ISSN (Print) : 2319-8613
ISSN (Online) : 0975-4024

An Automatic Content Based Video Ranking from Surveillance Videos

Ruchit Shah [#1], Prof. Tushar Ratanpara [#2]
[#1] Computer Engineering Department, Dharmsinh Desai University, Nadiad, India
[#2] Asst. Professor, Department Of Computer Engineering, Dharmsinh Desai University, Nadiad, India

***Abstract*— Surveillance systems become more popular in our daily life because of rapid development of the camera industry. However, it is very difficult to track specific person and active time slots from a mass of long duration surveillance videos. Automatic surveillance video ranking is also desirable in many applications like human activity detection, traffic monitoring, public safety, crime prevention etc. This research work proposed a method to extract an active time slot of a person. An approach is also introduced for ranking multiple surveillance videos based on person activity duration. First valid frames for synopsis are identified using Viola-Jones face detection algorithm and foreground object detection method. Video synopsis is generated combining all these valid frames. Object tracking is done by computing face feature using Speed-Up Robust Feature (SURF). Face detection and face feature computation are also applied to query image. Similarity with query image is found based on a match face features. Active time slot, entry and exit time of a query person is computed with respect to match feature value between query image and video synopsis. Multiple surveillance video ranking is done based on active time slot. The proposed approach is tested on various surveillance videos.**

Keyword - Viola-Jones face detection, Speed-Up Robust Feature, Video synopsis, Video ranking.

I. INTRODUCTION

Video processing is a particular case of signal processing, that often uses video filters where the input and output both signals are video streams or video files [1]. Video is made of multiple scenes. Scene is made of multiple shots [21]. Shot is made of multiple frames. Video is at first layer while frames are at last layer. Longer video means high number of frames at last level [22]. Video ranking is the process of arranging multiple videos in order based on some parameters. Main objective of video ranking is to classify the results and providing the best video in final result. Several parameters for sorting videos are considered like order by relevant content, order by length and order by number of views, order by user rating. Videos are stored in database. Video ranking system is fetched videos from database and arrange them in order based on user query and algorithm. Figure 1 shows Basic flow diagram of video ranking system.

