Redundancy Elimination in Video Summarization

Hrishikesh Bhaumik, Siddhartha Bhattacharyya and Susanta Chakraborty

Abstract Video summarization is a task which aims at presenting the contents of a video to the user in a succinct manner so as to reduce the retrieval and browsing time. At the same time sufficient coverage of the contents is to be ensured. A trade-off between conciseness and coverage has to be reached as these properties are conflicting to each other. Various feature descriptors have been developed which can be used for redundancy removal in the spatial and temporal domains. This chapter takes an insight into the various strategies for redundancy removal. A method for intra-shot and inter-shot redundancy removal for static video summarization is also presented. High values of precision and recall illustrate the efficacy of the proposed method on a dataset consisting of videos with varied characteristics.

Keywords Video summarization · Redundancy removal · Feature descriptors · Metrics for video summary evaluation · Three-sigma rule

1 Introduction

The ever growing size of online video repositories like DailyMotion, YouTube, MyVideo etc. have propelled the need for efficient Content Based Video Retrieval Systems. This has augmented research in several related fields such as, feature extraction, similarity/dissimilarity measures, video segmentation (temporal and semantic),

H. Bhaumik (✉) · S. Bhattacharyya
Department of Information Technology, RCC Institute of Information Technology, Kolkata, India
e-mail: hbhaumik@gmail.com

S. Bhattacharyya
e-mail: dr.siddhartha.bhattacharyya@gmail.com

S. Chakraborty
Department of Computer Science and Technology,
Indian Institute of Engineering Science and Technology, Shibpur, Howrah, India
e-mail: susanta.chak@gmail.com

A.I. Awad and M. Hassaballah (eds.), *Image Feature Detectors and Descriptors*,
Studies in Computational Intelligence 630, DOI 10.1007/978-3-319-28854-3_7

key-frame extraction, indexing, annotation, classification and retrieval of videos. Video Summarization is a task which stands at the intersection of these research issues. It reduces to a great extent the demand for computing resources in processing the huge volumes of video data. The basic objective of the video summarization process is to provide a concise, yet meaningful representation of the video to the user. The efficacy of a video summarization system depends on maximizing two conflicting features-coverage and succinctness. The important application areas of video summarization include content-based video retrieval [1–3], semantic indexing [4], Copied Video Detection (CVD) [5], video surveillance [6], generation of highlights for sports [7–10], movies and drama [11–14], bandwidth-constrained video processing applications [15] etc. The hierarchical levels of the composing units in a video may consist of scenes, shots or frames depending on the granularity intended as depicted in Fig. 1. The composing units share a temporal relationship with each other. Distortion of semantic content occurs if these temporally sequenced units are disordered. A video may be represented as $V = u_1 \bigotimes u_2 \bigotimes u_3 \ldots \bigotimes u_n$, where u_i is the ith composing unit. Depending on the summarization approach used, a Boolean decision is made for each constituent unit of the video as to whether it will be a part of the generated summary or not. The mechanisms for selecting these units determine the efficacy of the approach used. Video summarization can also be viewed as a task of amalgamating those video units which have the maximum entropy. The system generated summary (SGS) consists of a subset of V (extraction type) or a transformed set of the elements in V (abstraction type). In both cases the duration of SGS is far smaller than V. Static video summarization falls under extraction type where a set of key-frames are chosen to represent the video. This is particularly helpful in bandwidth constrained scenarios, where the user needs to get an overview of the contents of a video. On the other hand, a dynamic video summary may be produced by coalescing together the units which have greater significance. In such case, the summary generated may be either extraction based (e.g. sports highlights package) where the chosen units have the same time-sequence in which they occur in the original video or abstraction based (e.g. movie trailer) where the selected units may be intermingled in a manner so as to produce a meaningful abstract of the given

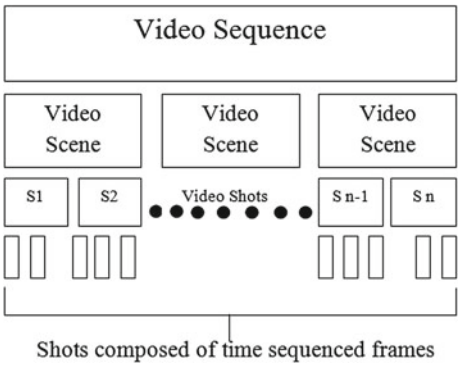

Fig. 1 Hierarchical representation of video units

video. The challenges for designing a video summarization system arise from the fact that the summary generated must be represented by the most significant composing units. The ranking of video units according to significance is the crux of the problem besetting researchers in this field. An equally considerable problem is to remove redundancy without diminishing the coverage made by the representative units. Elimination of visual redundancy is possible by extracting mid-level features such as interest points. However, removal of semantic redundancy is a problem of a different dimension. The task of semantic redundancy elimination is more complex as it encompasses fields like object recognition, tracking, gesture/action identification, event detection etc. to name a few. Hence, such a task requires extraction of high level features. In this chapter, we focus on approaches taken for reduction of visual redundancy.

The rest of the chapter is organized as follows. Section 2 details the related works in video summarization using redundancy elimination approaches. Section 3 presents an insight into the redundancy elimination problem in video summarization and why it needs to be tackled. Section 4 enumerates the role of interest points detection in video summarization. This section also elaborates the various feature descriptors in use. The proposed method is presented in Sect. 5. In Sect. 6 some widely used metrics for measuring the quality of summary is presented. The results of the summarization process using the proposed method along with the details of the dataset used is elaborated in Sect. 7. Some concluding remarks are presented in Sect. 8.

2 Survey of Related Works

Redundancy occurs due to the appearance of similar visual content at several points in a video. This invariably increases the size of the resulting summary as it is formed by coalescing together several video units taken from different points in the video. The task of redundancy removal refers to the elimination of the repeated content which conflicts with the objective of producing a concise summary. Several techniques for redundancy removal have been devised over the years. These methods can be broadly categorized into two groups:

1. Techniques using feature descriptors
2. Other redundancy elimination techniques.

2.1 Methods of Redundancy Elimination Using Feature Descriptors

Similarity in visual content may be aptly captured by using feature descriptors. The feature descriptors capture medium-level semantic content. This is intermediate

between low-level characteristics such as histogram comparisons [16, 17], statistical differences [18, 19], standard deviation of pixel intensities [20], frame-to-frame pixel intensity difference [21], gray level histograms of successive frames [17, 22], statistical information like mutual and joint entropy [19], mean and variance of pixel values, luminance etc. and high-level features like shapes of objects, edges in the frames, optical flow [23], motion vectors [24], event modeling [25], etc. The high-level features which are connected to the content of a video such as scenes, objects etc. are more natural to humans than the low-level features. As such the features which capture points on the objects rather than the whole semantic meaning come under mid-level features. The mid-level features are useful for detection and recognition of objects which have consistent low-level characteristics. However, these mid-level characteristics may not be useful for semantic analysis of the content in a video. Feature descriptors have been used in several video analysis problems. These include Shot Boundary Detection [18], Video Summarization [26], Object tracking [27] etc.

The various approaches related to Shot Boundary Detection aim at extracting feature descriptors from the time sequenced frames of a video. Depending on the feature descriptors extracted from the frames the similarity between consecutive frames are computed. The number of matched features is tracked to find abrupt discontinuities. The points of discontinuities are the shot boundaries in the video sequence. Gradual discontinuity patterns indicate smooth transition from one shot to another. These are achieved through fades, wipes and dissolves. Apart from detecting shot boundaries, the matched feature descriptors also serve as indicators to the amount of content match between two shots. This aspect is exploited by researchers in tasks related to summarization of a video. In [28], key-points are recognized for all the frames of each shot in a video. A set of unique key-points is built for the shots. The set of feature descriptors corresponding to the key-points are extracted. The representative set of key-frames is constructed such that minimum number of frames covers the entire pool of key-points. This ensures maximizing the coverage and minimizing the redundancy [29]. In [30], an approach for static video summarization using semantic information and video temporal segmentation is taken. The performance and robustness of local descriptors are also evaluated as compared to global descriptors. The work also investigates, as to whether descriptors using color information contribute to better video summarization than those which do not use it. Also the importance of temporal video segmentation is brought out in the work. The summarization process uses a bag of visual words concept where the feature descriptors are used to describe a frame. Visual word vectors are formed to cluster similar frames and finally filter out the representative frames. The performance of various feature detectors and descriptors in terms of tracking speed and effectiveness were evaluated in [31]. The work pertains to evaluation of these feature descriptors for face detection in real-time videos. Change in structure for non-rigid objects, sudden changes in object motion resulting in varied optical flow, change in manifestation of objects, occlusions in the scene and camera motion are some of the inherent challenges which have to be overcome for accurate tracking of objects. A proper amalgamation of these feature descriptors may serve to improve the overall tracking precision. The work concludes that a single feature detector may not provide enough accuracy for object tracking.

An event detection system has been proposed in [32] for field sport videos. The system runs two parallel threads for detecting text and scene in the video stream. The output of a text detector is provided to a scoreboard analyzer for notifying the user of an interesting event. The scene analyzer which runs parallel to the text analyzer takes input from the scene detector and provides output to the event notification system to tag the interesting sequences. Since the approach is designed for real-time videos, the authors stress the need for using feature detectors which are inherently fast and have an acceptable recognition rate.

