Pattern Recognition Letters 32 (2011) 56–68

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method

Sandra Eliza Fontes de Avila [a,*], Ana Paula Brandão Lopes [a,b], Antonio da Luz Jr. [a,c], Arnaldo de Albuquerque Araújo [a]

[a] Computer Science Department, Federal University of Minas Gerais, Av. Antônio Carlos, 6627, Pampulha 31270–901, Belo Horizonte, MG, Brazil
[b] Exact and Technological Sciences Department, State University of Santa Cruz, Rodovia Ilhéus-Itabuna, KM 16, Pavilhão Jorge Amado 45600-000, Ilhéus, BA, Brazil
[c] Federal Technical School of Palmas, Science and Technology of Tocantins, Setor Agroindustrial 77600–000, Paraíso, TO, Brazil

ARTICLE INFO

Article history:
Available online 24 August 2010

Keywords:
Video summarization
Static video summary
Evaluation method
Clustering

ABSTRACT

The fast evolution of digital video has brought many new multimedia applications and, as a consequence, has increased the amount of research into new technologies that aim at improving the effectiveness and efficiency of video acquisition, archiving, cataloging and indexing, as well as increasing the usability of stored videos. Among possible research areas, *video summarization* is an important topic that potentially enables faster browsing of large video collections and also more efficient content indexing and access. Essentially, this research area consists of automatically generating a short summary of a video, which can either be a *static summary* or a *dynamic summary*. In this paper, we present VSUMM, a methodology for the production of static video summaries. The method is based on color feature extraction from video frames and *k*-means clustering algorithm. As an additional contribution, we also develop a novel approach for the evaluation of video static summaries. In this evaluation methodology, video summaries are manually created by users. Then, several user-created summaries are compared both to our approach and also to a number of different techniques in the literature. Experimental results show – with a confidence level of 98% – that the proposed solution provided static video summaries with superior quality relative to the approaches to which it was compared.

1. Introduction

The recent advances in compression techniques, the decreasing cost of storage and the availability of high-speed connections have facilitated the creation, storage and distribution of videos. This leads to an increase in the amount of video data deployed and used in applications such as search engines and digital libraries, for example. Such situation puts not only multimedia data into evidence, but also leads to the requirement of efficient management of video data. Those requirements paved the way for new research areas, such as *video summarization*.

Generally, a *video summary* is defined as a sequence of still or moving pictures (with or without audio) presenting the content of a video in such away that the respective target group is rapidly provided with concise information about the content, while the essential message of the original video is preserved (Pfeiffer et al., 1996).

According to Truong and Venkatesh (2007), there are two fundamental types of video summaries: *static video summary* – also called *representative frames*, *still-image abstracts* or *static storyboard* – and *dynamic video skimming* – also called *video skim*, *moving-image abstract* or *moving storyboard*. Static video summaries are composed of a set of keyframes[1] extracted from the original video, while dynamic video summaries are composed of a set of shots[2] and are produced taking into account the similarity or domain-specific relationships among all video shots.

One advantage of a video skim over a keyframe set is the ability to include audio and motion elements that potentially enhance both the expressiveness and the amount of information conveyed by the summary. In addition, according to Li et al. (2001), it is often more entertaining and interesting to watch a skim than a slide show of keyframes. On the other hand, keyframe sets are not restricted by any timing or synchronization issues and, therefore, they offer much more flexibility in terms of organization for browsing and navigation purposes, in comparison to strict sequential

* Corresponding author. Tel.: +55 31 34095854; fax: +55 31 34095858.
E-mail addresses: sandra@dcc.ufmg.br (S.E.F. de Avila), paula@dcc.ufmg.br (A.P.B. Lopes), daluz@dcc.ufmg.br (A. da Luz Jr.), arnaldo@dcc.ufmg.br (A. de Albuquerque Araújo).

[1] A *keyframe* is a frame that represents the content of a logical unit, like a shot or scene, for example. This content must be the most representative as possible.

[2] A *shot* represents a spatio-temporally coherent frame sequence, which captures a continuous action from a single camera.

0167-8655/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2010.08.004

display of video skims, as demonstrated in (Yeung and Leo, 1997; Uchihashi et al., 1999; Ćalić et al., 2007; Wang et al., 2007). In this paper, we focus on the production of static video summaries.

Recently, video summarization has attracted considerable interest from researchers and as a result, various algorithms and techniques have been proposed in the literature, most of them based on clustering techniques (Hadi et al., 2006; Mundur et al., 2006; Chen et al., 2009; Herranz and Martinez, 2009; Furini et al., 2010). Comprehensive surveys of past video summarization results can be found in (Li et al., 2006; Truong and Venkatesh, 2007; Money and Agius, 2008).

In the case of clustering-based techniques, the basic idea is to produce the summary by clustering together similar frames/shots and then showing a limited number of frames per cluster (usually, one frame per cluster). For such approaches, it is important to select the features upon which the frames can be considered similar (e.g., color distribution, luminance, motion vector). Additionally, it is needed also to establish different criteria that can be employed to measure the similarity.

Although there are some techniques that produce summaries of acceptable quality, they typically intricate clustering algorithms that make the summarization process computationally expensive (Furini et al., 2010). For example, in (Mundur et al., 2006), the computation of the summaries takes around 10 times the video length. This means that a potential user would wait around 20 min to have a concise representation of a video that he/she could have watched in just two minutes.

In this paper, it is proposed a simple and effective approach for automatic video summarization, called *Video SUMMarization* (VSUMM). The method is based on the extraction of color features from video frames and unsupervised classification. In addition, a new subjective methodology to evaluate video summaries is developed, called *Comparison of User Summaries* (*CUS*). In this methodology, the video summaries are created by users and are compared with approaches found in the literature. The evaluation of VSUMM is performed both on videos from the Open Video Project[3] (OV) and also on videos from web sites (cartoons, news, sports, commercials, tv-shows and home videos). Experimental results show that the VSUMM approach produces video summaries with superior quality relative to the approaches to which it was compared.

The main contributions of this paper are (1) a mechanism designed to produce static video summaries, which presents the advantages of the main concepts of related work in the video summarization; (2) a new evaluation method of video summaries, which reduces the subjectivity in the evaluation task, quantifies the summary quality and allows more objective comparisons among different techniques; and (3) a statistically well-founded experimental evaluation of both the proposed summarization technique – contrasted to others in the literature – and the evaluation method.

This paper is organized as follows: in Section 2, some related works are described; our approach is presented in Section 3; the experimental results are discussed in Section 4; finally, some concluding remarks and future lines of investigation are derived in Section 5.

2. Related works

Some of the main approaches related to static summarization which can be found in the literature are discussed next.

Zhuang et al. (1998) proposed a method for keyframe extraction based on unsupervised clustering. In that work, the video is segmented into shots and then a color histogram (in the HSV color space) is computed from every frame. The clustering algorithm uses a threshold δ which controls the clustering density. Before a new frame is classified as pertaining to a certain cluster, the similarity between this node and the centroid of the cluster is computed first. If this value is less than δ, it means that this node is not close enough to be added into the cluster. The keyframe selection is employed only to the clusters which are big enough to be considered as keyclusters. In such case, a representative frame is extracted from this cluster as the keyframe. A keycluster is considered large enough if it is larger than the average cluster size. For each keycluster, the frame which is closest to the keycluster centroid is selected as the keyframe. According to Zhuang et al. (1998), the proposed technique is efficient and effective, however, no comparative evaluation is performed for validating such assertions.

Hanjalic and Zhang (1999) presented a method for producing a summary of an arbitrary video sequence. The method is based on cluster-validity analysis and is designed to work without any human supervision. The entire video material is first grouped into clusters. Each frame is represented by color histograms in the YUV color space. A partitional clustering is applied n times to all frames of a video sequence. The prespecified number of clusters starts at one and is increased by one each time the clustering is applied. Next, the system automatically finds the optimal combination(s) of clusters by applying the cluster-validity analysis. After the optimal number of clusters is found, each cluster is represented by one characteristic frame, which then becomes a new keyframe for that video sequence. Hanjalic and Zhang (1999) concentrated on the evaluation of the proposed procedure for cluster-validity analysis, instead of on evaluating the produced summaries.

Gong and Liu (2000) proposed a technique for video summarization based on Singular Value Decomposition (SVD). At first, a set of frames in the input video is selected (one from every ten frames) and then, color histograms in the RGB color space are used to represent video frames. To incorporate spatial information, each frame is divided into 3×3 blocks, and a 3D-histogram is created for each of the blocks. These nine histograms are then concatenated together to form a feature vector. Using this feature vector extracted from the frames, a feature-frame matrix A (usually sparse) is created for the video sequence. Therefore, SVD is performed on A to obtain the matrix V, in which each column vector represents one frame in the refined feature space. Next, the cluster closest to the origin of the refined feature space is found, the content value of this cluster is computed and this value is used as the threshold for clustering the remaining frames. From each cluster, the system selects the frame that is closest to the cluster center as keyframe. This method is not compared with other techniques.

Mundur et al. (2006) developed a method based on Delaunay Triangulation (DT), which is applied for clustering the video frames. The method starts by pre-sampling the frames of the original video. Each frame is represented by a color histogram in the HSV color space. This histogram is represented as a row vector and the vectors for each frame are concatenated into a matrix. To reduce the dimensions of this matrix, Principal Components Analysis (PCA) is applied. After that, the Delaunay diagram is built. The clusters are obtained by separating edges in the Delaunay diagram. Finally, for each cluster, the frame that is nearest to its center is selected as the keyframe. To evaluate the summaries, Mundur et al. (2006) defined three objective metrics: significance factor, overlap factor and compression factor. In spite of the fact that the proposed method has been designed to be fully automatic (i.e., with no user-specified parameters and well suited for batch processing), it requires between 9 and 10 times the video length to produce the summary. Furthermore, the method does not preserve the video temporal order.

