VISUC: VIdeo Summarization with User Customization

Rameswar Panda, Sanjay K. Kuanar, Ananda S. Chowdhury.
{rameswar183, sanjay.kuanar }@gmail.com, {aschowdhury@etce.jdvu.ac.in}
Department of Electronics and Telecommunication Engineering
Jadavpur University, Kolkata – 700032, India.

Abstract – Design of video storyboards, which enables a user to access any video in a friendly and meaningful way, has emerged as an important area of research in the multimedia community. In this paper, we propose a novel semi-automated method for construction of video storyboards based on Delaunay graphs. A robust edge pruning strategy, where the edge weights are assumed to follow a Gaussian distribution, is applied on an appropriately constructed Delaunay graph. The proposed method also takes into account two advanced user needs, namely the waiting time and the number of frames an user wants to see in the storyboard. Experimental results on some standard videos of different genre clearly indicate the superiority of the proposed method in terms of the $F_{0.5}$ measure.

Index Terms- **Video Storyboard, Delaunay Graphs, Edge Pruning, User customization, Gaussian distribution.**

I. INTRODUCTION

Due to the rapid advances in data storage, data compression, and data transmission, information pertaining to videos is massively entering our life. This growing availability of video information exposes consumers to video library with a very large number of videos. Hence, video browsing has become a common activity aimed at finding the right video. To facilitate such activity, each available video should be represented with a temporally reduced version so that browsing may be performed on the reduced version and the consumer can decide which video he/she wishes to watch. Since it would be too expensive to manually produce such reduced versions, it is necessary to develop mechanisms that automatically produce such versions. This has been the goal of a quickly evolving research area known as video summarization. Video summarization can be broadly classified into two categories, namely, Video Storyboard and Video Skimming. Storyboard is a set of static key frames which preserves the overall content of a video with minimum data. Video skimming refers to a set of images with audio and motion information. Though the technique of skimming provides important pictorial, audio and motion information, video storyboard summarizes the video content in a more rapid and compact manner [1].

Different clustering techniques have been proposed in the literature to address the problem of summarizing a video sequence [2-5]. Although these existing approaches produce summaries with acceptable quality, their performance heavily depends on user inputs and/or certain threshold parameters [3-5]. This type of dependency on threshold parameters makes the clustering process very expensive and time consuming. Some recent approaches use the notion of similarity between successive frames to obtain the key frames [6]. However, choice of similarity measures greatly influences the effective content representation of the key frame set. Avila et al. presented VSUMM (Video SUMMarization), an approach to cope with the video summarization problem in which the clustering step is obtained using the k-means algorithm [3]. Key frames produced by this algorithm fail to preserve large portion of the video content due to several limitations of the k-means algorithm such as circular polarization and high chance of trapping at local minima. Moreover, the number clusters produced by a shot boundary detection method is not accurate as several types of transitions are present within successive shots (e.g., fade in, fade out, abrupt cut). This makes the shot detection method computationally intensive for different genres of videos having large number of video frames. Another drawback of this scheme is the complete lack of user customization. In this paper, we present **VISUC (Video Summarization with User Customization)**, a novel Delaunay graph-based clustering algorithm with several improvements over [3]. A Delaunay Graph based clustering method, called Global standard deviation reduction (GSDR_DC) can be found in [2]. Our method splits the Delaunay graph using a better edge pruning strategy where selection of a proper edge is determined using standard deviation and average of edge lengths. Moreover, the proposed method was designed to offer user customization. In particular, users can specify the number of key frames they actually want to view and also specify the time they are willing to wait. Hence, visual dynamics of the frames are captured better and a more informative video summarization is achieved with better user perception.

II. THEORETICAL FOUNDATIONS

Our clustering strategy is based on efficient pruning of edges in a Delaunay graph to provide the number of key frames as indicated by a user. Some useful definitions related to this method are provided in this section.

Definition 1: ***Delaunay triangulation (DT)*** of a point set is the straight line dual of the Voronoi Diagram, used to represent the

978-1-4673-4700-6/12/$31.00 ©2012 IEEE

interrelationship between each data point in multi-dimensional space to its nearest neighboring points.

