J. Vis. Commun. Image R. 24 (2013) 1212–1227

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

# Video key frame extraction through dynamic Delaunay clustering with a structural constraint

Sanjay K. Kuanar, Rameswar Panda, Ananda S. Chowdhury *

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

ARTICLE INFO

*Article history:*
Received 27 November 2012
Accepted 5 August 2013
Available online 20 August 2013

*Keywords:*
Video summarization
Delaunay graphs
Edge pruning
Deviation ratio
Information-theoretic pre-sampling
Feature extraction
Key frame visualization
Clustering

ABSTRACT

Key frame based video summarization has emerged as an important area of research for the multimedia community. Video key frames enable an user to access any video in a friendly and meaningful way. In this paper, we propose an automated method of video key frame extraction using dynamic Delaunay graph clustering via an iterative edge pruning strategy. A structural constraint in form of a lower limit on the deviation ratio of the graph vertices further improves the video summary. We also employ an information-theoretic pre-sampling where significant valleys in the mutual information profile of the successive frames in a video are used to capture more informative frames. Various video key frame visualization techniques for efficient video browsing and navigation purposes are incorporated. A comprehensive evaluation on 100 videos from the Open Video and YouTube databases using both objective and subjective measures demonstrate the superiority of our key frame extraction method.



## 1. Introduction

With the recent advancement in video capture, storage and distribution technologies, the extent of video content accessible in the daily life has increased exponentially. To handle such huge amount of data, proficient video management systems are being developed to access the video information in a user-friendly way [1,2]. The problem of video summarization deals with succinct representation of a video [3]. Such a representation makes users aware of the content of any video without watching it entirely [4]. Video summarization refers to a class of nonlinear content-based video compression techniques which can efficiently represent most significant information in a video stream using a combination of still images, video segments, graphical representations and textual descriptors [5]. According to Truong and Venkatesh [3], there are two fundamental types of video summaries, namely, Video key frame extraction (static) and Video Skimming (dynamic). Video Storyboard is a set of static key frames (motionless images) which preserve the overall content of a video with minimum data. Video skimming is a set of images with audio and motion information. Video skim, unlike a video storyboard, includes both audio and motion elements that can potentially enhance the expressiveness and information of the summary. In contrast, video storyboard summarizes the video content in a more compact manner and the static key frames can be further organized for browsing and navigation purposes.

Various clustering methods are applied over the years to extract key frames from a video [6–9]. The main aim of these clustering-based techniques is to extract key frames by grouping video frames based on a set of features like color, motion, shape, and texture. After the clustering is complete, usually, one frame per cluster is selected as the key frame to produce the video summary. Performance of such clustering methods depends heavily on the user inputs and/or certain threshold parameters (e.g., number of clusters) [8–10]. In addition, different criteria that are used to measure the similarity between the video frames significantly influence the key frame set [8,11,16]. Furthermore, many of the existing video summarization methods use uniform sampling in the pre-processing stage which may result in exclusion of some informative frames [6–8].

Key frame based video summarization is modeled in [6] as a clustering problem on Delaunay graphs. In this paper, we present a novel and effective approach for video key frame extraction using improved Delaunay clustering. Both color and texture features are used in the clustering process. The main contributions of this paper are: (1) efficient splitting of the Delaunay graph using a dynamic edge pruning strategy where overall reduction in the global standard deviation of edge lengths is maximized and a structural constraint in form of a lower limit on the deviation ratio of the graph vertices is imposed *i.e.*, the constraint on deviation ratio is checked before removal of an edge such that the edges within a cluster are preserved to ascertain better content coverage in the summary; (2) better frame pre-sampling using a combination of fixed sampling and a

* Corresponding author. Tel.: +91 33 2457 2405; fax: +91 33 2414 6217.
*E-mail addresses:* sanjay.kuanar@gmail.com (S.K. Kuanar), rameswar183@gmail.com (R. Panda), aschowdhury@etce.jdvu.ac.in (A.S. Chowdhury).

1047-3203/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jvcir.2013.08.003

sampling based on mutual information between successive frames of the video leading to a more informative input to the actual clustering process; (3) incorporation of user perception in the performance evaluation process using three subjective measures in addition to three objective measures makes the comparisons comprehensive and unbiased. Performance comparison of the proposed method with three different state-of-the-art approaches [6–8], on 50 videos each from the Open Video Project and the YouTube using the above objective and subjective measures clearly indicate its superiority. A preliminary version of this work was published in [5], where, neither any structural constraint for splitting of the Delaunay graph nor any information-theoretic pre-sampling was used. Furthermore, experiments in [5] were restricted to 5 videos and results were compared only with [6] using solely the objective measures.

The rest of this paper is organized as follows. Section 2 discusses the related work and highlights our contribution. Section 3 provides the theoretical foundations of our proposed approach. Section 4 describe our proposed method. Video key frame visualization techniques are presented in Section 5. Section 6 reports experimental results with detailed analysis. Finally, Section 7 concludes the paper with an outline of future research directions.

## 2. Related work

A comprehensive review of video summarization approaches can be found in [3,4]. Only some representative works are discussed here. Hanjalic and Zhang [12] developed a technique for video key frame extraction by finding an optimal clustering through cluster-validity analysis. A partitional clustering is applied several times depending on the number of frames present in a video sequence. Though the above technique produce summaries of acceptable quality, the partitional clustering process makes the summarization computationally expensive. Gong and Liu [10] used Singular Value Decomposition (SVD) for the purpose of video summarization. Initially, a subset of all the available video frames (one from every ten frames) is selected using pre-sampling approach. SVD is applied on a feature-frame matrix formed using global color histogram. One problem with this approach is the clustering process is dependent on proper choice of a threshold. Mundur et al. [6] proposed a Delaunay triangulation-based clustering approach to automatically extract the key frames from a video. After an initial pre-sampling phase, each frame is represented by a 256 dimensional vector in HSV color space. Then, Principal Components Analysis (PCA) is applied to reduce the dimension of the feature vector. A Delaunay graph is constructed with these frames and the edges are classified into short edges and separating edges using average and standard deviation of edge lengths at each vertex. The separating edges are removed to form the distinct clusters. One major problem with this method is that the separating edges are removed only once. This type of static edge removal process is 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 limitation has 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 [6] often produces redundant frames with similar spatial concepts. Furini et al. [7] proposed STIMO (STIll and MOving Video Storyboard), a video summarization technique based on an improved version of the Furthest-Point-First (FPF) algorithm. Once a feature-frame matrix is constructed after pre-sampling and color histogram formation, similar frames are clustered together based on FPF algorithm. For obtaining the number of clusters, pair wise dissimilarity between consecutive frames is computed according to the Generalized Jaccard Distance (GJD). Though this method allows user customization in terms of length of the storyboard and maximum waiting time to get the key frame, implementation of fixed pre-sampling and selection of GJD based dissimilarity measure adversely affect the content representation of the key frame set. Avila et al. [8] presented VSUMM (Video SUMMarization), where key frames are extracted using the k-means algorithm. The estimation of the number of clusters is based on a simple shot boundary detection method, where the number of cluster is incremented for each sufficient content change in the video sequence. This type of estimation, based on shot boundary detection method, is computationally intensive for videos having large number of frames. Moreover, since only color histogram is used for shot boundary detection, this estimation is not accurate for different genres of video.

The proposed approach is designed to address some of the important limitations of the above-mentioned techniques. We aim at obtaining superior video summaries using improved Delaunay clustering and information-theoretic pre-sampling. The main advantage of Delaunay clustering, as indicated by Mundur et al. [6] lies in automatic extraction of key frames. Delaunay clustering has been improved in this paper through a dynamic edge pruning strategy where the overall reduction in the global standard deviation of edge lengths is maximized with imposition of a structural constraint in form of a lower limit on the deviation ratio of the graph vertices. This constraint on deviation ratio of graph vertices is checked before removal of the corresponding edge such that the edges within a cluster are preserved. We consider both color and texture features for the purpose of video summarization. Information-theoretic pre-sampling is applied during the pre-processing stage so that frames corresponding to the significant valleys in the mutual information profile between successive frames of any video are chosen. Moreover, we present various key frame visualization techniques that arrange the key frames in an organized manner to facilitate the user in efficient video browsing and navigation. Finally, a comprehensive performance evaluation and comparisons with three well-known existing summarization methods [6–8] are carried out over a collection of 100 videos with different genres as well as durations (downloaded from Open Video project and YouTube) using three subjective measures (Clarity, Conciseness, Overall quality) and three objective measures (Fidelity, Shot Reconstruction Degree, Compression Ratio).

## 3. Theoretical foundations

Our clustering strategy is based on efficient pruning of edges in a Delaunay graph. Some useful definitions pertaining to this method are provided in this section.

**Definition 1.** *Delaunay triangulation (DT)* of a point set is the straight line dual of famous Voronoi diagram, used to represent the inter-relationship between each data point in multi-dimensional space to its nearest neighboring points.

**Definition 2.** Under the standard assumption that no four points of $P$ are co circular, the Delaunay triangulation is indeed a triangulation [13] 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 $b$ [14]. The closed disc bounded by the circle contains no sites of $P$ other than $a$ and $b$. Fig. 1 graphically presents the relation between Voronoi diagram and its dual Delaunay triangulation.