[3] http://www.open-video.org.

Furini et al. (2010) introduced STIMO (STIll and MOving Video Storyboard), a summarization technique designed to produce on-the-fly video storyboards. STIMO is composed of three phases. First, the video is analyzed in order to extract the HSV color description. For each input frame, a 256-dimensional vector is extracted. Those vectors are then stored in a matrix and then, in the second phase, the clustering algorithm is applied to extracted data. The authors exploited the triangular inequality in order to filter out useless distance computations. To obtain the number of clusters, the pairwise distance of consecutive frames is computed. If the distance is greater than the threshold Γ, the number of clusters is incremented. The third and last phase aims at removing meaningless video frames from the produced summary. STIMO is evaluated through a comparison study with other approaches: the DT technique (Mundur et al., 2006) and the Open Video storyboards. Furini et al. (2010) asked a group of 20 people to evaluate the produced summaries, using the following procedure: the video is presented to the user, and just after that, the corresponding summary is also shown. The users are asked whether the summary is a good representation of the original video. The quality of the video summary is scored on a scale going from 1 (bad) to 5 (excellent), and the mean opinion score is considered as an indication of the summary quality.

Guironnet et al. (2007) proposed a method for video summarization based on camera motion. It consists in selecting frames according to the succession and the magnitude of camera motions. The method is based on rules to avoid temporal redundancy among the selected frames. The authors developed a subjective method to evaluate the proposed summary. In their experiments, 12 subjects are asked to watch a video and to create a summary manually. From the summaries of different subjects, an "optimal" one is built automatically. This "optimal" summary is then compared with the summaries obtained by different methods. The construction of an "optimal" summary is a complex stage, which requires various parameters to be fixed.

According to the analysis of the approaches found in literature, it can be noticed that the keyframe selection techniques can use several visual features and statistics, which affect both the computational complexity and the summary quality. Normally, the extraction of the video features may produce a high dimensional matrix. For this reason, dimensionality reduction techniques are used and this additional step requires even more processing time, as it can be seen in (Gong and Liu, 2000; Mundur et al., 2006), for example. Another issue that can be observed is the lack of trustworthy comparisons among existing techniques. In other words, a consistent evaluation framework is seriously missing in video summarization research.

The VSUMM approach, proposed in present work, draws on the advantages of the existing techniques and concepts presented in related works. A fully reproducible evaluation framework is proposed and applied for comparisons among VSUMM and three other proposals, indicating that VSUMM is able to provide better summaries, according to the defined metrics. In addition, a new collection of videos is created and evaluated, indicating the consistency of results across datasets with different characteristics.

3. VSUMM approach

Fig. 1 illustrates the steps of our method to produce static video summaries. Initially, the original video is split into frames (step 1). In next step (step 2), color features are extracted to form a color histogram in HSV color space. VSUMM does not consider all the video frames, but takes a sample instead. In addition, the meaningless frames found in the sample are removed. After that (step 3), the frames are grouped by *k*-means clustering algorithm. Then (step 4), one frame per cluster is selected (this selected frame is the keyframe). To refine the static video summary composed by the keyframes (step 5), the keyframes that are too similar are eliminated. Finally, the remaining keyframes are arranged in the original temporal order to facilitate the visual comprehension of the result. Each step is detailed in next subsections.

3.1. Video frames pre-sampling

Temporal video segmentation is the first step towards automatic video summarization. Its goal is to split the video stream into a set of meaningful and manageable basic elements (e.g., shots, frames) (Koprinska and Carrato, 2001). In literature, the *shot*

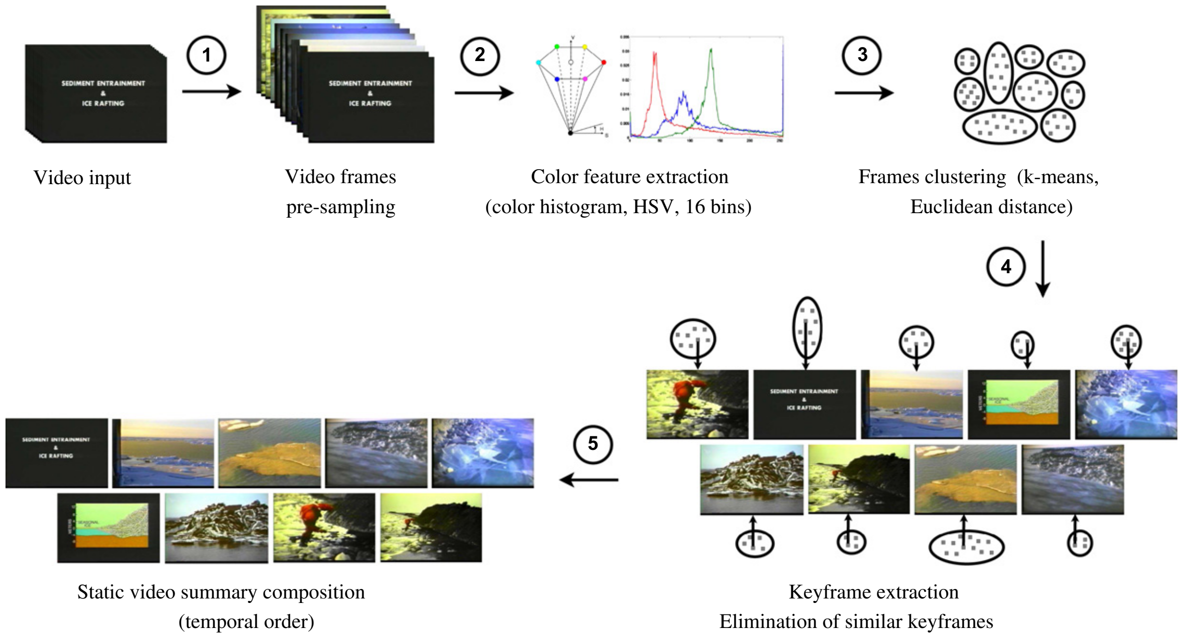

SEDIMENT ENTRAINMENT & ICE RAFTING
1
2
3
4
5
Video input
Video frames pre-sampling
Color feature extraction (color histogram, HSV, 16 bins)
Frames clustering (k-means, Euclidean distance)
Keyframe extraction
Elimination of similar keyframes
Static video summary composition (temporal order)

Fig. 1. VSUMM approach.

boundary detection (Cotsaces et al., 2006) is widely used as first step to produce summaries.

Most of the approaches (Zhu et al., 2004; Li et al., 2005; Cernekova et al., 2006; Hadi et al., 2006; Chang and Chen, 2007) rely in a way or another on detecting shot changes, and are therefore, dependent on having the shot detection correctly done. Detecting shot changes automatically is still a difficult problem, primarily due to the variety of transitions that can be used between shots (Lienhart, 1999).

Another type of video segmentation is the *extraction of video frames*, where there is no temporal analysis of the video. Each frame is treated separately, the video sequence is split into images. Several authors have used this approach (Gong and Liu, 2000; Yahiaoui et al., 2001; Mundur et al., 2006; Wang and Merialdo, 2009; Furini et al., 2010), and it is also used in this work. Moreover, VSUMM does not consider all the video frames, but takes only a subset taken at a predetermined sampling rate. In other words, the VSUMM uses the so-called *pre-sampling* approach.

By using a sampling rate, the number of video frames to be analyzed is reduced. The sampling rate assumes a fundamental importance, since the larger the sampling rate, the shorter the video summarization time. Nevertheless, very low sampling rates can lead to poor quality summaries, which could miss important information contained in the video.

Videos that have long shots tend to present an advantage with the pre-sampling approach, on the other hand, in those videos that present shorter shots, important parts of its content may not be represented. The relationship between the *loss of information* and the *shot size* is directly associated with the sample rates selected during the summarization process.

In VSUMM, the sampling rate is fixed on one frame per second, i.e., the number of frames to be extracted is given by the duration of each video in seconds. For example, for a two-minute-long video with a frame rate of 30 frames/s (i.e., 3600 frames), the total number of frames to be extracted is given by 120 (3600/30) frames.

3.2. Color feature extraction

Color is perhaps the most expressive of all the visual features (Trémeau et al., 2008). In VSUMM, color histogram (Swain and Ballard, 1991) is applied to describe the visual content of video frames. This technique is computationally trivial and is also robust to small changes of the camera position. Furthermore, color histograms tend to be unique for distinct objects. For these reasons, this technique is widely used in automatic video summarization (Zhuang et al., 1998; Hanjalic and Zhang, 1999; Gong and Liu, 2000; Cheung and Zakhor, 2003; Mundur et al., 2006; Furini et al., 2010).

Some key issues of histogram-based techniques are the selection of an appropriate color space and the quantization of that color space. In VSUMM, the color histogram algorithm is applied to the HSV color space, which is a popular choice for manipulating color. The HSV color space was developed to provide an intuitive representation of color and to be near to the way in which humans perceive and manipulate color. The VSUMM color histogram is computed only from the Hue component, which represents the dominant spectral component color in its pure form (Manjunath et al., 2001). Moreover, the quantization of the color histogram is set to 16 color bins, aiming at reducing significantly the amount of data without loosing important information. The color bins value was established through experimental tests (see Avila et al., 2008b).

