21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Video Storyboard Design using Delaunay Graphs

Ananda S. Chowdhury, Sanjay K. Kuanar, Rameswar Panda, Moloy N. Das
{aschowdhury@etce.jdvu.ac.in,}, {sanjay.kuanar, rameswar183, dasmoloy87}@gmail.com

Department of Electronics and Telecommunication Engineering
Jadavpur University, Kolkata – 700032, India.

Abstract

Design of video storyboards has emerged as a popular research area in the multimedia community. Different pattern clustering techniques are applied to extract the key frames from a video sequence to form a storyboard. In this paper, we propose an automatic method for the selection of key frames of a video sequence using Delaunay graphs. We prune certain edges from the Delaunay graph using an iterative strategy where overall reduction in the global standard deviation of edge lengths is maximized. Resulting connected components in the graph correspond to the separate clusters. The proposed algorithm also utilizes edge information in addition to the color histogram information to achieve semantic dependency between different video frames. Performance of our algorithm is evaluated using Fidelity, Shot Reconstruction Degree and Compression Ratio. Experiments on standard video datasets indicate the supremacy of the proposed method over a previous Delaunay clustering-based key frame extraction algorithm.

Keywords: Video storyboard, Delaunay graph, Edge pruning, Global standard deviation reduction.

1. Introduction

Video summarization is a nonlinear content-based video compression technique which efficiently represents most significant information in a video stream using a combination of still images, video segments, graphical representations and textual descriptors [1]. Video summarization can be broadly classified into two categories: Storyboard and Video Skimming [2]. Storyboard is a set of static key frames (motionless images) which preserves the overall content of a video with minimum data. Video skimming refers to a set of images with audio and motion information [3]. 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. Various clustering methods are applied over the years to design a video story board through extraction of key frames [3, 5-6]. Performance of such clustering methods heavily depends on user inputs and/or certain threshold parameters [3, 5]. Some recent clustering approaches use the notion of similarity between successive frames [4]. However, choice of similarity measures greatly influences the effective content representation of the key frame set. Mundur *et al.* used Delaunay triangulation-based clustering (DC) to automatically extract the key frames in a video [7]. The edges of a Delaunay graph are classified into short edges and separating edges using average and standard deviation of edge lengths at each vertex. Separating edges are removed only once. This type of static edge removal process is however incapable of properly detecting local variations in the input data, and it fails to give good results in situations where sparse clusters may be adjacent to high-density clusters. The above limitations have an adverse effect on the content representation of the video summary. Furthermore, since only color histogram is used to extract the key frames, the algorithm in [7] often produces redundant frames with similar spatial concepts. In this paper, we propose a Delaunay graph-based clustering algorithm with several improvements over [7]. A minimum spanning tree-based clustering method, called maximum standard deviation reduction (MSDR) can be found in [9]. Our method splits the Delaunay graph using a better edge pruning strategy where overall reduction in the global standard deviation of edge lengths is maximized. We call this GSDR_DC method. Secondly, unlike [7], our method is dynamic in nature (repeated until a threshold value is reached). So, visual dynamics of the frames are captured better and a more informative video summarization is achieved. Finally, the proposed algorithm utilizes edge information along with color histogram to achieve higher semantic dependency between different video frames. So, spatial redundancy between frames is eliminated.

2. The Proposed Algorithm

Delaunay triangulation of a point set is the dual of Voronoi Diagram, used to represent the interrelationship between each data point in multidimensional space to its nearest neighboring points. 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* [8]. For each vertex/point in the Delaunay graph, we calculate the local standard deviation using local mean length to highlight the local effects. To incorporate the global effects, global standard deviation reduction is chosen as the optimization criterion to obtain the disjoint clusters. Some useful definitions are given below [7]:

Definition 1. Local mean length of a point p_i $LML(p_i)$ in the Delaunay graph is defined as:

$$LML(p_i) = \frac{1}{d(p_i)} \sum_{j=1}^{d(p_i)} |e_j| \qquad (1)$$

where $d(p_i)$ denotes the number of edges incident to p_i and $|e_j|$ denotes the length of the j^{th} edge.

Definition 2. The local standard deviation of length of edges incident to p_i is denoted by $LSD(p_i)$ and is defined as:

$$LSD(p_i) = \sqrt{\frac{1}{d(p_i)} \sum_{j=1}^{d(p_i)} (LML(p_i) - |e_j|)^2} \qquad (2)$$

Definition 3. The global standard deviation for *DT* of *N* points is defined as:

$$GSD(DT) = \frac{1}{N} \sum_{i=1}^{N} LSD(p_i) \qquad (3)$$

Our algorithm removes an edge to obtain the clusters such that the overall global standard deviation reduction of the edges in the Delaunay graph is maximized. This edge removal process is repeated until a threshold is reached. Delaunay graph for a given point set is partitioned into *K* disjoint clusters $DT_K = \{C_1, C_2, \ldots, C_K\}$ such that the following objective function is satisfied:

$$DT_K = argmax\big(GSD(DT_0)\big) - GSD\big((DT_K)\big) \qquad (4)$$

$$\left|\Delta GSD(DT_K) - \Delta GSD(DT_K^*)\right| < \left|\alpha\left(\Delta GSD(DT_K) + 1\right)\right| \qquad (5)$$

