BETWEEN THE TRADITIONAL METHODS AND THE PRESENTED METHOD FOR ROAD VIDEO

Video/ methods		Frame-Diff	Histogram-Diff	SCV	Proposed method
Metro video sequence of 1000 frames	Video size before	25 MB	25 MB	25 MB	25 MB
	Video size after	7.09	6.60 MB	7.98	2.35 MB
	Number of frame before	1000	1000	1000	1000
	Number of frame after	284	262	318	92
	DCR	-2.5261	-2.7879	-2.1328	-10.1111
	S_{saving}	0.7172	0.7360	0.6808	0.9100
	CR	71%	73%	68%	90%

Figs. 3 and 4 (a) show the results of implementation of the presented method using combination of histogram differences and SCV similarity metrics between the successive frames in the metro video sequence. The red line indicates the automatic threshold value, where the x-axis represents the number of frames and y-axis represents the values of similarity metric. The frames with values of histogram differences that are greater than Thr_1 and values of SCV that are greater than Thr_2 are saved as summary sequence. The frames that have values of histogram differences smaller than Thr_1 and values of SCV greater than Thr_2 are removed. Figs. 3 and 4 (b) show the results of SCV without using histogram differencing.

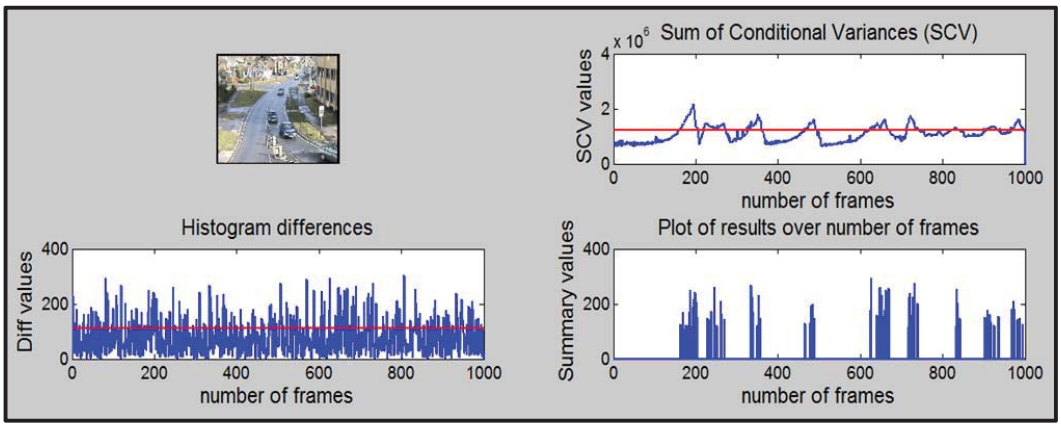


(a)

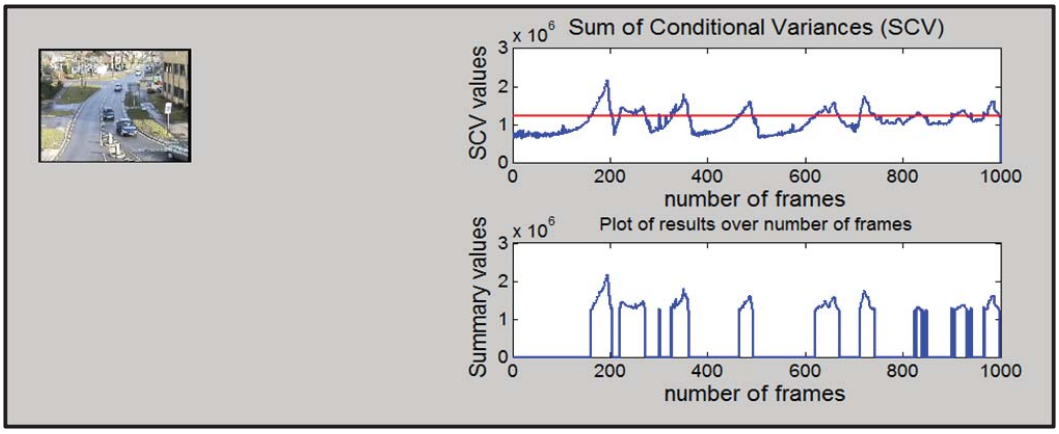


(b)

Fig. 3 (a) The results of the presented method for the metro video sample (b) The results of using SCV alone for the same video sample



(a)



(b)

Fig. 4 (a) The results of the presented method for the road video sample (b) The results of using SCV alone for the same video sample

Table II shows the results of the performance assessment of the proposed method compared with different temporal differencing for the metro video sequence of 1000 frames.

Traditional summarization techniques store frames that are up the automatic threshold Thr_1 and remove frames that are down of Thr_2. Meaning that they store frames containing motion and cancel frames containing no motion. While the presented method removes frames containing no motion and redundant frames containing motion that are not useful. In other words, it keeps only the most important motion information. For this reason, DCR and S_{saving} metrics give better results for the video sequence by implementing of the proposed method against the values of DCR and S_{saving} for the video sequence by implementing of the traditional technique. The CR metric, also gives good results (96%) using the presented method against the value of CR using the traditional technique.

Table III shows the performance assessment results after applying the traditional technique and the presented method on the road video sequence of 1000 frames. DCR, S_{saving}, and CR metrics give good results by applying the presented method compared with the traditional technique.

From the experimental results, the presented method based on the combination of frame differencing and SCV for data stream summarization works well compared with the existing summarization method based on temporal differencing. In addition, it has the ability to summarize the video stream with all important content in the original video sequence.