3.3. Elimination of meaningless frames

A *meaningless frame* is a monochromatic frame due to fade-in/fade-out effects. To remove possible meaningless frames, VSUMM computes the standard deviation of the frame feature vector. The standard deviation of monochromatic frames is equal to zero or a sufficiently small value close to zero.[4] This information is used by VSUMM to removes these frames.

This step is also employed by Furini et al. (2010). Unlike VSUMM, which removes meaningless frames as a preprocessing step, Furini et al. (2010) apply it as a post-processing step, after an initial summary is produced. Nevertheless, there is no point about using meaningless frames in their clustering step. The removal of such frames is performed before clustering in VSUMM, thus saving computation time.

3.4. Frames clustering

The k-means clustering algorithm (MacQueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem (Duda et al., 2001). In this work, the k-means algorithm is applied to cluster similar frames, although slightly modified in how it initially distributes the video frames among the k clusters. This modification is applied to improve k-means performance while producing more effective results.

The frames are initially grouped in sequential order, instead of randomly as in the original k-means algorithm. As an example, suppose $k = 5$ and a set of 50 frames sampled from a video. In the original k-means, the frames would be initially allocated randomly among the 5 clusters in order to start their iterative refinement. In case of VSUMM, the initial allocation is going to be done by associating the first 10 frames to the first cluster, the next 10 frames to the second one, and so on. This procedure is adopted based on the fact that consecutive frames typically already show some similarity among them, making it faster for k-means to converge.

One drawback of the k-means clustering algorithm is that it demands the number of clusters k to be fixed *a priori*. Nevertheless, k is related to the summary size, which is going to depend both on video length and on its dynamics. This means that different videos require different values for k. To overcome this difficulty imposed by k-means, a fast procedure to make a reasonable estimate of the number of clusters is implemented. VSUMM computes the pairwise distance of consecutive frames in the extracted sample, according to Euclidean distance. Then, the value selected for k is based on a threshold τ, which measures the sufficient content change in the video sequence. Every time the distance between two consecutive frames is greater than τ, then k is incremented. The threshold value applied in this work, established through experimental tests, is equal to 0.5.

Fig. 2 shows an example of how these distances are distributed along time. It is observed that there are points in time in which the distance between consecutive frames varies considerably (corresponding to peaks), while there are longer periods in which the variation is very small (corresponding to denser regions). Usually, peaks correspond to a sudden change in the video, while in dense regions frames are more similar to one another. Hence, frames between two peaks can be considered as a set of similar frames and therefore, the number of peaks provides a reasonable estimation to the number of clusters k.

It is worth noticing that our method for the estimation of the number of clusters is based on a simple shot boundary detection method (Guimaraes et al., 2003), whereas k is incremented for each sufficient content change in the video sequence.

3.5. Keyframe extraction

Once the clusters are formed by k-means, they can be further analyzed for keycluster selection. The strategy applied for

[4] There are frames that are not completely homogeneous in color, but can be regarded as meaningless frames.

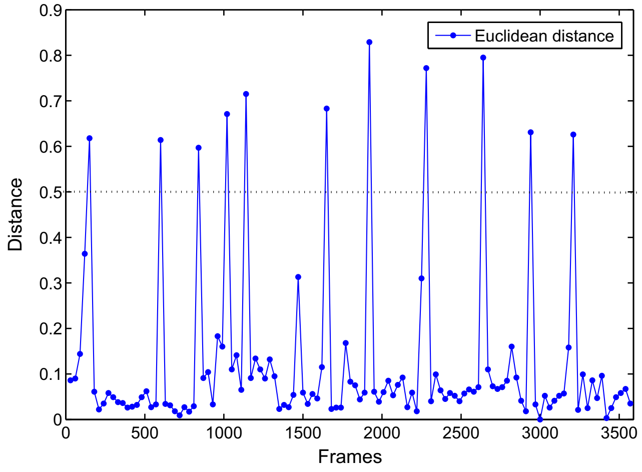


Fig. 2. Pairwise distances of sampled frames of the video *Drift Ice as a Geologic Agent, segment 8* (available at *Open Video Project*).

keyclusters selection is similar to the one proposed in (Zhuang et al., 1998). In VSUMM, a cluster is considered a *keycluster* if its size is larger than half the average cluster size (this value has shown to be more suitable as cut-off point than the average cluster size, as defined in (Zhuang et al., 1998)). For each keycluster, the frame which is closest to the keycluster centroid – measured by Euclidean distance – is selected as a keyframe. In the experiments described in Section 4, two different approaches are used: VSUMM$_1$ produces the summaries without performing keycluster selection and VSUMM$_2$ uses keycluster selection to produce its summaries.

3.6. Elimination of similar keyframes

The goal of this step is to avoid that keyframes too similar appear in the produced summaries. For this purpose, the keyframes are compared among themselves through color histogram. The similarity is based on a threshold τ, the same used to estimate the number of clusters. If the measured similarity is lower than τ, then the keyframe is removed from the summary.

In Fig. 3, it is possible to see an example of similar keyframes ($\tau < 0.5$) and non-similar keyframes ($\tau \geqslant 0.5$). It is interesting to notice that the frames do not need to be identical to be considered too similar.

Finally, the remaining keyframes are arranged in temporal order to make the produced summary more understandable.

3.7. Evaluation of video summary

In any knowledge area, to advance effectiveness and/or efficiency of new solutions to a particular problem, these need to be objectively evaluated, preferably against pre-existing ones. However, a consistent evaluation framework is seriously missing for video summarization research. Presently, every work has its own evaluation methodology, often presented without any performance comparison with previously existing techniques. To some extent, this happens because, unlike other research areas, such as object detection and recognition, the definition of what should be considered a "correct" summary is not a straightforward task, due to the lack of an objective ground-truth. The existing evaluation methods for video summarization are grouped into three different categories (Truong and Venkatesh, 2007): result description, objective metrics and user studies.

Result description is the most popular and simple form of evaluation, as it does not involve any comparison with other techniques. This category is also used to discuss the influence of the system parameters or visual dynamics of the video sequence on the keyframe set extracted (Hanjalic et al., 1998; Zhang et al., 2003; Yu et al., 2004). Some works may attempt, in descriptive form, to explain and illustrate advantages of the proposed technique compared with some existing methods (Joshi et al., 1998; Vermaak et al., 2002).

In objective metrics, for keyframe extraction techniques, the metric is often the fidelity function computed from the extracted keyframe set and original frame sequence. The metric is used to compare the keyframe set generated by different techniques, or by one underlying technique, but with different parameter sets. However, there is also no experimental justification for whether the metric maps well to human judgement regarding the quality of a keyframe set.

User studies are employed for evaluating keyframe extraction techniques in (Yahiaoui et al., 2001; Li et al., 2003; Wang et al., 2007; Avila et al., 2008a; Furini et al., 2010). These studies involve independent users judging the quality of generated video summaries, and are probably the most useful and realistic form of evaluation (especially when keyframes are extracted for user-based interactive tasks such as content browsing and navigation). Nevertheless, yet not widely employed due to the difficulty in setting them up.

In this work, it is proposed a novel evaluation method to evaluate video summaries. In this evaluation method, called *Comparison of User Summaries* (*CUS*), the video summary is built manually by a number of users from the sampled frames. The user summaries are taken as reference to be compared with the summaries obtained by different methods. In this way, the user summaries are the reference summaries, i.e., the ground-truth. Such comparisons are based on specific metrics, which are introduced in the following paragraphs.

The *CUS* evaluation method is based on Guironnet et al. (2007). In that evaluation method, an "optimal" summary is automatically built from user summaries. These summaries are then compared with the results of their summarization technique. Nevertheless, unlike that evaluation method, *CUS* compares each user summary directly with the automatic summaries, thus keeping the original opinion of every user. For comparing keyframes from different summaries, the same color histograms used in Section 3.2 are applied, while the distance among them is measured by Manhattan distance. Two keyframes are similar if the distance between them

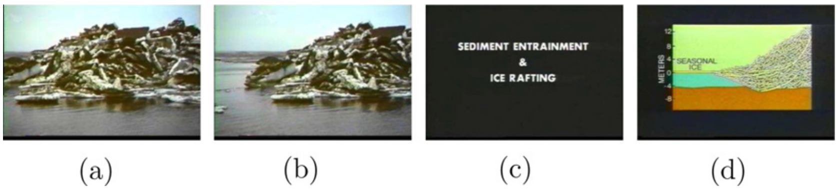


Fig. 3. Similar keyframes (a, b) and non-similar keyframes (c, d) of the video *Drift Ice as a Geologic Agent, segment 8* (available at *Open Video Project*).

is less than a predetermined threshold δ. Once two frames are matched, they are removed from the next iteration of the comparing procedure. The threshold value used, established through experimental tests, is equal to 0.5.

Fig. 4 illustrates our evaluation method. Firstly, the users are asked to watch the video and then manually create a summary for it. For the users to produce their summaries, the sampled frames are displayed to them (step 1). They are oriented to select a set of frames that, in their opinion, is able to summarize the original video content. The users are free to select any number of frames to compose their summaries. Next (step 2), the user summaries are compared with the automatically generated summary. The quality of the automatically generated summary is assessed (step 3) by two metrics, called accuracy rate CUS_A and error rate CUS_E, which are defined as follows:

$$CUS_A = \frac{n_{mAS}}{n_{US}}, \tag{1}$$

$$CUS_E = \frac{n_{\bar{m}AS}}{n_{US}}, \tag{2}$$

where n_{mAS} is the number of matching keyframes from automatic summary (AS), $n_{\bar{m}AS}$ is the number of non-matching keyframes from AS and n_{US} is the number of keyframes from user summary (US).