**Definition 3.** *Mean length of edges* incident to each point $p_i$ is denoted by $\mathrm{LML}(p_i)$ and is defined as

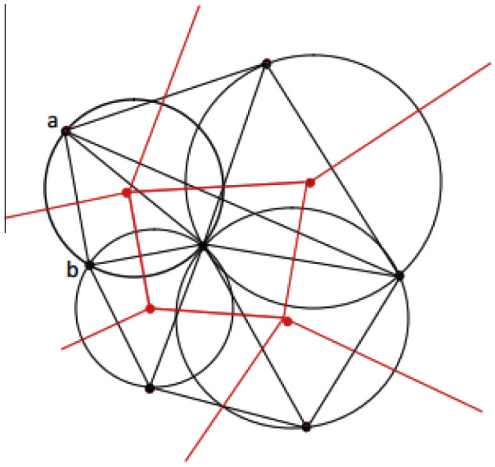



**Fig. 1.** Delaunay triangulation (in black) and Voronoi diagram (in red). ab represents a Delaunay path. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$$\mathrm{LML}(P_i) = \frac{1}{d(P_i)} \sum_{j=1}^{d(P_i)} |e_j| \tag{1}$$

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 4.** *Local standard deviation of length* of edges incident to $p_i$ is denoted by $\mathrm{LSD}(p_i)$ and is defined as:

$$\mathrm{LSD}(P_i) = \sqrt{\frac{1}{d(P_i)} \sum_{j=1}^{d(P_i)} \left(\mathrm{LML}(P_i) - |e_j|\right)^2} \tag{2}$$

**Definition 5.** *The global standard deviation* for DT of $N$ points is defined as:

$$\mathrm{GSD}(\mathrm{DT}) = \frac{1}{N} \sum_{i=1}^{N} \mathrm{LSD}(P_i) \tag{3}$$

**Definition 6.** *Deviation ratio* for each point $p_i$ in Delaunay graph is denoted by $\mathrm{DR}(p_i)$ and is defined as:

$$\mathrm{DR}(P_i) = \frac{\mathrm{LSD}(P_i)}{\mathrm{GSD}(\mathrm{DT})} \tag{4}$$

## 4. Proposed method

In Fig. 2, we illustrate the main steps of our proposed video key frame extraction method. The proposed method consists of four main steps: (1) video frames pre-sampling; (2) feature extraction; (3) Delaunay clustering; (4) key frame extraction.

### 4.1. Video frames pre-sampling

The first step towards key frame extraction is to split the video stream into a set of meaningful and manageable basic units by the process of temporal video segmentation. Most of these approaches [16,17] depend on shot detection, which become inaccurate due to the presence of different types of transitions (e.g., fade in, fade out, abrupt cut) between successive video frames. Another well-known approach of video segmentation is to divide the video stream into frames (still images). Several authors have used this approach [6–8] and it is also used by our proposed method.

In a pre-sampling approach, only a subset of frames, which potentially represents the overall content of the whole video stream, is usually considered. Sampling rate becomes a very important parameter which can directly influences the content coverage of the final key frame set. Very low sampling rate leads to poor quality of video summary and at the same time increases the time required to obtain the summary. In contrast, very high sampling rate could miss important information contained in the video. Hence, judicious selection of sampling rate is an important design parameter in the process of video summarization. Videos having long shots have an advantage with the fixed pre-sampling approach as more number of frames is selected for further processing. However, for the shots with short duration, there is a possibility that no frame gets selected. To handle this type of problem, we use a combination of fixed pre-sampling and information-theoretic pre-sampling based on mutual information. The fixed sampling rate is one frame per second (same as that in VSUMM). We additionally employ information-theoretic pre-sampling where frames corresponding to the significant valleys in the mutual information profile between successive frames of the entire video segment are chosen. Mutual information between two frames indicates the extent of similarity between those frames. In our method, estimation of mutual information is based on joint entropy calculation between successive frames [18]. Let $\mathrm{MI}(F_t, F_{t-1})$ represents the mutual information between one frame $F_t$ at time instant $t$ and another frame $F_{t-1}$ at time instant $t-1$. A significant valley is given by the following criterion:

$$\frac{\mathrm{MI}(F_t, F_{t-1})}{\mathrm{MI}(F_{t-1}, F_{t-2})} + \frac{\mathrm{MI}(F_t, F_{t-1})}{\mathrm{MI}(F_{t+1}, F_t)} \leqslant 2(1-\varepsilon) \tag{5}$$

In Eq. (5), $\varepsilon$ is a threshold for significant valley detection. A similar approach for significant peak detection can be found in [11]. For video shots having longer duration, more frames are selected even with fixed sampling rate. However, in case of videos having shorter duration shots, like cartoon videos, loss of information is inevitable with fixed pre-sampling rate. So, information-theoretic pre-sampling can select frames for these types of videos which could be missed during sampling with a fixed rate. Fig. 3 demonstrates the process of significant valley detection in mutual information change between successive frames for a cartoon video downloaded from YouTube. Symbols a, b, c, d indicate valleys.

Note that frames corresponding to those valleys are missed due to fixed pre-sampling rate of one frame per second. (i.e., frames having numbers as multiples of 30 are selected for videos with frame rate of 30 fps approximately).

### 4.2. Feature extraction

Feature extraction is an important step to efficiently represent the video frames in multi-dimensional space. We use both color and texture feature to represent the content of video frames in our proposed algorithm.

#### 4.2.1. Color feature extraction

Color is the most expressive low level feature. We represent each video frame by a 256-dimensional feature vector, obtained from a color histogram. This is a computationally efficient technique and is also robust to small changes of the camera position [11]. One key issue of such a histogram-based approach is the selection of an appropriate color space. In our case, it is important that the color model reflects the human perception of colors. So, we decide to obtain the color histogram using the HSV color space, which is also found to be more resilient to noise [11,19]. The HSV

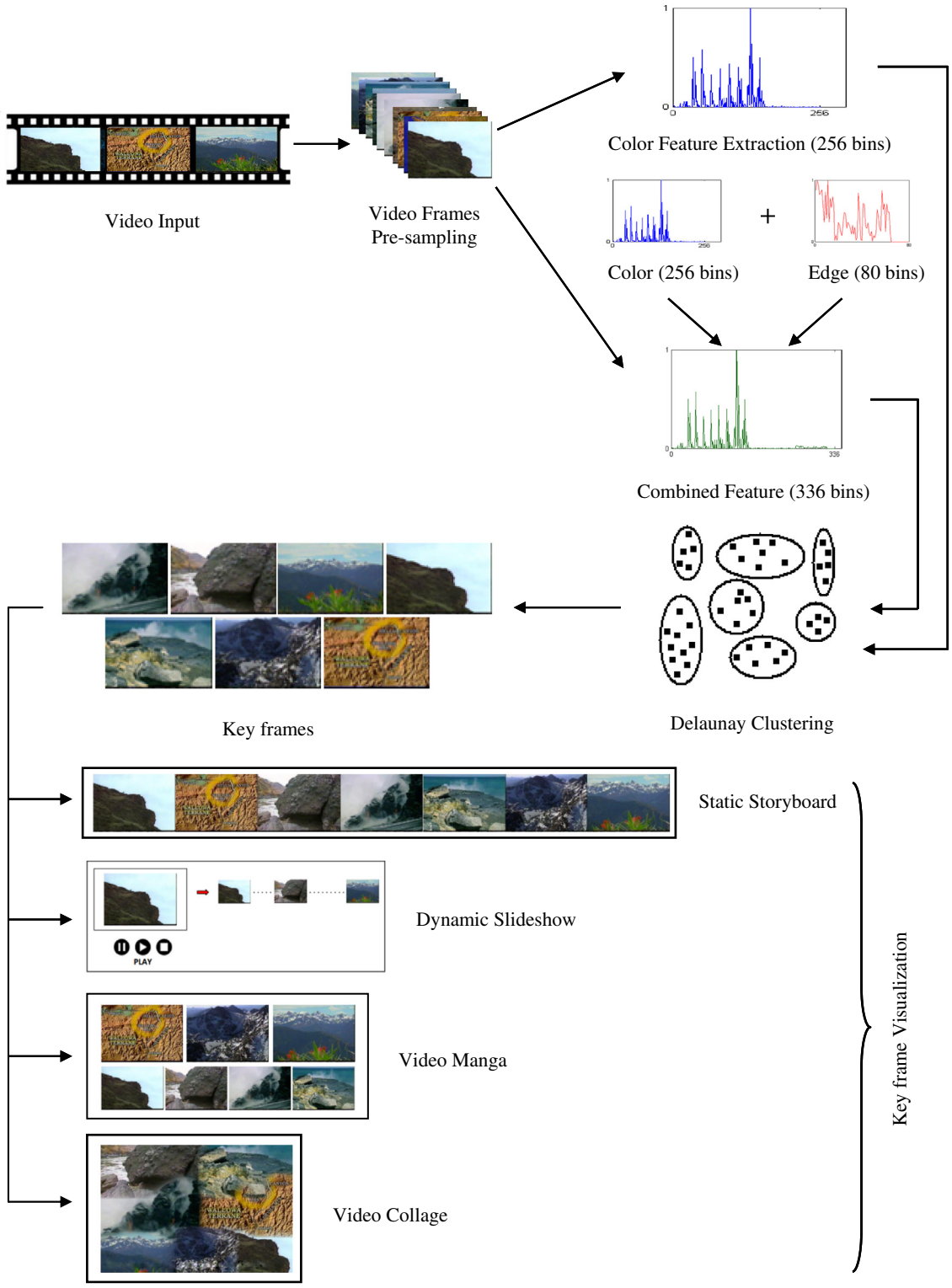



**Fig. 2.** Flowchart of the proposed method.

color space is divided into 256 color subspaces, using 16 ranges of $H$, 4 ranges of $S$, and 4 ranges of $V$ according to the MPEG-7 generic color histogram descriptor.

*4.2.2. Texture feature extraction*

In addition to color, texture feature is also extracted from the video frames using edge histogram descriptor [20]. A video frame is first sub-divided into $4 \times 4$ blocks, and then local edge histograms for each of these blocks are computed. Edges are broadly grouped into five categories: vertical, horizontal, 45° diagonal, 135° diagonal, and isotropic. Thus, each local histogram has five bins corresponding to the above five categories. Finally each frame is represented by a 80-dimensional feature vector corresponding to texture feature.