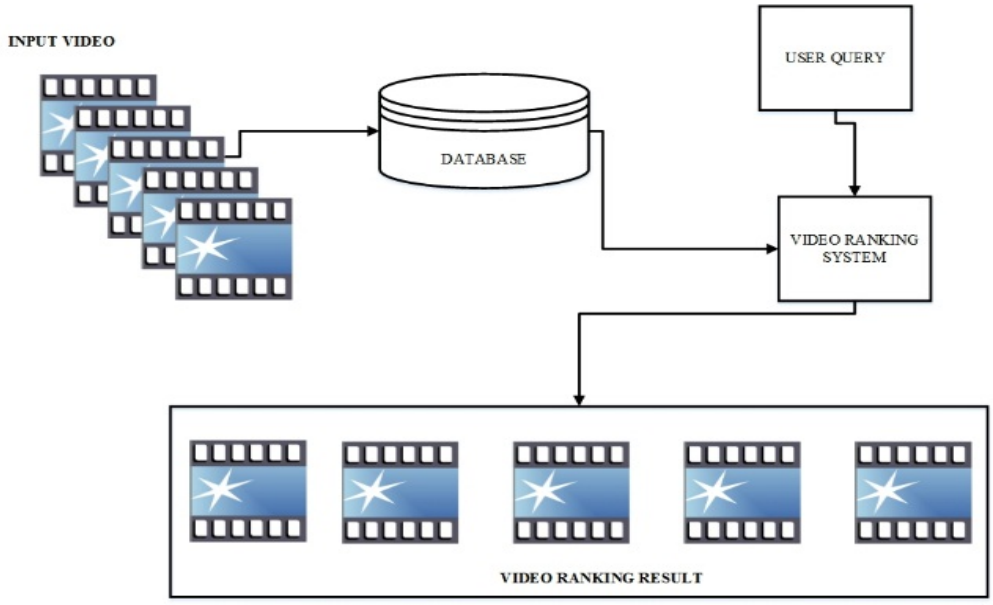

Fig. 1. Basic flow diagram video ranking system

Video ranking system for longer video has several drawbacks. More processing time is required for longer videos, ranking longer video requires more resources, high cost is required, ranking longer video is more complex process compare to short length video. A technique is required which generate short video from long video without changing the main information of video sequence and without collisions among the objects. Video synopsis is an approach to create a short video summary/abstraction of a long video. It analyses and tracks moving objects and converts video streams into a database of activities and objects. The technology has specific applications in the field of video surveillance. In video surveillance, viewing and analysis of recorded footage is still a costly, time-intensive and labour-intensive task [2] [3].

II. RELATED WORK

R. Karlsen, J. Morell, L. Luque, and V. Salcedo [4] have introduced a method for retrieval and re-rank of trustworthy health videos from YouTube. This algorithm identify source of video and re-ranks YouTube result. Therefore, videos coming from trusted sources are given higher priority in ranking list. This approach focus on four domains Hospital, Health Organization, Active User and other Users. First three domains (Hospital, Health Organization and Active User) are identified as a valid source by system automatically. These three domains are considered as white list. Video coming from the white list domain are given higher priority and higher rank. 2000 health videos related to diabetes diseases on YouTube is used as a dataset. Accuracy of this approach is 81%. L. Chen, K. Chin and H. Liao [5] have designed a method for a video retrieval system based on the integration of several visual cues and video ranking based on relevance feedback technique. In this approach all frames are analysed within a shot to generate a compact representation of video shot. Video matching is done by integrating the motion and colour features. This system is able to use the spatio-temporal information contained in video. Relevance feedback technique includes a human as part of video retrieval process. In the relevance feedback technique high level concepts are described by low level features. This technique provides the link between the two levels from a viewer feedback. The viewer only needs to specify that video clip is relevant to the query or not. Feedback is given by viewer in form of score like 3 highly relevant, 1 relevant, 0 no-opinion, −1 non-relevant and −3 highly non-relevant. Based on viewer feedback video rank is updated during each round of the retrieval process. This approach is feasible and effective in ranking and retrieving similar video clips. R. Hannane, A. Elboushaki and K. Afdel [6] have presented an approach for automatic video surveillance indexing based on face descriptors System. This approach have three key steps like video surveillance summarisation, face detection and video indexing. Video surveillance summarisation is done by key frame selection technique based on local foreground entropy. Face detection is done by skin colour based method. Facial feature descriptors extraction is done by Speed-Up Robust Feature (SURF) algorithm. Video indexing is done by using vocabulary tree. Chokepoint public dataset is used in this approach. Accuracy of this approach is 87.92%. V. Choudhary and A. Tiwari [3] have proposed an approach for synopsis of surveillance video by removing spatial and temporal redundancy. In this approach, spatial redundancy is removed by showing two actions/events in a single frame that happened in different frames at different spatial locations. Temporal redundancy is removed by detecting the frames having no/low actions and then discarding those frames. DBSCAN algorithm is used to track the path of the objects in video. Then using stroboscopic effect synopsis video is generated. This approach is integrated with a Media Player for indexing video. C. Chou, C. Lin, T. Chiang, H. Chen and S. Lee [7] have designed a method to develop an approach to avoid too much information in a synopsis frame may distract viewers' attention. In this method authors propose a novel system to alleviate the above issues using coherent event classification. First Object trajectories are extracted by background subtraction. Then Object trajectories are clustered together. In this approach longest common subsequence (LCS) algorithm is used to measure the similarity among trajectories. At last trajectories in each cluster are rescheduled. Then trajectories are stitched onto the background to generate synopsis videos with coherent events. 30 minutes videos of 4 indoor scenes in a building is used as a dataset in this approach. Advantages of longest common subsequence algorithm are clustering of similar trajectories with different rates and different lengths and no influence of abnormal trajectories. Y. Pritch, S. Ratovitch, A. Hendel and S. Peleg [8] have proposed a new method for the generation of coherent and short video summary. This summary is based on clustering of similar kind of activities. Objects with similar properties and activities are easy to watch simultaneously. In this approach, outliers are identified instantly. The main advantage of this approach is that summary is very clear and making video browsing very easy. SIFT descriptors are used as appearance features and to find out activity features. Appearance distance and motion distance are computed between different activities as well as play time is assigned to each cluster. Another advantage of this approach is clustered summaries is used for video browsing.