Definition 2: Under the standard assumption that no four points are co circular, the Delaunay triangulation satisfies the property of a triangulation [7] and the corresponding graph is called the ***Delaunay graph***. An edge *ab* in a Delaunay graph *D(P)* of a point set *P* connecting points *a* and *b* is constructed iff there exists an empty circle through *a* and *b* [7]. The closed disc bounded by the circle contains no sites of *P* other than *a* and *b*.

Definition 3: Weight of an edge in a Delaunay Graph *D(P)* is equal to its length, i.e., the distance between the two vertices constituting the edge. ***Average weight*** corresponds to the mean length of all the edges present in *D(P)* and is defined as:

$$\hat{w}(D(P)) = \frac{1}{N}\sum_{j=1}^{N}|e_j| \quad (1)$$

Where $\hat{w}(D(P))$ denotes average weight of *N* edges in *D(P)*.

Definition 4: A ***connected component*** of an undirected graph is a subgraph in which any two vertices are connected to each other by paths, and which is connected to no additional vertices of the original graph.

III. THE PROPOSED ALGORITHM

There exist several measures for describing the dispersion of the dataset. In VISUC, we use standard deviation as a measure of dispersion of the data set [8]. Given a point set *S* in multi-dimensional space and the desired number of clusters *k*, our method starts by constructing a Delaunay graph from the points in *S*. We assume the edge weights *W* of the Delaunay graph to follow a normal distribution:

$$W \sim N(\hat{w}, \sigma) \quad (2)$$

In equation (2), $\hat{w}$ *and* σ respectively represents the mean and the standard deviation of the normal distribution. Average weight $\hat{w}$ of the edges in the entire Delaunay Graph and its standard deviation σ are first computed (see Definition 3). It is a well-known property of the normal distribution that two standard deviations from mean account for 95.45% of the total population. Following this property, any edge with a weight $w > \hat{w} + 2\sigma$ is removed from the graph. This is because such an edge is most likely to be an inter-cluster edge which connects two clusters. This leads to a set of disjoint connected components of the graph $DG_K = \{C_1, C_2, \ldots, C_K\}$. Each connected component C_i is treated as a cluster, which has a centroid s_i. This edge removal process is repeated until we get the desired number of clusters. The modeling and separating processes are shown in Fig. 1. VISUC allows advanced user customization. The users can specify the number of key frames they actually want to view and can also specify the time they are willing to wait to get the summary. This feature is offered because the production time of a video summary depends on the video length. The longer the duration, the longer is the production time. Several observational studies on user behavior have shown that waiting time is critical [9]. For instance, up to 5secs to get a complete web page is considered a good waiting time, whereas more than 10secs is considered to be poor. However, if the web page loads incrementally, waiting time up to 39secs is considered acceptable [10].

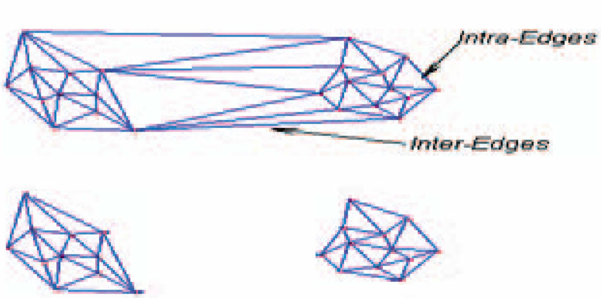

Figure 1. Delaunay Edge Removal Process (Top showing the original Delaunay Graph. The connected components after the edge removal process is shown below)

In order to provide advanced user customization in terms of waiting time, we synchronize the sampling rate in order to reduce the total number of processed frames. We have varied the sampling rate between 10 and 60 according to the input waiting time. In other words, for maximum waiting time of 39secs, we have taken the sampling rate to be 10. It may also be noted that the shorter the waiting time, the poorer the quality of the summary, as many frames need to be discarded. Moreover, several experiments on user behavior have shown that the number of key frames is also very crucial in the production of video summaries. Through experiments over our test video collection, we have found that values between 1% and 2% of the total sampled frames as good choices. Various steps of the proposed algorithm are summarized in figure 2.

