Video Summarization using clustering

Mr. Satvik Khara[1], Mr. Bhargav Modi[2], Mr. Darshil J. Shah[3], Mr. Rikin Thakkar[4]
[1]*H.O.D. Computer Department, SOCET*
[2]*A/Prof. Computer Department, SOCET*
[3]*A/Prof. Computer Department, SOCET*
[4]*A/Prof. Computer Department, SOCET*

***Abstract*— In this paper, we present modified approach for generating static video summary. This approach is based on extracting feature using color, texture and shape feature of video frames and then generate summary using modified DBSCAN algorithm. Video summary generate by this method are compared to other methods.**

***Index Terms*- Video summarization, Color, Texture and Shape feature, clustering**

I. INTRODUCTION

Video summarization is a technique to get a still or moving sequence of images from given original video as a summary for that video [6].

With the increase of daily usage of the recording technology, video has become a popular part of daily life. This situation led to a mass of raw information that needs evaluating before everything at all [4].

In today's world development of computation, communications in between, and storage infrastructures, are rapidly increase and contributing to a large and steadily increase availability of video content [5]. Wide investment in digital video is done after that, the capabilities of an average user to handle, interact with and manage videos are still not achieve what average users can do with other types of media such as text or images. The Main reason behind this is the temporal storage nature of video and the size of the video.

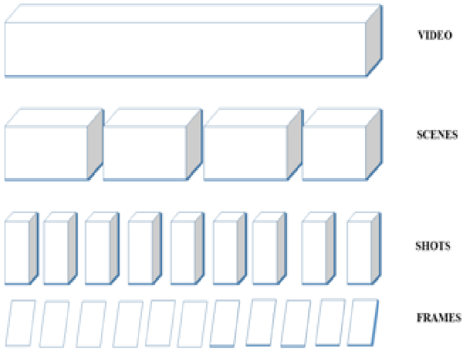

Figure 1 hierarchy structure of video [7]

A video is nothing but a synchronous sequence of a number of frames, each frame being a 2-D image. So the basic unit in a video is a frame. The video can also be thought of as a collection of many scenes, where a scene is a collection of shots that have the same context. This is shown in figure-1. A frame is single still image from video. It defines as fps (frame per second).

From over past years, different approach has been proposed for video summarization. These approaches have number of drawbacks. First, almost all approaches are based on single visual descriptor like color of video. Second, clustering algorithm used in that approaches could not detect noise frame automatically. Third, current approaches are basis of some special input like required specified input for clustering.

Basically video summarization method divided in two parts (1) Static video summarization and (2) Dynamic video summarization [8], [9], [10], [18].

II. RELATED WORK

Some of previously used techniques for static video summary are given below.

In Keyframe-based video summarization using Delaunay clustering [1], an approach is developed using Delaunay clustering (DT). Firstly, pre-sampling step is executing. Then video frames are presented using color histogram in HSV space. After that, the Delaunay diagram is built and clusters are formed by separating edges in the Delaunay diagram. Finally, for each cluster, the frame that is closest to its center is selected as the key frame.

In VSUMM [16], a video summarization approach is explained. In first step, pre-sampling is executed on video frames using one frame per seconds. Then color feature is extracted from video frames in HSV space. Then clustering is applied to frames and from every cluster key frame is selected. Finally, one other step occurs in which the key frames are compared with each other through color histogram to eliminate those similar key frames in the produced summaries.

In VSCAN [5], an approach is used density based clustering. In first step, video frames are pre-sampled using one frame per second rate. In second step, color feature and texture feature of video frames are extracted. Color feature extraction is done using color histogram in HSV color space. Texture feature extraction is done using Discrete Wavelet Transformation (DWT) [19]. Then DBSCAN is applying for format the clustering. Finally select key frames from each cluster.

III. PROPOSED METHOD

In figure 2, steps for proposed approach are given. First, original video is pre-sampled. In Second step, feature extraction is done using different techniques. Color feature are extracted from video using color histogram in HSV space. Texture feature are extracted using Gabor method. Shape feature are extracted using Fourier techniques.

Input: static video, video dataset

Output: Summary of given input video.

Procedure: Procedure is as follows:-

Step 1: Pre-sampling

- Reduce the number of frames to be processed from original video.

Step 2: Features extraction

In these step features are extracted from video frame.