III. PROPOSED APPROACH

Main goal of our system is to extract an activity duration, entry and exit time of a person in a surveillance video and ranking of multiple surveillance videos based on query person activity duration. Figure 2 shows abstract model of proposed approach.

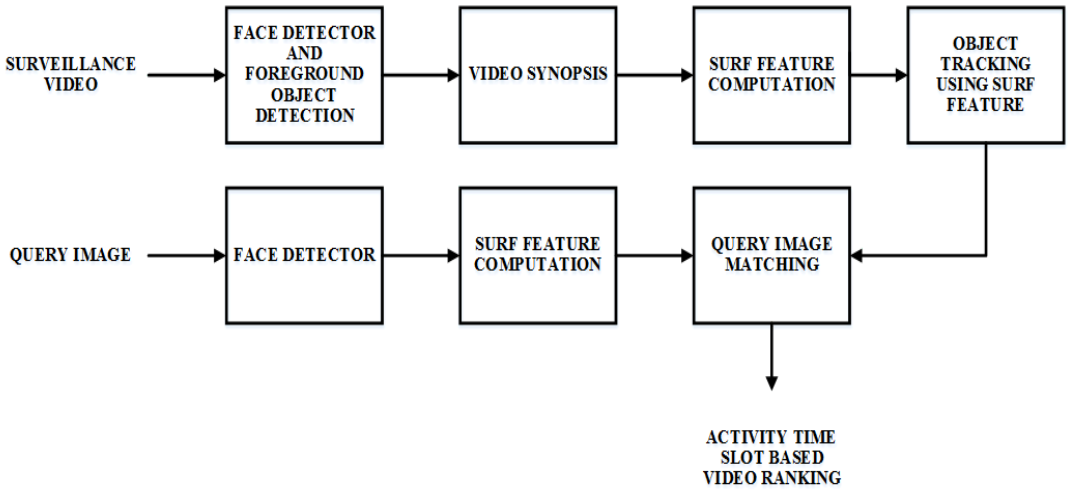

Fig. 2. Abstract model of proposed approach

A. Face Detector and Foreground Object Detection

A computer needs accurate constraints and instructions to detect a face. Main goal of this stage is to detect all faces in given video as well as determine the exact position of the faces [9]. This step is based on Viola-Jones face detection algorithm. There are various advantages of Viola-Jones face detection algorithm like scale and location invariant detector and efficient feature selection. Detected face is represented by bounding box that contains four parameters x, y, Height, Width. Face part is cropped for further processing using BOUNDBOX property of Viola-Jones face detection algorithm. Figure 3 shows face detection using Viola-Jones.

Fig. 3. Face detection using Viola-Jones

Next step is removal of frames with no objects or contain less part of object. Foreground is detected by taking XOR of each frame with background frame. This is illustrated by the following equation:

$$I = \text{bitxor}(F_0, Fx) \quad (1)$$

Where background frame F_0 (time t = 0), current frame Fx (time t = x), final image I. If the foreground pixel count value (Image I) is smaller than threshold value (T_f) that frame is not considered for final video synopsis. Figure 4 shows foreground object detection with white pixels count are 64722, 193852 and 170091 respectively.