V. Conclusion

In this paper, we developed a method for video summarization using the combination between the histogram differencing and Sum Conditional Variant similarity metrics. The performance assessment has been achieved for the presented method compared with some traditional similarity functions. The results indicated that the improved video summarization method gives compact video summary with the useful data. This method has one limitation, it is sensitive to distant motion objects. This problem can be solved by decreasing the threshold which is used to detect motion objects in video sequence.

References

[1] Gao, Yue, and Qiong-Hai Dai. "Shot-based similarity measure for content-based video summarization." In Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on, pp. 2512-2515. IEEE, 2008.

[2] Mahapatra, Ansuman, Pankaj K. Sa, Banshidhar Majhi, and Sudarshan Padhy. "MVS: A Multi-view video synopsis framework." Signal Processing: Image Communication (2016).

[3] Kuanar, Sanjay K., Rameswar Panda, and Ananda S. Chowdhury. "Video key frame extraction through dynamic Delaunay clustering with a structural constraint." Journal of Visual Communication and Image Representation 24.7 (2013): 1212-1227.

[4] Ejaz, Naveed, Irfan Mehmood, and Sung Wook Baik. "Efficient visual attention based framework for extracting key frames from videos." Signal Processing: Image Communication 28.1 (2013): 34-44.

[5] Ejaz, Naveed, et al. "Video Summarization by Employing Visual Saliency in a Sufficient Content Change Method." International Journal of Computer Theory and Engineering 6.1 (2014): 26.

[6] Al-Musawi, Nada and Saad Talib Hasson. "Improving Video Streams Summarization Using Synthetic Noisy Video Data." (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 6, No. 12, 2015.

[7] Gao, Yue, and Qiong-Hai Dai. "Shot-based similarity measure for content-based video summarization." Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on. IEEE, 2008.

[8] Gygli, Michael, Helmut Grabner, and Luc Van Gool. "Video summarization by learning submodular mixtures of objectives." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[9] Bhaumik, Hrishikesh, Siddhartha Bhattacharyya, and Susanta Chakraborty. "Redundancy Elimination in Video Summarization." Image Feature Detectors and Descriptors. Springer International Publishing, 2016. 173-202.

[10] Yang, Xue, and Zhicheng Wei. "Video segmentation and summarization based on Genetic Algorithm." Image and Signal Processing (CISP), 2011 4th International Congress on. Vol. 1. IEEE, 2011.

[11] Lei, Shaoshuai, Gang Xie, and Gaowei Yan. "A Novel Key-Frame Extraction Approach for Both Video Summary and Video Index." The Scientific World Journal 2014 (2014).

[12] Song, Xinhui, Li Sun, Jie Lei, Dapeng Tao, Guanhong Yuan, and Mingli Song. "Event-based large scale surveillance video summarization." Neurocomputing (2015).

[13] Wang, Yu, and Jun Kato. "A distance metric learning based summarization system for nursery school surveillance video." Image Processing (ICIP), 2012 19th IEEE International Conference on. IEEE, 2012.

[14] Li, Xuelong, Zhigang Wang, and Xiaoqiang Lu. "Surveillance Video Synopsis via Scaling Down Objects." (2016).
[15] Ren, Jinchang, Jianmin Jiang, and Yue Feng. "Activity-driven content adaptation for effective video summarization." Journal of Visual Communication and Image Representation 21.8 (2010): 930-938.
[16] astelo-Fernández, César, and Guillermo Calderón-Ruiz. "Automatic Video Summarization Using the Optimum-Path Forest Unsupervised Classifier." Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Springer International Publishing, 2015. 760-767.
[17] Lee, Yuan-Shan, Chia-Yung Hsu, Po-Chuan Lin, Chia-Yen Chen, and Jia-Ching Wang. "Video summarization based on face recognition and speaker verification." In Industrial Electronics and Applications (ICIEA), 2015 IEEE 10th Conference on, pp. 1821-1824. IEEE, 2015.
[18] Luo, Chu. "Video Summarization for Object Tracking in the Internet of Things." Next Generation Mobile Apps, Services and Technologies (NGMAST), 2014 Eighth International Conference on. IEEE, 2014.
[19] Shaikh, Soharab Hossain, Khalid Saeed, and Nabendu Chaki. "Moving Object Detection Approaches, Challenges and Object Tracking." Moving Object Detection Using Background Subtraction. Springer International Publishing, 2014. 5-14.
[20] Petersohn, Christian. Temporal video segmentation. Jörg Vogt Verlag, 2010.
[21] Delabarre, Bertrand, and Eric Marchand. "Visual servoing using the sum of conditional variance." Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on. IEEE, 2012.
[22] Maki, Atsuto, and Riccardo Gherardi. "Conditional variance of differences: A robust similarity measure for matching and registration." Structural, Syntactic, and Statistical Pattern Recognition. Springer Berlin Heidelberg, 2012. 657-665.
[23] Richa, R., Souza, M., Scandaroli, G., Comunello, E., & von Wangenheim, A. "Direct visual tracking under extreme illumination variations using the sum of conditional variance." Image Processing (ICIP), 2014 IEEE International Conference on. IEEE, 2014.
[24] Kevin Roebuck, “Data Deduplication: High-impact Strategies -What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors”, Emereo Publishing, 2012.
[25] Bendraou, Y., Essannouni, F., Aboutajdine, D., & Salam, A. "Video cut detection method based on a 2D luminance histogram using an appropriate threshold and a post processing." Wseas Transactions on Signal Processing, 2015.