The CUS_A values range from 0 (the worst case, when none of the keyframes from AS match with the keyframes from US, or vice versa) to 1 (the best case, when all the keyframes from US match with the keyframes from AS). It is important to notice that $CUS_A = 1$ does not necessarily mean that all the keyframes from AS and US are matched. That is, if $n_{US} < n_{AS}$ (n_{AS} is the number of keyframes from AS) and $CUS_A = 1$, then some keyframes from AS did not match.

For CUS_E, the values range from 0 (the best case, when all the keyframes from AS matches with the keyframes from US) to n_{AS}/n_{US} (the worst case, when none of the keyframes from AS match with the keyframes from US, or vice versa).

This means that the CUS_A and CUS_E metrics are complementary, the highest summary quality being when $CUS_A = 1$ and $CUS_E = 0$, meaning that all keyframes from AS and US are exactly matched.

The goals of this method are: (1) to reduce the subjectivity in the evaluation task; (2) to quantify the summary quality and; (3) to allow more objective comparisons among different techniques.

4. Experimental results

The experiments are performed into two parts: (1) preliminary experiments, aimed at analyzing the VSUMM parameters that have the strongest impact on results and to identify possible problems; and (2) refined experiments, aimed at improving those previous results. The preliminary results are published in (Avila et al., 2008a,b). In this paper, only the refined results are presented.

Ideally, in order to compare different approaches to video summarization, each one should be tested on the same data sets and measured using the same metrics. However, almost all previous papers related to our work (i.e., dealing with static summarization) present experimental results based on different data sets. In addition, most of them did not make available either the data sets or the algorithm implementations, which makes direct comparisons almost impossible.

For this reason, we focus on evaluating our approach under 50 videos selected from the Open Video Project (OV). Those videos are the same ones used by Mundur et al. (2006) and Furini et al. (2010), thus the usage of the OV collection makes a comparative evaluation possible.

In addition, to verify the quality of VSUMM for videos with different characteristics of those of OV, we created a new database composed of videos collected from web sites like YouTube. The videos in this new collection differ in color, length, motion and subject (e.g., cartoons, news, sports, commercials, tv-shows and home videos).

4.1. Results for the Open Video database

All videos are in MPEG-1 format (30 fps, 352 × 240 pixels). The selected videos are distributed among several genres (documentary, educational, ephemeral, historical, lecture) and their duration varies from 1 to 4 min.

The user summaries were created by 50 users, each one dealing with 5 videos, meaning that each video has 5 video summaries created by 5 different users. In other words, 250 video summaries were created manually. All user summaries can be seen at http://www.npdi.dcc.ufmg.br/VSUMM.

As stated earlier (Section 3.5), two slightly different approaches were applied to produce the automatic summaries: VSUMM$_1$ and

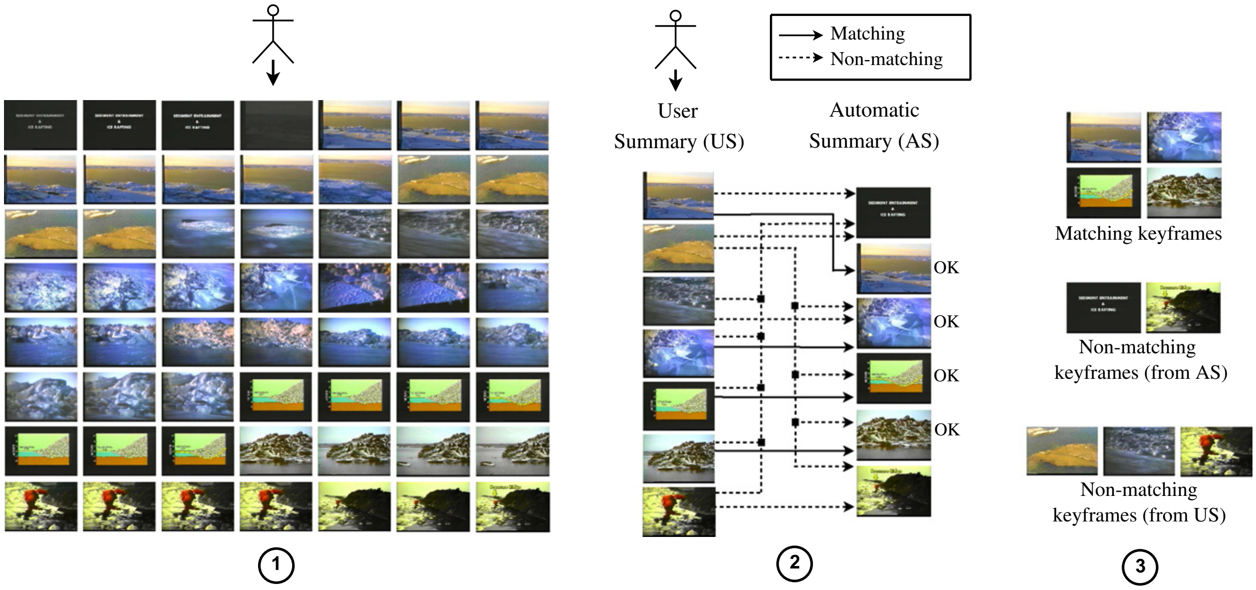


Fig. 4. *CUS* evaluation method.

Table 1
Mean accuracy rate CUS_A and mean error rate CUS_E achieved by different approaches.

	OV	DT	STIMO	VSUMM$_1$	VSUMM$_2$
CUS_A	0.70	0.53	0.72	**0.85**	0.70
CUS_E	0.57	0.29	0.58	0.38	**0.27**

The bold values indicates the best results for CUS_A and CUS_E.

Table 2
Difference between mean accuracy rates CUS_A at a confidence of 98%.

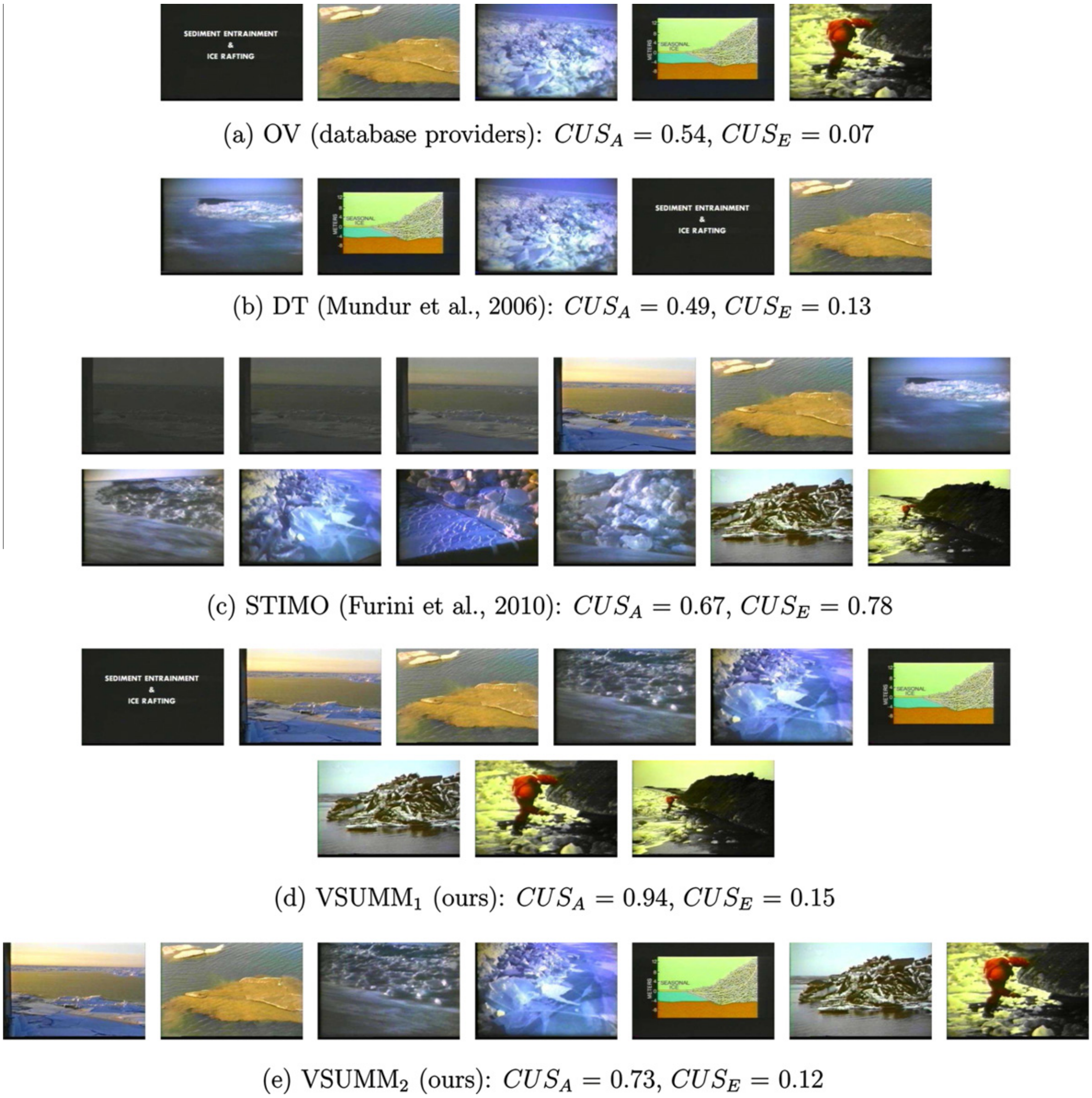
Difference	Confidence interval (98%)	
	Min.	Max.
VSUMM$_1$ – OV	0.08	0.22
VSUMM$_1$ – DT	0.26	0.38
VSUMM$_1$ – STIMO	0.07	0.20
VSUMM$_1$ – VSUMM$_2$	0.11	0.18

Table 3
Difference between mean error rates CUS_E at a confidence of 98%.