In sports videos, the event is usually covered by a fixed set of cameras on stands. As such the coverage is free from rotational variance. Also there are long shots and close-ups which need to be distinguished. The algorithms designed for such purposes may therefore ignore scale and rotation invariance strategies. The BRIEF descriptor was chosen for this work as it satisfies these considerations and is computationally efficient. BRIEF has been reported to be almost sixty times faster than SURF, while ensuring an acceptable recognition rate [33]. In [34] an elaborate comparison of the various descriptor extraction techniques is presented. The work reviews techniques like SIFT, DIFT, DURF and DAISY in terms of speed and accuracy for real-time visual concept classification. A number of high speed options have been presented for each of the components of the Bag-of-Words approach. The experiments consist of three phases i.e. descriptor extraction, word assignment to visual vocabulary and classification. The outcome of this work can be extended for designing robust methods for redundancy elimination based on visual concept classification.

Li [35] employs SIFT as the basis for computing content complexity and frame dissimilarity. This allows detection of video segments and merging of the shots based on similarity. Key-frames are then extracted from these merged shots. In [36], Compact Composite Descriptors (CCDs) [37] and the visual word histogram are extracted for each image. The object descriptor used is based on SURF. The CCD consists of four descriptors i.e. the Color and Edge Directivity Descriptor (CEDD) [38], the Fuzzy Color and Texture Histogram (FCTH) [39], the Brightness and Texture Directionality Histogram (BTDH) [40] and the Spatial Color Distribution Descriptor (SpCD) [41]. A Self-Growing and Self-Organized Neural Gas (SGONG) network is used for frame clustering. The main aspect of this method is the ability to determine the appropriate number of clusters. As in some of the other methods, the cluster centers are chosen to generate the summary. Redundancy elimination is carried out in [42] by extracting the SURF and GIST features from the representative frames obtained by generating a Minimal Spanning Tree for each shot. The duplicate frames in the representative set are eliminated using a threshold based on the three-sigma rule in accordance with the number of descriptor matches for each pair of frames in the representative set. A comparison of the summaries after redundancy elimination using SURF and GIST are also elaborated.

2.2 Other Methods of Redundancy Elimination

Apart from using mid-level features in the form of interest points for removal of content duplication, several other methods have been proposed by the researchers. A few of the important approaches are presented in this section.

The approaches using key-frame selection for static video summarization aim to summarize the video by selecting a subset of frames from the original set of decomposed frames. In order to remove redundancy from the set of selected frames, clustering is applied on the set of selected key-frames by extracting features. One such method is presented in [43] which includes a feature extraction phase required for clustering. Duplicates from the selected key-frames are removed using a combination of local and global information. In [44] an exploration framework for video summarization is proposed. Key-frames are selected from each shot based on the method described in [45]. The redundant frames are eliminated using a self-organizing map. The redundancy eliminated set of key-frames are connected in a network structure to allow the users to browse through the video collection. The power and simplicity of color histograms have been exploited in several works for finding the similarity between frames and thereby remove duplication. In [46] the main low-level feature used is a color histogram. The given video is first segmented into shots and clustering is performed on the set of frames based on color histogram extracted from each frame. The frame at the centroid of each cluster forms a part of the final key-frame set. Although color histogram is a very elegant low-level feature, however, the computational complexity involved for extraction and comparison is high as it represents a vector of high dimensionality. In order to eliminate the components having lower discrimination power, singular value decomposition (SVD) is used in [47]. In [48] principal component analysis (PCA) is applied on the color histogram to reduce the dimensionality of the feature vector. Delaunay clustering is used to group the frames using the reduced feature vector. The center of each cluster represents a key-frame of the storyboard. PCA has also been used in [49, 50] to reduce the elements in a histogram. Further, in [49], Fuzzy C-means and frame difference measures are used to detect shot boundaries in the video. The use of Fuzzy-ART and Fuzzy C-Means is also proposed in [50] to extract shots from the given video by identifying the number of clusters without any *apriori* information. A cost-benefit analysis of using PCA has not yet been done.

Furini et al. proposed a tool called STIMO (Still and Moving Video Storyboard) in [51] which was capable of generating still and moving storyboards on the fly. The tool also enabled users to specify the length of summary and waiting time for summary generation. A clustering algorithm is executed on the HSV color descriptors extracted from each of the frames. A representative frame is selected from each cluster to produce the static storyboard. A similar approach is used to cluster the shots and choose sequences from the clusters to produce a moving storyboard. A similar approach is used in [52] where the K-means clustering algorithm is used on the HSV color features. The final storyboard is formed by choosing a frame from each cluster. In [53] an approach for summarization of news videos is discussed. The extracted

key-frames are clustered together using affinity propagation. A vector space model approach is then used to select shots having high information content. This ensures that the key-frames having discrimination power are retained and visual redundancy is removed. Liu et al. [54] and Ren et al. [55] are works which aim to summarize the rushes video [56]. Liu et al. [54] implements a multi-stage clustering algorithm to remove redundant shots. A value for frame significance is computed based on change of content with time and spatial image salience. The most important parts of the video are extracted based on the frame significance value. Using formal language technique, [55] introduces a hierarchical model to remove unimportant frames. An adaptive clustering is used to remove redundancy while summarizing the rushes video. In [57], a pair of clips is modeled as a weighted bipartite graph. The similarity between the clips of a video is computed based on max-weighted bipartite matching algorithm. The clustering process is based on affinity propagation algorithm and serves to remove redundancy. In [58] a method for video object segmentation is presented which removes redundancy from the spatial, temporal and content domains. A 3D graph-based algorithm is used to extract video objects. These objects are clustered based on shapes using the K-means algorithm. Key objects are identified by selecting objects from the clusters for obtaining intended summarization. A joint method for shot boundary detection and key frame extraction is presented in [59] wherein a method based on three probabilistic components is considered. These are the prior of the key frames, the conditional probability of shot boundaries and the conditional probability of each video frame. Gibbs sampling algorithm [60] is used for key frame extraction and generation of the storyboard. This also ensures that duplication is removed from the final summary.

3 Redundancy Elimination in Video Summarization

Duplication in video content occurs when the same scene or objects are covered by a set of multiple cameras. This duplication of visual content may occur within a given shot (intra-shot level) or between several shots (inter-shot level). Removal or retention of such redundant content is contextual and depends on the genre of the video. Duplication of content holds a different perspective for a sports video like soccer than for news video or a documentary. It is still different for video surveillance applications where only the frames containing some event or activity might be of interest. This emphasizes the point that different approaches to redundancy removal are required in different situations and the same algorithm may not work in all cases. The basic objective of the video summarization task is to provide maximum coverage of the contents while attempting to select the minimum number of video units possible. It can be easily perceived that the two objectives are inversely proportional and conflicting to each other. Redundancy elimination aims to achieve the later objective without affecting the former. Hence it is seen as one of the most important steps in the summarization task. Since redundancy removal is a phase where an attempt is made to eliminate visually redundant units of the video, it assumes

vital importance in bandwidth constrained scenarios where the user perception has to be maximized with minimum amount of data transmission between the source and destination. Redundancy elimination thus helps the user to get an insight into the contents of the video in least possible time. This is also important for video indexing applications where the non-redundant frames can be viewed as the features of the video. This characterization helps to symbolize the video in order to facilitate content based video retrieval. Visual redundancy is removed through the use of one or more members from the family of feature descriptors like SIFT, SURF, DAISY, GIST, BRIEF, ORB etc. (described in later sections). The interest points extracted by using these feature descriptors serve as mid-level features necessary for finding the overlap in visual content between the composing units of the video. Setting a threshold for permitting overlap is another important task in this process. A stringent threshold ensures that there is almost no overlap in visual content. This is sometimes necessary for a storyboard representation of the video. For duplication removal in video skimming applications, the amount of similarity between shots may be computed from the number of matching interest points in the frames composing the shots. A decision on elimination is taken on a threshold computed on the similarity values of these features. Figure 2 depicts removal of duplicate frames in a video. An elaboration on the various feature descriptors used for redundancy control in video summarization tasks is presented in the next section.