Fig. 4. Foreground object detection

Frame for synopsis (valid frame) is selected using following two conditions.

1. Human face is detected in frame.
2. Number of foreground pixels are greater than threshold value (T_f).

B. Video Synopsis

Video synopsis is generated by combining all the valid frames which is shown in figure 5. Frame rate of video synopsis is 15 fps (frame per second). As synopsis contains frames which contains human face and large number of foreground pixels, video synopsis length is small compare to original video.

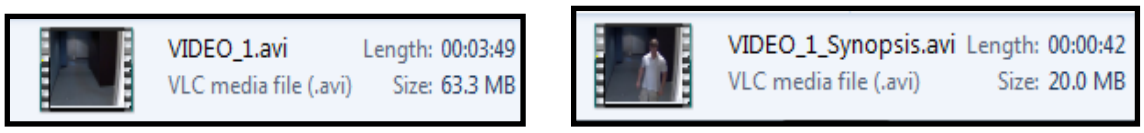

Fig. 5. Original video (VIDEO_1) and video synopsis

C. SURF Feature Computation

First frames are extracted from video synopsis. Face detection is applied using Viola-Jones face detection algorithm. Face is extracted using skin colour approach. Facial feature extraction process is based on Speed-Up Robust Feature (SURF). SURF is a scale as well as in-plane rotation invariant detector. 64 dimensions is used to reduce the time cost for both feature computation and matching. SURF focuses only on facial interest points (SURF descriptors) instead of focusing all the information of face [11]. SURF features are computed for every synopsis frame. Figure 6 shows SURF feature computation with SURF points 50, 61 and 62 respectively, three strongest points are shown.

Fig. 6. SURF feature computation

D. Object Tracking Using SURF Feature

SURF features are stored for each synopsis frame. Object tracking is done by taking absolute difference of SURF feature of each frame with previous frame. If the absolute difference value is larger than threshold value (Ts) than new object is considered. This is illustrated by the following equation:

$$S = \text{abs}(Sc - Sp) \quad (2)$$

Where current frame SURF value is Sc (time t = 0), previous frame SURF value is Sp (time t = x) and S is the final SURF value. Figure 7 shows frame number 23, 24, 25, 26 from VIDEO_1 Synopsis. Table I shows about object tracking for frame number 23, 24, 25, 26. Where Frame No. is a frame number, SURF Value (Sc) is current frame SURF value, Difference(S) is SURF difference of current and previous frame, Object Number is unique object id. SURF threshold (Ts) value is 50. Therefore, new object is identified at frame number 25 (Difference(S) value is 57).

Fig. 7. Frame no. 23 to 26 from VIDEO_1 Synopsis

TABLE I. Object Tracking Frame No. 23 to 26 from VIDEO_1 Synopsis

Frame No.	SURF Value (Sc)	Difference (S)	Object Number
23	137	6	1
24	90	47	1
25	**33**	**57**	**2**
26	28	5	2

E. Query Image (Face Detector, SURF Feature Computation)

Query image is used to search the content of the image rather than metadata such as keywords, tags, or descriptions associated with the image [10]. Face detection, skin extraction and face feature computation using SURF are also applied to query image which is shown in figure 8. Query image has 93 SURF points.

Fig. 8. Query image A (face detection, SURF based face feature computation)

F. Query Image Matching

For matching, SURF features are compared between query image and all synopsis frame. Object number and whole object trajectory are found where highest surf points matches from the object tracking table. Activity time slot/activity duration of query person, entry and exit time of person are computed for input video using equation 3 and 4.

Query person enter time = Object start frame number/Video synopsis frame rate (3)

Query person exit time = Object end frame number/ Video synopsis frame rate (4)

Query person activity duration (in frames) = (Object end frame number – Object start frame number) + 1 (5)

Table II shows SURF match of query image A with object 1 and object 3 for VIDEO_10. Highest match (Match SURF value 25) is found at frame 6 for query image A. Object 1 appears from frame 2 to 145. Object starts with frame number 2 and ends with frame number 145. According to formula (3) and (4) enter time is 00:00 minutes and exit time is 00:10 minutes (frame rate 15 fps).

