Hindawi Publishing Corporation
EURASIP Journal on Image and Video Processing
Volume 2007, Article ID 60245, 12 pages
doi:10.1155/2007/60245

Research Article

Video Summarization Based on Camera Motion and a Subjective Evaluation Method

M. Guironnet, D. Pellerin, N. Guyader, and P. Ladret

Laboratoire Grenoble Image Parole Signal Automatique (GIPSA-Lab) (ex. LIS), 46 avenue Felix Viallet, 38031 Grenoble, France

Received 15 November 2006; Revised 14 March 2007; Accepted 23 April 2007

Recommended by Marcel Worring

We propose an original method of 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 between the selected frames. We also develop a new subjective method to evaluate the proposed summary and to compare different summaries more generally. Subjects were asked to watch a video and to create a summary manually. From the summaries of the different subjects, an "optimal" one is built automatically and is compared to the summaries obtained by different methods. Experimental results show the efficiency of our camera motion-based summary.

1. INTRODUCTION

During this decade, the number of videos has increased with the growth of broadcasting processes and storage devices. To facilitate access to information, various indexing techniques using "low-level" features such as color, texture, or motion have been developed to represent video content. It has led to the emergence of new applications such as video summary, classification, or browsing in a video database. In this paper, we will introduce two methods required to study video summary: the first one explains how to create a video summary and the second one how to evaluate it and to compare different summaries.

A video summary is a short version of the video and is composed of representative frames, called keyframes. The selection of keyframes has to be done with the aim of both representing the whole video content and suppressing the redundancy between frames. As we said, videos are usually described by "low-level" features to which it is difficult to give a meaning. On the contrary, a semantic meaning can be deduced from camera motions. For example, an action movie contains many scenes with strong camera motions: a zoom-in will focus the spectator's gaze on a particular location in a scene. In this paper, we exploit the information provided by camera motion to describe the video content and to choose the keyframes.

In the literature, some video summary methods were proposed from camera motion. The first family uses camera motion to segment the video but not to select the keyframes. The keyframe selection is based on other features. In [1], the camera motion is used to detect moving objects and this information is used to build the summary. In [2], camera motion is used to partition the shots in segments and keyframe selection is carried out with other indexes (4 basic measures, i.e., visually pleasurable, representative, informative, and distinctive). A shot is, by definition, a portion of video filmed continuously without special effects or cuts, and a segment is a set of successive frames having the same type of motion. In [3], shots are segmented according to camera motions. Then, MPEG motion vectors, that contain the camera and object motions, are used to define the motion intensity per frame and select the keyframes. Nevertheless, these approaches do not select keyframes directly according to camera motion. In fact, the camera motion is used more to segment the video than to create the summary itself.

The second family is based mainly on the presence or the absence of motion. Cherfaoui and Bertin [4] detect the shots, then determine the presence or the absence of camera motion. The shots with a camera motion are represented by three keyframes, whereas the shots with fixed camera have only one. Peker and Divakaran [5] work out a summary method by selecting the segments with large motions in order to capture the dynamic aspects of video. In this case they used camera motion and also object motion. In [6], the segments with a camera motion provide keyframes which are added to the summary. Nevertheless, these approaches are

based on simple considerations which exploit little information contributed by camera motion.

The third family uses camera motion to define a similarity measure between frames; this similarity is then used to select the keyframes. In [7], a similarity measure between two frames is defined by calculating the overlap between them. The greater the overlap is, the closer the content is and the fewer keyframes are selected. In the same way, Fauvet et al. [8] determine from the estimation of the dominant motion, the areas between two successive frames which are lost or appear. Then, a cumulative function of surfaces which appear between the first frame of the shot and the current frame is used to determine the keyframes. Nevertheless, these approaches are based on a low-level description which measures the overlap between frames. They are based on geometrical and local properties (number of pixels which appear or which are lost between two frames) and do not select frames according to the type of motion detected.