As global color histogram alone is incapable of preserving spatial information present in the video frames, our method utilizes

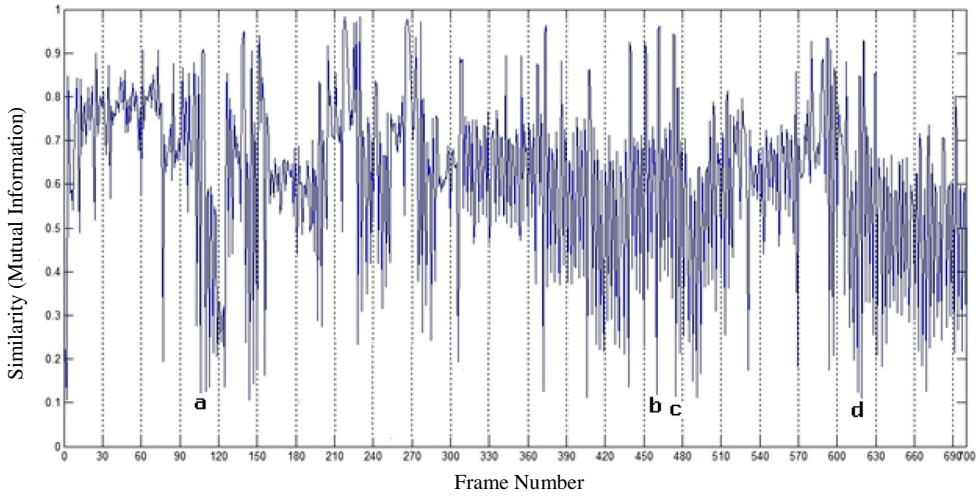



**Fig. 3.** Significant valley detection in mutual information change between successive frames. a, b, c, d indicate valleys.

texture feature along with color histogram to achieve higher semantic dependency between different video frames. So, spatial redundancy between frames is eliminated. After combining color and texture features using serial feature fusion strategy [30], each frame is represented by a 336-dimensional feature vector. Apart from serial fusion, various methods like parallel fusion [30], Canonical correlation analysis based fusion [31], KL transform based fusion [32], Multi-modality learning based fusion [33] are developed for efficient feature fusion in recognition tasks. However, for very long datasets, like a video, feature fusion for each frame is computationally prohibitive. On the other hand, for small sample size problems, these complex fusion strategies provide superior result compared to the serial feature fusion [30]. We next stack such combined feature vectors for each frame into the frame-feature matrix.

#### 4.2.3. Elimination of meaningless frames

A meaningless frame is a monochromatic frame which may be present due to different transitions (e.g., fade in, fade out) between successive frames. There exist some situations where these monochromatic frames are selected due to pre-sampling. Hence, these frames need to be discarded before clustering. That is why we compute the normalized variance for both color and edge histogram of sampled frames. Fig. 4 illustrates the behavior of those histograms for different types of video frames. Notice that, monochromatic frames have a high variance between histogram bins as they follow homogenous distribution [15]. Thus, we discard a selected frame if one of its histograms has a normalized variance greater than a predetermined threshold of 0.5.

### 4.3. Clustering on Delaunay graphs

Since the feature extraction process tends to generate a sparse matrix, we apply Principal Component Analysis (PCA) [21] to reduce the dimensions of the matrix without affecting the overall video content. After applying PCA, each frame in the $m$-dimensional ($m = 336$ in our case) feature space is projected on a $d$-dimensional refined feature space where $d$ is the number of the selected Principal Components (PCs). We choose $d$ depending on the variance of the video [6] (see Section 6.7).

We then construct the Delaunay graph using the data points in the refined feature space as its vertices. Each edge in the Delaunay graph represents spatial proximity between the corresponding vertices (or frames). In the Delaunay graph, the edges can be grouped into intra-cluster edges (edges whose end points are in the same cluster) and inter-cluster edges (edges whose end points are in different clusters). Note that the vertices lying on the boundary of any cluster exhibit greater variation in the lengths of edges incident on them. This is because some of the edges incident on such vertices are inter-cluster edges while the rest can be intra cluster edges. So deviation ratio for these vertices is $\geqslant 1$ whereas for the vertices which lie inside the clusters, deviation ratio is <1 (see Definition 6).

Our objective is to preserve intra-cluster edges and remove inter-cluster edges which connect the individual clusters in an efficient manner. In our method, the problem of edge pruning in the Delaunay graph is posed as a constraint optimization problem. We remove an edge $e$ such that the overall global standard deviation reduction of the edges in the Delaunay graph is maximized provided the edge joining the vertices in the Delaunay graph have deviation ratio $\geqslant 1$. At each step, after selecting the edge according to the maximum reduction in the global standard deviation criteria, the constraint on deviation ratio is checked to ensure that the edges within a cluster are preserved. This edge removal process is repeated until a threshold is reached. Delaunay graph for a given point set is partitioned into $K$ disjoint clusters $\mathrm{DT}_K = \{C_1, C_2, \ldots, C_K\}$ such that the following objective function is satisfied:

$$\mathrm{DT}_K = \mathrm{argmax}(\mathrm{GSD}(\mathrm{DT}_0)) - \mathrm{GSD}((\mathrm{DT}_K)) \tag{6}$$

$$|\Delta\mathrm{GSD}(\mathrm{DT}_K) - \Delta\mathrm{GSD}(\mathrm{DT}_K^*)| < |\alpha(\Delta\mathrm{GSD}(DT_K) + 1)| \tag{7}$$

$$DR\{Vertices(e)\} \geqslant 1 \tag{8}$$

In Eq. (6), $\mathrm{DT}_0$ denotes the original Delaunay triangulation, $\mathrm{GSD}(\mathrm{DT}_0)$ denotes the global standard deviation of $\mathrm{DT}_0$ and $\mathrm{GSD}(\mathrm{DT}_K)$ represents the global standard deviation after the end of edge removal process. The term $\Delta\mathrm{GSD}(\mathrm{DT}_K)$ denotes maximum global standard deviation reduction that leads to final clusters whereas the term $\Delta\mathrm{GSD}(\mathrm{DT}_K^*)$ denotes maximum global standard deviation reduction in the penultimate stage, i.e., $\mathrm{DT}_K^* = \{C_1, C_2, \ldots, C_{K-1}\}$. The constant $\alpha$ in Eq. (7) has a small positive value which determines the termination criterion of this iterative algorithm. Eq. (8) represents the constraint on deviation ratio of vertices containing an edge selected for removal. Remaining connected components of the final Delaunay graph $\mathrm{DT}_K$ represent individual clusters.

We now provide a justification of using deviation ratio as a structural constraint. The edge ab in the Delaunay graph of Fig. 5(a) is longer than the edge cd. So removal of the edge ab will

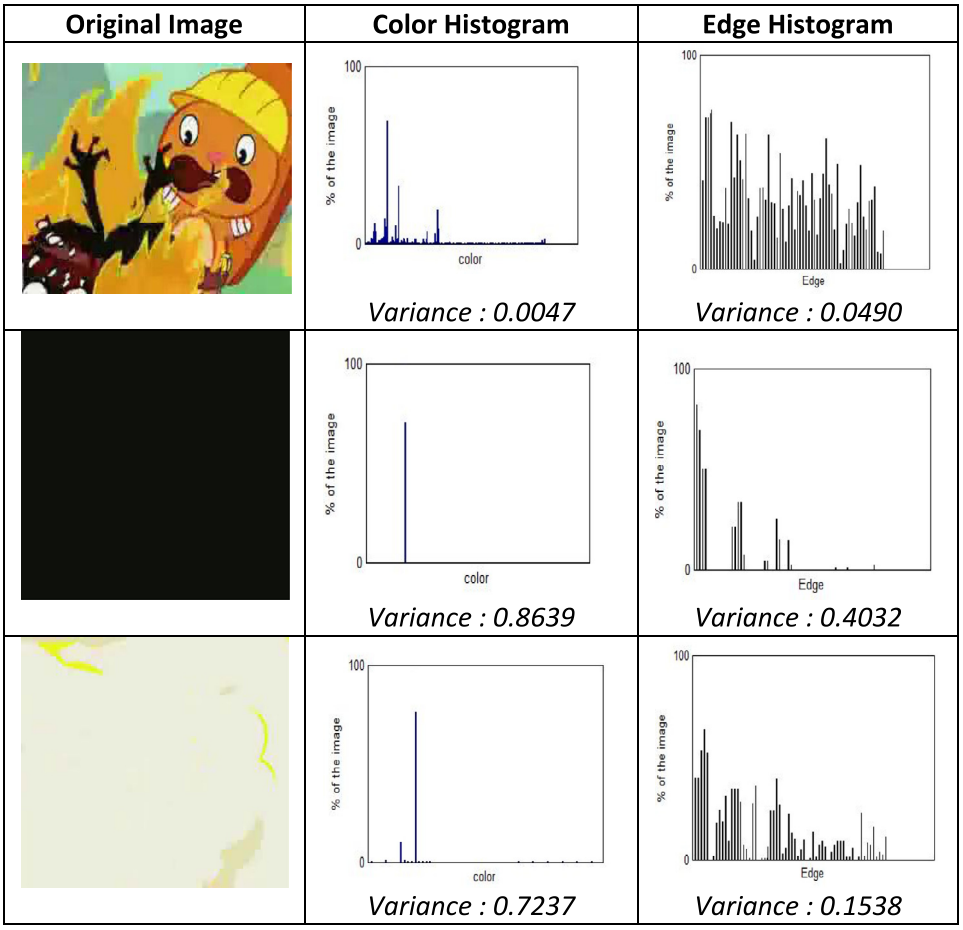



**Fig. 4.** Characteristics of color histograms (second column) and edge histograms (third column) for different frames (first column): normal frames (first row), fade-in frames (second row), and transition frames (third row).

lead to maximum global standard deviation reduction as compared to removal of the edge cd. Without imposition of the constraint on deviation ratio, the edge ab will be deleted which is actually an intra-cluster edge (as shown in Fig. 5(b)). In contrast, incorporation of the deviation ratio constraint will ensure removal of the edge cd and not the edge ab (as shown in Fig. 5(c)). So, we can conclude that incorporation of a constraint on deviation ratio of frames removes inter-cluster edges more effectively as compared to the case where only global standard deviation reduction is minimized. The proposed method, as a result, leads to more natural clusters of the video frames (see Fig. 5).

We also check the imposition of this structural constraint (DR) from purely a clustering perspective. This is analyzed using a graph-clustering fitness measure. As shown in Section 6.6, incorporation of this structural constraint yields a superior clustering performance.