1. Sampling: Sample the input video sequence to get the selected frames according to the user input waiting time. (Waiting time in the range of 5s to 39s).
2. Feature Extraction: Extract color histogram from each selected frame to form the frame-feature vector. For our problem, each frame is represented by a 256 (16 ranges of H, 4 ranges of S, and 4 ranges of V) dimensional feature vector. HSV color space is chosen as it captures human perception of color better and is more robust to noise [3].
3. Dimensionality Reduction: Use principal component analysis (PCA) to reduce the dimension of the above feature vector. Through experimental tests, we have found that the number of principal components between 5 and 7 is sufficient to capture 90% or more of the total variation for most of the videos in our test collection.
4. Delaunay Graph Construction: Generate *DG* for the 5-7 dimensional feature vectors. Compute the average weight $\hat{w}$ and standard deviation σ of the edges in the entire DG.
5. Edge Removal Process: Assuming the edge weights follow a Gaussian distribution, any edge with a weight $w > \hat{w} + 2\sigma$ is selected for removal from the graph. Find the remaining connected components after removal of the selected edge to obtain the individual clusters.
6. Stopping Criteria: Repeat step 5 until we get the desired number of clusters as input by the user.
7. Key Frame Selection: The frames which are closest to the centroid of each cluster (obtained from the final Delaunay graph) are deemed as the key frames. Finally, the key frames are arranged in a temporal order to make the video storyboard more understandable.

Figure 2. VISUC Algorithm

IV. Experimental Results

Unlike other research areas, evaluating a video summary is not a straightforward task due to the lack of an objective ground-truth. A consistent evaluation framework is seriously missing for video summarization research. In this work, we adopted a subjective evaluation method to assess the quality of video summaries, known as Comparison of User Summaries (CUS) [3]. It incorporates the judgment of the user in evaluating the quality of a video summary. Initially, the subjects are asked to watch the whole video. Then, they are asked to select a subset of frames which they think is able to summarize the video content. Each subject is free to select any number of frames to compose his/her summaries. Finally, their summaries are compared with the summaries provided by various algorithms.

The standard measures Precision and Recall can then be used to evaluate the automatic summary. Precision is the ratio of the number of matching frames to the total number of frames in the automatic summary.

$$Precision = \frac{n_{m1}}{n_{TAS}} \qquad (3)$$

In equation (3), n_{m1} is the number of matching frames and n_{TAS} is the total number of frames in the automatic summary. Recall is the ratio of the number of matching frames to the total number of frames in the user summary.

$$Recall = \frac{n_{m2}}{n_{TUS}} \qquad (4)$$

In equation (4), n_{m2} and n_{TUS} respectively represent the number of matching frames in automatic summary and the total number of frames in the user summary. However, there is a trade-off between precision and recall. Greater precision decreases recall and greater recall leads to decreased precision. So, we choose the $F_{0.5}$ as the metric used for assessing the quality of the automatic summaries. The $F_{0.5}$ combines both precision and recall into a single measure by a harmonic mean [11]:

$$F_{0.5} = \frac{2 \times Precision \times Recall}{Precision + Recall} \qquad (5)$$

We have so far experimented with 5 test video segments belonging to different genres and having different durations (40 sec. to 2 min) from the Open Video (OV) projects [13]. Each test video is in MPEG-1 format with a frame rate of 29.97 and the frames having dimensions of 352x240 pixels. Long videos are avoided due to limitation of annotation by a subject. The parameters used to produce the video summaries using VISUC are: waiting time = 22s (average waiting time) corresponding to sampling rate of 35 (same as the sampling rate adopted in VSUMM); number of key frames = 1% of the video length. Results of the proposed method for all 5 videos can be seen at https://sites.google.com/site/ivprgroup/home/research/vsuc. All the videos, the user summaries, and the storyboards produced by the VSUMM approach are available at http://www.sites.google.com/site/vsummsite/. Table 1 presents the mean $F_{0.5}$ achieved by both the approaches for several video categories. The results indicate that VISUC performs better than the VSUMM for all of the videos in our collection. In order to verify the statistical significance of those results, the confidence intervals for the differences between paired means of VISUC and VSUMM are computed. If the confidence interval includes zero, the difference is not significant at that confidence level. If the confidence interval does not include zero, then the sign of the mean difference indicates which alternative is better [12]. Since the confidence intervals (with a confidence of 98%) do not include zero, the results presented in Table 2 confirms that the betterment of VISUC over VISUMM (higher $F_{0.5}$ values) is statistically significant. Fig. 3 presents the video summaries produced by both these approaches for the video ***Drift Ice as a Geologic Agent, segment 07***. The user summaries for the same video are presented in Fig. 4. From Fig. 4, it can be noted that for most of the user summaries, our proposed method VISUC achieves higher $F_{0.5}$ value as compared to VSUMM. Notice that the summary produced by VSUMM has less information content as compared to our VISUC approach. The summary with the highest quality is achieved by our approach, which can also be confirmed by a visual comparison with the user summaries.