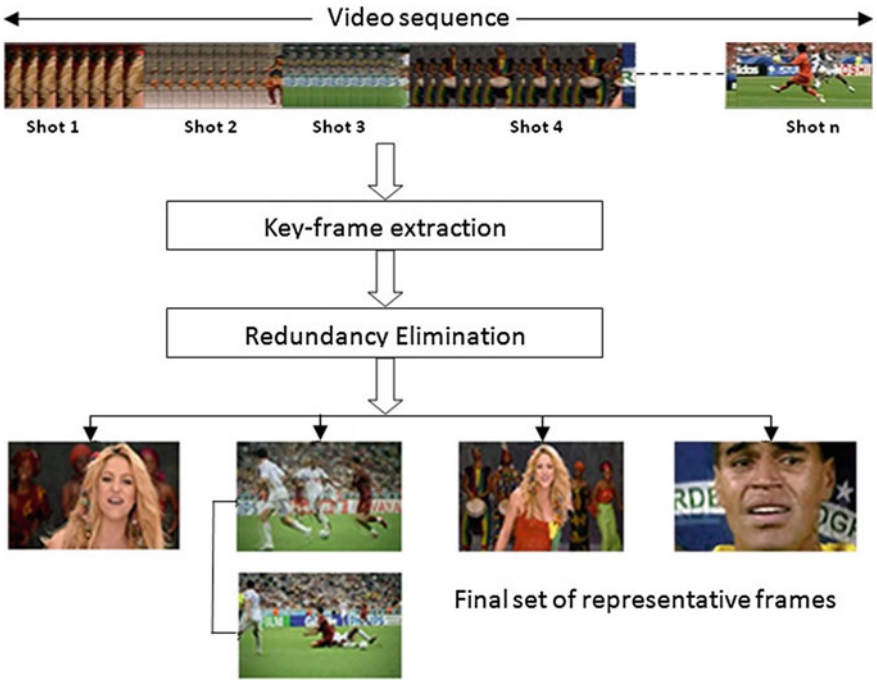

Fig. 2 Redundancy removal

4 Role of Interest Point Detection in Video Summarization

Interest point detection is a field of computer vision that refers to the identification of points which serve as features to the contents in an image. Interest point generation is characterized by a definite position in the spatial domain of an image and is defined by a strong mathematical model. This ensures a high degree of reproducibility in images under different transformations. An interest point descriptor is used to describe the texture around the point. Detection of image features has been the focal point of research in the field of computer vision over the last few decades. Image features include edges, corners, ridges, blobs, textures, interest points etc. The application areas of image feature extraction encompass object identification and tracking [61, 62], video surveillance [6, 63], image similarity/dissimilarity metrics [64], content-based image and video retrieval [1–3, 65, 66], image and video mosaicing [67, 68], video stabilization [69], 3D image construction [70], video summarization [71] etc., to name a few. The matching of a pair of images using feature points involves three stages i.e. detection of the feature points, description of these points using an n-dimensional vector and matching these feature vectors. In this chapter, we focus mainly on interest point detectors and descriptors which can be used for elimination of redundant frames in a video summarization task.

Initially, interest point detectors were developed with the motivation of extracting robust and stable features which could reliably represent salient points in the image and serve as identifiers to it. As research progressed in this field, the focus was on developing algorithms which extracted feature points immune to variations in light intensity, scale, rotation etc. Further advances in the field centered on development of methods which could reliably extract feature points in lesser time by eliminating information around the chosen points which would not degrade the performance of the interest point detector. Interest point detectors are based on well-substantiated mathematical models. Interest points are illustrated by a distinct position in the image space and are usually represented by a multi-dimensional representative vector. These vectors encompass local information content of that point which would help to discriminate it from other points and would also distinctly identify that point in a perturbed image. It is important to note that the change in relative position of the selected interest points can be used to estimate the amount of geometric transform in the objects of a given set of images. The noise points or outliers are detected by tracking huge change in the estimated transform for the objects in a given image with respect to the original scene. The interest points corresponding to an object in an image have mid-level semantic features for describing and identifying it. A majority of the interest points detected lie on the high frequency regions of the image.

Feature point descriptors have been used by researchers to boost the algorithms designed for video summarization. This is in contrast to summarization methods which use visual descriptors [72, 73]. Computer vision algorithms aim to extract the semantic meaning of images composing the video. This augments the target of video summarization algorithms to provide a content revealing summary through a concise representation. Semantic understanding of videos is still a far-fetched reality. In the

further sections, the various interest point detectors and descriptors are presented which have been used for the video summarization task in various ways.

4.1 Scale Invariant Feature Transform (SIFT)

SIFT is a feature point detector and descriptor method, proposed by Lowe [74] in 1999. The goal here is to extract certain key-points corresponding to objects in an image. These key-points are represented by a set of low-level features, necessary for identification of the objects in an experimental image containing other objects. The feature points so taken are immune to various translations (such as rotation and scaling) and also to changes in light intensity. The points are chosen from high contrast regions, rendering them to be detected under several types of perturbations. The four steps involved in this method include:-

1. Scale-space Extrema Detection
2. Key-point Localization
3. Orientation Assignment
4. Key-point Description

The SIFT detector and descriptor is designed to be fully immune to changes in scale and orientation. It is also partially immune to affine distortion and changes in light. It can be used to identify an object from a group of objects in a given scene. SIFT feature points are described by feature vectors having 128 elements. Given a pair of images, the feature points are first detected from both images and the corresponding descriptors are computed. Euclidean distance between the two set of feature vectors is then calculated to find the initial set of candidate matches. A subset of the feature point matches for an object is taken which agree on the scale, orientation and location is taken to separate out the superior matches. A hash table based on the generalized Hough transform is used to find the consistent clusters. A cluster must contain at least three feature points to be considered for the next stage of model verification. The probability for presence of an object is computed based on the set of features given. The matches that pass these checks are recognized as true matches with high confidence. Figure 3 depicts the detection of SIFT interest points on an image. A number of variants for SIFT such as PCA-SIFT (based on Principal Component Analysis) [75], Harris-SIFT (based on Harris interest points) [76] etc. with different characteristics have been designed for various uses. Various approaches to video summarization use SIFT or its variants. In [77] a video summarization method is presented where web images are used as prior input for summarizing videos containing similar set of objects. SIFT Flow [78] is used to define frame distance in order to determine the similarity of one frame with another frame. As mentioned previously, the SIFT descriptor has been used vastly in computer vision for its robustness and ability to handle intensity, rotation, scale variations despite its high computational cost. In [79], SIFT and SURF (described in the next section) feature descriptors have been used to detect forgery in images where copy-move

Fig. 3 An image marked with SIFT interest points

technique has been used. An approach to video summarization where the semantic content is preserved has been presented in [80]. The video is segmented into shots and SIFT features for each shot are extracted. The latent concepts are detected by spectral clustering of bag-of-words features to produce a visual word dictionary.

4.2 Speeded-Up Robust Features (SURF)

Speeded-Up Robust Features (SURF) [81] was proposed by Herbert Bay et. al. It was inspired from SIFT. The main advantage of SURF over SIFT is its low execution speed and computational complexity over the latter. It is claimed to be more robust than SIFT for different image transformations. It provides reliable matching of the detected interest points by generating a 64 element vector to describe the texture around each point of interest. The generated vector for each interest point are designed to be immune to noise, scaling and rotation. SURF has been used widely in object detection and tracking. Determinant of the Hessian blob detector is used for the detection of interest points. To detect scale-invariant features, a scale-normalized second order derivative on the scale space representation is used. SURF approximates this representation using a scale-normalized determinant of the Hessian (DoH) operator. The feature descriptor is computed from the sum of the Haar wavelet [82] response around the point of interest. To find the similarity between a pair of images, the interest points detected are matched. The amount of similarity between the images is the ratio of descriptor matches to the total number of interest points detected. Figure 4 illustrates the SURF correspondences on two similar video frames. Research article [83] deals with identifying faces in CCTV cameras installed for surveillance purposes. A database of human faces is created as new faces appear in front of the camera. A Haar classifier is used for recognizing human faces in images. SURF descriptors provide a match between the detected face and existing faces in the database. In case the faces in the database do not match, the new

Fig. 4 SURF correspondences between two similar images

face is updated in the database. Bhaumik et al. [84] presents a technique for static video summarization in which key-frames detected in the first phase from each shot are subjected to redundancy elimination at the intra-shot and inter-shot levels. For removal of redundant frames, SURF and GIST feature descriptors were extracted for computing the similarity between the frames. The work also compares the quality of summary obtained by using SURF and GIST descriptors in terms of precision and recall.

4.3 DAISY

The DAISY feature descriptor was proposed by Tola et al. [85]. It was inspired by the SIFT and GLOH [86] feature descriptors and is equally robust. DAISY forms 25 sub-regions of 8 orientation gradients, resulting in a 200 dimensional vector. The sub-regions are circular in nature and can be computed for all pixels in an image. A Gaussian kernel is used in DAISY as opposed to a triangular kernel for SIFT and GLOH. In this descriptor, several Gaussian filters are used on the convolution of the gradients in definite directions. This is in contrast to the weighted sums of gradient norms used in SIFT and GLOH. DAISY provides very fast computation of feature descriptors in all directions and is therefore appropriate for dense-matching. According to [34], DAISY is 2000 % faster than SIFT 4×4, when sampling each pixel.

4.4 GIST

GIST [87] feature descriptor was proposed by Oliva et al. in 2001 to represent the dominant spatial structure of a scene. This low-level representation is done using a set of five perceptual dimensions i.e. naturalness, openness, roughness, expansion and ruggedness. The spectral components at different spatial locations of the spatial envelope is computed by using a function called the Windowed Discriminant Spectral

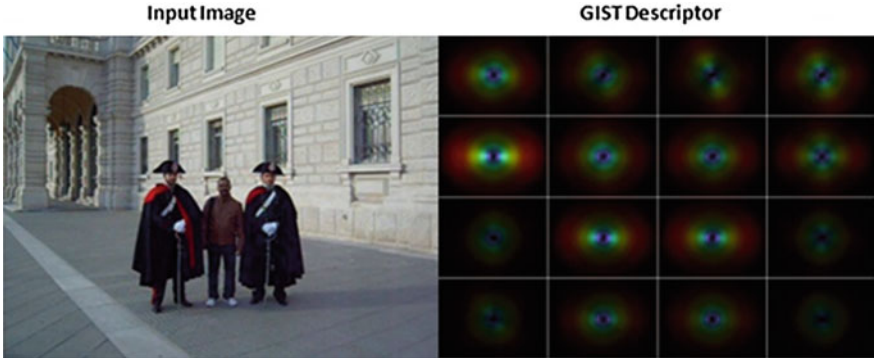

Fig. 5 GIST descriptors for an image

Template (WDST). The perceptual dimensions can be reliably computed by using spectral and coarsely localized information. GIST has been used by researchers for various applications such as finding similarity in images for redundancy removal [42, 84], similar image retrieval [88] and 3D modeling [89], scene classification [90], image completion [91] etc. Different approaches have been developed by Torralba et al. [92, 93] to compress the GIST descriptor. Figure 5 depicts GIST descriptors for an image.