Difference	Confidence interval (98%)	
	Min.	Max.
VSUMM$_1$ – OV	−0.38	−0.01
VSUMM$_1$ – DT	0.01	0.17
VSUMM$_1$ – STIMO	−0.32	−0.09
VSUMM$_1$ – VSUMM$_2$	0.07	0.15

VSUMM$_2$. The only difference between them is that in VSUMM$_1$, one keyframe is selected per cluster, and in VSUMM$_2$, one keyframe is selected per keycluster. These approaches were compared with two other approaches found in the literature for automatic summarization – DT (Mundur et al., 2006) and STIMO (Furini et al., 2010). Additionally, the summaries produced by VSUMM$_1$ and VSUMM$_2$ were compared with the OV summaries, which are generated using the algorithm from DeMenthon et al. (1998) added to some manual intervention to refine the produced summaries. All static video summaries for the aforesaid approaches (OV, DT, STIMO, VSUMM$_1$, VSUMM$_2$) can be seen at http://www.npdi.dcc.ufmg.br/VSUMM.

The summaries quality is evaluated by the accuracy rate CUS_A (Eq. (1)) and error rate CUS_E (Eq. (2)). The results are shown in Table 1.



(a) OV (database providers): $CUS_A = 0.54$, $CUS_E = 0.07$

(b) DT (Mundur et al., 2006): $CUS_A = 0.49$, $CUS_E = 0.13$

(c) STIMO (Furini et al., 2010): $CUS_A = 0.67$, $CUS_E = 0.78$

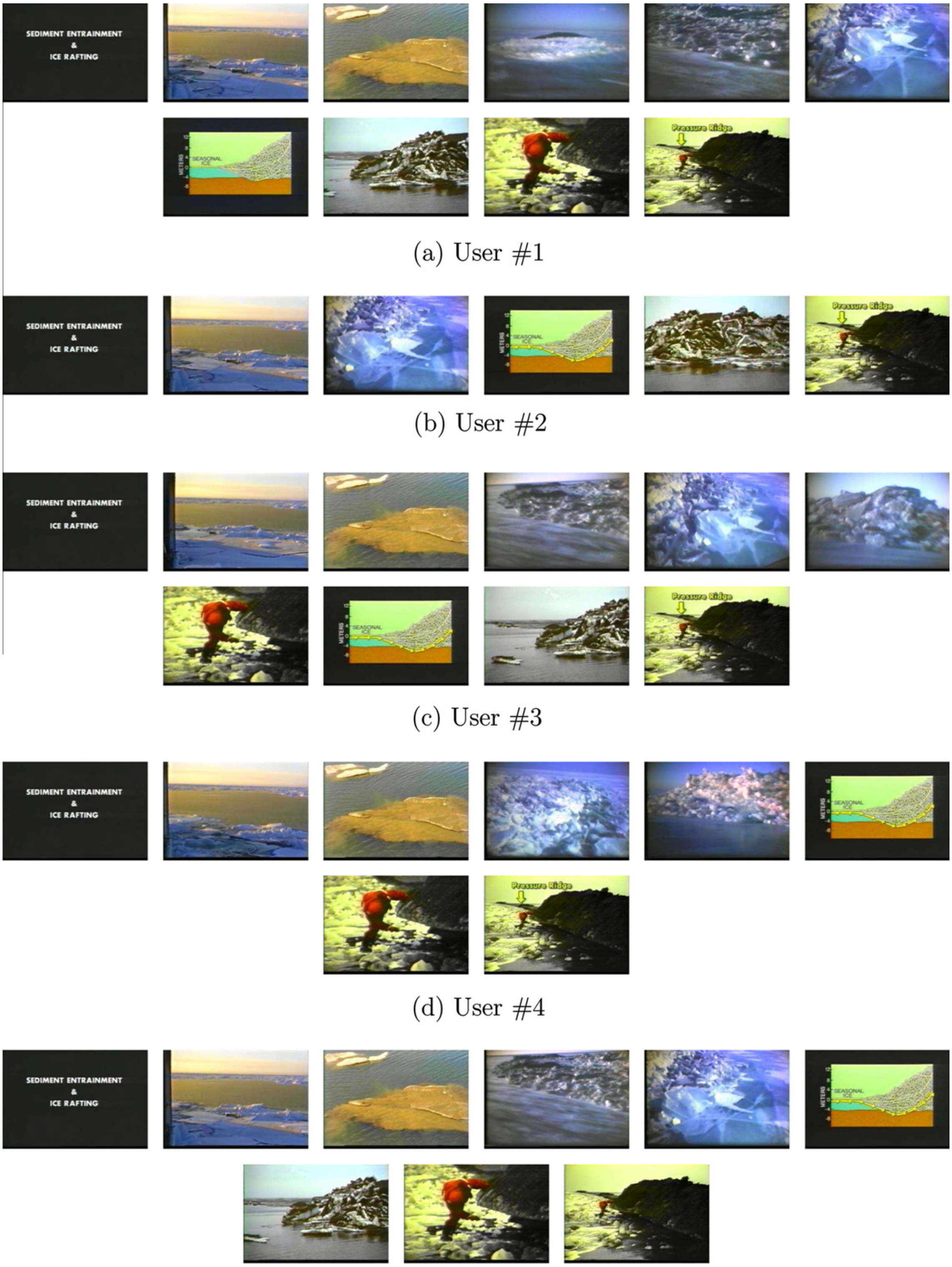
(d) VSUMM$_1$ (ours): $CUS_A = 0.94$, $CUS_E = 0.15$

(e) VSUMM$_2$ (ours): $CUS_A = 0.73$, $CUS_E = 0.12$

Fig. 5. Video summaries of different approaches of the video *Drift Ice as a Geologic Agent, segment 8* (available at *Open Video Project*).

The results indicated that VSUMM_1 achieved the highest accuracy rate and VSUMM_2 achieved the lowest error rate. To verify the statistical significance of these results, the confidence intervals for the differences between paired means were computed to compare every pair of approaches. 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 (Jain, 1992).

Tables 2 and 3 show the results of such comparisons between VSUMM_1 and the other approaches considered. These tables show the accuracy rates and the error rates, respectively.



(a) User #1

(b) User #2

(c) User #3

(d) User #4

(e) User #5

Fig. 6. User summaries of the video *Drift Ice as a Geologic Agent, segment 8* (available at *Open Video Project*).

Since the confidence intervals – with a confidence of 98% – do not include zero in any case, the results presented in Tables 2 and 3 confirm that VSUMM$_1$ approach provides results with superior quality (highest accuracy rate) relative to the approaches to which it was compared. In addition, it is possible to say that the VSUMM$_1$ summaries are closer to the summaries created by users.

Moreover, also with 98% confidence, the results confirm that VSUMM$_1$ approach presents a lower error rate than OV and STIMO approaches. However, VSUMM$_1$ presents a higher error rate than DT and VSUMM$_2$ approaches.

In DT approach, this "better" result was expected because the DT approach produces much smaller summaries than the summaries created by users. Consequently, the DT summaries present a low error rate at a cost of a low accuracy rate. In other words, this result can be disregarded, since as explained in Section 3.7, the most interesting summaries are those that present low error rate and, at the same time, high accuracy rate.

In the case of VSUMM$_2$ approach, the analysis is similar to the DT approach. The VSUMM$_2$ summaries show at most the same size of the VSUMM$_1$ summaries, but eventually smaller, since some clusters are disregarded in the keycluster refinement step. As VSUMM$_2$ produces smaller summaries, it tends to miss less, but also to hit less frames, as can be seen in Table 1, where the accuracy rate achieved by VSUMM$_2$ approach is significantly smaller than the accuracy rate achieved by VSUMM$_1$ approach.

Considering these observations, it is possible to conclude that VSUMM$_1$ approach provides better results relative to the approaches to which it was compared. Nevertheless, for applications which require lower error rate, the VSUMM$_2$ approach can be a better choice.

Fig. 5 shows the video summaries produced by all different approaches considered for comparison (OV, DT, STIMO, VSUMM$_1$, VSUMM$_2$). The video under consideration is *Drift Ice as a Geologic Agent, segment 8* and Fig. 6 displays the user summaries. For the *CUS* values reported, the OV and DT approaches exhibit similar low rates. Furthermore, it is possible to note that the STIMO summary contains keyframes that are very similar to each other, while VSUMM$_1$ provides a more concise summary for the video. VSUMM$_2$ achieves a low error rate ($CUS_E = 0.12$), but its accuracy rate is also low, even so, its accuracy rate is better than the OV, DT and STIMO rates. The highest summary quality ($CUS_A = 0.94$ and $CUS_E = 0.15$) is achieved by VSUMM$_1$ approach, which can be confirmed by a visual comparison with the user summaries that can be seen in Fig. 6.