- Extract Color feature using color histogram techniques
- Extract Texture feature using Gabor techniques
- Extract Shape feature using Fourier techniques

Step 3: Video Frame clustering

- For clustering modified DBSCAN algorithm is used.
- For dissimilarity measure Bhattacharyya distance equation is used.

$$\text{Bhattacharyya Distance} = \sum_{i=0}^{n} \sqrt{\sum Pi \bullet \sum Qi} \quad (1)$$

Step 4: Steps for clustering

1) Select an arbitrary frame p.
2) Retrieve all frames that are indirectly-similar from p.
3) If p is a core frame, a video cluster is formed.
4) If p is a border frame, no frames are indirectly-similar from p and the next frame of the database is visited.
5) Continue the process until all the frames have been processed.

Step 5: Key Frame extraction

Select the key frames from the video clusters and Noise frames are discarded.

- For each cluster the middle core frame in the ordered frames sequence is selected to construct the video summary.

Step 6: Make video from Selected Key Frames.

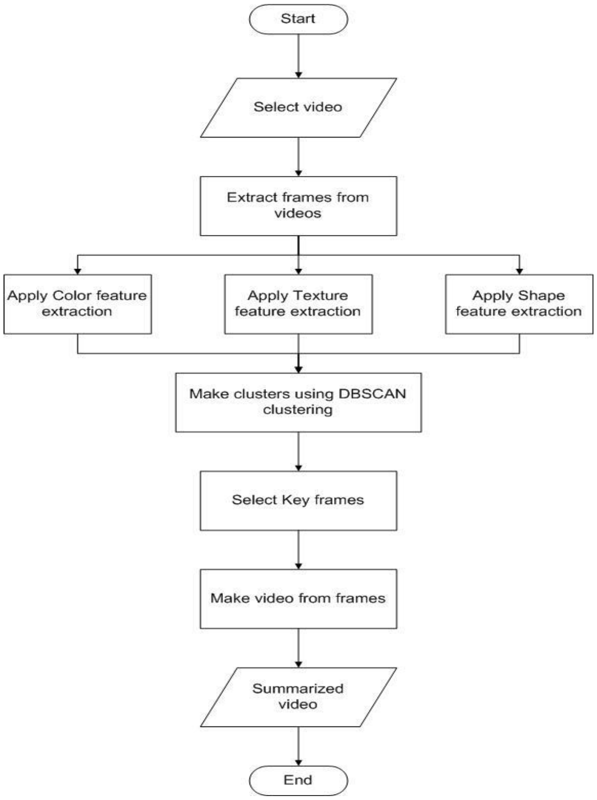
Start
Select video
Extract frames from videos
Apply Color feature extraction
Apply Texture feature extraction
Apply Shape feature extraction
Make clusters using DBSCAN clustering
Select Key frames
Make video from frames
Summarized video
End

Figure 2 Proposed Model

Description of Procedure:

The first step for video summarization is pre-sampling to reduce the number of frame. Sampling rate is important for summary. In this approach, the sampling rate used is selected to be one frame per second.

The second step for summary is feature extraction. In that first color feature is extracted from video frame. For that color histogram [15] is used to describe visual content of video image. The color histogram[13] is computed from HSV color space where 32 bins of H, 4 bins of S and 2 bins of V. HSV also defines the brightness and intensity of images.

Then texture feature is extracted from video frames. Texture is a powerful low-level feature for representing images. Gabor is one of the good methods to extract texture feature from frames. Here we use the Gabor method to extract the texture feature. The input to Gabor method[14] is image. Then the function applied on that image takes four other parameters. They are as follows:

- u: No. of scales (mostly set to 5)
- v: No. of orientations (mostly set to 8)
- m: No. of rows in a 2-D Gabor filter (an odd integer number mostly set to 39)
- n: No. of columns in a 2-D Gabor filter (an odd integer number mostly set to 39)

Output to this function is a gaborArray: A u by v array, element of which is m by n matrices; each matrix being a 2-D Gabor filter.

Shape of an object is the characteristic surface which is represented by the outline or contour. Shape recognition or detection is one of the modes through which human perception of the environment is executed. Here we use the Fourier transformation method to extract the Shape feature. The input to this method is image. It detects the boundary of that given input image.

Y = fft2(X) returns the two-dimensional discrete Fourier transform (DFT) of X. The DFT is computed with a fast Fourier transform (FFT) algorithm. The result, Y, is the same size as X. If the dimensionality of X is greater than 2, the fft2 function returns the 2-D DFT for each higher dimensional slice of X. For example, if size (X) = [100 100 3],then fft2 computes the DFT of X (:,:,1) , X(:,:,2) and X(:,:,3). Y = fft2(X,m,n) truncates X, or pads X with zeros to create an m-by-n array before doing the transform. The result is m-by-n.