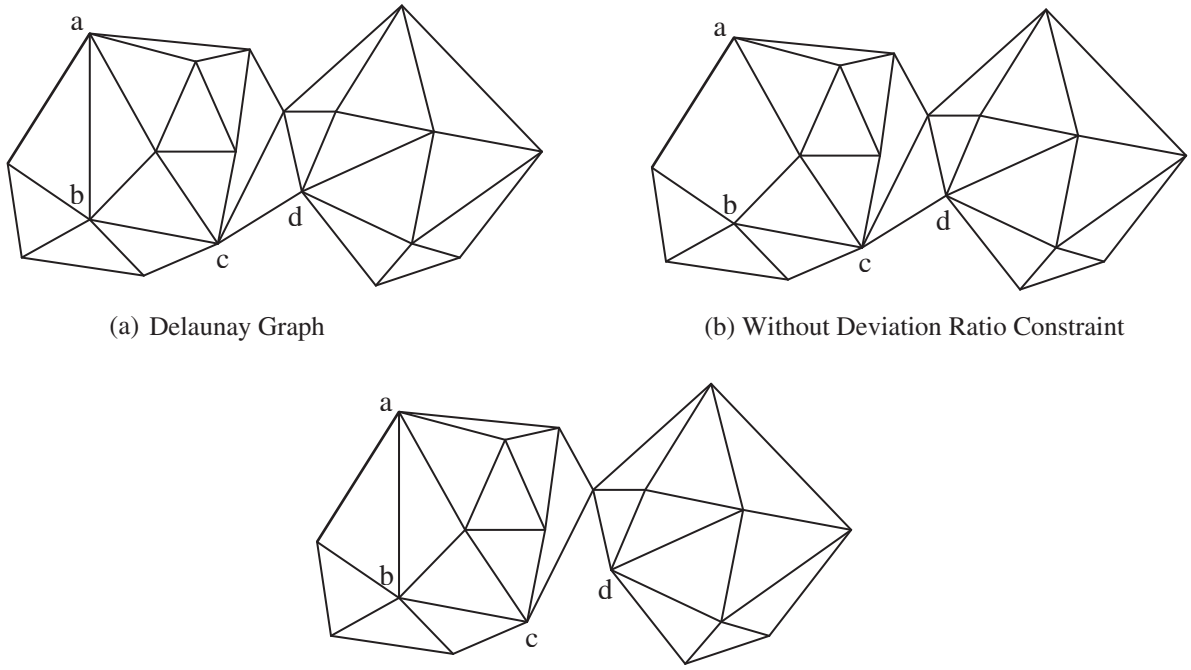
### 4.4. Key frame extraction

Extraction of connected components from the Delaunay graph is performed using Dulmage–Mendelsohn decomposition of the adjacency matrix after the dynamic edge pruning process is complete [22]. This decomposition finds a maximum-size matching in the bipartite graph of the matrix and the diagonal blocks of the adjacency matrix represent the connected components of the Delaunay graph. The frames which are closest to the centroids of each cluster are deemed as the key frames. Finally, the key frames are arranged in an organized manner to make the video summary more understandable.

## 5. Video key frame visualizations

Once the key frames are extracted, they need to be presented in an organized manner for facilitating the user in efficient video browsing and navigation purposes. Video visualization methods aim to present the key frames in some meaningful way which allows the user to grasp the content of a video without watching it entirely [3]. The two most common approaches for key frame visualization are static storyboard display and dynamic slideshow. The former arranges the extracted key frames in lines with maintaining temporal order while the later deals with sequential display of key frames in which the user has no control over the viewing rate. Although screen space is an issue with static storyboard display but it is still the preferred method over the dynamic slideshow [34]. Apart from these two basic forms of key frame visualization methods, there exist another group of methods which present the video summary using a single image. Video poster [35], Video Manga [36], Stained glass [37], Video mosaic [38], VideoSpaceIcon [38], Blocked recursive image composition [39], Video collage [40,41] are the most popular form of key frame visualization methods using a single image.

We have presented four different key frame visualization methods such as static storyboard, dynamic slideshow, video Manga [36] and video collage [40,41] using the extracted key frames. Fig. 6 presents the different key frame visualizations for the video Exotic Terrane, segment 03. Video Manga and video collage are generated using the methods described in [36,41] respectively. We have considered only the duration of clusters as the dominance/importance score in generating video Manga.




(a) Delaunay Graph

(b) Without Deviation Ratio Constraint

(c) With Deviation Ratio Constraint

**Fig. 5.** Edge pruning strategy under different circumstances.

## 6. Experimental results

In this section, the proposed video summarization method is analyzed and the results are compared with three well known approaches [6–8] presented in the literature. In addition, some information about performance measures and evaluation datasets are also provided.

### 6.1. Performance measures

Unlike other research areas, a consistent evaluation framework for video analysis and summarization is somewhat lacking, possibly due to the absence of well-defined objective ground truth. In order to do a comprehensive evaluation of the proposed method, we use three objective and three subjective measures. The objective measures used are Fidelity [24], Shot Reconstruction Degree (SRD) [25] and Compression Ratio (CR) [26]. These measures are preferred because they employ two different approaches. Fidelity provides a global description of the visual content of the video summary, while the Shot Reconstruction Degree uses a local evaluation of the key frames. Compression Ratio is additionally used to examine the compactness of the video summary [26]. However, [27] points to the limitation of using only objective measures for video summarization. As video summarization is a subjective task to a large extent, subjective evaluation becomes necessary in addition to the objective evaluation. In this paper, subjective evaluation using clarity, conciseness, and overall quality [43] is also carried out to judge the perception of users towards the video summaries. All the above measures are now discussed below.

*A. Fidelity*: The fidelity measure is based on the 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:

$$\mathrm{DIST}(F, KF) = \mathrm{Min}\{\mathrm{Diff}(F, F_{K_j})\}, \quad j = 1 \text{ to } M \tag{9}$$

In Eq. (9), Diff () is a suitable frame difference measure. For this work, we use color histogram intersection-based dissimilarity measure [28]. The distance between the video sequence $V_{seq}$ and set of key frames $KF$ can be defined as:

$$\mathrm{DIST}(V_{seq}, KF) = \mathrm{Max}\{\mathrm{DIST}(F_i, KF)\}, \quad i = 1 \text{ to } N \tag{10}$$

$$\mathrm{FIDELITY}(V_{seq}, KF) = \mathrm{MaxDiff} - \mathrm{DIST}(V_{seq}, KF) \tag{11}$$

MaxDiff in Eq. (11) 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:

$$\mathrm{SRD}(V_{seq}, KF) = \sum_{i=1}^{N} \mathrm{Sim}(F_i, F_i') \tag{12}$$

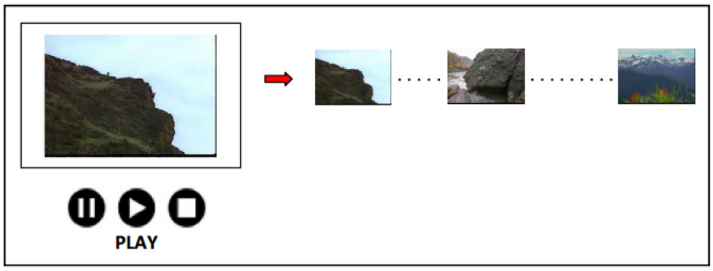
Sim () is the similarity measure between two frames, $F_i$ is the $i$th frame and, $F_i'$ is the $i$th reconstructed frame obtained using suitable frame interpolation technique. We have considered an inertia-based frame interpolation algorithm (IMCI) [29] and color histogram intersection-based similarity function to calculate SRD. High SRD provides more detailed information about local behavior of key frames.

*C. Compression Ratio (CR):* Compression Ratio for a video sequence with $N$ frames having a key frame set of cardinality $M$ is defined as:

$$\mathrm{CR}(V_{seq}) = 1 - (M/N) \tag{13}$$

High Compression Ratio is desirable for a good quality video summary.

(a) Static Storyboard




(b) Dynamic Slideshow


(c) Video Manga


(d) Video Collage

**Fig. 6.** Key frame visualizations for the video Exotic Terrane, segment 03.

*D. Clarity:* Frames within the summary should be clearly visible. In other words, the video summary should not contain transition frames that are not clearly discernible to the users.

*E. Conciseness:* Any frame selected for the video summary should contain only necessary information. Thus, the video summary should be as short as possible provided that it captures all the essential information of a video stream.

*F. Overall quality:* Overall quality of a video summary is evaluated by taking into consideration the factors like coverage, coherence and amiability.

We evaluated 50 videos each from the Open Video Project [44] and the YouTube [45]. All the experiments were performed on a machine with Intel(R) core(TM) i5-2400 processor and 8 GB of DDR2-memory.

### 6.2. Performance analysis with Information theoretic pre-sampling

We first evaluate the effect of information theoretic pre-sampling over fixed sampling for video key frame extraction. Fig. 7 presents the video summaries produced by OURS(C + E) method and VSUMM. From the figure, it can be seen that the second and sixth frames present in the output of our proposed technique is due to the information-theoretic pre-sampling. These two frames are not visually similar to the other frames present in the video summary. So, the presence of these two frames increases the content coverage of the generated summary to a great extent. In fixed sampling rate of one frame per second, this frame would not have been selected for processing. So, the combination of fixed sampling and information-theoretic sampling is shown to be more useful. It can be noticed that though our technique produces much shorter summaries as compared to VSUMM, the quality of summary obtained using the proposed method outperforms VSUMM in terms of other objective and subjective measures.

### 6.3. Performance analysis with deviation ratio constraint

Fig. 8 presents the video summaries produced by our proposed method with and without presence of the deviation ratio constraint. It may be noted that the appropriate selection of key frames plays a major role in maximizing the content coverage or entropy information of a video summary. From Fig. 8(a), it can be seen that both sixth and seventh frames are missing due to improper edge removal process in absence of deviation ratio constraint whereas addition of this constraint makes the clustering process more significant which in turn helps to increase the content coverage of the produced video summary. Presence of these frames makes the video summary more meaningful because it increases the overall content coverage (maximizes the entropy information).

### 6.4. Performance comparison with state-of-the-art methods

In this section we make a comparative performance analysis to evaluate the results of the proposed method for both Open Video and YouTube database.