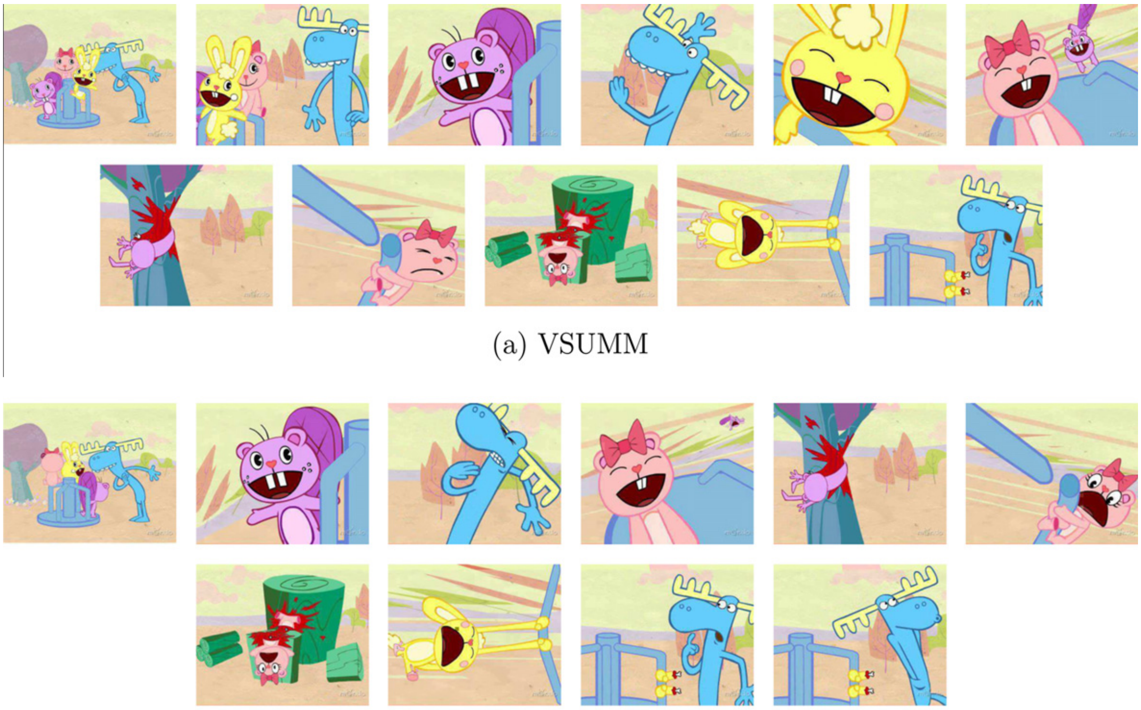
4.1.1. Discussion

It is interesting to compare the accuracy rates of VSUMM$_1$ and VSUMM$_2$ approaches. On the contrary to what could be expected at first sight, the VSUMM$_1$ approach provided results with superior quality relative to the VSUMM$_2$ approach. Once VSUMM$_2$ selects the keyframes from the keyclusters, eliminating the clusters that, in theory, would not be too important – because they are composed of a small number of frames –, then it was expected that the accuracy rate achieved by it would be higher than the accuracy rate achieved by VSUMM$_1$ approach.

This result is mainly due to the high number of keyframes of the video summaries created by users. Before performing the evaluation process, it was informed to the users that they should select the frames which, in their opinion, would better represent the original video content concisely. Thus, it was expected to obtain user summaries consisting only of the most relevant frames (keyframes). Nevertheless, the experiments showed that the users preferred to create more extensive summaries that represent all the various video segments, regardless of the segment size.

4.2. Results for the new database

The next 50 videos were collected from web sites, like YouTube. These videos are distributed among several genres (cartoons, news, sports, commercials, tv-shows and home videos) and their duration varies from 1 to 10 min.



(a) VSUMM

(b) User #1

Fig. 7. VSUMM summary and one user summary of a cartoon video.

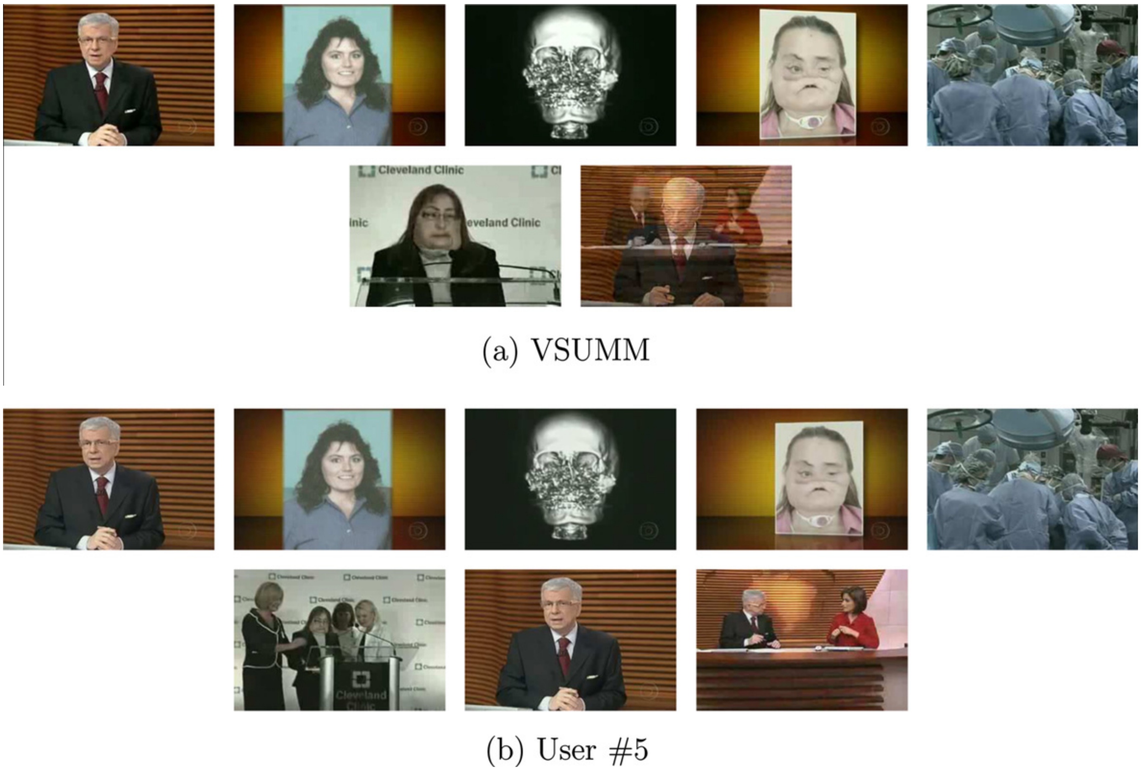
Table 4
Mean accuracy rate CUS_A and mean error rate CUS_E achieved by different video categories.

	# Videos	CUS_A	CUS_E
Cartoons	10	0.87	0.22
News	15	0.88	0.32
Sports	17	0.76	0.65
Commercials	2	0.93	0.06
TV-shows	5	0.91	0.33
Home	1	0.85	0.23
Weighted average	50	0.84	0.40

Since it was observed in the previous experiment that $VSUMM_1$ produces a better result in terms of summary quality, only $VSUMM_1$ is applied in this new set of experiments.

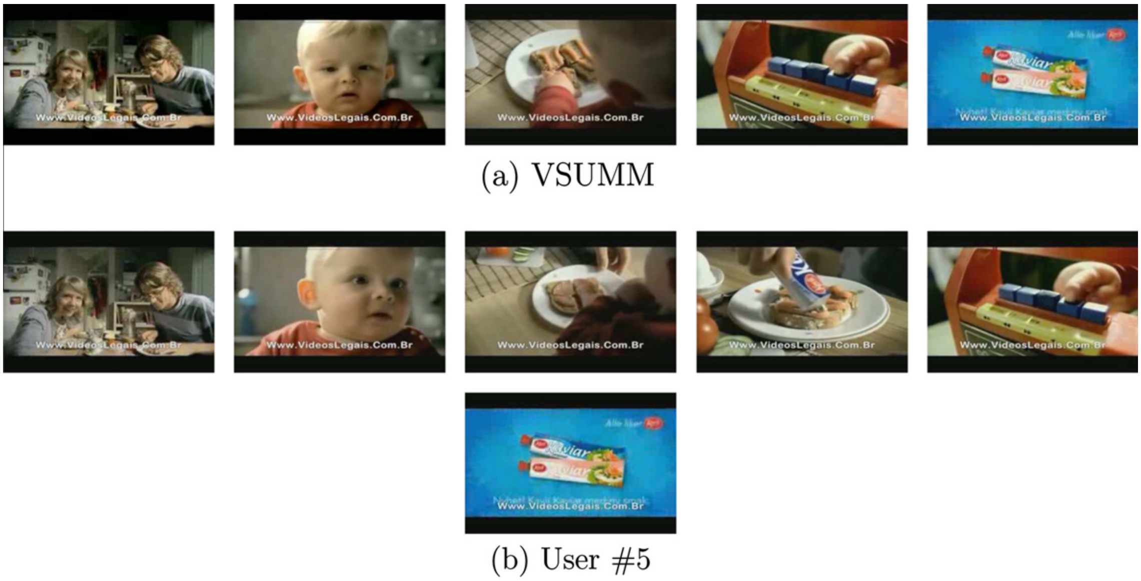
Following the same experimental protocol as before, we invited 50 users to manually create static summaries for the videos in the new database. Five user summaries were produced for each video and all of them can be seen at http://www.npdi.dcc.ufmg.br/VSUMM.

The summaries quality is evaluated by the accuracy rate CUS_A and error rate CUS_E. The results are shown in Table 4.


(a) VSUMM

(b) User #5

Fig. 8. VSUMM summary and one user summary of a news video.


