VIDEO SUMMARIZATION USING CONTOURLET TRANSFORM AND CLUSTERING

Ms. Mithila Metri[1], Ms. Razia Sardinha[2]
[1, 2] *Information Technology, Goa University, Goa, India*

Abstract
With the arrival of digital media, a tremendous content of movies, news, television shows and sports is widely available everywhere. However, the user might not want to view the entire video or have the time to view lengthy video content. In these cases, viewing only the summary of the video is preferred instead of watching the entire video.
Summarization of video, such as news clip or movie clip or surveillances, is to find all the segments that covers important information present in that video, hence allowing the user to get the overall message of the video by just viewing the summarized video clips instead of watching the entire lengthy video.
In my work, I have summarized videos by extracting all the available key frames needed to represent the content of each shot. All the shots were determined, followed by clustering the similar frames for the given shot to extract one representative key frame for that shot. The key frames give the summary of the given video which can then be used for different applications as required for various analysis. I have used Contourlet transforms to detect the shot boundaries and k-Means Clustreing to extract the key frames. Repeated key frames were discarded and the number of people present in the summary of the video were counted. The summarised video can be used for further analysis.

Keywords: *Contourlet transform, Shot Boundary detection, Clustering, Key Frame Extraction.*

I. INTRODUCTION

[1][2][4][14]There has been a tremendous growth in the video contents over past years. Thisincrease in video content causes problem of overloading and management of video content. In order to manage the growing videos on the web and also to extract an efficient and valid information from the videos, more attention has to be paid towards video and image processing technologies. Video summaries provide condensed representations of the content of a video stream by combining the still images, video segments and graphical representations .Although a lot of digital content such as movies, news, television shows and sports is widely available, the user may not have sufficient time to watch the entire video or the whole of video content may not be of interest to the user. In such cases, the user may just want to view the summary of the video. Thus, the summary must convey as much information about the occurrence of various incidents present in the video.

Video summarization is a fundamental process in the video processing. It is used mainly in variety of applications, such as video databases, video indexing, video skimming, video retrieval and so on.The main task in video summarization is to detect the shots and extract the video frames from the original video that would be provide the appropriate information and representation of the whole video. Such frames are referred as key frames

[14]To process and analyse the video based on content of the video, we first analyse the structure of the video. A video can be segmented into different units, such as frames, shots, or scenes. The structure of a video is shown in Figure 2.1. The complete moving picture in a video can be discretised to a finite sequence of still images, called a “frame”, which is the basic unit of the video. The image sequence has the frame number. All the frames in specific video have a same size and equal time between each two frames. A video shot is defined as a series of interrelated consecutive frames taken by a single camera that represents a continuous action. In general, shots are joined together to produce a video.

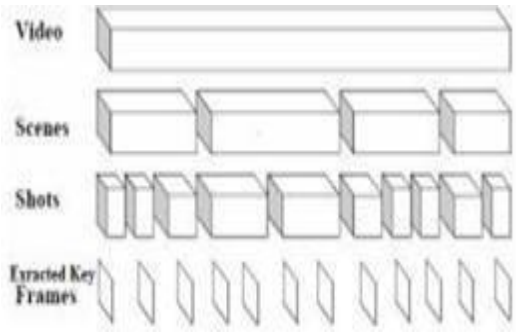

Fig. 1Structural hierarchy of a video

A. Related Work

RaviKansagara, DarshakThakore, MahaswetaJoshi [1] proposed study of various techniques for key frames based video summarization available in the literature. It created summary of video to enable a quick browsing of a collection of large video database. Mainly two types of video summarization techniques are available in the literature, viz. key frame based and video skimming. For key frame based video summarization, selection of key frames plays important role for effective, meaningful and efficient summarizing process.

Sachan Rajendra, Dr. Keshaveni N [2], proposed some of the recent work on content-based multimedia information retrieval and discussed their role in current research directions which include browsing and search paradigms, user studies, effective computing, learning, semantic queries, new features and media types, high performance indexing, and evaluation techniques. Based on the current state of the art, they also discuss the major challenges for the future.

Zaynab El khattabi, Youness Tabii, Abdelhamid Benkaddour [3], focus on approaches to video summarization. The video summaries can be generated in many different forms. However, two fundamentals ways to generate summaries are static and dynamic. We present different techniques for each mode in the literature and describe some features used for generating video summaries.