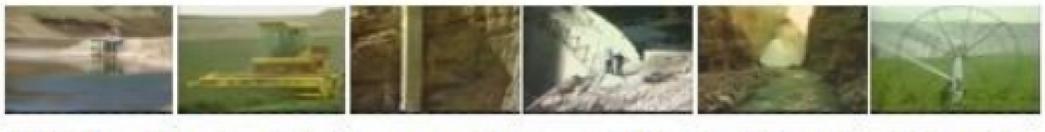
#### 6.4.1. Results for the Open Video database

First, we discuss the results on videos downloaded from the Open Video Project [44]. We evaluate our approach on 50 test video segments belonging to different genres (e.g., documentary, educational, and lecture) and having different durations (30 s to 4 min). Each test video is in MPEG-1 format with a frame rate of 29.97 and the frames having dimensions of 352 × 240 pixels. Long videos are avoided due to limitation of annotation by a subject. For comparison, we used the summarization results on same videos, as reported by three other techniques, namely, DT [6], STIMO [7], and VSUMM [8]. All 50 videos along with the summaries produced by the above techniques are available at <http://www.sites.google.-com/site/vsummsite/>. We apply the clustering method on two different sets of features, denoted as OURS(C) and OURS(C + E). In OURS(C), only color feature is used whereas in OURS(C + E), both color and edge (texture) features are used. The reason behind separately taking only color feature is that it makes comparisons more unbiased as the other summarization techniques use only a color feature. So, we can separately show the impact of our improved clustering strategy as well betterment due to use of both color and texture features. For objective evaluation, Fidelity, Shot Recon-

(a) VSUMM [8]: Fidelity = 0.697, Shot Reconstruction Degree = 4.373, Compression ratio = 0.986, Clarity = 3.76, Conciseness = 3.76, Overall Quality = 3.82.

(b) OURS(C+E) : Fidelity = 0.778, Shot Reconstruction Degree = 4.536, Compression ratio = 0.991, Clarity = 4.18, Conciseness = 3.52, Overall Quality = 4.04.

**Fig. 7.** Summary produced by different approaches for the video from YouTube database (Cartoon Category).



(a) Without Deviation Ratio Constraint: Fidelity = 0.721, Shot Reconstruction Degree = 6.758, Compression ratio = 0.996, Clarity = 4.10, Conciseness = 3.18, Overall Quality = 2.87.

(b) With Deviation Ratio Constraint: Fidelity = 0.862, Shot Reconstruction Degree = 7.682, Compression ratio = 0.996, Clarity = 4.14, Conciseness = 4.02, Overall Quality = 3.96.

**Fig. 8.** Summary produced by under different instances of deviation ratio constraint for the video "A New Horizon, segment 08". Top row → GSDR_DC [5] without deviation ratio constraint. Bottom row → Proposed method with deviation ratio constraint.

struction Degree and Compression Ratio are used. For subjective evaluation, users are asked to rate all the summarized results on a scale of 1–5 (1 corresponds to worst and 5 corresponds to best) in Clarity, Conciseness, and Overall Quality categories. Altogether 25 subjects are involved and each user rated 10 videos. So, summary of each video is evaluated by five different subjects. A sample sheet of user survey and our results for all the 50 videos are available at: <https://sites.google.com/site/ivprgroup/home/research/video-story-board-design>. The parameters used to obtain the video summaries using our method are $\varepsilon = 0.75$ and $\alpha = 0.00010$ (see Section 5.4). Table 1 presents the average value for both objective and subjective measures achieved by different approaches for several video categories. The results indicate that both OURS(C) and OURS(C + E) perform better than all the competing methods. For DT approach, the average Compression Ratio measure is more as it produces much smaller summaries at a cost of poor quality of key frames. From Table 1, it can be concluded that the summary produced using combined feature space has more user satisfaction as compared to using only color feature. The proposed OURS(C + E) strategy eliminates redundant frames with similar spatial concepts.

To judge the relative performance of OURS(C + E) with respect to the other four algorithms ([6–8], OURS(C)), the following relative improvement ($\Delta Q$) measure is employed [26]:

$$\Delta Q(X) = \frac{(\text{Measure\_Alg}(\text{OURS}(C+E)) - \text{Measure\_Alg}(X))}{\text{Measure\_Alg}(X)} \quad (14)$$

where Measure_Alg corresponds to the average values for different measures (both objective and subjective metrics), and X{DT, STIMO, VSUMM and OURS(C)}. Table 2 shows the average relative improvement for different measures achieved by OURS(C + E) approach on the 50 videos from OV database.

It can be seen that the relative improvement on the subjective measures are more as compared to the objective measures which indicates that the video summary produced using OURS(C + E) method has more user satisfaction as compared to others. Note that the average Shot Reconstruction Degree measure for OURS(C) is more as compared to OURS(C + E). This happens because for Documentary and Lecture videos, key frames selected using only color feature are more accurate to reconstruct the whole video sequence as compared to frames selected using combined feature space. Fig. 9 presents the video summaries produced by different ap-

**Table 1**
Average value for different measures for OV database.

| Measures | Category | #Videos | DT | STIMO | VSUMM | OURS(C) | OURS(C + E) |
|---|---|---|---|---|---|---|---|
| Fidelity | Documentary | 44 | 0.504 | 0.522 | 0.562 | 0.567 | 0.586 |
| | Educational | 2 | 0.478 | 0.468 | 0.504 | 0.554 | 0.555 |
| | Lecture | 4 | 0.586 | 0.605 | 0.614 | 0.621 | 0.636 |
| | Weighted average | 50 | 0.509 | 0.527 | 0.564 | 0.571 | 0.584 |
| Shot Reconstruction Degree | Documentary | 44 | 3.671 | 3.754 | 3.944 | 4.079 | 4.080 |
| | Educational | 2 | 3.203 | 2.989 | 2.889 | 3.265 | 3.236 |
| | Lecture | 4 | 4.205 | 4.212 | 4.155 | 4.262 | 4.245 |
| | Weighted average | 50 | 3.695 | 3.760 | 3.919 | 4.061 | 4.059 |
| Compression Ratio | Documentary | 44 | 0.997 | 0.996 | 0.997 | 0.997 | 0.997 |
| | Educational | 2 | 0.996 | 0.996 | 0.996 | 0.996 | 0.996 |
| | Lecture | 4 | 0.997 | 0.996 | 0.997 | 0.997 | 0.997 |
| | Weighted average | 50 | 0.997 | 0.996 | 0.997 | 0.997 | 0.997 |
| Clarity | Documentary | 44 | 3.269 | 3.394 | 3.634 | 3.900 | 3.999 |
| | Educational | 2 | 3.430 | 3.480 | 3.680 | 3.830 | 4.200 |
| | Lecture | 4 | 3.305 | 3.255 | 3.410 | 3.535 | 3.875 |
| | Weighted average | 50 | 3.278 | 3.384 | 3.618 | 3.868 | 3.990 |
| Conciseness | Documentary | 44 | 3.439 | 3.555 | 3.774 | 3.866 | 3.969 |
| | Educational | 2 | 3.780 | 4.000 | 4.000 | 4.000 | 4.210 |
| | Lecture | 4 | 3.350 | 3.540 | 3.740 | 3.800 | 3.950 |
| | Weighted average | 50 | 3.452 | 3.572 | 3.780 | 3.866 | 3.977 |
| Overall quality | Documentary | 44 | 3.430 | 3.495 | 3.836 | 4.005 | 4.135 |
| | Educational | 2 | 3.800 | 4.000 | 4.200 | 4.250 | 4.410 |
| | Lecture | 4 | 3.885 | 3.710 | 4.230 | 4.190 | 4.355 |
| | Weighted average | 50 | 3.481 | 3.532 | 3.882 | 4.029 | 4.163 |

**Table 2**
Relative improvements of OURS(C + E) over DT, STIMO, VSUMM and OURS(C).

| OV project videos | Fidelity | Shot Reconstruction Degree | Compression Ratio | Clarity | Conciseness | Overall quality |
|---|---|---|---|---|---|---|
| DT | 15.64 | 9.92 | 0 | 21.92 | 15.44 | 19.69 |
| STIMO | 11.94 | 8.03 | 0.96 | 18.03 | 11.38 | 17.91 |
| VSUMM | 4.44 | 3.68 | 0 | 10.49 | 5.20 | 7.29 |
| OURS(C) | 3.14 | −0.04 | 0 | 4.83 | 2.87 | 3.50 |

proaches for the video Exotic Terrane, segment 03. The figure clearly shows some redundancy in the output of OURS(C) method (inclusion of both the fourth and the fifth frame) being removed in the video summary obtained from the OURS(C + E) method.

Presence of redundant frames in the video summary decreases the overall quality of the summary. The highest summary quality in terms of both objective and subjective measures is achieved by our OURS(C + E), which can also be confirmed by a visual comparison with the video summaries, obtained from other methods.

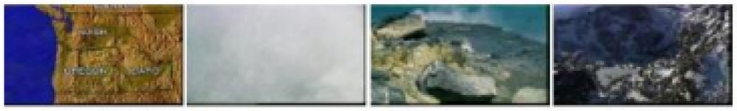
*6.4.2. Results for the YouTube database*

We also evaluate our proposed techniques over 50 videos collected from YouTube website [45]. These videos also belong to different genres (e.g., sports, news, TV-shows, commercials, and home videos) and their durations vary from 1 to 10 min. Since the results of DT and STIMO on YouTube database are not available, we have compared our results with only VSUMM for the videos downloaded from the YouTube. All the videos along with the video summaries produced by VSUMM can be seen at <http://www.sites.google.com/site/vsummsite/>. Since, we already demonstrated that on videos from OV database, OURS(C + E) approach yielded better results as compared to that of OURS(C), only OURS(C + E) is applied on this new set of videos.

Once again, the same 25 subjects were invited to manually rate the summaries for the videos and each video summary has received five different user evaluations. The parameters used to obtain the video summaries using our method are $\varepsilon = 0.85$ and $\alpha = 0.00015$ (see Section 6.7). All the summarization results are available at: <https://sites.google.com/site/ivprgroup/home/research/video-story-board-design>. Table. 3 presents the compara-

(a) DT [6]: Fidelity = 0.607, Shot Reconstruction Degree = 6.232, Compression ratio = 0.998, Clarity = 3.18, Conciseness = 3.72, Overall Quality = 3.52.