In this paper, we propose a new method of video summary based on camera motions (translation and zoom) or on static camera. We think that camera motion carries important information on video content. For example, a zoom in makes it possible to focus spectator attention on a particular event. In the same way, a translation indicates a change of place. Therefore, keyframes were selected according to camera motion characteristics. More precisely, the method consists in studying the succession and the magnitude of camera motions. From these two criteria, various rules are worked out to build the summary. For example, the keyframe selection will be different according to the magnitude and the succession of the motions detected. The advantage of this method is to avoid a direct comparison between frames (similarity measure or overlap between frames on pixel level) and it is based only on camera motion classification.

Video summarization methods must be evaluated to verify the relevance of the selected keyframes. As already mentioned, video summarization methods are widely studied in the literature. Nevertheless, there is no standard method to evaluate the various video summaries. Some authors [9, 10] propose objective (mathematical) measures that do not take human judgment into account. To overcome this problem, other authors propose subjective evaluation methods. Three families of subjective evaluation can be distinguished to judge video summarization methods.

The first family of methods compares two summaries. For example, in [11], people view the entire video and choose between two summaries the one which best represents the video viewed. One summary results from a video summarization method to be tested and the other comes from another method developped by other researchers (a regular sampling of the video or a simplified version of the summarization method to be tested). The aim is to show that the summary suggested by one method is better than another method.

The second family creates a summary manually, a kind of "ground truth" of video, that is used for the comparison with the summary obtained by its automatic method. The comparison is made with some indices (recall and precision). The comparison is carried out either manually or by computing distances. For example, Ferman and Tekalp [12] evaluate their summary by requiring a neutral observer to announce the forgotten keyframes and the redundant ones. The criteria of evaluation are thus the number of forgotten and redundant keyframes.

In the third family, subjects are asked to measure the level of meaning of the proposed summary. A subject views a video, then he is asked to judge the summary according to a given scale. The subjects can be asked questions also to measure the degree of performance of the proposed summary. In [13], the quality of the summary is evaluated by asking subjects to give a mark between one and five for four criteria: clarity, conciseness, coherence, and overall quality. In [14], the subject must initially give an appreciation for each shot on the single selected keyframe (good, bad, or neutral) then he must give appreciations on the number of keyframes per shot (good, too many, too few). In [15], three questions are asked about the summary: who, what, and coherence. Ngo et al. [16] propose two criteria of evaluation to judge the summary: informativeness and enjoyability. The first criterion reveals the ability of the summary to represent all the information in the video by avoiding redundancy, and the second evaluates the performance of the algorithm in giving enjoyable segments.

The evaluation method that we propose belongs to the second family. It consists in building an "optimal" summary, called the reference summary, from the summaries obtained by various subjects. Next, an automatic comparison is carried out between the reference summary and the summaries provided by various methods. This evaluation technique provides a method to test different summaries quickly.

The camera motion-based method to create a video summary is explained in Section 2. Then, in Section 3, the subjective method to evaluate the proposed summary is presented. Finally, Section 4 concludes the paper.

2. VIDEO SUMMARIZATION METHOD FROM CAMERA MOTION

The principle of the summarization method consists in cutting up each video shot in segments of homogeneous camera motion, then in selecting the keyframes according to the succession and the magnitude of camera motions. The method requires the parameters extracted from the camera motion recognition and described in [17] to be known. A short recall of the camera motion recognition method is presented followed by an explanation of the keyframe selection method.

2.1. Recognition of camera motion

This recognition consists in detecting translation (pan and/or tilt), zoom and static camera in a video. The system architecture, depicted in Figure 1, is made up of three phases: motion parameter extraction, camera motion classification (e.g., zoom), and motion description (e.g., zoom with an enlargement coefficient of five). The extraction phase consists in estimating the dominant motion between two successive

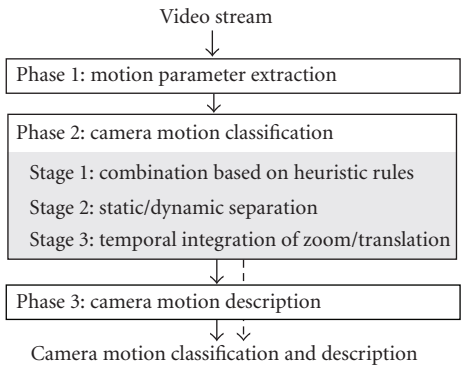


Figure 1: System architecture for camera motion classification and description.

frames by an affine parametric model. The core of the work is the classification phase which is based on transferable belief model (TBM) and is divided into three stages.