(a) VSUMM

(b) User #5

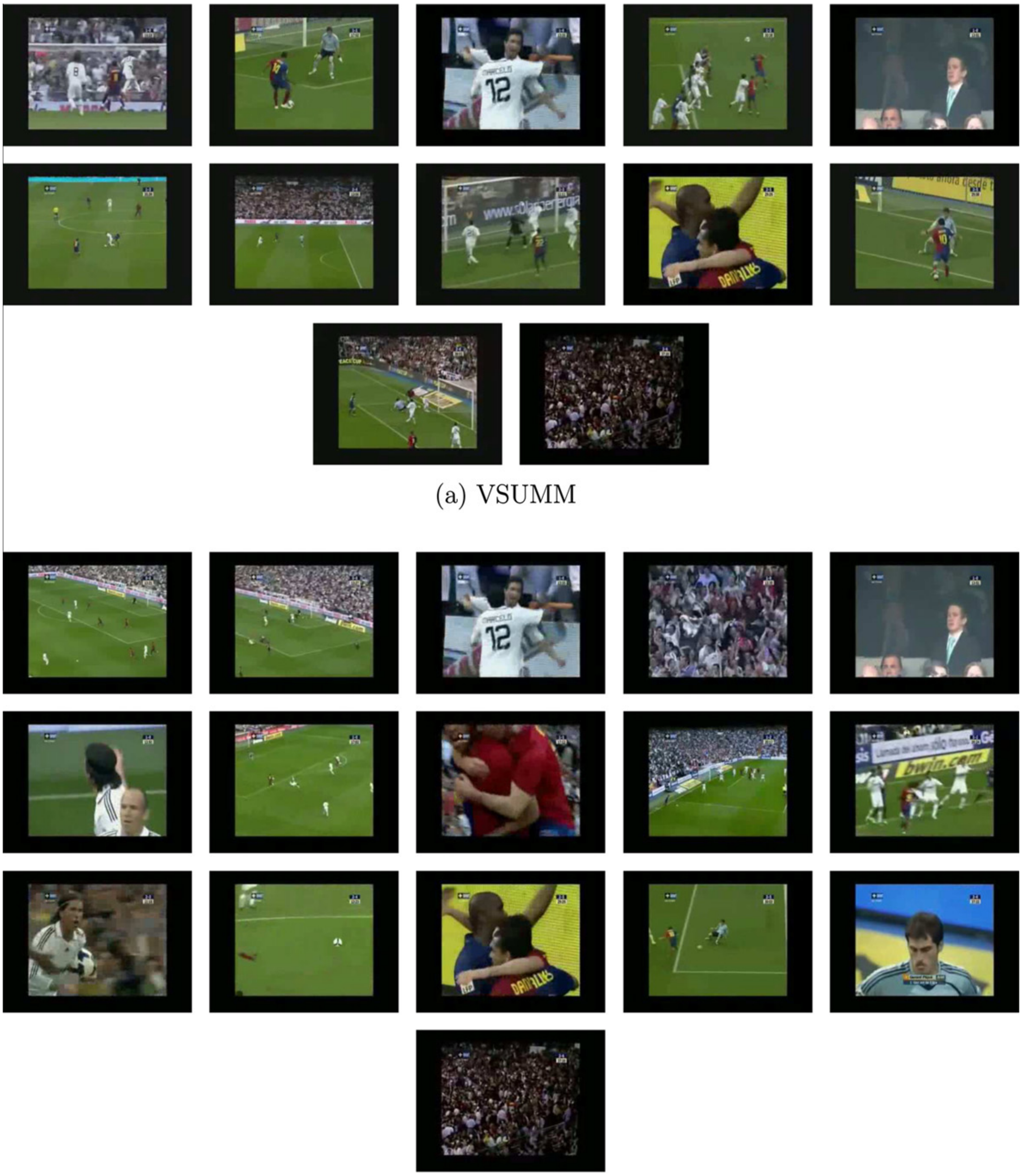
Fig. 9. VSUMM summary and one user summary of a commercial video.

According to those results, we can notice that VSUMM presented a better result for the two videos in the category of commercials, for which the CUS_A had the highest value and the CUS_E presented the lowest value among all categories. Figs. 7–11 show the static summary produced by VSUMM and one user summary for the considered categories of videos.

In case of sports videos (mostly soccer videos), VSUMM presented a CUS_E fairly high. That happened because, while VSUMM is aimed at producing very concise summaries, the users preferred to show entire sequences of moves. This result indicates that in cases like sports, domain-specific summarization techniques as the one presented in (Ekin et al., 2003) can be a better choice.

A similar effect could be found in the tv-shows and news videos, but this time with a less detrimental effect on the overall summary quality, since in those cases the CUS_A values were also quite high. Again, the users seemed to regard that preserving some information about the sequence of events was important for the summary. For example, in news videos, several appearances of the same anchors were represented in the users summaries, although they are practically identical from the visual point of view.

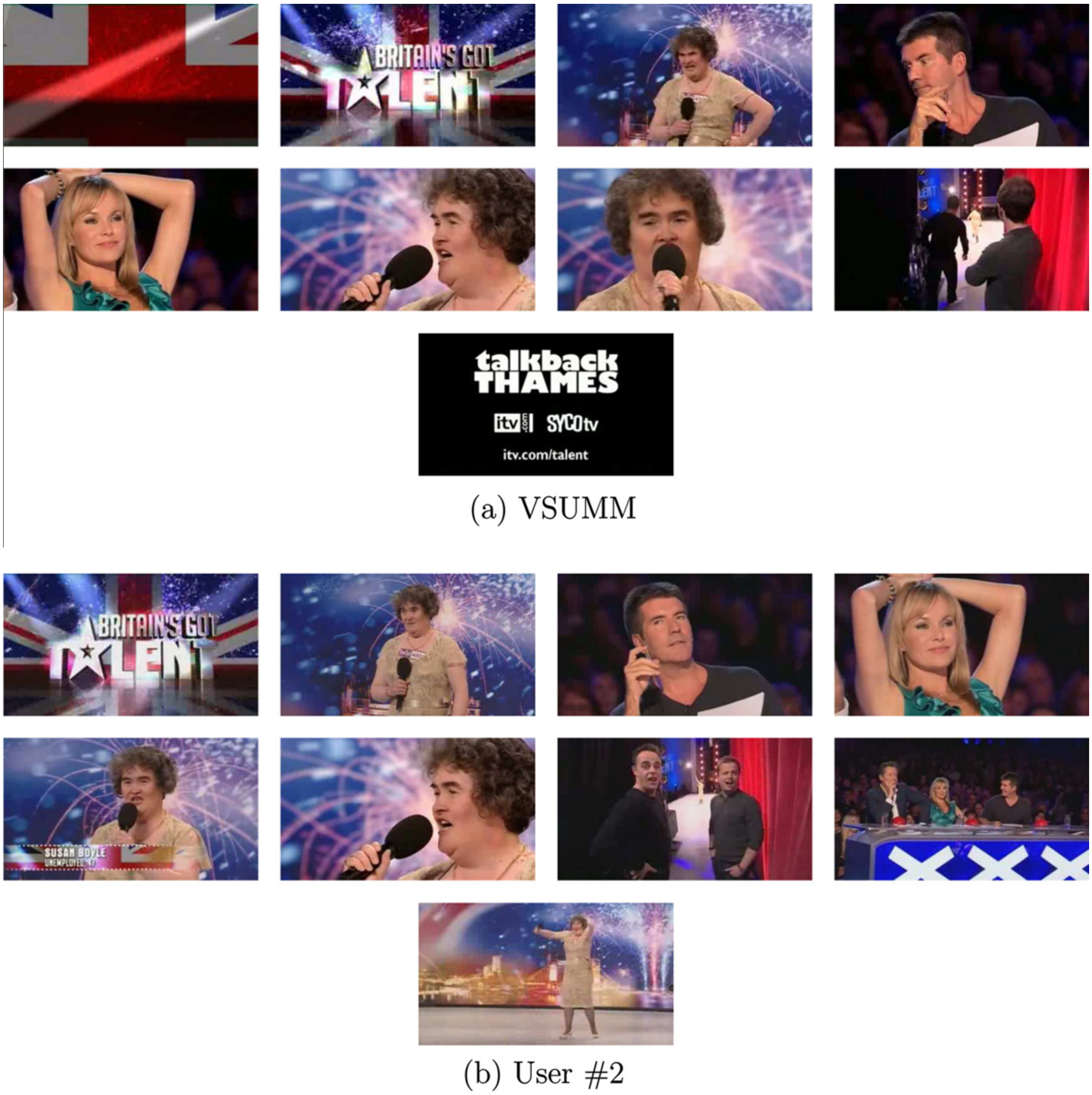
Finally, it is important to notice that in the experiments with this new database, the average CUS_A and CUS_E values were very similar to those of the previous experiment. Such result indicates



(a) VSUMM

(b) User #3

Fig. 10. VSUMM summary and one user summary of a sport video.



(a) VSUMM

(b) User #2

Fig. 11. VSUMM summary and one user summary of a tv-show video.

that the overall quality of VSUMM summaries can be sustained in video collections with different characteristics.

5. Conclusions

Automatic video summarization has been receiving growing attention from the scientific community. This attention can be explained by several factors, for example, (1) the advances in the computing and network infrastructure, (2) the growth of the number of videos published on the Internet, (3) scientific challenges, (4) practical applications as search engines and digital libraries, (5) inappropriate use of traditional video summarization techniques to describe, represent and perform search in large video collections. As examples, video search engines like Google[5] and Yahoo[6] usually represent entire videos by a single keyframe.