(b) STIMO [7]: Fidelity = 0.612, Shot Reconstruction Degree = 6.311, Compression ratio = 0.997, Clarity = 3.24, Conciseness = 3.72, Overall Quality = 3.32.



(c) VSUMM [8]: Fidelity = 0.601, Shot Reconstruction Degree = 6.198, Compression ratio = 0.998, Clarity = 3.54, Conciseness = 3.88, Overall Quality = 3.82.

(d) OURS(C) : Fidelity = 0.642, Shot Reconstruction Degree = 6.639, Compression ratio = 0.997, Clarity = 3.96, Conciseness = 4.12, Overall Quality = 4.22.

(e) OURS(C+E) : Fidelity = 0.645, Shot Reconstruction Degree = 6.648, Compression ratio = 0.997, Clarity = 4.16, Conciseness = 4.12, Overall Quality = 4.30.

**Fig. 9.** Summary produced by different approaches for the video Exotic Terrane, segment 03.

tive results between OURS(C + E) and VSUMM for different categories.

It is interesting to note that OURS(C + E) attains lower value in terms of subjective measures for the videos in the category of TV-shows. It seems that both the approaches have a low performance for the videos in the TV-shows category as users want to see several appearances of the same anchor in the video summaries which are practically identical from the visual point of view. Table 4 shows relative improvement in the performance of our algorithm over VSUMM on these videos from the You Tube database. These relative improvements are of the same order as in [26].

In addition to relative improvements, we verify the statistical significance of all the results, the confidence intervals for the differences between paired means were computed to compare every pair of methods. 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 [23]. Since the confidence intervals (with a confidence of 98%) do not include zero in 28 out of 30 comparisons in terms of both objective and subjective measures, the results presented in Tables 5 and 6 confirm that our approach produces summaries with superior quality in relation to the compared methods.

It is important to mention that in the experiments with YouTube database, the average values of the objective and subjective measures for our method are similar to those in the Open Video database. So, we can conclude that the proposed method produces video summaries of acceptable quality for video collections with quite different characteristics.

### 6.5. Performance comparison with K-means clustering

In this section, we provide a comparative analysis between our proposed method OURS(C + E) and key frames generated using classical K-means clustering [42]. We choose K-means because of its low computational overhead in clustering of high dimensional data. On the other hand, the major drawback of K-means clustering is to decide an optimal number of clusters (key frames) to obtain the required content coverage of the produced summary. The combination of both color and edge feature are used in K-means clustering to make a fair comparison with OURS(C + E). We have chosen six videos (3 from OV and rest 3 from YouTube) randomly from the evaluation dataset. Detailed information about the six videos are given in Table 7. Table 8 presents the average value for both objective and subjective measures achieved by both approaches for the selected videos. We set the value of K as same as the number of key frames produced by the method OURS(C + E). The results indicate that proposed OURS(C + E) perform better than K-means clustering for all the video segments.

Fig. 10 shows the key frames produced by both K-means clustering and OURS(C + E) for the video A New Horizon, segment 02. From Figure, it can be noticed that there exist a lot more redundant

**Table 3**
Average value for different measures for YouTube database.

| Measures | Category | #Videos | VSUMM | OURS(C + E) |
|---|---|---|---|---|
| Fidelity | Sports | 17 | 0.438 | 0.451 |
| | Cartoons | 10 | 0.476 | 0.482 |
| | Commercials | 2 | 0.487 | 0.481 |
| | News | 15 | 0.425 | 0.441 |
| | TV-shows | 5 | 0.548 | 0.565 |
| | Home | 1 | 0.409 | 0.429 |
| | Weighted average | 50 | 0.454 | 0.466 |
| Shot Reconstruction Degree | Sports | 17 | 4.325 | 4.631 |
| | Cartoons | 10 | 3.724 | 3.821 |
| | Commercials | 2 | 4.385 | 4.396 |
| | News | 15 | 4.379 | 4.481 |
| | TV-shows | 5 | 2.902 | 2.907 |
| | Home | 1 | 4.218 | 4.228 |
| | Weighted average | 50 | 4.079 | 4.234 |
| Compression Ratio | Sports | 17 | 0.995 | 0.998 |
| | Cartoons | 10 | 0.991 | 0.993 |
| | Commercials | 2 | 0.996 | 0.996 |
| | News | 15 | 0.997 | 0.997 |
| | TV-shows | 5 | 0.998 | 0.998 |
| | Home | 1 | 0.995 | 0.994 |
| | Weighted average | 50 | 0.995 | 0.996 |
| Clarity | Sports | 17 | 3.804 | 4.022 |
| | Cartoons | 10 | 3.606 | 3.784 |
| | Commercials | 2 | 3.490 | 3.740 |
| | News | 15 | 3.449 | 3.697 |
| | TV-shows | 5 | 3.053 | 3.125 |
| | Home | 1 | 3.840 | 3.940 |
| | Weighted average | 50 | 3.571 | 3.774 |
| Conciseness | Sports | 17 | 3.888 | 4.067 |
| | Cartoons | 10 | 3.628 | 3.816 |
| | Commercials | 2 | 3.685 | 3.935 |
| | News | 15 | 3.584 | 3.941 |
| | TV-shows | 5 | 2.621 | 2.752 |
| | Home | 1 | 3.660 | 3.680 |
| | Weighted average | 50 | 3.605 | 3.834 |
| Overall Quality | Sports | 17 | 3.965 | 4.292 |
| | Cartoons | 10 | 3.694 | 3.950 |
| | Commercials | 2 | 3.900 | 3.990 |
| | News | 15 | 3.721 | 4.043 |
| | TV-shows | 5 | 2.921 | 3.075 |
| | Home | 1 | 3.740 | 3.740 |
| | Weighted average | 50 | 3.726 | 4.004 |

**Table 4**
Relative improvements of OURS(C + E) over VSUMM.

| YouTube Videos | Fidelity | Shot Reconstruction Degree | Compression Ratio | Clarity | Conciseness | Overall quality |
|---|---|---|---|---|---|---|
| VSUMM | 2.74 | 3.65 | 0.14 | 5.66 | 6.37 | 7.39 |

**Table 5**
Difference between mean of different measures at a confidence of 98% for OV database.

| Measures | Difference | Confidence interval (98%) | |
|---|---|---|---|
| | | Min. | Max. |
| Fidelity | OURS (C + E) – DT | 0.14 | 0.32 |
| | OURS (C + E) – STIMO | 0.09 | 0.21 |
| | OURS (C + E) – VSUMM | 0.11 | 0.19 |
| | OURS (C + E) – OURS(C) | 0.06 | 0.15 |
| Shot Reconstruction Degree | OURS (C + E) – DT | 0.29 | 0.38 |
| | OURS (C + E) – STIMO | 0.12 | 0.23 |
| | OURS (C + E) – VSUMM | 0.14 | 0.26 |
| | OURS (C + E) – OURS(C) | −0.046 | 0.081 |
| Compression Ratio | OURS (C + E) – DT | 0.10 | 0.25 |
| | OURS (C + E) – STIMO | 0.008 | 0.142 |
| | OURS (C + E) – VSUMM | 0.002 | 0.056 |
| | OURS (C + E) – OURS(C) | 0.08 | 0.08 |
| Clarity | OURS (C + E) – DT | 0.36 | 0.64 |
| | OURS (C + E) – STIMO | 0.09 | 0.36 |
| | OURS (C + E) – VSUMM | 0.12 | 0.25 |
| | OURS (C + E) – OURS(C) | 0.07 | 0.20 |
| Conciseness | OURS (C + E) – DT | 0.42 | 0.68 |
| | OURS (C + E) – STIMO | 0.24 | 0.45 |
| | OURS (C + E) – VSUMM | −0.112 | 0.008 |
| | OURS (C + E) – OURS(C) | 0.09 | 0.15 |
| Overall Quality | OURS (C + E) – DT | 0.47 | 0.72 |
| | OURS (C + E) – STIMO | 0.34 | 0.51 |
| | OURS (C + E) – VSUMM | 0.25 | 0.33 |
| | OURS (C + E) – OURS(C) | 0.16 | 0.31 |

**Table 6**
Difference between mean of different measures at a confidence of 98% for YouTube database.

| Measures | Difference | Confidence interval (98%) | |
|---|---|---|---|
| | | Min. | Max. |
| Fidelity | OURS(C + E) – VSUMM | 0.09 | 0.17 |
| Shot Reconstruction Degree | OURS(C + E) – VSUMM | 0.31 | 0.54 |
| Compression Ratio | OURS(C + E) – VSUMM | 0.16 | 0.34 |
| Clarity | OURS(C + E) – VSUMM | 0.36 | 0.68 |
| Conciseness | OURS(C + E) – VSUMM | −0.80 | 0.31 |
| Overall quality | OURS(C + E) – VSUMM | 0.21 | 0.54 |

**Table 7**
Dataset information.

| Video ID | Video Segment Title | Source | Frames | Genre |
|---|---|---|---|---|
| 1 | A New Horizon, segment 02 | OV | 1797 | Documentary |
| 2 | Drift Ice as a Geologic Agent, segment 03 | OV | 2742 | Educational |
| 3 | Drift Ice as a Geologic Agent, segment 10 | OV | 1407 | Lecture |
| 4 | Cartoon video | YouTube | 1424 | Cartoon |
| 5 | Sports video | YouTube | 8728 | Sports |
| 6 | Home video | YouTube | 1206 | Home |

frames (presence of both 7th and 8th frame) in the summary generated by K-means clustering. This type of redundancy is eliminated in our clustering scheme because there is no fixed number of clusters that the content needs to be distributed to as in K-means clustering. Moreover, the produced key frames lack clarity (presence of both 1st and 3rd frame) as compared to key frames produced by our proposed method. Comparing the two results, we conclude that the advantage of our proposed method over K-means is its suitability to automatic batch processing with no user specified parameters such as the number of clusters.