Sandip T. Dhagdi, Dr. P.R. Deshmukh[4], discusses the importance of key frame extraction; and proposed a new approach for key frame extraction based on the block based Histogram difference and edge matching rate.

Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman [5], presented a video summarization approach for egocentric or "wearable" camera data. Given hours of video, the proposed method produces a compact storyboard summary of the camera wearer's day.

Pamarthy Chenna Rao[6] proposed that Contourlet transform is good enough to extract the key frames and also proved that the method is having high accuracy rate and low error rates with selected feature vector and distance measure technique on all kinds of videos.

Minh N. Do,Martin Vetterli [7], pursued a "true" twodimensional transform that can capture the intrinsic geometrical structure that is key in visual information.They showed some numerical experiments demonstrating the potential of contourlets in several image processing applications

Lin-Bo Cai, Zi-Lu Ying [8], introduces the theory of Contourlet Transform and proposes a new approach of facial expression recognition based on Contourlet Transform. Experimental results show that the proposed approach gets better recognition rate than both Wavelet Transform and Principal Component Analysis. The facial expression recognition based on Contourlet Transform is an effective and feasible algorithm.

Wei-shi Tsai[9],have shown that the contourlet transform can be used for edge detection and performed a subjective analysis because an objective analysis should be done based on how the extracted features are used.

Pamarthy Rao, Ramesh Patnaik [10], proposed that the video summarization in terms of key frames are extracted using feature vectors, which are obtained from features calculated for each sub-band in the contour let transform. Where the energy and standard deviation features of each sub - band are used to form a feature vector. The experimental results proved that this novel method had more accuracy rate and low error rate.

WenzhuXu, Lihong Xu [11], presented a novel shot boundary detection algorithm based on K-means clustering. At first the feature of color is extracted, the dissimilarity of video frames is defined .Then the video frames are divided into several different sub-clusters through performing K-means clustering. It can detect cut and gradual shot by the adaptive double threshold of different sub-clusters. The efficiency of the proposed algorithm is extensively tested on movie, news and other videos. The experiments results indicated that the method had a high accurate rate in both cut shot detection and gradual shot detection.

Victor Shen, Hsin-Yi Tseng, Chan-Hao Hsu [12], focused on the shot change, a part of the video summarization, to conduct an experimental sample on news programs. Moreover, a high-level fuzzy Petri net model is presented to describe boundary frames combination which indicates a shot boundary used for video frame sequence to detect both cut transitions and gradual transitions. The experimental results manifested that it saves a lot of time and reduces the occurrence of improper shot change caused by the motions of objects and cameras when comparing this method with human labor.

Nikita Sao, Ravi Mishra [13], Presents a brief survey on the shot detection techniques. Processing of video and image provides an understanding of the scene that it describe. It is an essential component of a number of technologies including video surveillance, robotics and multimedia. It represents an area of research with huge growth in the recent past. Video shot boundary detection is one of the research works in the field of video processing. Many researchers are trying to put forward different algorithm in this respect. Here they presented a brief literature survey that depicts the work done till date.

A.V.Kumthekar, Prof.Mrs.J.K.Patil [14], proposed a method for video key frame extraction based on color histogram and edge detection, the purpose was to remove the redundant frames, reduce the computational complexity and improve recognition efficiency. The compression ratio is 98%.

Azra Nasreen, Kaushik Roy, Kunal Roy, Shobha G[15], proposed and implemented a novel robust key frame extraction and foreground isolation method using k-

means clustering and mean squared error method for variable frame rate videos. They also isolated foreground objects in the video whilst eliminating the noise generated in the recording. The flickering of the frames caused as a result of variable frame rate in a recorded video is reduced by a considerable degree using this method. Also, the k-means clustering is performed on Apache's hadoop infrastructure to make the results of the computation faster. They have implemented this method and obtained results to be clear enough to extract meaningful detail from the frames. The results of the method have been compared to similar results obtained using well- known techniques such as the Gaussian Mixture Model and have been shown to be better.

Supriya Kamoji [16],proposed system that provide with a summary of a video by utilizing and capturing the motion throughout it. It was found out that precision and summarization factor are important parameters in this process and the idea was to maximize both. However, as per the above observations different categories of video produced different results. The summarization proves effective in situations having limited area and definite objects as it eases the formation of motion activity descriptors. The block matching technique used affects the process which can be seen from the results. Diamond Search has an advantage over Three Step Search where it achieves higher precision.