TABLE II. SURF Match with Object 1 and 3 for Query Image A

Frame No.	SURF (Sc)	Difference (S)	Obj. No.	Query Image SURF	Match SURF
2	42	42	1	93	8
3	41	1	1	93	14
4	56	15	1	93	3
5	51	5	1	93	23
6	**67**	**16**	**1**	**93**	**25**
7	60	7	1	93	15
8	65	5	1	93	14
9	65	0	1	93	14
10	65	0	1	93	11
150	90	52	3	93	6
151	89	1	3	93	12
152	97	8	3	93	9
153	73	24	3	93	8
154	27	46	3	93	0
155	29	2	3	93	8

Multiple surveillance video ranking is done based on activity duration frame count. Video having more activity duration is assigned higher priority/rank compare to less activity duration video. Table III shows video ranking results for query image A. Input videos are VIDEO_1, VIDEO_10 and VIDEO_11. VIDEO_10 has maximum frame count (144).

TABLE III. Video Ranking results for Query Image A

Video	Entry Time (Minutes)	Exit Time (Minutes)	Active Duration Frame Count	Rank
VIDEO_1	00:00	00:02	23	2
VIDEO_10	00:00	00:10	144	1
VIDEO_11	00:00	00:00	0	3

IV. EXPERIMENTAL RESULTS

Different types of surveillance videos are used for performance evaluation which are publicly available. Faces in such videos are different in terms of pose, sharpness and illumination conditions. Dataset contains both type of videos indoor and outdoor [6]. Frame rate for all video is 10 fps (frame per second) and image resolution is 800 * 600 pixels. Table IV shows dataset information before and after synopsis with compression ratio.

TABLE IV. Dataset Information Before and After Synopsis

Video	Before Synopsis			After Synopsis			Compression Ratio (%)
	Frame	Length (Minutes)	Size (MB)	Frame	Length (Minutes)	Size (MB)	
VIDEO_1	2292	3:49	63.3	644	00:42	20.0	71.90
VIDEO_2	2292	3:49	68.2	511	00:34	16.9	77.71
VIDEO_3	2292	3:49	61.8	690	00:46	19.7	69.90
VIDEO_4	2370	3:56	65.7	402	00:26	13.0	83.04
VIDEO_5	2370	3:56	71.8	348	00:23	12.2	85.32
VIDEO_6	2370	3:56	64.5	672	00:44	19.8	71.65
VIDEO_7	2470	4:06	69.0	781	00:52	23.9	68.38
VIDEO_8	2470	4:06	75.2	292	00:19	10.4	88.18
VIDEO_9	2360	3:55	65.9	681	00:45	21.5	71.14
VIDEO_10	2360	3:55	71.4	534	00:35	18.0	77.37
VIDEO_11	2166	3:36	71.9	89	00:05	3.43	95.90
VIDEO_12	760	1:16	29.9	256	00:17	10.5	66.32
VIDEO_13	2166	3:36	89.1	542	00:36	22.8	74.98
VIDEO_14	760	1:16	29.2	436	00:29	17.2	42.63
VIDEO_15	760	1:16	30.6	197	00:13	7.99	74.08
VIDEO_16	2180	3:37	73.4	659	00:43	23.1	69.77
VIDEO_17	738	1:13	29.2	323	00:21	13.4	56.23
VIDEO_18	736	1:13	28.2	424	00:28	16.6	42.39
VIDEO_19	736	1:13	30.1	136	00:09	5.76	81.52
VIDEO_20	788	1:18	25.0	319	00:21	10.6	59.52

Table V shows actual and proposed approach entry, exit time and active frame count with video rank for query image B (Figure 9). Input videos are VIDEO_1, VIDEO_2, VIDEO_3, VIDEO_4, VIDEO_5, VIDEO_6, VIDEO_7, VIDEO_8, VIDEO_9 and VIDEO_12. Maximum frame count (312) is found in VIDEO_7 whereas minimum frame count (0) is observed in VIDEO_12.