The first stage is designed to convert the motion model parameters into symbolic values. This representation aims at facilitating the definition of rules to combine data and to provide frame-level "mass functions" for different camera motions. The second stage carries out a separation between static and dynamic (zoom, translation) frames. In the third stage, the temporal integration of motions is carried out. The advantage of this analysis is to preserve the motions with significant magnitude and duration. Finally, a motion is associated with each frame and a video is split into segments (i.e., set of successive frames having the same type of motion).

The description phase is then carried out by extracting different features on each video segment containing an identified camera motion type. For example, a zoom segment (see Figure 2(a)) is represented by the enlargement coefficient *ec* and the direction of the zoom (in or out). A translation segment (see Figure 2(b)) is described by the distance traveled noted *dt* and the total displacement noted *td*. The total displacement *td* corresponds to the displacement along the straight line between the initial and the final positions, whereas the distance traveled *dt* is the original path and corresponds to the integration of all displacements between sampling times.

Consequently, this method is used to identify and describe camera motion segments inside each video shot. The parameters extracted to describe translation and zoom segments will be used to create the summary.

2.2. Keyframe selection according to camera motions

Keyframe selection depends on camera motions in each video shot. As mentioned before, each shot is first cut into segments of homogenous camera motion. The keyframe selection is divided into two steps. First, some frames are chosen to be potential keyframes to describe each segment: one at the beginning and one at the end, and in some cases one in the middle. In practice, even for long segments, we noted that three keyframes are enough to describe each segment. Then, some of the keyframes are kept and others removed according to certain rules. We will present the keyframe selection first according to the succession of motions, second the magnitude of motions and finally by the combination of both.

2.2.1. Keyframe selection according to succession of camera motions

To select the keyframes, we define heuristic rules. Because of the compactness of the summary, only two frames are selected to describe the succession of two camera motions. If one of the two successive segments is static, the two frames are selected at the beginning and at the end of the segment with motion. One of these frames is also used to represent the static segment. If the two successive segments have camera motions, a frame is selected at the beginning of each segment. Figure 3 recapitulates how the keyframes are selected. The process is repeated iteratively for all the motion segments of the shot.

This technique processes two consecutive motions at a time. Let us suppose that three consecutive motions are detected in a shot: static, translation, and static. By applying the rules defined in Figure 3, we obtain the results shown in Figure 4. Each iteration corresponds to the process of two consecutive segments. By superposition of the iterations, the result obtained is two selected frames: one at the end of the static segment (or at the beginning of the translation segment) and one at the end of the translation segment (or at the beginning of the last segment).

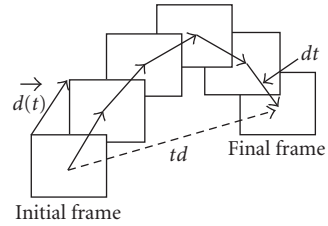
2.2.2. Keyframe selection according to magnitude of camera motions

Keyframe selection also has to take into account the magnitude of camera motions. For example, a translation motion with a strong magnitude requires more keyframes to be described than a static segment, since the visual content is more dissimilar from one frame to the following one. In the same way, a zoom segment is described by a number of keyframes linked to its enlargement coefficient.

For a translation segment, the coefficient $c_r = (dt - td)/dt$ is calculated in order to determine if the trajectory is rectilinear. This coefficient c_r lies between 0 and 1 and describes the motion trajectory. The smaller c_r is, the more rectilinear the motion is. Consequently, if coefficient c_r is lower than a threshold δ_r, the motion is considered rectilinear. In this case, if the total displacement *td* is large, that is, higher than threshold δ_{td}, the first and the last frames of the segment are selected. Only the last frame is selected if the total displacement *td* is weak (lower than threshold δ_{td}). On the other hand, if coefficient c_r is higher than δ_r, the motion changes direction. If the total displacement *td* is higher than threshold δ_{td}, the frames of the beginning, the middle, and the end of the segment are selected. If not, the last frame of the segment is selected.