Padmavathi Mundur[17], proposed an automatic video summarization technique based onDelaunay Triangulation. They presented meaningful comparisons of the results from the proposed DT algorithm to OV storyboard by defining metrics such as significance factor, overlap factor, and the compression factor all of which evaluate the representational power of the proposed DT summarization technique. We also demonstrated the DT advantage for batch processing over K-means clustering where the optimal number of clusters needs to be predefined.

N. Dalal,B. Triggs [18] , introduced HOG features. He developed and tested several variants of HOG descriptor with differing spatial organization,gradient computation and normalization method.

Rakesh Kumar, Tapesh Parashar, Gopal Verma [19], presented a multiple people counting using only single camera, entering or exiting in a region of interest. Sigma-Delta background modeling and subtraction was used to segment the people from region. Experiments showed that proposed method was robust and provides approximate accurate counts.

Shafraz Subdurally, Devin Dya, Sameerch and Pudaruth [20], proposed two systems for counting people from images. Their proposed methods are based on the observation that heads are significantly more visible than any other features and are thus more easily distinguishable. The proposed systems use blobs and contour detection respectively to count the number of people. The results obtained from each system are very reliable. The average head detection rate of the systems is 82 and 84 percent respectively.

Cansın Yıldız[21], implemented a Histogram of Oriented Gradients (HOG) detector for pedestrians. Although this new implementation gave good result on performance tests, it has some drawbacks when it comes to real application.

II. PROPOSED WORK

Video summarization is creating a summary of a given digital video, which needs to satisfy the following properties

(1) The video summary must include all the important contents and the events from the video,

(2) Summary should maintain continuity to be understandable

(3) The summary must contain unique frames , hence there should not be free from the duplicates

The main objective of the proposed work is to summarize a given video. In this we first segment a given video into different frames, apply contourlet transform and determine the beginning and end of each shot. This is then used to extract a single frame called as keyframe, representing the main content of each shot by k-means clustering. The extracted key frames from each detected video shot are then used for summarization by discarding the duplicate ones.The additional proposed work is to use the summary to detect and count the number of people.

[6]The proposed work checks if the approaches used are good enough to detect the shot boundary and the summary. Also it checks accuracy rate and error rates on all kinds of videos. It uses contourlet transform to detect the shot boundaries and kmeans clustering algorithm to extract the keyframes. To detect the people, it uses HOG descriptor.

The video summarization process involves, taking video stream as input followed by framing the input video and apply the Contourlet transform on each input frame and form a feature vector. Now these feature vectors are used to decide whether the given frame is boundary frame or not. To compare these feature vectors of two consecutive frames computes the distance between two feature vectors. Finally, the resulted distance is compared with the predetermined threshold to check whether the frame is boundary frame or not. If the distance is above threshold then the frame is classified as key frame, otherwise the frame belongs to same shot.

Further, we propose a method to extract key frames from each of the shots using a kmeans clustering. Sometimes it may results in redundant key frames. Therefore, we summarize the extracted keyframes by eliminating the duplicates. Now these summarized keyframes are further processed using HOG descriptors to detect and count the number of people. The block diagram for video summarization is shown in figure 2.

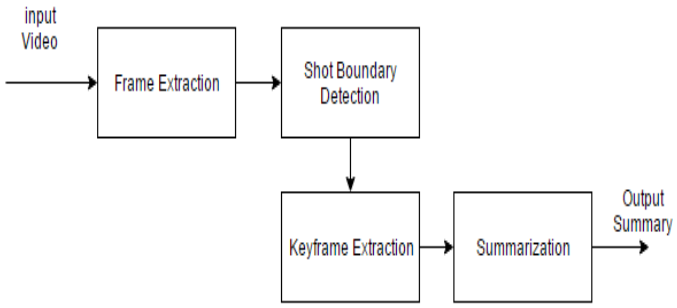

Fig. 2 Block Diagram of Video summarization

A. Contourlet Transforms

[6][7][10]The Contourlet construction was chosen to obtain a sparse expansion of the images to produce a piecewise smooth contours. This Contourlet has the advantage that wavelet couldn't achieve because wavelets are lack of directionality where as Contourlet can find directionality. Wavelets are only good at catching point discontinuities but don't capture geometrical smoothness of the contours [6]. Due to this inefficiency in wavelet transform, Contourlets were developed. The Contourlet transform offers a high degree of directionality and anisotropy in addition to wavelet properties like multi scale and time frequency localization property. Contourlet transform has its basis functions oriented at power of two's number of directions. The basic block diagram is shown below. The basic block diagram is shown in fig 3.