4.5 Binary Robust Independent Elementary Features (BRIEF)

BRIEF [94] was proposed by Calonder et al. in 2010. It is a feature point descriptor which can be used with any available feature detector. Commonly used feature detectors like SIFT and SURF generate long vectors of 128 and 64 dimensions respectively. Generation of such features for a large number of points not only takes a fair amount of computation time but also consumes a lot of memory. A minimum of 512 and 256 bytes are reserved for storing a feature point in SIFT and SURF respectively. This is because of using floating point numbers to store the dimension values. As a result an appreciable time is taken to match the feature descriptors due to large number of elements in the descriptor vectors. Since all the elements are not required for matching, methods like PCA or LDA may be used to find the more important dimensions. Local Sensitive Hashing (LSH) may be used to convert the floating point numbers to string of binary values. Hamming distance between the binary strings is used to compute the distance by performing the XOR operation and finding the number of ones in the result. BRIEF provides a shorter way to find the binary strings related to an interest point without finding the descriptor vectors. As BRIEF is a feature descriptor, feature detectors like SIFT, SURF etc. have to be used

to find the interest points. BRIEF is thus a quicker method for computing the feature descriptor and matching the feature vectors. Subject to moderate in-plane rotation, BRIEF provides a high recognition rate.

4.6 Oriented FAST and Rotated BRIEF (ORB)

ORB [95] is a fast and robust feature point detector, proposed by Rublee et al. in 2011. Many tasks in computer vision like object identification, 3D image reconstruction, image similarity analysis etc. can be done using ORB. It is based on the FAST feature point detector and BRIEF (Binary Robust Independent Elementary Features) visual descriptor. It is invariant to rotation and noise resistant. ORB provides a fast and efficient alternative to SIFT and has been shown to be two orders of magnitude faster than SIFT. A method to detect moving objects during camera motion is presented in [96]. To compensate the camera motion, Oriented FAST and Rotated BRIEF (ORB) is used for the feature matching task. The mismatched features between two frames are rejected for obtaining accuracy in compensation. The work also evaluates SIFT and SURF against the presented method to estimate performance in terms of speed and precision.

5 Proposed Methodology

A flow diagram of the proposed method is given in Fig. 6. The various steps of the method are detailed in further sub-sections.

5.1 Extraction of Time Sequenced Image Frames from a Video

The first step towards the video summarization process is to disintegrate the video into a set of time-sequenced image frames to facilitate the process of extracting key-frames from it. This is done by using a standard codec, corresponding to the file type i.e. MP4, MPEG, AVI etc. The images thus obtained are stored as bitmaps for further processing.

5.2 Detection of Video Segments

The transition between two shots is usually classified into two categories i.e. abrupt and gradual. The abrupt transitions are also referred to as hard cuts, whereas, gradual

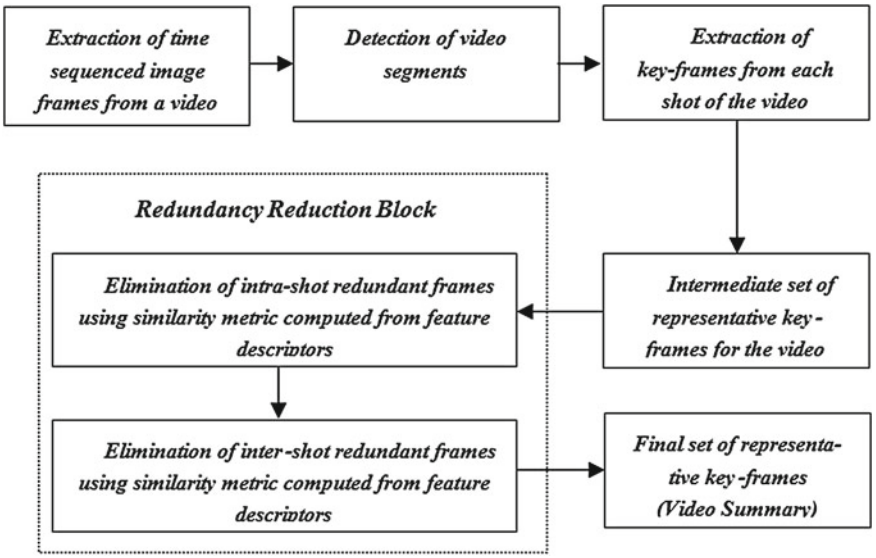

Fig. 6 Flow diagram of proposed method

transitions include dissolves, fade in and fade out. 95 % of these transitions are hard cuts. The decomposed video frames in the previous step are analyzed for detection of shot boundaries. An effective mechanism for video segmentation has been developed by the authors in [97], where a spatio-temporal fuzzy hostility index was used. The same mechanism is employed for detection of shot boundaries in this work.

5.3 Shot-Wise Extraction of Key-Frames and Formation of Representative Set

The key-frames in a video are the representative frames which aptly represents its contents. Given a video, $V = s_1 \bigotimes s_2 \bigotimes s_3 \ldots \bigotimes s_n$ where s_i is a composing shot of V, the task of static video summarization is to assign a Boolean value to the pair (f_{ij}, rs_j) where f_{ij} is the ith frame of the jth shot and rs_j is the representative set of the jth shot. Thus, the initial summary generated after the shot-wise extraction of key-frames is $RS = \{rs_1, rs_2, rs_3, \ldots, rs_n\}$. Initially, the frame having the highest average Fuzzy Hostility Index (FHI) [98] within a shot is chosen as the first key-frame. A search is conducted in both directions of the chosen key-frame such that a frame is reached which has dissimilarity more than $(\mu + 3\sigma)$ where μ is the mean of the average FHIs of the frames in the shot and σ is the standard deviation. The key-frame extraction method has been depicted in Fig. 7. To ensure proper content coverage,

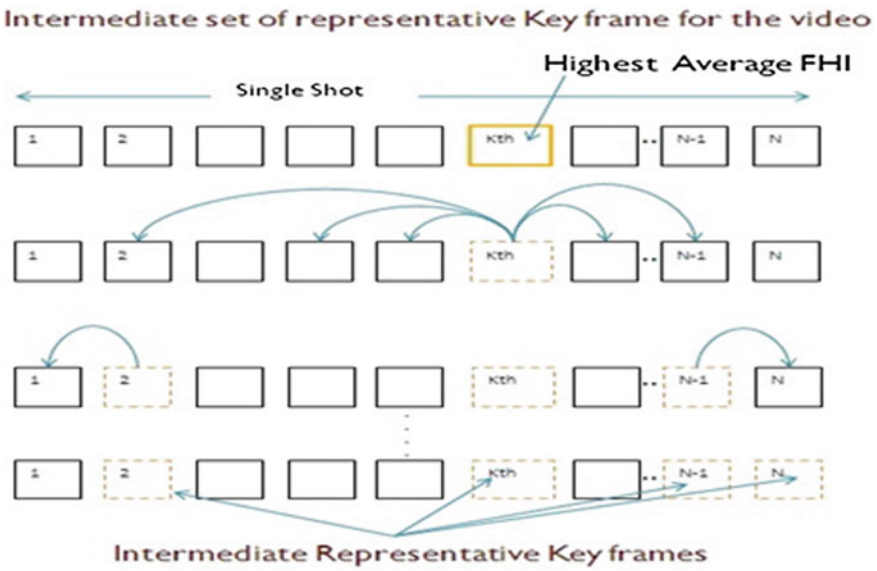

Fig. 7 Key-frame selection process

representative frames are chosen from each shot. The set of key-frames which are extracted from each shot of the video form the representative set.

5.4 Redundancy Reduction to Generate the Final Video Summary

Redundancy of content may occur at the intra-shot and inter-shot levels. Intra-shot content duplication takes place when multiple frames containing the same visual content are chosen as key-frames from within a particular shot. This occurs when there are enough discriminating features between the frames to render a conclusion that the frames are dissimilar in spite of same visual content. It may also occur in cases where the similarity metric or function chosen for the purpose, yields a value below a pre-determined threshold. Inter-shot redundancy occurs when shots with similar content are intermingled with other shots. This leads to similar frames being chosen from multiple shots. The process of intra-shot redundancy reduction on the set can be viewed as a task of eliminating a set of frames $F_i = \{f_1, f_2, f_3, \ldots, f_k\}$ from the representative set rs_i of the ith shot. The same operation is performed on all the shots and the set obtained may be referred to as reduced representative set (*RRS*). Thus, $RRS = \{rs_1 - F_1, rs_2 - F_2, \ldots, rs_n - F_n\}$. The inter-shot redundancy reduction is elimination of a key-frame set $F_R = \{f_1, f_2, f_3, \ldots, f_m\}$ such that the final representative set *FRS* or final summary generated is $FRS = RRS - F_R$. The result

of such elimination process ensures that the similarity (δ) between two elements in *FRS* is less than a pre-determined threshold (γ). Thus, if we consider a set $T = \{y : \delta(x, y) > \gamma, x \in FRS, y \in FRS\}$, then $T = \phi$.