Category	#Videos	VSUMM	VISUC
Documentary	2	0.649	0.811
Educational	1	0.842	0.907
Lecture	2	0.665	0.828
Weighted average	**5**	**0.694**	**0.837**

Table 1. MEAN $F_{0.5}$ ACHIEVED BY BOTH VSUMM AND VISUC FOR SEVERAL VIDEO CATEGORIES

Method	Confidence Interval (98%)	
	Min.	Max.
VISUC - VSUMM	**0.143**	**0.278**

Table 2. DIFFERENCE BETWEEN MEAN $F_{0.5}$ AT A CONFIDENCE OF 98%

V. Conclusion and Future work

We propose a video summarization technique based on novel edge pruning in Delaunay graphs which provides advanced user customization. Experimental results show that our technique VISUC outperforms the work described in [3]. Future work will focus on the evaluation of more complex forms of user interaction, such as the use of the user feedback to refine the video summary. Another direction of future research is to produce multi-view video summaries for real world surveillance systems. We also plan to work on the personalized video summaries with a focus on unobtrusively sourced user-based information.

References

[1]. B.T. Truong, S. Venkatesh, Video abstraction: a systematic review and classification, *ACM Transactions on Multimedia Computing, Communications, and Applications*, 3 (1) pp. 1–37, 2007.

[2]. A.S. Chowdhury, S. Kuanar, R. Panda and M.N. Das. Video Storyboard Design using Delaunay Graphs, *Twenty First IAPR/IEEE Int'l. Conf. on Pattern Recognition (ICPR);* pp. 3108-3111, 2012.

[3]. S.E.F. Avila, A.P.B. Lopes, A. Jr. Luz and A.A. Araujo. VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters*, 32 (1), pp. 56–68, 2011.

[4]. Y. Gong and X. Liu. Video summarization and Retrieval using Singular Value Decomposition. *ACM Multimedia Systems Journal*, 9(2), pp. 157-168, 2003.

[5]. D. Q. Zhang, C. Y. Lin, S. F. Chang, and J. R. Smith. Semantic video clustering across sources using bipartite spectral clustering. *Proc. IEEE Conference on Multimedia and Expo (ICME),* 117-120, 2004.

[6]. J. Almeida, N.J. Leite and R.S. Torres. VISON: VIdeo Summarization for ONline applications. *Pattern Recognition Letters,* pp. 397-409, 2012.

[7]. Joseph O' Rourke, Computational Geometry in C, *Cambridge University Press*, New York, 2005.

[8].Ying Xia and Xi Peng, A Clustering Algorithm based on Delaunay Triangulation. *Proc. of the 7th World Congress on Intelligent Control and Automation,* Chongqing, China, 2008.

[9]. C.W .Johnson, M.D.Dunlop, Subjectivity and notions of time and value in interactive information retrieval. *Interact. Comput.* 10 (1), 67–75, 1998.

[10]. A. Bouch, A.Kuchinsky, N.T. Bhatti, Quality is in the eye of the beholder: Meeting users' requirements for internet quality of service. *In: ACM Internat. Conf. Human Factors in Comput. Syst.,* pp. 297–304, 2000.

[11]. H. Blanken, A.P.Vries, H.E. Blok, L. Feng. Multimedia Retrieval. *Springer-Verlag*, Inc, 2007.