For a zoom segment, the keyframes are selected according to the enlargement coefficient *ec*. If the enlargement is

(a) Definition of the enlargement coefficient ec



(b) Definition of the distance traveled dt and the total displacement td from displacement $d(t)$ between 2 successive frames

Figure 2: Example of parameters extracted to describe each segment of a video for (a) a zoom and (b) a translation.

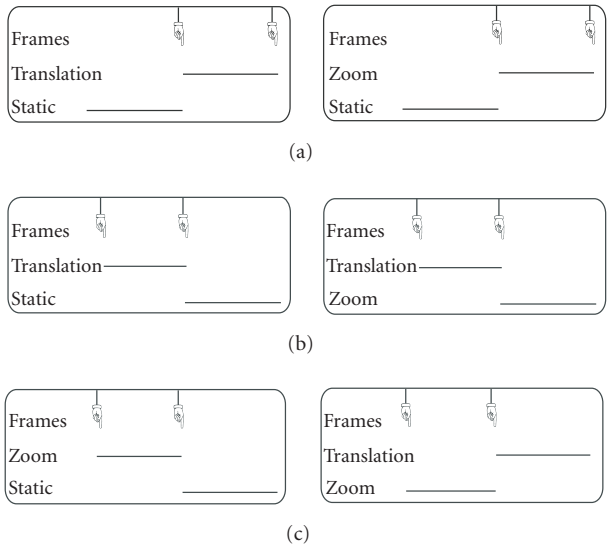


Figure 3: Rules for keyframe selection according to two consecutive camera motions. Cases: (a) translation and static, (b) zoom and static, (c) translation and zoom. For example, if a static segment is followed by a translation segment (Figure (a) left), the first frame of the translation segment (or the last frame of the static segment) is selected as well as the last frame of the translation segment.

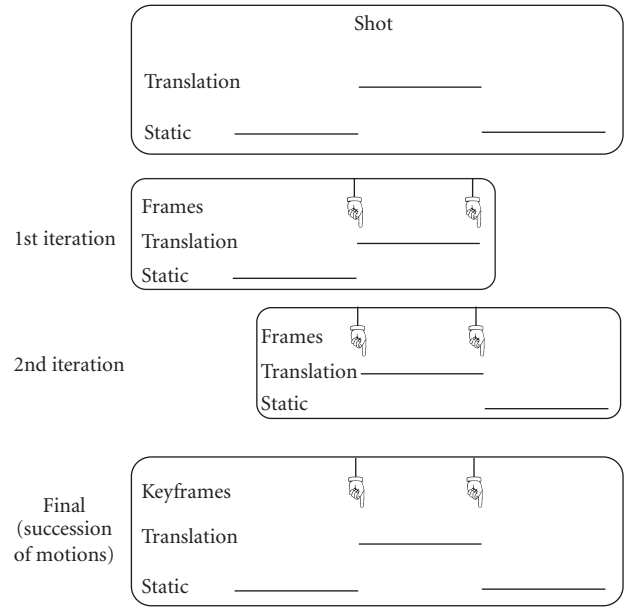


Figure 4: Illustration of keyframe selection. The first iteration corresponds to the process of segments 1 and 2. In the same way, the second iteration corresponds to the succession of segments 2 and 3. Keyframe selection is one frame at the end of the static segment (or beginning of the translation segment) and one frame at the end of the translation segment (or at the beginning of the last segment).

great (i.e., higher than threshold δ_{ec}), the first and the last frames of the segment are selected. In the opposite case, only the last frame is selected.

After an experimental study, we chose the following thresholds: $\delta_r = 0.5$, $\delta_{td} = 300$, and $\delta_{ec} = 5$. Keyframe selection according to camera motion magnitude is summarized in Figure 5.