The pre-determined threshold may be computed in accordance with an empirical statute in statistics, called the three-sigma rule. According to this rule (refer Fig. 8), 68.2 % values in a normal distribution lie in the range $[M - \sigma, M + \sigma]$, 95.4 % values in $[M - 2\sigma, M + 2\sigma]$ and 99.6 % in the range $[M - 3\sigma, M + 3\sigma]$, where M denotes the arithmetic mean and σ denotes the standard deviation of the normally distributed values. This rule can be effectively utilized for computing the threshold (γ) used for redundancy elimination. A set of p feature point descriptors are extracted from an image frame I_1. The same set of descriptors are matched in another image frame I_2. Assuming that q out of p descriptors match, the similarity between the two image frames, $\delta(I_1, I_2) = \frac{p}{q}$. It can easily be seen that the extent of similarity between the two images is expressed as a real number in the range [0, 1]. Values closer to 1 denote a high similarity. It may further be noted that since δ is calculated on the basis of feature point descriptors, the metric used is closely related to the visual content of an image rather than other low level descriptors such as color model, histogram, statistical measures on pixel values etc. Therefore, for a shot $S_i = \{I_1, I_2, I_3, \ldots, I_n\}$ the mean and standard deviation of the similarity values is computed as:

$$\mu = \frac{\sum_{i,j=1}^{n} \delta(I_i, I_j)}{\binom{n}{2}} \tag{1}$$

$$\sigma = \sqrt{\frac{\sum_{i,j=1}^{n} (\delta(I_i, I_j) - \mu)^2}{\binom{n}{2}}}, i \neq j \tag{2}$$

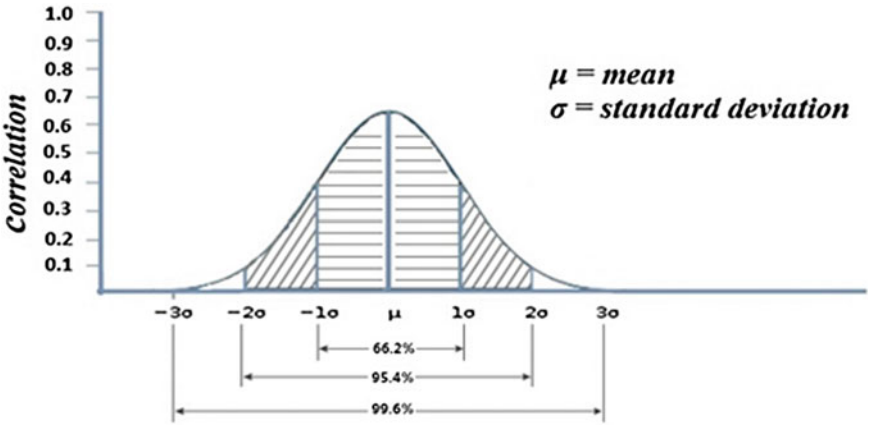

Fig. 8 Normal distribution with three standard deviations from mean

If the similarity value for a pair of frames is greater than $\mu + 3\sigma$, one of the two is eliminated. After intra-shot redundancy is eliminated from the composing shots of the video, the set *RRS* is obtained. Although intra-shot redundancy elimination ensures coverage, it compromises with conciseness of representation. The technique described above may be applied as a whole on the image frames of the set to generate the set after intra-shot redundancy reduction. The proposed method is able to tackle duplication of visual content not only at the intra-shot level but also on the video as a whole. In addition, the user can exercise control over the amount of redundancy by selecting a threshold above or below $\mu + 3\sigma$ which is based on statistical measure.

6 Metrics for Video Summary Evaluation

The evaluation of a video summary is not a simple task due to the unavailability of ground truth for the videos in the dataset under consideration. Moreover, the quality of a summary is based on human perception. It is sometimes difficult for humans to decide as to which summary is the better one. This has rendered difficulties for researchers in designing the different metrics necessary for both evaluation of the summaries and comparison of the different approaches. A brief explanation of the various approaches followed for video summary evaluation is presented in further sub-sections.

6.1 Key-Frame Evaluation

This method was proposed in [99] and focuses on an indirect evaluation of the system generated summary. The key-frames selected by the system are rated on a scale of 5 by independent evaluators [100, 101]. A score of 1 denotes least significance and 5 denote that the chosen key-frame is most significant and relevant for the summary. Appraisal of the video summary is also done by the evaluators to ensure the quality of the summary. The quality of a summary depends on two important factors:

1. Amount of information content (entropy)
2. Coverage of the video content through the key-frames

The mean score of the key-frames is computed to quantify the quality of summary. The formula used in [99] is:

$$score = \frac{sum\ of\ keyframe\ score}{number\ of\ keyframes} \tag{3}$$

6.2 Shot Reconstruction Degree

The extent to which a set of key-frames is able to reconstruct a shot by means of interpolation is called shot reconstruction degree (SRD) [102]. SRD represents the ability of a key-frame set to reconstruct the original contents. Maximizing the SRD ensures that the motion dynamics of the shot content is preserved. The remaining frames of the video are generated from the key-frame set by employing an interpolation function. The summarization capability of the system is judged by the extent to which the original shot is reconstructed. A similarity function is used to compute the distance between the frames of the original video and those generated by interpolating the key-frames. Different schemes involving SRD have been proposed in [103, 104].

6.3 Coverage

The coverage of a set of key-frames extracted from the original video is defined as the number of frames which are represented by the key-frame set. In [42] a Minimal Spanning Tree (MST) is constructed from the frames of a shot. An adaptive threshold is calculated separately for each shot based on the mean and standard deviation of the edge weights of the MST. The density of a node is the number of frames lying within a disc, the radius of which is equal to the computed threshold. A greedy method is used to choose frames from the list with maximum density. Frames represented by the chosen key-frame are eliminated from the list. This ensures that the most appropriate representatives are chosen as key-frames. It can easily be seen that the chosen key-frames provide a full coverage of the shot. In [105] the coverage has been defined as the number of visually similar frames represented by a chosen key-frame. Hence, coverage may be computed by the following formula:

$$coverage = \frac{number\ of\ frames\ represented}{total\ number\ of\ frames} \tag{4}$$

In [67], the coverage is based on the number of feature points covered by a frame from the unique pool of feature points created from the composing frames of a shot. Initially all the feature points are part of the set $K_{uncovered}$. The coverage of a frame is computed using the formula:

$$C = \eta(K_{uncovered} \bigcap FP_i) \tag{5}$$

The redundancy of a frame is given by:

$$R = \eta(K_{covered} \bigcap FP_i) \tag{6}$$

where, $\eta(X)$ is the cardinality of set X and FP_i is the set of feature points in a frame. Coverage is thus another metric which reveals the quality of a video summary.

6.4 Recall, Precision and F_1 Score

The output of a video summarizer is referred to as the System Generated Summary (SGS). It is essential to evaluate the quality of this summary. The appropriate way for appraisal of the SGS is to compare it with a ground truth. Since the SGS is generated for users, it is natural to bring the ground truth as close as possible to human perception. The ground truth has been referred to as User Summary [30, 84] in the literature. The User Summary (US) is generated by a group of users. The videos under consideration are browsed by the users after disintegrating into constituent frames. The important frames according to user perception are chosen in order to form the US. The Final User Summary (FUS) is formed by an amalgamation of the user summaries. The amount of overlap between the FUS and SGS portrays the efficacy of the summary. The recall and precision are computed as follows:

$$recall = \frac{\eta(FUS \bigcap SGS)}{\eta(FUS)} \tag{7}$$

$$precision = \frac{\eta(FUS \bigcap SGS)}{\eta(SGS)} \tag{8}$$

FUS: Set of frames in user summary
SGS: Set of frames in system generated summary
$\eta(X)$: Cardinal no. of set X
The harmonic mean of precision and recall is taken for computing the F_1 score. It provides a consistent measure for determining the overall efficiency of an information retrieval system. The following expression is used to calculate the F_1 score:

$$F_1 = 2\frac{precision \times recall}{precision + recall} \tag{9}$$

The F_1 score varies in the range [0, 1] where a score of 1 indicates that the system is most efficient.

6.5 Significance, Overlap and Compression Factors

Mundur et al. [48] introduces three new factors for determining the quality of a summary. The *Significance Factor* denotes the importance of the content represented by a cluster of frames. The significance of the ith cluster is given as:

$$Significance\_Factor(i) = \frac{C_i}{\sum_{j=1}^{k} C_j} \tag{10}$$

where C_i is the total number of frames in the ith cluster and k is the total number of clusters.

The *Overlap Factor* determines the total significance of the overlapped clusters found in two summaries. In other words, we compute the cumulative significance of those clusters which have a common key-frame set with the ground-truth summary. This is an important metric for comparing two summaries. This factor is computed as:

$$Overlap\_Factor = \frac{\sum_{p \in Common\ keyframe\ clustes}^{C_p}}{\sum_{j=1}^{k} C_j} \tag{11}$$

A higher value of the *Overlap Factor* denotes a better representative summary with respect to the ground-truth.

The *Compression Factor* for a video denotes the size of the summary with respect to the original size of the video. It is defined as:

$$Compression\_Factor = \frac{No\ of\ keyframes\ in\ summary}{Total\ number\ of\ keyframes} \tag{12}$$

7 Experimental Results and Analysis

The proposed method for storyboard generation was tested on a dataset consisting of nine videos. The dataset is divided into two parts. The first part (Table 1) consists of short videos having average length of 3 min and 21 s. The second part (Table 2) consists of longer videos of average length 53 min and 34 s.