### 6.6. Clustering performance analysis

To compare the performance of our clustering approach with deviation ratio constraint with different clustering methods based on Delaunay graph, we use the mean density-based cluster fitness measure [46]. Density-based cluster fitness measure ($\mathcal{F}_D$) is the product of local densities and relative densities of the clusters of a given graph $G$. The relative density is the probability that a randomly chosen edge incident on the cluster is an internal edge. The local density is the probability that two randomly chosen clus-

**Table 8**
Average values for different measures for both OURS(C + E) and K-means summary.

| | Video ID | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| K-means clustering | Key frames | 8 | 8 | 5 | 12 | 10 | 7 |
| | Fidelity | 0.496 | 0.498 | 0.524 | 0.561 | 0.432 | 0.432 |
| | SRD | 2.667 | 1.896 | 2.778 | 4.223 | 5.325 | 4.221 |
| | Comp. Ratio | 0.9955 | 0.9970 | 0.9964 | 0.9915 | 0.9989 | 0.9942 |
| | Clarity | 2.98 | 3.53 | 3.81 | 3.74 | 4.12 | 3.95 |
| | Conciseness | 3.06 | 3.45 | 3.52 | 3.70 | 3.68 | 3.62 |
| | Overall quality | 3.27 | 3.78 | 4.50 | 3.84 | 4.10 | 3.74 |
| OURS(C + E) | Key frames | 8 | 8 | 5 | 12 | 10 | 7 |
| | Fidelity | 0.694 | 0.617 | 0.682 | 0.778 | 0.518 | 0.427 |
| | SRD | 3.372 | 2.205 | 3.007 | 4.536 | 5.672 | 4.228 |
| | Comp. Ratio | 0.9955 | 0.9970 | 0.9964 | 0.9915 | 0.9989 | 0.9942 |
| | Clarity | 3.94 | 4.30 | 4.00 | 4.18 | 4.20 | 3.94 |
| | Conciseness | 3.86 | 4.22 | 4.14 | 3.52 | 4.14 | 3.68 |
| | Overall quality | 4.04 | 4.38 | 4.68 | 4.04 | 4.30 | 3.74 |


(a) K-means: Fidelity = 0.496, Shot Reconstruction Degree = 2.667, Compression ratio = 0.9955, Clarity = 2.98, Conciseness = 3.06, Overall Quality = 3.27.


(b) OURS(C+E): Fidelity = 0.694, Shot Reconstruction Degree = 3.372, Compression ratio = 0.9955, Clarity = 3.94, Conciseness = 3.86, Overall Quality = 4.04.

**Fig. 10.** Summary produced by both K-means and OURS(C + E) approaches for the video A New Horizon, segment 02.

ter members are connected by an edge. A high value of average $\mathcal{F}_D$ indicates a good clustering [46]. The mean $\mathcal{F}_D$ measure for a graph clustering with $k$ clusters $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$, is given by Eq. (15):

$$\mathcal{F}_D(\mathrm{G}|\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k) = \frac{1}{k}\sum_{i=1}^{k}\mathcal{F}_D(\mathcal{C}_i) \tag{15}$$

In the above equation, $\mathcal{F}_D(\mathcal{C}_i)$ represents the cluster fitness measures of the $i$th cluster. (For details, see the Appendix A). In Table 9, we present the mean $\mathcal{F}_D$ measure obtained from three DT-based clustering methods for a five video segments randomly selected from OV and YouTube. Table 9 clearly demonstrates the superiority of the proposed method from purely clustering point of view over its two competitors.

### 6.7. Tuning of the parameters

In this section, we first show how the values of different parameters, used in the proposed method, are obtained. We choose $d$ as the minimum number of principal components which together contribute close to 90% of the total variation [6]. A maximum value of 7 can be used for $d$ in MATLAB implementations of DT [6]. From experiments, we find that a value between 5 and 7 is sufficient to capture 90% or more of the total variation for most of the videos in our test collection. Table 10 shows the variance of different video segments for different values of $d$. out of 100 video segments in our test video collection only 14 video (8 from OV and 6 from YouTube database) segments have variances less than 80% even for the maximum value of $d$(=7).

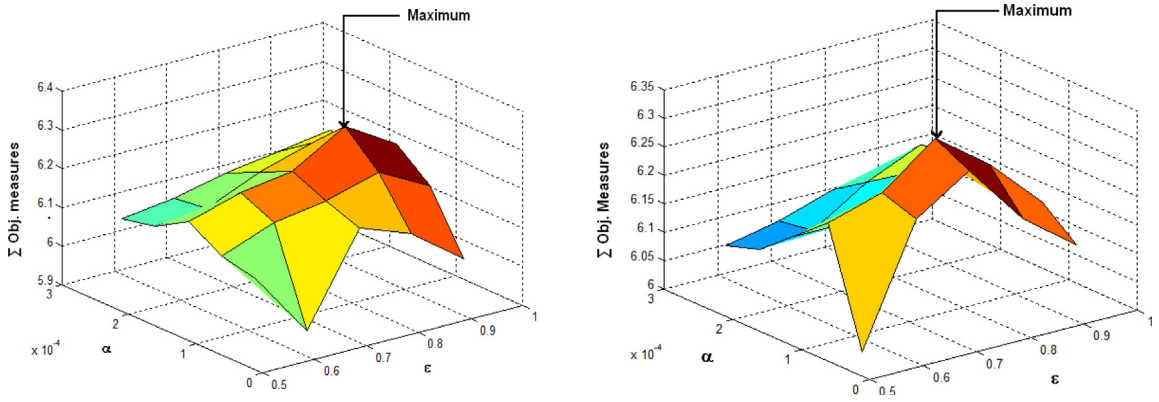
We next discuss evaluation of the parameters: $\varepsilon$ and $\alpha$ in a manner similar to [11]. The results are shown in Fig. 11, where the $x$- and $y$-axes represent the variation in the parameters $\varepsilon$ and $\alpha$, respectively. The values in the $z$-axis represent the average of the sum of objective measures achieved by each combination of those parameters. The arrows point to best combination of parameters for each database (i.e., values for $\varepsilon$ and $\alpha$ that maximize the sum of objective measures). These values are: $\varepsilon = 0.75$ and $\alpha = 0.0001$ for the Open Video database; and $\varepsilon = 0.85$ and $\alpha = 0.00015$, for the YouTube database.

**Table 9**
Mean $\mathcal{F}_D$ measure for different clustering methods.

| Video segment title | Mean $\mathcal{F}_D$ measure | | |
|---|---|---|---|
| | DT [6] | GSDR_DC [5] | With DR constraint |
| The Voyage of the Lee, Segment 05 (OV) | 0.32 | 0.29 | 0.37 |
| Drift Ice as Geologic Agent, Segment 10 (OV) | 0.22 | 0.41 | 0.43 |
| A New Horizon, Segment 08 (OV) | 0.28 | 0.36 | 0.47 |
| Cartoons Video #1 (YouTube) | 0.47 | 0.56 | 0.59 |
| News Video #12 (YouTube) | 0.33 | 0.37 | 0.41 |

**Table 10**
Variance of different video segments.

| Video segment title | #Principal components ($d$) | Variance (%) |
|---|---|---|
| The Great Web of Water, Segment 1 (OV) | 7 | 90 |
| The Great Web of Water, Segment 2 (OV) | 5 | 98 |
| A New Horizon, Segment 6 (OV) | 7 | 89 |
| Exotic Terrane, Segment 6 (OV) | 7 | 88 |
| Senses And Sensitivity, Introduction to Lecture 2 (OV) | 5 | 97 |
| America's New Frontier, Segment 4 (OV) | 5 | 90 |
| The Future of Energy Gases, Segment 5 (OV) | 7 | 92 |
| Ocean Floor Legacy, Segment 2 (OV) | 7 | 86 |
| Sports video #7 (YouTube) | 5 | 96 |
| Cartoons Video #1 (YouTube) | 7 | 89 |
| Home Video #1 (YouTube) | 7 | 83 |
| News Video #12 (YouTube) | 7 | 93 |
| Commercials Video #2 (YouTube) | 5 | 95 |




(a) Open Video (b) YouTube

**Fig. 11.** Parameter estimation strategy for $\varepsilon$ and $\alpha$.

### 6.8. Time complexity analysis

Time-complexity of our approach (in terms of number of frames $n$ and dimensionality of feature vector $d$) is O($n$ log $n$). Time-complexity for the construction of DT is O($n$ log $n$) [6] and that for the dynamic edge pruning strategy is O($kn$), $k \ll n$, $k$ is the number of iteration. So, the total complexity of the proposed method is O($n$ log $n$). This complexity is exactly same as that of the [6]. It is important to note that we obtain a better video summary as compared to [6] without affecting the time complexity.

## 7. Conclusion and future work

In this paper, we present a novel automatic video summarization technique using improved Delaunay clustering and information-theoretic pre-sampling. A combination of fixed pre-sampling and information- theoretic pre-sampling is employed for selecting the input frames for the clustering process. Information-theoretic sampling is based on detection of global valleys in the mutual information profile between successive frames of a video sequence. This approach considerably reduces the chance of loss of information as compared to only fixed pre-sampling. Improved Delaunay clustering is achieved through a dynamic edge pruning strategy via maximum global standard deviation reduction of edge lengths along with imposition of a structural constraint in form of a lower limit on the deviation ratio of the graph vertices. We undertake a comprehensive evaluation of the proposed method on 100 videos from the Open Video Project and the YouTube using three subjective and three objective measures. The detailed experimental results clearly demonstrate qualitatively and quantitatively that the proposed method produces video summaries with high quality and high user satisfaction as compared to three state-of-the-art techniques.

In future, we will focus on implementation of higher order Delaunay graphs for production of both static and dynamic video summaries using different graph centrality measures. Another direction of future research will be to use a more extensive set of features like color, motion, shape and texture along with an efficient feature fusion strategy to obtain more meaningful video summaries.