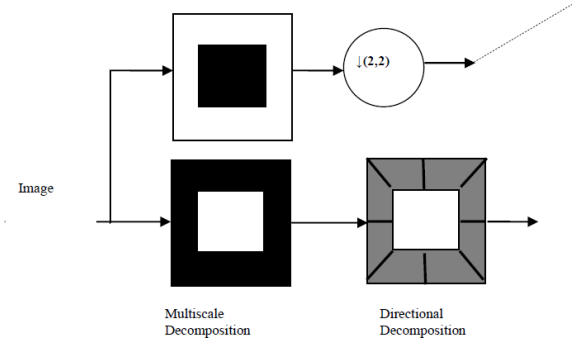

Fig. 3 Contourlet Filter Bank

[6][7][8][10] The Contourlet transform is implemented using directional filter banks which decompose the image into several directional sub-bands at multiple scales. Contourlet achieves this by combining directional filter banks with Laplacian pyramid at each scale. This cascade structure, multi scale and directional decomposition are independent of each other. The Fig 4 shows an example frequency partition of the Contourlet transform

[6][10]Two feature vectors similarity can be measured by using different techniques by measuring distance between two feature vectors. In this work, we choose the Euclidean distance as a metric to measure distance between two feature vectors.Euclidean distance between two feature vectors of two consecutive frames is computed. This distance is then compared against predetermined threshold. If the distance is above threshold, then the given frame is classified as belonging to the next shot hence the shot boundary is detected, otherwise it is classified as a frame belongs to same shot.

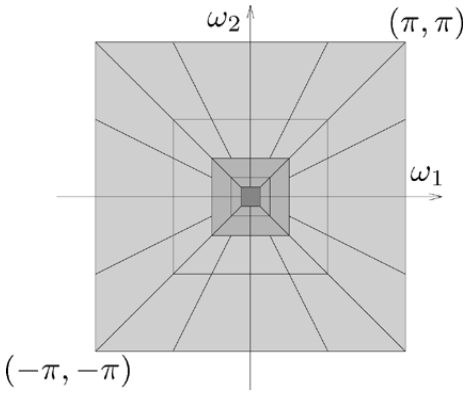

Fig. 4. Example frequency partition by the contourlet transform.

B. K Means Clustering

[15]K-means clustering refers to the grouping of similar dataset representing some data into the defined number of clusters. This grouping is done by calculating the measure of the average of all the points representing the data when mapped into a co-ordinate system. In this proposed work, the datasets ate the frames of the video. Once the desired clusters are obtained, the characteristics that define a particular cluster can be used to further processing of data.

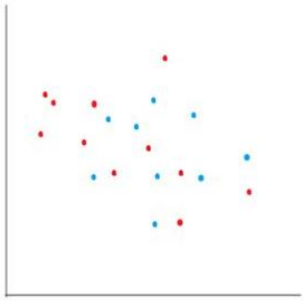

Fig. 5 Unstructured data

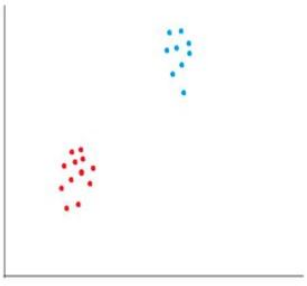

Fig. 6 Clustered data

Figure 5 shows unstructured data when plotted on the co-ordinate axis. Figure 6 shows the results of clustering. After grouping into clusters in this manner meaningful common characteristics can be identified among data points and processed.

III. RESULTS AND IMPLEMENTATION

A video (hiker.mpg) is selected to test the results. Firstly, extract all the frames from the selected video. For Each Frame we apply the contourlet transforms which returns the contourlet coefficients. Using these we use the Euclidean distance to find the difference between the frames and detect the shot boundaries. For each shot, extraction the frames using Kmeans algorithm. Further process the extracted frames to get the unique keyframes. The unique extracted keyframes are then used to detect the people.