2.2.3. *Keyframe selection according to succession and magnitude of camera motions*

Keyframe selection takes into account both the succession and the magnitude of camera motions. We will combine the different rules explained above. First, the identified motions which have a weak magnitude or a weak duration are processed as static segments. If a translation motion of duration T with a total displacement td is detected, the standardized total displacement $td_s = td/T$ is calculated. This is regarded as a static segment if the duration T is shorter than threshold δ_T and if the standardized total displacement td_s is shorter than threshold δ_t. In the same way, a zoom of duration T with an enlargement ec is regarded as a static segment if the duration T is shorter than threshold δ_T and if the enlargement ec is lower than δ_e. In our experiment, the thresholds

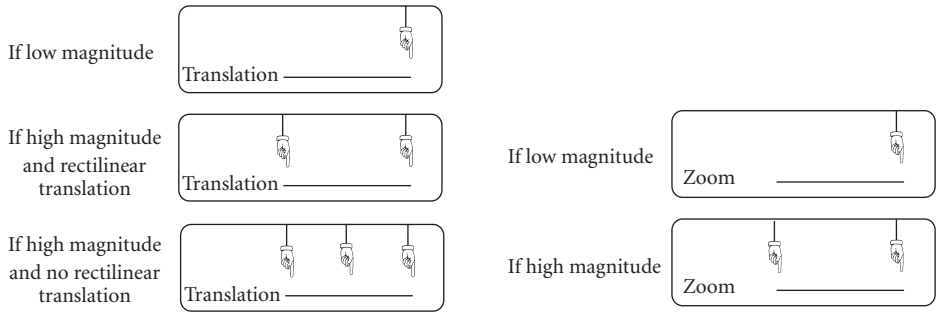


Figure 5: Keyframe selection according to the type and magnitude of camera motions.

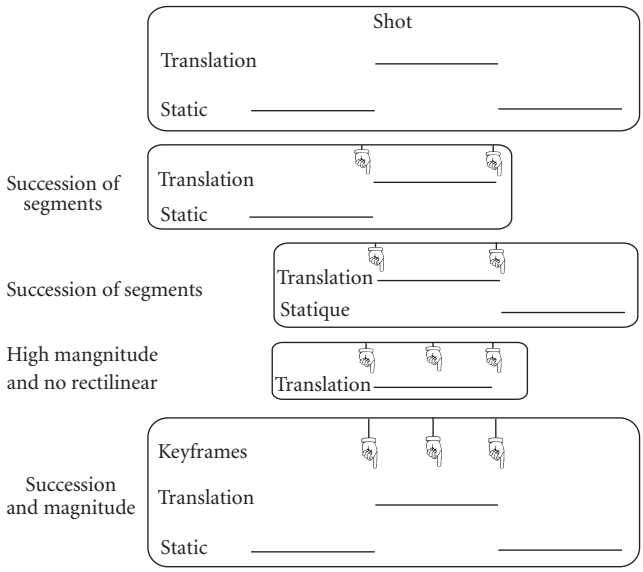
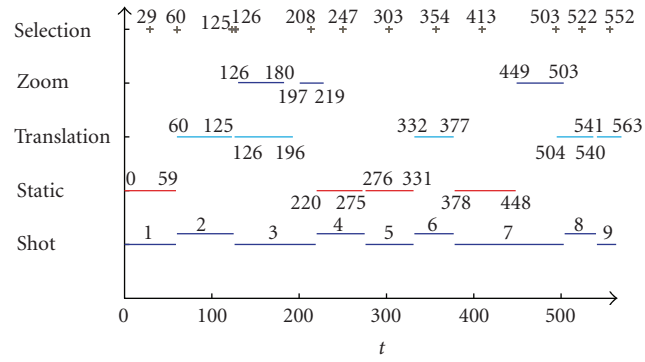


Figure 6: Illustration of keyframe selection according to succession and magnitude of motions.

(a) Sampling of the "baseball" video (1 frame out of 25)



(b) Keyframe selection according to succession and magnitude of motions

(c) Summary of the video "baseball" according to succession and magnitude of motions

Figure 7: Example of video summary made by camera motion-based method.

were fixed in an empirical way at $\delta_t = 1.5$, $\delta_e = 1.8$, and $\delta_T = 50$.