## Appendix A

In a graph G = (V,E), a cluster candidate is a set of vertices $\mathcal{C} \subseteq V$. The order of the cluster is the number of vertices included in the cluster, denoted by $|\mathcal{C}|$. The internal degree and external degree of a cluster $\mathcal{C}$ are defined as follows:

$$\deg_{\text{int}}(\mathcal{C}) = |\{\{v,u\} \in E | v,u \in \mathcal{C}\}| \tag{A.1}$$

$$\deg_{\text{ext}}(\mathcal{C}) = |\{\{v,u\} = E | v \in \mathcal{C}, u \in V \setminus \mathcal{C}\}| \tag{A.2}$$

Relative density is the ratio of the internal degree to the number of edges incident to the cluster,

$$\rho_r(\mathcal{C}) = \frac{\deg_{\text{int}}(\mathcal{C})}{\deg_{\text{int}}(\mathcal{C}) + \deg_{\text{ext}}(\mathcal{C})} = \frac{\sum_{v \in \mathcal{C}} \deg_{\text{int}}(v,\mathcal{C})}{\sum_{v \in \mathcal{C}} \deg_{\text{int}}(v,\mathcal{C}) + 2\deg_{\text{ext}}(v,\mathcal{C})} \tag{A.3}$$

which favors connected components with few connections to other parts of the graph.

The internal degree of a vertex can be defined as

$$\deg_{\text{int}}(v,\mathcal{C}) = |\Gamma(v) \cap \mathcal{C}| \tag{A.4}$$

To measure how densely $v$ is connected to $\mathcal{C}$, we need to scale this by the maximum number of neighbors that a vertex could have in $\mathcal{C}$, to obtain a measure in $[0,1]$:

$$\delta(v,\mathcal{C}) = \frac{\deg_{\text{int}}(v,\mathcal{C})}{|\mathcal{C}|-1} \tag{A.5}$$

The local density measure would be a scaled sum of vertex densities given by Eq. (A.6)

$$\delta_{\text{l}}(\mathcal{C}) = \frac{1}{|\mathcal{C}|}\sum_{v\in\mathcal{C}}\delta(v,\mathcal{C}) = \frac{1}{|\mathcal{C}|(|\mathcal{C}|-1)}\sum_{v\in\mathcal{C}}\deg_{\text{int}}(v,\mathcal{C}) \tag{A.6}$$

The sum of the internal degrees of vertices in $\mathcal{C}$ is twice the internal degree of the cluster, as each internal edge is counted independently by both of its endpoints. This simplifies the above equation into

$$\delta_{\text{l}}(\mathcal{C}) = \frac{1}{|\mathcal{C}|(|\mathcal{C}|-1)}\cdot 2\deg_{\text{int}}(\mathcal{C}) = \frac{\deg_{\text{int}}(\mathcal{C})}{\binom{|\mathcal{C}|}{2}} \tag{A.7}$$

Finally, the $\mathcal{F}_D$ measure for individual cluster is given by

$$\mathcal{F}_D(\mathcal{C}) = \delta_{\text{l}}(\mathcal{C})\cdot\rho_{\text{r}}(\mathcal{C}) = \frac{2\deg_{\text{int}}(\mathcal{C})^2}{|\mathcal{C}|(|\mathcal{C}|-1)(\deg_{\text{int}}(\mathcal{C})+\deg_{\text{ext}}(\mathcal{C}))} \tag{A.8}$$

## References


[1] S.-F. Chang, W. Chen, H.J. Meng, H. Sundaram, D. Zhong, A fully automated content-based video search engine supporting spatio-temporal queries, IEEE Transactions on Circuits Systems for Video Technology 8 (5) (1998) 602–615.

[2] D.B. Ponceleon, S. Srinivasan, A. Amir, D. Petkovic, D. Diklic, Key to effective video retrieval: effective cataloging and browsing, in: Proceedings of the ACM International Conference on Multimedia, 1998, pp. 99–107.

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

[4] A.G. Money, H.W. Agius, Video summarization: a conceptual framework and survey of the state of the art, Journal of Visual Communication and Image Representation 19 (2) (2008) 121–143.

[5] A.S. Chowdhury, S. Kuanar, R. Panda, M.N. Das, Video Storyboard Design using Delaunay Graphs, in: Twenty First IAPR/IEEE International Conference on Pattern Recognition (ICPR), Tsukuba City, Japan, 2012, pp. 3108–3111.

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

[7] M. Furini, F. Geraci, M. Montangero, M. Pellegrini, STIMO: STIll and Moving video storyboard for the web scenario, Multimedia Tools and Application 46 (1) (2010) 47–69.

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

[9] L. Herranz, J. Martinez, An efficient summarization algorithm based on clustering and bitstream extraction, in: IEEE International Conference on Multimedia and Expo, 2009, pp. 654–657.

[10] Y. Gong, X. Liu, Video summarization and retrieval using singular value decomposition, ACM Multimedia Systems Journal 9 (2) (2003) 157–168.

[11] J. Almeida, N.J. Leite, R.S. Torres, VISON: video summarization for online applications, Pattern Recognition Letters 33 (4) (2012) 397–409.

[12] A. Hanjalic, H. Zhang, An integrated scheme for automated video abstraction based on unsupervised cluster- validity analysis, IEEE Transactions on Circuits Systems for Video Technology 9 (8) (1999) 1280–1289.

[13] F.P. Preparata, M.I. Shamos, Computational Geometry: An Introduction, Springer-Verlag, New York, 1985.

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

[15] J. Almeida, N.J. Leite, R.S. Torres, Online video summarization on compressed domain, Journal of Visual Communication and Image Representation 24 (6) (2013) 729–738.

[16] Z. Cernekova, I. Pitas, C. Nikou, Information theory-based shot cut/fade detection and video summarization, IEEE Transactions on Circuits Systems for Video Technology 16 (1) (2006) 82–91.

[17] Z. Li, G.M. Schuster, A.K. Katsaggelos, Minmax optimal video summarization, IEEE Transactions on Circuits Systems for Video Technology 15 (10) (2005) 1245–1256.

[18] Z. Cernekova, C. Nikou, I. Pitas, Entropy metrics used for video summarization, in: Proceedings of the 18th Spring Conference on, Computer Graphics, 2002, pp. 73–82.

[19] G. Paschos, Perceptually uniform color spaces for color texture analysis: an empirical evaluation, IEEE Transactions on Image Processing 10 (6) (2001) 932–937.

[20] B.S. Manjunath, J.R. Ohm, V.V. Vasudevan, A. Yamada, MPEG-7 color and texture descriptors, IEEE Transactions on Circuits and Systems for Video Technology 6 (11) (2000).

[21] E. Sahouria, A. Zakhor, Content analysis of video using principal components, IEEE Transactions on Circuits and Systems for Video Technology 9 (8) (1999).

[22] A. Pothen, C.J. Fan, Computing the block triangular form of a sparse matrix, ACM Transactions on Mathematical Software 16 (4) (1990) 303–324.

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

[24] H.S. Chang, S. Sull, Sang Uk Lee, Efficient video indexing scheme for content-based retrieval, IEEE Transactions on Circuits and Systems for Video Technology, 9 (8), 1999, pp. 1269–1279.

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

[26] G. Ciocca, R. Schettini, A innovative algorithm for key frame extraction in video summarization, Journal of Real-Time Image Processing 1 (1) (2006) 69–88.

[27] M. Slaney, Precision-recall is wrong for multimedia, IEEE Multimedia 18 (3) (2011) 4–7.

[28] M. Swain, D. Ballard, Color indexing, International Journal of Computer Vision 7 (1) (1991) 11–32.

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

[30] J. Yang, J.Y. Yang, D. Zhang, J.F. Lu, Feature fusion: parallel strategy vs. serial strategy, Pattern Recognition 36 (6) (2003) 1369–1381.

[31] Q.S. Sun, S.G. Zeng, Y. Liu, P.A. Heng, D.S. Xia, A new method of feature fusion and its application in image recognition, Pattern Recognition 38 (2005) 2437–2448.

[32] J. Yang, J.-Y. Yang, Generalized K–L transform based combined feature extraction, Pattern Recognition 35 (1) (2002) 295–297.

[33] H. Tong, J. He, M. Li, C. Zhang, W. Ma. Graph Based Multi-Modality Learning, in: ACM Conference on Multimedia, 2005, pp. 862–871.

[34] A. Komlodi, G. Andmarchionini, Key frame preview techniques for video browsing, in: Proceedings of ACM Conference on Digital Libraries, 1998, pp. 118–125.

[35] M.M. Yeung, B.L. Andleo, Video visualization for compact representation and fast browsing of pictorial content, IEEE Transaction on Circuit System for Video Technology 7 (5) (1997).

[36] S. Uchiashi, J. Foote, A. Girgensohn, J. Andboreczky, Video manga: generating semantically meaningful video summaries, in: Proceedings of the ACM Multimedia Conference, 1999, pp. 383–392.

[37] P. Chiu, A. Girgensohn, Q. Andliu, Stained-glass visualization for highly condensed video summaries, in: Proceedings of the International Conference on Multimedia and Expo, 2004.

[38] Y. Tonomura, A. Akutsu, K. Otsuji, T. Andsadakata, Video Map and video SpaceIcon: Tools for anatomizing video content, in: Proceedings of the INTERCHI Conference, 1993, pp. 131–136.

[39] C. Brian Atkins, Blocked Recursive Image Composition, in: Proceedings of ACM Conference on Multimedia, 2008, pp. 821–824.

[40] T. Wang, T. Mei, X.S. Hua, X.L. Liu, H. Q. Zhou, Video collage: a novel presentation of video sequence, in: International Conference on Multimedia & Expo, 2007, pp. 1479–1482.

[41] C. Rother, L. Bordeaux, Y. Hamadi, A. Blake, Autocollage, in: ACM SIGGRAGPH, 2006.

[42] S. Bow, Pattern Recognition and Image Processing, Publisher Marcel Dekker, Inc., New York, 2002.

[43] J.C.S. Yu, M.S. Kankanhalli, P. Mulhen, Semantic video summarization in compressed domain MPEG video, in: IEEE International Conference on Multimedia and Expo, 2003, pp. 329–332.

[44] The Open Video Project: <http://www.open-video.org>.

[45] Youtube Database: <http://www.youtube.com/>.

[46] S.E. Scaeffer, Graph clustering, Computer Science Review 1 (1) (2007) 27–64.