Fig. 9. Query image B

TABLE V. Actual and Proposed Approach Entry Exit Time and Active Frame Count for Query Image B

Video	Actual		Proposed Approach			
	Entry Time (Minutes)	Exit Time (Minutes)	Entry Time (Minutes)	Exit Time (Minutes)	Active Duration Frame Count	Rank
VIDEO_1	00:02	00:05	00:02	00:05	47	6
VIDEO_2	00:01	00:04	00:01	00:04	46	7
VIDEO_3	00:02	00:04	00:02	00:06	63	5
VIDEO_4	00:13	00:17	00:00	00:04	66	4
VIDEO_5	00:10	00:14	00:09	00:14	67	3
VIDEO_6	00:25	00:28	00:09	00:10	8	9
VIDEO_7	**00:13**	**00:16**	**00:15**	**00:36**	**312**	**1**
VIDEO_8	00:05	00:08	00:10	00:12	29	8
VIDEO_9	00:12	00:16	00:08	00:16	127	2
VIDEO_12	00:00	00:00	00:00	00:00	0	10

Table VI shows actual and proposed approach entry, exit time and active frame count with video rank for query image C (Figure 10). Input videos are VIDEO_13, VIDEO_14, VIDEO_15, VIDEO_17, VIDEO_18, VIDEO_19 and VIDEO_20. Maximum frame count (250) is observed in VIDEO_14.

Fig. 10. Query image C

TABLE VI. Actual and Proposed Approach Entry Exit Time and Active Frame Count for Query Image C

Video	Actual		Proposed Approach			
	Entry Time (Minutes)	Exit Time (Minutes)	Entry Time (Minutes)	Exit Time (Minutes)	Active Duration Frame Count	Rank
VIDEO_13	00:00	00:00	00:12	00:19	111	3
VIDEO_14	**00:00**	**00:05**	**00:06**	**00:23**	**250**	**1**
VIDEO_15	00:00	00:02	00:00	00:03	39	6
VIDEO_17	00:02	00:06	00:03	00:12	138	2
VIDEO_18	00:04	00:11	00:05	00:12	100	4
VIDEO_19	00:01	00:03	00:03	00:03	8	7
VIDEO_20	00:04	00:10	00:00	00:04	66	5

V. CONCLUSION

In this proposed work, an approach is introduced for person active time slot extraction using video synopsis generation, face detection and face feature extraction. Video synopsis is generated by identifying valid frames using Viola-Jones face detection algorithm and foreground detection. Object tracking is done by computing face feature using Speed-Up Robust Feature and finding absolute difference of consecutive frames. Active time slot, entry and exit time of a query person is computed with respect to match feature value between query image and video synopsis. Multiple surveillance video ranking is done based on query person active time frame count. Proposed approach worked well for ranking of multiple surveillance videos. In future, algorithm is extended for multiple appearance of query person and system should give multiple activity time slots of single query person in single video. Approach is also extended for query by video instead of query by image.