All the videos in the dataset have a resolution of 640 × 360 pixels at 25 fps (except video V7 which is at 30 fps). The videos are in MP4 file format (ISO/IEC 14496-14:2003), commonly named as MPEG-4 file format version 2.

The efficacy of the proposed method is evaluated by computing the recall, precision and F_1 score of the system generated summary (*SGS*) against the final user summary (*FUS*) as explained in Sect. 6.4. A frame to frame comparison is performed between the *SGS* and *FUS* by an evaluator program written for the purpose. A pair of frames is considered to be matched if the correlation is more than 0.7. It has been seen that the frames are visually similar when the correlation exceeds 0.7. This is significantly higher than the threshold used in [30], where the match threshold was considered as 0.5.

Table 1 Test video dataset-I

Video	V1	V2	V3	V4	V5
Duration (mm:ss)	02:58	02:42	04:10	03:27	03:31
No. of frames	4468	4057	6265	4965	5053
No. of hard cuts	43	70	172	77	138
Average no. of frames in each shot	101.54	57.14	36.21	63.65	36.35

Table 2 Test video dataset-II

Video	V6	V7	V8	V9
Duration (mm:ss)	44:14	52:29	58:06	59:29
No. of frames	66339	94226	87153	89226
No. of hard cuts	626	543	668	1235
Average no. of frames in each shot	105.80	173.21	130.27	72.18

7.1 The Video Dataset

The video dataset considered for testing comprised of videos of short and long duration. The first video (V1) is the Wimbledon semifinal match highlights between Djokovic and Del Potro. The video consists of small duration shots and rapid movement of objects. The second video (V2) is a Hindi film song "Dagabaaz" from the movie "Dabangg2". It consists of shots taken in the daylight and night time. The third video (V3) is another song "Chammak Challo" from the Hindi film "Ra.One". This video consists of shots taken indoors, as well as some digitally created frames intermingled with real life shots. A violin track by Lindsey Stirling forms the fourth video (V4) of the data set. Simultaneous camera and performer movements are observed in the video. Also there are quick zoom-in and zoom-out shots which are taken outdoors. The last of the small videos (V5) is the official song of the FIFA world cup called "Waka Waka". It consists of shots with varied illumination and background.

The videos in Table 2 are four documentaries (V6–V9) from different TV channels. The videos V6–V9 are four documentaries of longer duration from different TV channels. The videos are "Science and Technology developments in India", "Under the Antarctic Ice", "How to build a satellite" and "Taxi Driver". All the videos in the dataset are available on YouTube.

7.2 Experimental Results

The initial storyboard generated by the proposed method is called the representative set (RS). It is formed by extracting the key-frames as described in Sect. 5.3. The key-frames in RS are compared with user summary prior to redundancy removal and the

results are presented in Table 3. The results show high precision, recall and F_1 scores indicating the efficacy of the proposed system. In the next step, intra-shot redundancy reduction is carried out using both the SURF and GIST feature descriptors on both *RS* and user summary. The amount of reduction achieved is summarized in Table 4. The recall and precision are again computed and the results are presented in Table 5. In the final step, redundancy is further removed from *RS* and user summary at the inter-shot level using SURF and GIST descriptors. The amount reduction achieved is enumerated in Table 6. The recall and precision values computed after the inter-shot redundancy phase are presented in Table 7. It can be easily seen from the results that elimination of duplicate frames does have effect on the precision and recall. In certain cases the post-redundancy metric values are better than the pre-redundancy phase.

Table 3 Comparison between user and system generated summary prior to redundancy removal

Video	Precision (%)	Recall (%)	F_1 Score (%)
V1	98.43	92.64	95.45
V2	90.85	97.54	94.08
V3	99.10	95.67	97.35
V4	94.69	93.98	94.33
V5	96.59	91.89	94.18
V6	99.65	99.55	99.59
V7	98.88	97.12	97.99
V8	98.39	97.18	97.78
V9	98.52	100	99.25

Table 4 Intra-shot redundancy reduction

Video	% reduction (SURF)	% reduction (GIST)
V1	23	28.12
V2	44.57	55.14
V3	21.645	28.76
V4	26.51	42.10
V5	17.83	23.24
V6	53.18	68.68
V7	57.83	74.19
V8	50.36	72.91
V9	40.38	55.24

Table 5 Comparison after intra-shot redundancy removal

Video	Precision (SURF) (%)	Recall (SURF) (%)	Precision (GIST) (%)	Recall (GIST) (%)
V1	98.03	98.03	100	100
V2	97.42	99.47	99.03	100
V3	100	100	100	100
V4	97.93	98.95	98.70	97.43
V5	100	100	100	100
V6	97.45	97.24	98.65	99.20
V7	98.30	98.85	96.45	97.22
V8	98.34	98.85	98.67	98.66
V9	96.56	95.55	98.36	98.42

Table 6 Inter-shot redundancy reduction

Video	% reduction (SURF)	% reduction (GIST)
V1	27	37.5
V2	70.28	78
V3	35.49	39.38
V4	41.66	54.13
V5	35.67	39.45
V6	63.07	73.60
V7	64.94	77.59
V8	48.17	73.11
V9	51.05	68.62

Table 7 Comparison after inter-shot redundancy removal

Video	Precision (SURF) (%)	Recall (SURF) (%)	Precision (GIST) (%)	Recall (GIST) (%)
V1	97.67	95.45	100	100
V2	100	100	100	100
V3	98.65	97.35	100	97.81
V4	96.10	96.10	95.23	96.77
V5	97.39	100	98.19	97.32
V6	98.35	97.63	98.75	100
V7	99.4	98.55	97.65	98.25
V8	97.68	98.44	99.32	99.74
V9	95.36	96.45	99.24	98.86

8 Discussions and Conclusion

Redundancy removal remains an important step in the task of video summarization. The proposed method is able to illustrate that the quality of the generated summary is not degraded by removing duplicate frames having nearly the same visual content. An additional contribution of this work is the determination of an automatic threshold for elimination of redundant frames based on the three-sigma rule. The results illustrate the efficacy of the threshold used. The experimental results leads us to conclude that the prominent features of a video may be represented in a succinct way, thereby saving the retrieval and browsing time of a user. This is particularly useful for low bandwidth scenarios.

Although the problem of removing visual redundancy has been tackled to a great extent by the use of feature descriptor, yet there is a long way to go in terms of semantic understanding of the video. For semantic understanding, development of semantic descriptors need to be designed which in turn require extraction of high level features. These high level features need to be presented in a manner which provides comparison and matching between the high level feature vectors. This would propel research in abstraction based representation of the video contents.

References

1. Zhang, H.J., Wu, J., Zhong, D., Smoliar, S.W.: An integrated system for content-based video retrieval and browsing. Pattern Recogn. **30**(4), 643–658 (1997)
2. Chang, S.F., Chen, W., Meng, H.J., Sundaram, H., Zhong, D.: A fully automated content-based video search engine supporting spatiotemporal queries. IEEE Trans. Circuits Syst. Video Technol. **8**(5), 602–615 (1998)
3. Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: state of the art and challenges. ACM Trans. Multimedia Comput. Commun. Appl. (TOMM) **2**(1), 1–19 (2006)
4. Papadimitriou, C.H., Tamaki, H., Raghavan, P., Vempala, S.: Latent semantic indexing: a probabilistic analysis. In: Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pp. 159–168. ACM (1998)
5. Kim, H.S., Lee, J., Liu, H., Lee, D.: Video linkage: group based copied video detection. In: Proceedings of the 2008 International Conference on Content-Based Image and Video Retrieval, pp. 397–406. ACM (2008)
6. Kim, C., Hwang, J.N.: Object-based video abstraction for video surveillance systems. IEEE Trans. Circuits Syst. Video Technol. **12**(12), 1128–1138 (2002)
7. Ekin, A., Tekalp, A.M., Mehrotra, R.: Automatic soccer video analysis and summarization. IEEE Trans. Image Proc. **12**(7), 796–807 (2003)
8. Babaguchi, N., Kawai, Y., Ogura, T., Kitahashi, T..: Personalized abstraction of broadcasted American football video by highlight selection. IEEE Trans. Multimedia **6**(4), 575–586 (2004)
9. Pan, H., Van Beek, P., Sezan, M.I.: Detection of slow-motion replay segments in sports video for highlights generation. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 3, pp. 1649–1652 (2001)
10. Tjondronegoro, D.W., Chen, Y.P.P., Pham, B.: Classification of self-consumable highlights for soccer video summaries. In: 2004 IEEE International Conference on Multimedia and Expo, 2004. ICME'04, vol. 1, pp. 579–582. IEEE (2004)