As seen in table 2, for the successful extraction of relevant keyframes, the proposed work gives higher precision and an average of 79.25% recall. The average accuracy of the application in extracting the keyframes is also 79.25%. The average error rate is calculated as 20.75%

The experimental result is as follows

Fig. 6a. Example of a video sequence(hiker.mpg)

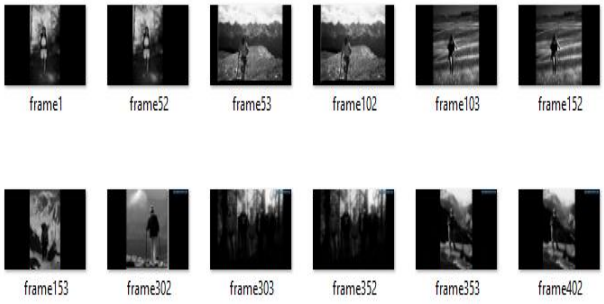

Fig. 6b. Shot Boundary Detection

Fig. 6c. Extracted Keyframes

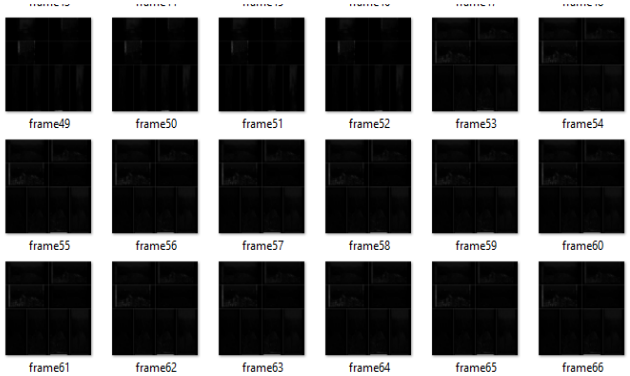

Fig. 6d. Example of Contourlet Coefficients of the frames.

Fig. 6e. People detected.

For the Proposed application, we tested 7 sample test videos. The results of the tested videos is summarized in the following table 1

Sr. no.	Name of the Video (*.mpg)	Total No. of the frames in the video	No. of the shots detected	No. of the Keyframes Extracted	No. of the People detected
1	Hikers	402	6	5	6
2	Sportsmatch1	252	4	3	15
3	Color_video	303	8	8	18
4	PedestrianCrossing	502	4	4	13
5	RandomPplWalking	377	5	5	12
6	Survillence_2	240	4	4	8
7	Road_scene	302	5	5	17

Table 1

Accuracy, Recall and Precision for extracting the keyframe was calculated and compared with the manually identified keyframes for the following 7 videos. The analysis is given in the following table 2.

Sr.No.	Name of the Video	No. of Keyframes Extracted		Recall (%)	Precision (%)	Accuracy (%)	Error Rate (%)
		VSCC	Manaully Identified				
1	Hikers	5	7	71.4	100	71.4	28.57
2	Sportsmatch1	3	5	60	100	60	40
3	Color_video	8	10	80	100	80	20
4	PedestrianCrossing	4	5	80	100	80	20
5	RandomPplWalking	5	5	100	100	100	0
6	Survillence_2	4	5	80	100	80	20
7	Road_scene	5	6	83.33	100	83.33	16.67

Table 2

Accuracy, Recall and Precision for detecting the people in the video summary (extracted keyframes) was calculated and compared with the manually identified people for the following 6 videos. The analysis is given in the following table 3

Sr.No.	Name of the Video	No. of People Detected		TP	FP	FN	Recall (%)	Precision (%)	Accuracy (%)
		VSCC output	Manaully Identified						
1	Hikers	6	9	6	-	3	66.67	100	66.67
2	Sportsmatch1	15*	21	17	-	4	80.95	100	80.95
3	PedestrianCrossing	13	12	11	2	1	91.67	84.62	78.57
4	RandomPplWalking	12	10	9	3	1	90	75	69.23
5	Survillence_2	8	9	6	2	3	66.67	75	54.55
6	Road_scene	17*	30	24	2	6	80	92.3	75

*When 2 people are very close to each other, the detecting rectangle overlaps.This overlapping rectangles is counted as one.

Table 3

IV. CONCLUSION

In this proposed work, shot boundaries were detected using contourlet transform and then extracting the keyframes using clustering was implemented and experimentally proved. The number of clusters in K-means clustering is determined by the number of the shot boundaries detected. This work show good accuracy rate and error rate in key frame extraction. It gives 100% of accuracy for the video having video shots with no similarity among others. The proposed work was tested on 7 different videos (gray as well as colour video).It successfully summarized the given video by extracting the keyframes from the detected shots. The accuracy for summarizing the video is an average of 79.25% with the average error rate of 20.75%. It was observed that the result of the summary for the video taken using a single camera and constant angle gives less accurate summary as the shots are not defined accurately.