[12]. R. Jain. The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling. *John Wiley and Sons*, Inc.,1991.

[13].The Open Video Project: http://www.open-video.org.

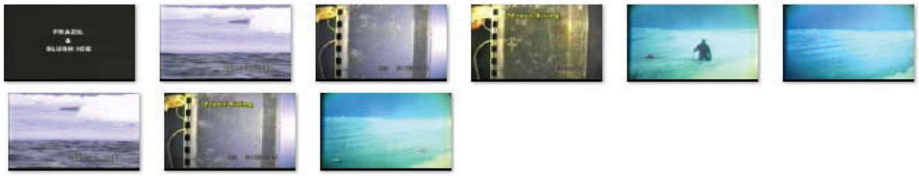

Figure 3. Summarization results for the video "*Drift Ice as a Geologic Agent, segment 07*": Top row -> VISUC (Mean $F_{0.5}$ measure: 0.839), bottom row -> VSUMM [3] (Mean $F_{0.5}$ measure: 0.591)

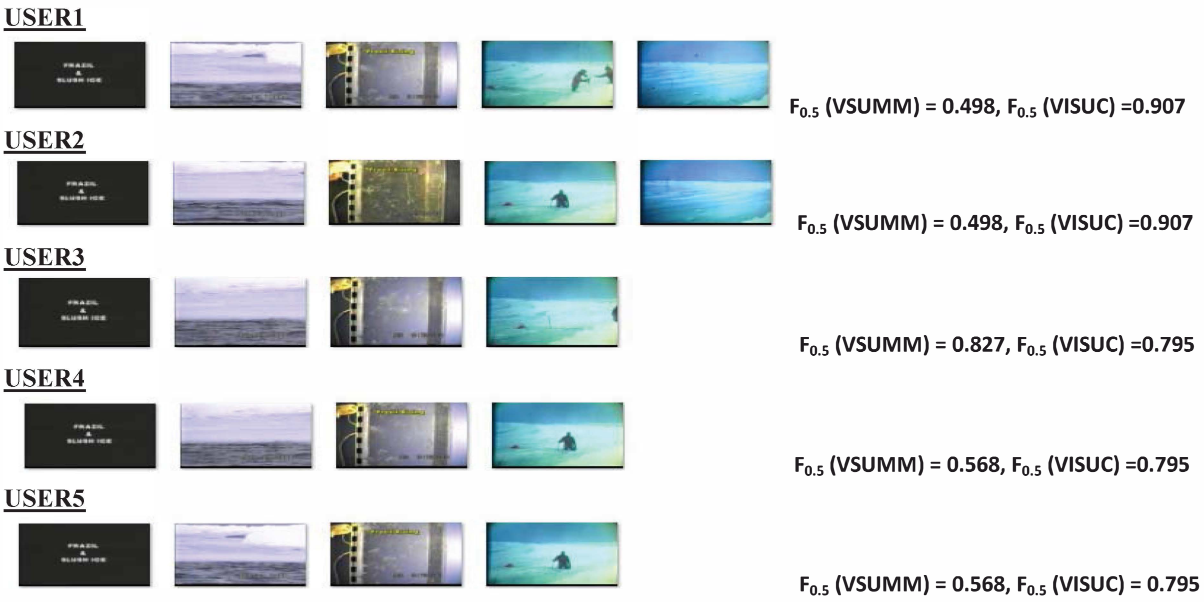
USER1
$F_{0.5}$ (VSUMM) = 0.498, $F_{0.5}$ (VISUC) =0.907
USER2
$F_{0.5}$ (VSUMM) = 0.498, $F_{0.5}$ (VISUC) =0.907
USER3
$F_{0.5}$ (VSUMM) = 0.827, $F_{0.5}$ (VISUC) =0.795
USER4
$F_{0.5}$ (VSUMM) = 0.568, $F_{0.5}$ (VISUC) =0.795
USER5
$F_{0.5}$ (VSUMM) = 0.568, $F_{0.5}$ (VISUC) = 0.795

Figure 4. User summaries for the video "*Drift Ice as a Geologic Agent, segment 07*"