11. Nam, J., Tewfik, A.H.: Dynamic video summarization and visualization. In: Proceedings of the Seventh ACM International Conference on Multimedia (Part 2), pp. 53–56. ACM (1999)
12. Pfeiffer, S., Lienhart, R., Fischer, S., Effelsberg, W.: Abstracting digital movies automatically. J. Vis. Commun. Image Represent. **7**(4), 345–353 (1996)
13. Yeung, M.M., Yeo, B.L.: Video visualization for compact presentation and fast browsing of pictorial content. IEEE Trans. Circuits Syst. Video Technol. **7**(5), 771–785 (1997)
14. Moriyama, T., Sakauchi, M.: Video summarisation based on the psychological content in the track structure. In: Proceedings of the 2000 ACM Workshops on Multimedia, pp. 191–194. ACM (2000)
15. Yeung, M.M, Yeo, B.L.: Video content characterization and compaction for digital library applications. In: Electronic Imaging'97, pp. 45–58 (1997)
16. Lienbart, R., Pfeiffer, S., Effelsberg, W.: Scene determination based on video and audio features. In: IEEE International Conference on Multimedia Computing and Systems, 1999, vol. 1, pp. 685–690. IEEE (1999)
17. Thakore, V.H.: Video shot cut boundary detection using histogram. Int. J. Eng. Sci. Res. Technol. (IJESRT) **2**, 872–875 (2013)
18. Baber, J., Afzulpurkar, N., Dailey, M.N., Bakhtyar, M.: Shot boundary detection from videos using entropy and local descriptor. In: 2011 17th International Conference on Digital Signal Processing (DSP), pp. 1–6. IEEE (2011)
19. Cernekova, Z., Nikou, C., Pitas, I.: Shot detection in video sequences using entropy based metrics. In: 2002 International Conference on Image Processing. 2002. Proceedings, vol. 3, p. III-421. IEEE (2002)
20. Hampapur, A., Jain, R., Weymouth, T.E.: Production model based digital video segmentation. Multimedia Tools Appl. **1**(1), 9–46 (1995)
21. Zhang, H., Kankanhalli, A., Smoliar, S.W.: Automatic partitioning of full-motion video. Multimedia Syst. **1**(1), 10–28 (1993)
22. Tonomura, Y.: Video handling based on structured information for hypermedia systems. In: International conference on Multimedia Information Systems' 91, pp. 333–344. McGraw-Hill Inc. (1991)
23. Barron, J.L., Fleet, D.J., Beauchemin, S.S.: Performance of optical flow techniques. Int. J. Comput. Vis. **12**(1), 43–77 (1994)
24. Wang, T., Wu, Y., Chen, L.: An approach to video key-frame extraction based on rough set. In: International Conference on Multimedia and Ubiquitous Engineering, 2007. MUE'07, pp. 590–596. IEEE (2007)
25. Li, B., Sezan, M.I.: Event detection and summarization in sports video. In: IEEE Workshop on Content-Based Access of Image and Video Libraries, 2001. (CBAIVL 2001), pp. 132–138. IEEE (2001)
26. Potapov, D., Douze, M., Harchaoui, Z., Schmid, C.: Category-specific video summarization. In: Computer Vision-ECCV 2014, pp. 540–555. Springer (2014)
27. Wang, F., Ngo, C.W.: Rushes video summarization by object and event understanding. In: Proceedings of the International Workshop on TRECVID Video Summarization, pp. 25–29. ACM (2007)
28. Guan, G., Wang, Z., Lu, S., Deng, J.D., Feng, D.D.: Keypoint-based keyframe selection. IEEE Trans. Circuits Syst. Video Technol. **23**(4), 729–734 (2013)
29. Panagiotakis, C., Pelekis, N., Kopanakis, I., Ramasso, E., Theodoridis, Y.: Segmentation and sampling of moving object trajectories based on representativeness. IEEE Trans. Knowl. Data Eng. **24**(7), 1328–1343 (2012)
30. Cahuina, E.J., Chavez, C.G.: A new method for static video summarization using local descriptors and video temporal segmentation. In: 26th SIBGRAPI-Conference on Graphics, Patterns and Images (SIBGRAPI), 2013, pp. 226–233. IEEE (2013)
31. Patel, A., Kasat, D., Jain, S., Thakare, V.: Performance analysis of various feature detector and descriptor for real-time video based face tracking. Int. J. Comp. Appl. **93**(1), 37–41 (2014)
32. Kapela, R., McGuinness, K., Swietlicka, A., Oconnor, N.E.: Real-time event detection in field sport videos. In: Computer Vision in Sports, pp. 293–316. Springer (2014)

33. Khvedchenia, I.: A battle of three descriptors: surf, freak and brisk. Computer Vision Talks
34. Uijlings, J.R., Smeulders, A.W., Scha, R.J.: Real-time visual concept classification. IEEE Trans. Multimedia **12**(7), 665–681 (2010)
35. Li, J.: Video shot segmentation and key frame extraction based on sift feature. In: 2012 International Conference on Image Analysis and Signal Processing (IASP), pp. 1–8. IEEE (2012)
36. Papadopoulos, D.P., Chatzichristofis, S.A., Papamarkos, N.: Video summarization using a self-growing and self-organized neural gas network. In: Computer Vision/Computer Graphics Collaboration Techniques, pp. 216–226. Springer (2011)
37. Lux, M., Schöffmann, K., Marques, O., Böszörmenyi, L.: A novel tool for quick video summarization using keyframe extraction techniques. In: Proceedings of the 9th Workshop on Multimedia Metadata (WMM 2009). CEUR Workshop Proceedings, vol. 441, pp. 19–20 (2009)
38. Chatzichristofis, S.A., Boutalis, Y.S.: Cedd: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval. In: Computer Vision Systems, pp. 312–322. Springer (2008)
39. Chatzichristofis, S., Boutalis, Y.S., et al.: Fcth: Fuzzy color and texture histogram-a low level feature for accurate image retrieval. In: Ninth International Workshop on Image Analysis for Multimedia Interactive Services, 2008. WIAMIS'08, pp. 191–196. IEEE (2008)
40. Chatzichristofis, S.A., Boutalis, Y.S.: Content based radiology image retrieval using a fuzzy rule based scalable composite descriptor. Multimedia Tools Appl. **46**(2–3), 493–519 (2010)
41. Chatzichristofis, S.A., Boutalis, Y.S., Lux, M.: Spcd-spatial color distribution descriptor-a fuzzy rule based compact composite descriptor appropriate for hand drawn color sketches retrieval. In: ICAART (1), pp. 58–63 (2010)
42. Bhaumik, H., Bhattacharyya, S., Das, M., Chakraborty, S.: Enhancement of Perceptual Quality in Static Video Summarization Using Minimal Spanning Tree Approach. In: 2015 International Conference on Signal Processing, Informatics, Communication and Energy Systems (IEEE SPICES), pp. 1–7. IEEE (2015)
43. Liu, D., Shyu, M.L., Chen, C., Chen, S.C.: Integration of global and local information in videos for key frame extraction. In: 2010 IEEE International Conference on Information Reuse and Integration (IRI), pp. 171–176. IEEE (2010)
44. Qian, Y., Kyan, M.: Interactive user oriented visual attention based video summarization and exploration framework. In: 2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE), pp. 1–5. IEEE (2014)
45. Qian, Y., Kyan, M.: High definition visual attention based video summarization. In: VISAPP, vol. 1, pp. 634–640 (2014)
46. Zhuang, Y., Rui, Y., Huang, T.S., Mehrotra, S.: Adaptive key frame extraction using unsupervised clustering. In: 1998 International Conference on Image Processing, 1998. ICIP 98. Proceedings, vol. 1, pp. 866–870. IEEE (1998)
47. Gong, Y., Liu, X.: Video summarization and retrieval using singular value decomposition. Multimedia Syst. **9**(2), 157–168 (2003)
48. Mundur, P., Rao, Y., Yesha, Y.: Keyframe-based video summarization using delaunay clustering. Int. J. Digit. Libr. **6**(2), 219–232 (2006)
49. Wan, T., Qin, Z.: A new technique for summarizing video sequences through histogram evolution. In: 2010 International Conference on Signal Processing and Communications (SPCOM), pp. 1–5. IEEE (2010)
50. Cayllahua-Cahuina, E., Cámara-Chávez, G., Menotti, D.: A static video summarization approach with automatic shot detection using color histograms. In: Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV), p. 1. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp) (2012)
51. Furini, M., Geraci, F., Montangero, M., Pellegrini, M.: Stimo: still and moving video storyboard for the web scenario. Multimedia Tools Appl. **46**(1), 47–69 (2010)