In equation (4), DT_0 denotes the original Delaunay triangulation, $GSD(DT_0)$ denotes the global standard deviation of DT_0 and $GSD(DT_K)$ represents the global standard deviation after the end of edge removal process. The term $\Delta GSD(DT_K)$ denotes maximum global standard deviation reduction that leads to final clusters whereas the term $\Delta GSD(DT_K^*)$ denotes maximum global standard deviation reduction in the penultimate stage, *i.e.*, $DT_K^* = \{C_1, C_2, \ldots, C_{K-1}\}$. The constant α in equation (2) has a small positive value which determines the termination criterion of this iterative algorithm. Individual clusters $C_1, C_2, \ldots, C_K$ are obtained from the final Delaunay graph DT_K. Various steps of the proposed algorithm are summarized in figure 1.

1. Sampling: Sample the input video sequence to get the selected frames.
2. Feature Extraction: Extract color histogram and edge histogram from each selected frame to form a composite feature vector. For our problem, each frame is represented by a 336 (256 elements for color histogram and 80 elements for edge histogram) dimensional feature vector.
3. Dimensionality Reduction: Use principal component analysis (PCA) to reduce the dimension of the above feature vector. Depending on the variance of the video, 5-7 dimensional feature vectors are obtained.
4. Delaunay Graph Construction: Generate *DT* for the 5-7 dimensional feature vectors. Calculate the overall global standard deviation (GSD) of edge lengths in the corresponding graph. Assign $DT_K = DT_0$, and set $\alpha = 0.0001$.
5. Edge Removal Process: Choose an edge that leads to maximum *GSD* reduction once it is removed from DT_K.
6. Stopping Criteria: Repeat step 5 until: $\left|\Delta GSD(DT_K) - \Delta GSD(DT_K^*)\right| < \left|\alpha\left(\Delta GSD(DT_K) + 1\right)\right|$
7. Final Cluster Generation: Find the remaining connected components from the final DT_K to obtain individual clusters.
8. Key Frame Selection: The frames which are closest to the centroids of each cluster are deemed as the key frames.

Figure 1. GSDR_DC Algorithm

Time-complexity of GSDR_DC(in terms of number of frames *n* and dimension of feature vector *d*) is *O(n log*

n) (construction of DT: *O(n log n)* + Dynamic edge pruning strategy: *O(kn), k << n*, *k* is the number of iteration; total complexity: *O(n log n)*). Note that this complexity is same as that of the DC method [7].

3. Performance Measures

Evaluation of video summaries using key frame extraction techniques remains a challenging task. We choose two well-known objective measures, namely, Fidelity [10] and Shot Reconstruction Degree (SRD) [11] to evaluate the performance of the proposed method. Compression ratio [12] is additionally used to examine the compactness of the video summary. A brief description of these measures is given below.

A. Fidelity: The fidelity measure is based on semi-Hausdorff distance to compare each key frame in the summary with the other frames in the video sequence. Let $V_{seq} = \{F_1, F_2, \ldots, F_N\}$ be the frames of the input video sequence and $KF = \{F_{K1}, F_{K2}, \ldots, F_{KM}\}$ be the extracted key frame set. The distance between the set of key frames and a frame F belonging to V_{seq} can be computed as:

$$DIST(F, KF) = Min\left\{Diff\left(F, F_{K_j}\right)\right\}, j = 1 \; to \; M \quad (6)$$

In equation (6), Diff ( ) is a suitable frame difference measure. For this work, we use HD descriptor, a combination of color histogram intersection and edge histogram-based dissimilarity measure [12]. The distance between the video sequence V_{seq} and set of key frames KF can be defined as:

$$DIST\left(V_{seq}, KF\right) = Max\{DIST(F_i, KF)\}, i = 1 \; to \; N \quad (7)$$

$$FIDELITY\left(V_{seq}, KF\right) = MaxDiff - DIST\left(V_{seq}, KF\right) \quad (8)$$

MaxDiff is the largest possible value that Diff ( ) can assume. High Fidelity provides a good global description of the visual content of the video summary.

B. Shot Reconstruction Degree (SRD): This measure indicates how accurately we can reconstruct the whole video sequence from the extracted set of key frames using a suitable frame interpolation technique. SRD can be defined as:

$$SRD(\text{Vseq}, \text{KF}) = \sum_{i=1}^{N} Sim(F_i, F_i') \quad (9)$$

Sim( ) is the similarity measure between two frames, F_i is the i^{th} frame and F_i' is the i^{th} reconstructed frame obtained using an inertia-based frame interpolation algorithm (IMCI) [13]. HD descriptor-based similarity function is used to calculate SRD. High SRD provides more detailed information about local behavior of key frames.