It was observed that detecting the people using HOG Descriptor, depends the quality of the video and also the distance of the person from the camera. It does not detect the people who is too close or too far from the camera. To detect the person successfully, it searches for the entire body of the person. The people were detected with the average accuracy of 70.83%, the average recall of 79.33% and the average precision of 87.82%.

Lastly, the processing time increases as the length of the video increases.

REFERENCES

[1] RaviKansagara, DarshakThakore, MahaswetaJoshi," A Study on Video Summarization Techniques", International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 2, Issue 2, February 2014

[2] Sachan Priyamvada Rajendra and Dr. Keshaveni N," A Survey of Automatic Video Summarization Techniques", IJEECS ISSN 2348-117X Vol 3,Issue 1 April 2014

[3] Zaynab El khattabi, Youness Tabii, Abdelhamid Benkaddour," Video Summarization: Techniques and Applications", International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:9, No:4, 2015

[4] Mr. Sandip T. Dhagdi, Dr. P.R. Deshmukh," Keyframe Based Video Summarization Using Automatic Threshold & Edge Matching Rate", International Journal of Scientific and Research Publications, Volume 2, Issue 7, July 2012 1 ISSN 2250-3153

[5] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman," Discovering Important People and Objects for Egocentric Video Summarization", To appear, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012

[6] Pamarthy Chenna Rao and M. Ramesh Patnaik, "Contourlet Transform Based Shot Boundary Detection", ijsip Image Processing and Pattern Recognition, Vol.7, No.4 (2014)

[7] Minh N. Do, and Martin Vetterli, "The Contourlet Transform: An Efficient Directional Multiresolution Image Representation", IEEE Trans. Image Proc., vol. 14, no. 12, December 2005, pp. 2091-2106.

[8] Lin-Bo Cai, Zi-Lu Ying,"A New Approach of Facial Expression Recognition Based on Contourlet Transform", In. IEEE conference on wavelet analysis and pattern recognition, pp 275-280 (2009)

[9] Wei-shi Tsai," Contourlet Transforms for Feature Detection"

[10] Pamrthy Rao, Ramesh Patnaik,"Video Summarization using Contourlet Transform and ecludian Distance", International Journal of Electronics and Communication Engineering (IJECE) ISSN(P): 2278-9901; ISSN(E): 2278-991X Vol. 2, Issue 5, Nov 2013, 225-230

[11] WenzhuXu, Lihong Xu, "A Novel Shot Detection Algorithm Based on Clustering ", 2010 2nd International Conference on Education Technology and Computer (ICETC)

[12] Victor Shen, Hsin-Yi Tseng, Chan-Hao Hsu, "Automatic Video Shot Boundary Detection of News Stream Using a High-Level Fuzzy Petri Net ",2014 IEEE International Conference on Systems, Man, and Cybernetics ,October 5-8, 2014

[13] Nikita Sao, Ravi Mishra," A survey based on Video Shot Boundary Detection techniques", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 4, April 2014

[14] A.V.Kumthekar, Prof.Mrs.J.K.Patil," Key frame extraction using color histogram method", International Journal of Scientific Research Engineering & Technology (IJSRET), Volume 2, Issue 4, pp 207-214 ,July 2013 , ISSN 2278 – 0882

[15] Azra Nasreen, Kaushik Roy, Kunal Roy, Shobha G," Key Frame Extraction and Foreground Modelling Using K-Means Clustering", 7th International Conference on Computational Intelligence, Communication Systems and Networks

[16] Supriya Kamoji, Rohan Mankame, Aditya Masekar, Abhishek Naik," KeyFrame Extraction for Video Summarization using motion activity", IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308

[17] Padmavathi Mundur, Yong Rao, Yelena Yesha," Keyframe-based Video Summarization using Delaunay Clustering", International Journal on Digital Libraries Volume 6 Issue 2, April 2006 Pages 219 - 232

[18] N. Dalal,B. Triggs, "Histograms of oriented gradients for human detection", IEEE Computer Society Conference on computer vision and Pattern Recognition, 2005

[19] Rakesh Kumar, Tapesh Parashar, Gopal Verma, "Background Modeling and Subtraction Based People Counting for Real Time Video Surveillance", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-5, November 2012

[20] Shafraz Subdurally, Devin Dya, Sameerchand Pudaruth, "Counting People Using Blobs and Contours", International Journal of Computer Vision and Image Processing, 3(2), 1-16, April-June 2013

[21] Cansın Yıldız, "An Implementation on Histogram of Oriented Gradients for Human Detection"