In next step, DBSCAN clustering is density based clustering algorithm [20], [21] is applied to detect the cluster for given inputs. Using DBSCAN clustering algorithm has many advantages. First, it does not require specifying the number of clusters in the data a priori, as opposed to partitioning algorithms like k-means [12]. Second, DBSCAN can find arbitrarily shaped clusters. Third, it has a notion of noise. Finally, DBSCAN requires minimal number of input parameters.

In this approach, we apply a dual feature space DBSCAN algorithm. The proposed clustering algorithm used in this approach aims at adapting and modifying DBSCAN to be used by a video summarization system that utilizes all three colors, texture and shape features. Instead of accepting only one input dataset as in the original DBSCAN, the clustering algorithm in this approach accepts all three color, texture and shape features of video frames as input datasets, with the Bhattacharya distance [2] as a dissimilarity measure.

As in equation (1) Bhattacharyya distance is used. Selecting the Bhattacharyya distance as dissimilarity measure has many advantages [3]. First, the Bhattacharyya measure has a self-consistency property, as by using the Bhattacharyya measure all Poisson errors are forced to be constant therefore ensuring the minimum distance between two observations points is indeed a straight line. The second advantage is the independence between Bhattacharyya measure and the histogram bin widths, as for the Bhattacharyya metric the contribution to the measure is the same irrespective of how the quantities are divided between bins; therefore it is unaffected by the distribution of data across the histogram. Third advantage is that the Bhattacharyya measure is dimensionless; as it is not affected by the measurement scale used [2].

Finally frame is selected from each cluster to make video.

IV. EXPERIMENTS AND RESULTS

In this paper, a modified version of an evaluation method Comparison of User Summaries (CUS) described in [3] is used to evaluate the quality of video summaries. In CUS method, the video summary is built manually by a number of users from the sampled frames and the user summaries are taken as reference (i.e. ground truth) to be compared with the automatic summaries obtained by different methods.

$$F - measure = \frac{2 \times \Pr ecision \times \operatorname{Re} call}{\Pr ecision + \operatorname{Re} call} \qquad (2)$$

In order to evaluate the automatic video summary, the F-measure is used as a metric. The F-measure consolidates both Precision and Recall values into one value using the harmonic mean [11], and it is defined as: The Precision measure of video summary is defined as the ratio of the total number of color-based similar frames and texture-based similar frames to the total number of frames in the automatic summary; and the Recall measure is defined as the ratio of the total number of color-based similar frames and texture based similar frames to the total number of frames in the user summary.

In this approach is evaluated on a set of 50 videos selected from the Open Video Project 1. All videos are in MPEG-1 format (30 fps, 352 240 pixels). They are distributed among several genres (documentary, historical, lecture, educational) and their duration varies from 1 to 4 min. Also, we use the same user summaries used in [16] as a ground-truth data. These user summaries were created by 50 users, each one dealing with 5 videos, meaning that each video has 5 summaries created by five different users. So, the total number of video summaries created by the users is 250 summaries and each user may create different summary. For comparing VSCAN approach with other approaches, we used the results reported by three approaches: VSCAN [5], VSUMM [16], STIMO [17], and DT [1]. In addition to that, the automatic video summaries generated by our approach were compared with the OV summaries generated by the algorithm in [12]. All the videos, user summaries, and automatic summaries are available publicly.

Table 1 Mean F-measure achieved by different approaches

Approach	Mean F-Measure
OV	0.67
DT	0.61
STIMO	0.65
VSUMM	0.72
VSCAN	0.77
DB-Color	0.74
VSUC	**0.81**

Table 1 shows the mean F-measure achieved by the different video summarization approaches. The results indicate that this approach performs better than all other approaches. Also, we notice that combining all three features color, texture and shape features together more done in this approach gives better results than using color features only as in DB-Color. However, DB-Color achieved better results if compared to the other four approaches (OV, DT, STIMO, VSCAN and VSUMM), which indicates that using DBSCAN clustering algorithm is efficient for generating static video summary.

V. CONCLUSION

Proposed approach uses color, texture and shape features of the video frames for summarize the video content. Combining all color, texture and shape features enabled this to overcome the drawback of using single features only as in other approaches. Also, as an advantage of using a density-based clustering algorithm, this reduce extra step needed for estimating the number of clusters is avoided. It also removes noisy frames form video.