In this paper, we presented VSUMM, a mechanism designed to produce static video summaries. It presents the advantages of the concepts of related work in the video summarization area; on a single method, VSUMM includes the main contributions of previously proposed techniques. As an additional contribution, we proposed a new evaluation method better suited to compare competing summarization techniques, because (1) reduces the subjectivity in the evaluation task, (2) quantifies the summary quality and (3) allows more objective comparisons among different techniques.

Future work includes the evaluation of other visual features and their fusion. Furthermore, techniques to estimate the number of clusters can be exploited, for example, Akaike's Information Criterion (AIC) (Akaike, 1974) or Minimum Description Length (MDL) (Rissanen, 1978). Other clustering algorithms can also be investigated, for example, DBSCAN (Ester et al., 1996), a density-based clustering method.

Also, the investigation of techniques for introducing sequential information in the summaries in order to better match user expectations can be valuable to some application scenarios. Finally, VSUMM can be extended to produce video skims. This can be done from keyframes by joining fixed-size segments, subshots, or the whole shot that encloses them, as employed in (Hanjalic and Zhang, 1999).

[5] http://video.google.com.

[6] http://video.search.yahoo.com.

Acknowledgments

The authors are grateful to CNPq, CAPES and FAPEMIG, Brazilian research funding agencies, for the financial support to this work.

References

Akaike, H., 1974. A new look at statistical model identification. IEEE Trans. Automat. Control 19, 716–723.

Avila, S.E.F., da Luz Jr., A., Araújo, A.A., 2008a. VSUMM: A simple and efficient approach for automatic video summarization. In: 15th Internat. Conf. on Systems, Signals and Image Processing (IWSSIP), Bratislava, Slovakia, pp. 449–452.

Avila, S.E.F., da Luz Jr., A., Araújo, A.A., Cord, M., 2008b. VSUMM: An approach for automatic video summarization and quantitative evaluation. In: Proc. XXI Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI), pp. 103–110.

Ćalić, J., Gibson, D.P., Campbell, N.W., 2007. Efficient layout of comic-like video summaries. IEEE Trans. Circuits Systems Video Technol. 17 (7), 931–936.

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

Chang, I.-C., Chen, K.-Y., 2007. Content-selection based video summarization. Digest of Technical Papers International Conference on Consumer Electronics (ICCE), pp. 1–2.

Chen, B.-W., Wang, J.-C., Wang, J.-F., 2009. A novel video summarization based on mining the story-structure and semantic relations among concept entities. IEEE Trans. Multimedia 11 (2), 295–312.

Cheung, S.-S., Zakhor, A., 2003. Efficient video similarity measurement with video signature. IEEE Trans. Circuits Systems Video Technol. 13 (1), 59–74.

Cotsaces, C., Nikolaidis, N., Pitas, L., 2006. Video shot detection and condensed representation: A review. IEEE Signal Process. Mag. 23 (2), 28–37.

DeMenthon, D., Kobla, V., Doermann, D., 1998. Video summarization by curve simplification. In: Proc. ACM Internat. Conf. on Multimedia. NY, USA, pp. 211–218.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Unsupervised Learning and Clustering. Pattern Classification. Springer-Verlag New York, Inc.. p. 654.

Ekin, A., Tekalp, A., Mehrotra, R., 2003. Automatic soccer video analysis and summarization. IEEE Trans. Image Process. 12 (7), 796–807.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proc. Internat. Conf. on Knowledge Discovery and Data Mining (KDD), pp. 226–231.

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

Gong, Y., Liu, X., 2000. Video summarization using singular value decomposition. In: Proc. IEEE Internat. Conf. on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, Los Alamitos, CA, USA, pp. 2174–2180.

Guimaraes, S.J.F., Couprie, M., Araújo, A.A., Leite, N.J., 2003. Video segmentation based on 2D image analysis. Pattern Recognition Lett. 24 (7), 947–957.

Guironnet, M., Pellerin, D., Guyader, N., Ladret, P., 2007. Video summarization based on camera motion and a subjective evaluation method. EURASIP J. Image Video Process.. Article ID 60245, 12 p..

Hadi, Y., Essannouni, F., Thami, R.O.H., 2006. Video summarization by *k*-medoid clustering. In: Proc. ACM Symposium on Applied Computing (SAC), New York, NY, USA, pp. 1400–1401.

Hanjalic, A., Lagendijk, R.L., Biemond, J., 1998. A new method for key frame based video content representation. In: Image Databases and Multi-media Search. World Scientific, Singapore.

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

Herranz, L., Martinez, J., 2009. An efficient summarization algorithm based on clustering and bitstream extraction. In: IEEE Internat. Conf. on Multimedia and Expo (ICME), pp. 654–657.

Jain, R., 1992. The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling. John Wiley and Sons, Inc..

Joshi, A., Auephanwiriyakul, S., Krishnapuram, R., 1998. On fuzzy clustering and content based access to networked video databases. In: Proc. Internat. Workshop on Research Issues in Data Engineering (RIDA) Conference, pp. 42–43.

Koprinska, I., Carrato, S., 2001. Temporal video segmentation: A survey. Signal Process.: Image Comm. 16 (5), 477–500.

Li, Y., Lee, S.-H., Yeh, C.-H., Kuo, C.-C., 2006. Techniques for movie content analysis and skimming: Tutorial and overview on video abstraction techniques. Signal Process. Mag., IEEE 23 (2), 79–89.

Li, Y., Zhang, T., Tretter, D., 2001. An overview of video abstraction techniques. Tech. Rep., HP Laboratory, HP-2001-191.

Li, Z., Katsaggelos, K., Gandhi, B., 2003. Temporal rate-distortion based optimal video summary generation. In: Proc. IEEE Internat. Conf. on Multimedia and Expo (ICME), Washington, DC, USA, pp. 693–696.

Li, Z., Schuster, G.M., Katsaggelos, A.K., 2005. Minmax optimal video summarization. IEEE Trans. Circuits Systems Video Technol. 15 (10), 1245–1256.

Lienhart, R., 1999. Comparison of automatic shot boundary detection algorithms. In: Proc. IS&T/SPIE Conf. on Storage and Retrieval for Image and Video Databases, pp. 290–301.

MacQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations. In: Cam, L.M.L., Neyman, J. (Eds.), Proc. The Berkeley Symposium on Mathematical Statistics and Probability, vol. 1. University of California Press, pp. 281–297.

Manjunath, B.S., Ohm, J.R., Vinod, V.V., Yamada, A., 2001. Color and texture descriptors. IEEE Trans. Circuits Systems Video Technol. 11 (6), 703–715.

Money, A.G., Agius, H., 2008. Video summarisation: A conceptual framework and survey of the state of the art. J. Visual Comm. Image Represent. (JVCIR) 19 (2), 121–143.

Mundur, P., Rao, Y., Yesha, Y., 2006. Keyframe-based video summarization using Delaunay clustering. Internat. J. Dig. Libr. 6 (2), 219–232.

Pfeiffer, S., Lienhart, R., Fischer, S., Effelsberg, W., 1996. Abstracting digital movies automatically. Tech. Rep., University of Mannheim.

Rissanen, J., 1978. Modelling by shortest data description. Automatica 14, 465–471.

Swain, M.J., Ballard, D.H., 1991. Color indexing. Internat. J. Comput. Vision 7 (1), 11–32.

Trémeau, A., Tominaga, S., Plataniotis, K.N., 2008. Color in image and video processing: Most recent trends and future research directions. EURASIP J. Image Video Process. 2008 (3), 1–26.

Truong, B.T., Venkatesh, S., 2007. Video abstraction: A systematic review and classification. ACM Trans. Multimedia Comput. Comm. Appl. 3 (1).

Uchihashi, S., Foote, J., Girgensohn, A., Boreczky, J.S., 1999. Video manga: Generating semantically meaningful video summaries. In: Proc. ACM Internat. Conf. on Multimedia (Part 1), New York, NY, USA, pp. 383–392.

Vermaak, J., Pérez, P., Gangnet, M., Blake, A., 2002. Rapid summarisation and browsing of video sequences. In: Proc. British Machine Vision Conference (BMVC), vol. 1.

Wang, F., Merialdo, B., 2009. Multi-document video summarization. In: IEEE Internat. Conf. on Multimedia and Expo (ICME), pp. 1326–1329.

Wang, T., Mei, T., Hua, X.-S., Liu, X.-L., Zhou, H.-Q., 2007. Video collage: A novel presentation of video sequence. In: Proc. IEEE Internat. Conf. on Multimedia and Expo, pp. 1479–1482.

Yahiaoui, I., Mérialdo, B., Huet, B., 2001. Automatic video summarization. In: Multimedia Content-based Indexing and Retrieval (MCBIR).

Yeung, M.M., Leo, B.-L., 1997. Video visualization for compact representation and fast browsing of pictorial content. IEEE Trans. Circuits Systems Video Technol. 7 (5), 771–785.

Yu, X.-D., Wang, L., Tian, Q., Xue, P., 2004. Multi-level video representation with application to keyframe extraction. In: Proc. Internat. Multimedia Modelling Conference, pp. 117–121.

Zhang, X.-D., Liu, T.-Y., Lo, K.-T., Feng, J., 2003. Dynamic selection and effective compression of key frames for video abstraction. Pattern Recognition Lett. 24 (9–10), 1523–1532.

Zhu, X., Wu, X., Fan, J., Elmagarmid, A.K., Aref, W.G., 2004. Exploring video content structure for hierarchical summarization. Multimedia Systems 10 (2), 98–115.

Zhuang, Y., Rui, Y., Huang, T.S., Mehrotra, S., 1998. Adaptive key frame extraction using unsupervised clustering. In: Proc. IEEE Internat. Conf. on Image Processing (ICIP), vol. 1, pp. 866–870.