52. de Avila, S.E.F., Lopes, A.P.B., da Luz, A., de Albuquerque Araújo, A.: Vsumm: a mechanism designed to produce static video summaries and a novel evaluation method. Pattern Recogn. Lett. **32**(1), 56–68 (2011)
53. Xie, X.N., Wu, F.: Automatic video summarization by affinity propagation clustering and semantic content mining. In: 2008 International Symposium on Electronic Commerce and Security, pp. 203–208. IEEE (2008)
54. Liu, Z., Zavesky, E., Shahraray, B., Gibbon, D., Basso, A.: Brief and high-interest video summary generation: evaluating the at&t labs rushes summarizations. In: Proceedings of the 2nd ACM TRECVid Video Summarization Workshop, pp. 21–25. ACM (2008)
55. Ren, J., Jiang, J., Eckes, C.: Hierarchical modeling and adaptive clustering for real-time summarization of rush videos in trecvid'08. In: Proceedings of the 2nd ACM TRECVid Video Summarization Workshop, pp. 26–30. ACM (2008)
56. Dumont, E., Merialdo, B.: Rushes video summarization and evaluation. Multimedia Tools Appl. **48**(1), 51–68 (2010)
57. Gao, Y., Dai, Q.H.: Clip based video summarization and ranking. In: Proceedings of the 2008 International Conference on Content-Based Image and Video Retrieval, pp. 135–140. ACM (2008)
58. Tian, Z., Xue, J., Lan, X., Li, C., Zheng, N.: Key object-based static video summarization. In: Proceedings of the 19th ACM International Conference on Multimedia, pp. 1301–1304. ACM (2011)
59. Liu, X., Song, M., Zhang, L., Wang, S., Bu, J., Chen, C., Tao, D.: Joint shot boundary detection and key frame extraction. In: 2012 21st International Conference on Pattern Recognition (ICPR), pp. 2565–2568. IEEE (2012)
60. Casella, G., George, E.I.: Explaining the gibbs sampler. Am. Stat. **46**(3), 167–174 (1992)
61. Yilmaz, A., Javed, O., Shah, M.: Object tracking: a survey. ACM Comput. Surv. (CSUR) **38**(4), 13 (2006)
62. Aggarwal, A., Biswas, S., Singh, S., Sural, S., Majumdar, A.K.: Object tracking using background subtraction and motion estimation in mpeg videos. In: Computer Vision-ACCV 2006, pp. 121–130. Springer (2006)
63. Pritch, Y., Ratovitch, S., Hende, A., Peleg, S.: Clustered synopsis of surveillance video. In: Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, 2009. AVSS'09, pp. 195–200. IEEE (2009)
64. Kokare, M., Chatterji, B., Biswas, P.: Comparison of similarity metrics for texture image retrieval. In: TENCON 2003. Conference on Convergent Technologies for the Asia-Pacific Region, vol. 2, pp. 571–575. IEEE (2003)
65. Liu, Y., Zhang, D., Lu, G., Ma, W.Y.: A survey of content-based image retrieval with high-level semantics. Pattern Recogn. **40**(1), 262–282 (2007)
66. Belongie, S., Carson, C., Greenspan, H., Malik, J.: Color-and texture-based image segmentation using em and its application to content-based image retrieval. In: Sixth International Conference on Computer Vision, 1998, pp. 675–682. IEEE (1998)
67. Szeliski, R.: Foundations and trends in computer graphics and vision. Found. Trends Comput. Graphics Vis. **2**(1), 1–104 (2007)
68. Marzotto, R., Fusiello, A., Murino, V.: High resolution video mosaicing with global alignment. In: Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer
69. Matsushita, Y., Ofek, E., Ge, W., Tang, X., Shum, H.Y.: Full-frame video stabilization with motion inpainting. IEEE Trans. Pattern Anal. Mach. Intell. **28**(7), 1150–1163 (2006)
70. Zitova, B., Flusser, J.: Image registration methods: a survey. Image Vis. Comput. **21**(11), 977–1000 (2003)
71. Hu, W., Xie, N., Li, L., Zeng, X., Maybank, S.: A survey on visual content-based video indexing and retrieval. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **41**(6), 797–819 (2011)
72. Lee, J.H., Kim, W.Y.: Video summarization and retrieval system using face recognition and mpeg-7 descriptors. In: Image and Video Retrieval, pp. 170–178. Springer (2004)

73. Fatemi, N., Khaled, O.A.: Indexing and retrieval of tv news programs based on mpeg-7. In: International Conference on Consumer Electronics, 2001. ICCE, pp. 360–361. IEEE (2001)
74. Lowe, D.G.: Object recognition from local scale-invariant features. In: The Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999, vol. 2, pp. 1150–1157. IEEE (1999)
75. Ke, Y., Sukthankar, R.: Pca-sift: a more distinctive representation for local image descriptors. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004, vol. 2, p. II-506. IEEE (2004)
76. Azad, P., Asfour, T., Dillmann, R.: Combining harris interest points and the sift descriptor for fast scale-invariant object recognition. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, 2009. IROS 2009, pp. 4275–4280. IEEE (2009)
77. Khosla, A., Hamid, R., Lin, C.J., Sundaresan, N.: Large-scale video summarization using web-image priors. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2698–2705. IEEE (2013)
78. Liu, C., Yuen, J., Torralba, A.: Sift flow: dense correspondence across scenes and its applications. IEEE Trans. Pattern Anal. Mach. Intell. **33**(5), 978–994 (2011)
79. Pandey, R.C., Singh, S.K., Shukla, K., Agrawal, R.: Fast and robust passive copy-move forgery detection using surf and sift image features. In: 2014 9th International Conference on Industrial and Information Systems (ICIIS), pp. 1–6. IEEE (2014)
80. Yuan, Z., Lu, T., Wu, D., Huang, Y., Yu, H.: Video summarization with semantic concept preservation. In: Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia, pp. 109–112. ACM (2011)
81. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Comput. Vis. Image Underst. **110**(3), 346–359 (2008)
82. Struzik, Z.R., Siebes, A.: The haar wavelet transform in the time series similarity paradigm. In: Principles of Data Mining and Knowledge Discovery, pp. 12–22. Springer (1999)
83. Sathyadevan, S., Balakrishnan, A.K., Arya, S., Athira Raghunath, S.: Identifying moving bodies from cctv videos using machine learning techniques. In: 2014 First International Conference on Networks & Soft Computing (ICNSC), pp. 151–157. IEEE (2014)
84. Bhaumik, H., Bhattacharyya, S., Dutta, S., Chakraborty, S.: Towards redundancy reduction in storyboard representation for static video summarization. In: 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 344–350. IEEE (2014)
85. Tola, E., Lepetit, V., Fua, P.: Daisy: an efficient dense descriptor applied to wide-baseline stereo. IEEE Trans. Pattern Anal. Mach. Intell. **32**(5), 815–830 (2010)
86. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. Pattern Anal. Mach. Intell. **27**(10), 1615–1630 (2005)
87. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. Int. J. Comput. Vis. **42**(3), 145–175 (2001)
88. Pass, G., Zabih, R.: Histogram refinement for content-based image retrieval. In: Proceedings 3rd IEEE Workshop on Applications of Computer Vision, 1996. WACV'96., pp. 96–102. IEEE (1996)
89. Li, X., Wu, C., Zach, C., Lazebnik, S., Frahm, J.M.: Modeling and recognition of landmark image collections using iconic scene graphs. In: Computer Vision-ECCV 2008, pp. 427–440. Springer (2008)
90. Sikirić, I., Brkić, K., Šegvić, S.: Classifying traffic scenes using the gist image descriptor (2013). arXiv preprint arXiv:1310.0316
91. Hays, J., Efros, A.A.: Scene completion using millions of photographs. ACM Trans. Graphics (TOG) **26**(3), 4 (2007)
92. Torralba, A., Fergus, R., Weiss, Y.: Small codes and large image databases for recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008, pp. 1–8. IEEE (2008)
93. Weiss, Y., Torralba, A., Fergus, R.: Spectral hashing. In: Advances in neural information processing systems, pp. 1753–1760 (2009)

94. Calonder, M., Lepetit, V., Strecha, C., Fua, P.: Brief: binary robust independent elementary features. Comput. Vis.-ECCV **2010**, 778–792 (2010)
95. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: an efficient alternative to sift or surf. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 2564–2571. IEEE (2011)
96. Xie, S., Zhang, W., Ying, W., Zakim, K.: Fast detecting moving objects in moving background using orb feature matching. In: 2013 Fourth International Conference on Intelligent Control and Information Processing (ICICIP), pp. 304–309. IEEE (2013)
97. Bhaumik, H., Bhattacharyya, S., Chakraborty, S.: Video shot segmentation using spatio-temporal fuzzy hostility index and automatic threshold. In: 2014 Fourth International Conference on Communication Systems and Network Technologies (CSNT), pp. 501–506. IEEE (2014)
98. Bhattacharyya, S., Maulik, U., Dutta, P.: High-speed target tracking by fuzzy hostility-induced segmentation of optical flow field. Appl. Soft Comput. **9**(1), 126–134 (2009)
99. De Avila, S.E., da Luz, A., de Araujo, A., Cord, M.: Vsumm: an approach for automatic video summarization and quantitative evaluation. In: XXI Brazilian Symposium on Computer Graphics and Image Processing, 2008. SIBGRAPI'08, pp. 103–110. IEEE (2008)
100. De Avila, S.E., da Luz Jr, A., De Araujo, A., et al.: Vsumm: A simple and efficient approach for automatic video summarization. In: 15th International Conference on Systems, Signals and Image Processing, 2008. IWSSIP 2008, pp. 449–452. IEEE (2008)
101. Liu, X., Mei, T., Hua, X.S., Yang, B., Zhou, H.Q.: Video collage. In: Proceedings of the 15th international conference on Multimedia, pp. 461–462. ACM (2007)
102. Liu, T., Zhang, X., Feng, J., Lo, K.T.: Shot reconstruction degree: a novel criterion for key frame selection. Pattern Recogn. Lett. **25**(12), 1451–1457 (2004)
103. Lee, H.C., Kim, S.D.: Iterative key frame selection in the rate-constraint environment. Sign. Process. Image Commun. **18**(1), 1–15 (2003)
104. Liu, R., Kender, J.R.: An efficient error-minimizing algorithm for variable-rate temporal video sampling. In: 2002 IEEE International Conference on Multimedia and Expo, 2002. ICME'02. Proceedings, vol. 1, pp. 413–416. IEEE (2002)
105. Chang, H.S., Sull, S., Lee, S.U.: Efficient video indexing scheme for content-based retrieval. IEEE Trans. Circuits Syst. Video Technol. **9**(8), 1269–1279 (1999)