Future work includes combining other features to this approach like edge and motion descriptors. Also, another interesting future work could be generating video skims (dynamic key frames, e.g. movie trailers) from the extracted key frames.

REFRENCES

[1] Mundur, P., Rao, Y., Yesha, Y.: Keyframe-based video summarization using Delaunay clustering. International Journal on Digital Libraries-(Springer) 6(2), 219–232 (2006)

[2] Kailath, T.: The divergence and bhattacharyya distance measures in signal selection. IEEE Transactions on Communication Technology 15(1), 52–60 (1967)

[3] Aherne, F.J., Thacker, N.A., Rockett, P.I.: The bhattacharyya metric as an absolute similarity measure for frequency coded data. Kybernetika 34(4), 363–368 (1998).

[4] Ebrahim Asadi*, Nasrolla Moghadam Charkari," Video Summarization Using Fuzzy C-Means Clustering", 20th Iranian Conference on Electrical Engineering, (ICEE2012), May 15-17, Tehran, Iran

[5] Mahmoud, K. M., Ismail, M. A., & Ghanem, N. M. (2013). VSCAN: An Enhanced Video Summarization Using Density-Based Spatial Clustering. In*Image Analysis and Processing–ICIAP 2013* (pp. 733-742). Springer Berlin Heidelberg.

[6] Ajmal, M., Ashraf, M. H., Shakir, M., Abbas, Y., & Shah, F. A. (2012). Video summarization: techniques and classification. In *Computer Vision and Graphics*(pp. 1-13). Springer Berlin Heidelberg.

[7] Rajendra, S. P., & Keshaveni, N. A Survey of Automatic Video Summarization Techniques.

[8] Taskiran, C., & Delp, E. (2005). Video summarization. *Digital Image Sequence Processing, Compression, and Analysis*, 215-231.

[9] Truong, B. T., & Venkatesh, S. (2007). Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, *3*(1), 3.

[10] Ravi Kansagara , Darshak Thakore, Mahasweta Joshi. A study on video Summarization Techniques. *A International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization)Vol. 2, Issue 2, February 2014.*

[11] Blanken, H.M., De Vries, A., Blok, H.E., Feng, L.: Multimedia retrieval. Springer,Heidelberg (2007).

[12] DeMenthon, D., Kobla, V., Doermann, D.: Video summarization by curve simplification.In: Proceedings of the Sixth ACM International Conference on Multimedia, pp. 211–218. ACM Press (1998)

[13] Chary, R., Lakshmi, D. R., & Sunitha, K. V. N. (2012). Feature extraction methods for color image similarity. *arXiv preprint arXiv:1204.2336*.

[14] ping Tian, D. (2013). A Review on Image Feature Extraction and Representation Techniques. *International Journal of Multimedia and Ubiquitous Engineering*.

[15] Singha, M., & Hemachandran, K. (2012). Content based image retrieval using color and texture. *Signal & Image Processing: An International Journal (SIPIJ)*,*3*(1), 39-57.

[16] de Avila, S. E. F., & Lopes, A. P. B. (2011). VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters*, *32*(1), 56-68.

[17] Furini, M., Geraci, F., Montangero, M., & Pellegrini, M. (2010). STIMO: STIll and MOving video storyboard for the web scenario. *Multimedia Tools and Applications*, *46*(1), 47-69.

[18] Li, Y., Merialdo, B., Rouvier, M., & Linares, G. (2011, November). Static and dynamic video summaries. In *Proceedings of the 19th ACM international conference on Multimedia* (pp.1573-1576)ACM

[19] Stanković, R. S., & Falkowski, B. J. (2003). The Haar wavelet transform: its status and achievements. *Computers & Electrical Engineering*, *29*(1), 25-44.

[20] Yang, X., & Cui, W. (2008, December). A novel spatial clustering algorithm based on delaunay triangulation. In *International Conference on*

Earth Observation Data Processing and Analysis (pp. 728530-728530). International Society for Optics and Photonics.

[21] Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (1998). Density-based clustering in spatial databases: The algorithm gdbscan and its applications. *Data mining and knowledge discovery*, *2*(2), 169-194.

[22] Zhang, G., Ma, Z. M., Tong, Q., He, Y., & Zhao, T. (2008, August). "*Shape feature extraction using Fourier descriptors with brightness in content-based medical image retrieval*." In Intelligent Information Hiding and Multimedia Signal Processing, 2008. IIHMSP'08 International Conference on. ISBN: 978-0-7695-3278-3 pp. 71-74. IEEE