C. Compression Ratio measure: Compression ratio for a video sequence with *N* frames having a key frame set of *M* frames is defined as:

$$CR\left(V_{seq}\right) = 1 - \left({}^{M}/_{N}\right) \quad (10)$$

High Compression ratio indicates less redundancy.

4. Experimental Results

We have so far experimented with 5 test video segments belonging to different genres and having different durations (30 sec. to 2 min) from the Open Video (OV) projects [14]. 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. Performance comparison with OV storyboard and DC [7] algorithm is summarized in Table 1. Note that the OV storyboard cannot be considered as exact ground-truths because it may contain redundant frames due to temporal order arrangement. Table 1 shows that there is relative improvement in both fidelity and SRD over DC for all the five test video segments. The average relative improvement in fidelity is 3.65% and the average relative improvement in SRD is 5.06%. Maximum improvement of 6.14% in fidelity and 6.97% in SRD are achieved for the video stream ***A New Horizon, Segment 08***. In Figure 2, the key frames obtained from DC and GSDR_DC methods for the above video stream are arranged according to their cluster significance factor. As can be seen from fig. 2, redundancy in the output of the DC method (inclusion of both the fifth and the sixth frame) is removed in the video summary obtained from the proposed GSDR_DC method due to inclusion of edge information. Table 1 also demonstrates that the values of CR are comparable for DC and GSDR_DC methods. So, we can conclude that our method simultaneously captures detailed dynamics, provides a good global description, and, preserves compactness for all the test video segments.

5. Conclusion and Future work

We proposed a novel automatic video summarization technique based on Delaunay graphs with a better edge pruning strategy. Experimental results show that our algorithm outperforms the work described in [7] without incurring any additional computational costs. In future, we will focus on implementation of higher order Delaunay graphs for better clustering. Another direction of future research

is to produce personalized video summaries with unobtrusively sourced user-based information.

References

[1]. A.G. Money and H.W. Agius. Video summarization: A conceptual framework and survey of the state of the art. *J. Visual Commun. Image Represent*., 19 (2): 121–143, 2008.

[2]. B. T. Truong, and S.Venkatesh. Video abstraction: A systematic review and classification. *ACM Trans. Multimedia Comput. Commun.* Appl., 3(1): Article 3,2007.

[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): 56–68, 2011.

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

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

[6]. 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.

[7]. Padmavathi Mundur, Yong Rao, and Yelena Yesha. Keyframe-based video summarization using Delaunay clustering. *International Journal on Digital Libraries,* 6(2): 219 – 232, 2006.

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

[9]. O. Gryorash, Y. Zhou and Z, Jorgenssn. Minimum spanning tree-based clustering algorithms, *Proc. IEEE Conference on Tools with Artificial Intelligence,* 73-81, 2006.

[10]. H. S. Chang, S. Sull and Sang Uk Lee. Efficient Video Indexing Scheme for Content-Based Retrieval. *IEEE Trans. on Circuits and Systems for Video Tech.,* 9(8):1269-1279, 1999.

[11]. Tieyan. Liu, X. Zhang, J. Feng and K.T. Lo. Shot reconstruction degree: a novel criterion for key frame selection. *Pattern Recognition Letters,* 25:1451–1457, 2004.

[12]. G.Ciocca and R.Schettini. A innovative algorithm for key frame extraction in video summarization. *J. of Real-Time Image Processing* 1(1): 69-88, 2006.

[13]. T.Y. Liu, K.T.Lo, X.D. Zhang and J. Feng. Frame interpolation scheme using inertia motion prediction. *Signal Process.: Image Comm.* 18 (3): 221–229, 2003.

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

Table 1. Performance comparison of OV Storyboard, DC[7] and GSDR_DC methods

Video Segment Title(# frames)	OV # Key Frames	DC #Cluster (# Key Frames)	GSDR_DC #Cluster (# Key Frames)	CR DC	CR GSDR_DC	Fidelity DC	Fidelity GSDR_DC	Relative improvement Fidelity (%)	SRD DC	SRD GSDR_DC	Relative improvement SRD (%)
Anatomy of Hurricane, Segment 01(287)	2	2	2	0.9903	0.9903	0.375	0.392	4.533	3.588	3.786	5.518
New Indians Segment 08(707)	5	5	6	0.9929	0.9915	0.626	0.644	2.875	6.649	6.856	3.113
A New Horizon, Segment 08(1815)	7	7	6	0.9961	0.9966	0.733	0.778	6.139	7.277	7.784	6.967
A New Horizon, Segment 06(1944)	5	7	5	0.9963	0.9974	0.851	0.862	1.292	5.591	5..826	4.203
Exotic Terrene , Segment 03(2670)	14	5	5	0.9981	0.9981	0.907	0.938	3.417	6.238	6.584	5.546
Average Relative Improvement in Fidelity = 3.65%								**Average Relative Improvement in SRD = 5.06%**			

Figure 2. Summarization results for the video "A New Horizon, Segment 08": (a) OV Storyboard (top row), (b) DC[7] (middle row), and (c) GSDR_DC (bottom row).