REFERENCES

[1] Y. Wang, J. Ostermann and Y. Zhang, "Video Processing and Communications", 2002.
[2] Y. Nie and C. Xiao. "Compact video synopsis via global spatiotemporal optimization", IEEE transactions on visualization and computer graphics, vol. 19, issue 10, pp. 1664-1676, 2013.
[3] V. Choudhary and A. Tiwari, "Surveillance Video Synopsis", Sixth Indian Conference on Computer Vision, Graphics & Image Processing, pp. 207 - 212, 2008.
[4] R. Karlsen, J. Morell, L. Luque, and V. Salcedo, "Retrieval of trustworthy health videos from YouTube", IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), pp 25 – 28, 2014.
[5] L. Chen, K. Chin and H. Liao, "An Integrated Approach to Video Retrieval", Nineteenth Australasian Database Conference (ADC2008), Vol. 75, 2008.
[6] R. Hannane, A. Elboushaki and K. Afdel, "An automatic video surveillance indexing based on facial feature descriptors", Fifth International Conference on Information & Communication Technology and Accessibility (ICTA), pp. 1 - 6, 2015.
[7] C. Chou, C. Lin, T. Chiang, H. Chen and S. Lee, "Coherent event-based surveillance video synopsis using trajectory clustering", IEEE International Conference on Multimedia & Expo Workshops (ICMEW), pp. 1 - 6, 2015.
[8] Y. Pritch, S. Ratovitch, A. Hendel and S. Peleg, "Clustered Synopsis of Surveillance Video", Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 195 - 200, 2009.
[9] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features", Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol 1, pp. 511-518, 2001.
[10] J. Eakins and M. Graham, "Content-based Image Retrieval", University of Northumbria at Newcastle, 2004.
[11] H. Bay, T. Tuytelaars and L. Gool, "Speeded-Up Robust Features", Journal Computer Vision and Image Understanding, Vol 110, Issue 3, pp. 346-359, June 2008.
[12] R. Kansagara, D. Thakore and M. Joshi. "A study on video summarization techniques", International Journal of Innovative Research in Computer and Communication Engineering, vol. 2, 2014.
[13] T. Archana and T. Venugopal. "Face recognition: A template based approach", International Conference on Green Computing and Internet of Things (ICGCIoT), pp. 966 – 969, 2015.
[14] A. Mahapatra, P. Sa and B. Majhi, "A multi-view video synopsis framework", IEEE International Conference on Image Processing (ICIP), pp. 1260 - 1264, 2015.
[15] C. Huang, H. Chen and P. Chung, "Online surveillance video synopsis", IEEE International Symposium on Circuits and Systems, pp. 1843 - 1846, 2012.
[16] T. Yao, M. Xiao, C. Ma, C. Shen and P. Li, "Object based video synopsis", IEEE Workshop on Advanced Research and Technology in Industry, pp. 1138 - 1141, 2014.
[17] C. Hsia, J. Chiang, C. Hsieh and L. Hu, "A complexity reduction method for video synopsis system", International Symposium on Intelligent Signal Processing and Communication Systems, pp. 163 - 168, 2013.
[18] S. Raikwar, C. Bhatnagar, and A. Jalal, "A framework for key frame extraction from surveillance video", International Conference on Computer and Communication Technology (ICCCT), pp. 297 - 300, 2014.
[19] Sujatha C and U. Mudenagudi, "Gaussian Mixture Model for summarization of surveillance videos", Fifth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), pp. 1 - 4, 2015.
[20] M. Deza and E. Deza, "Encyclopedia of Distances", p. 94, 2009.
[21] Gandhi, Shefali, and Tushar V. Ratanpara. "Object-Based Surveillance Video Synopsis Using Genetic Algorithm." Applied Video Processing in Surveillance and Monitoring Systems. IGI Global, 2017. 193-219.
[22] Chandni Dhamsania, and Tushar Ratanpara. "Human Action Recognition Using Trajectory-Based Spatiotemporal Descriptors." Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications, Springer, pp. 1-9, 2017

AUTHOR PROFILE

Ruchit Shah have completed master of technology in computer engineering from Dharmsinh Desai University, Nadiad, India. His main area of interest is in image and video processing.

Prof. Tushar Ratanpara is working as an Assistant Professor in computer engineering department at Dharmsinh Desai University, Nadiad, India. He is also working as a counselor of CSI-DDU student branch since 2011. He has published more than 15 research papers in the international journals and conferences like IEEE, Springer, IGI-Global, Inderscience and ACM. His main research areas includes Image Processing, Video Processing, Audio Processing and Internet of Things.