Then, keyframes are selected by applying the rules according to the succession of motions. From the magnitude of motions, frames can be added for the summary. Let us have a look at the previous example with three consecutive detected motions in a shot: static, translation with a strong magnitude and static. Figure 6 illustrates the keyframe selection.

Moreover, in the case of a motion included in another one, if the motion included is of strong magnitude, then the segment containing this motion is described by the frame in the middle of this segment. Lastly, if a shot contains only one camera motion, then the keyframe selection is obtained by applying the rules according to the magnitude of the motions.

Figure 7 illustrates the different steps of the summarization method proposed. It concerns a video sequence named "Baseball," an extract from a baseball match, which has 9 shots (see Figure 7(a)). In Figure 7(b), from the bottom upwards on the y-axis, we have, respectively, the position of the shots, the identification of static segment (absence of motion), translation segment and zoom segment, and finally the selection of the keyframes. For example, $n°1$ shot (from frame 0 to frame 59) is identified as static and the keyframe corresponds to frame 29. In the same way, $n°7$ shot (from frame 378 to frame 503) contains two segments: a static segment (from frame 378 to frame 448) followed by a zoom segment (from frame 449 to frame 503). The keyframe selections for this shot are frames 413 and 503. Figure 7(c)

shows the keyframes used for the summary of the "Baseball" video.

For each shot of the "Baseball" video, the summary created from the succession and the magnitude of camera motions seems visually acceptable and presents little redundancy.

We developed a summary method which exploits the information provided by camera motion. In order to validate this method, we have designed an evaluation method.

3. EVALUATION METHOD OF VIDEO SUMMARIES

Video summarization methods must be evaluated to verify the relevance of the selected keyframes. However, the quality of a video summary is based on subjective considerations. Only the "user" can judge the quality of a summary. In this part, we propose a method to create an "optimal" summary based on summaries created by different people. This "optimal" summary, also called the reference summary, is used as a reference for the evaluation of the summaries provided by various approaches. The construction of a reference summary is a difficult stage which requires the intervention of subjects, but once this summary has been obtained, the comparison with another summary is rapid.

Our evaluation method is similar to that of Huang et al. [18]. Nevertheless, although their evaluation occurs on the video level, their method of building the reference summary is carried out on the shot level. The evaluation method that we propose was developed within a more general framework and provides (i) a reference summary with keyframes selected per shot and (ii) a hierarchical reference summary that takes into account the "importance" of each shot to add weight to the keyframes of the corresponding shot. As the summary from camera motions is proposed on the shot level, we only present the evaluation method on the level of each shot. We will present successively the manual creation of a summary, then the creation of the reference summary and finally the comparison between the reference summary and the automatic summary provided by our camera motion-based method.

3.1. Creation of a video summary by a subject

The goal of the experiment is to design a summary for different videos. We asked subjects to watch a video then to create a summary manually. From the various summaries, a method is proposed to generate the reference summary in order to compare it with the summaries provided by various algorithms.

3.1.1. Video selection

Video selection is an important stage which can influence the results. Two criteria were taken into account: the content and the duration of the video. We chose three videos with varied content and different durations: a sports documentary (called "documentary") with 20 shots and 3271 frames, "the avengers" series with 27 shots and 2412 frames and TV news (called "TV news") with 42 shots and 6870 frames. Each video is made up of color frames (288×352 pixels) displayed at a frequency of 25 frames per second.

It should be noted that these videos are of short duration. The longest lasts approximately 5 minutes. In comparison, the longest video used in [18] has 3114 frames and has a maximum number of 20 shots. The fact of not choosing long videos is linked to the duration of annotation by a subject. It is thus a question of finding a good compromise between a sufficient duration and a reasonable duration for the experiment. In our experiment, the manual creation of a video summary requires between 20 and 35 minutes.

3.1.2. Subjects

12 subjects participated in the experiment. They did the experiment three times (for the three videos). The order of video presentation is random from one subject to another. All the subjects had a normal or corrected to normal vision and they knew the aim of the experiment—the creation of a video summary—but they were not aware of our video summarization method based on camera motion.

3.1.3. Experimental design

The subjects did the experiment individually in front of a computer screen. The experiment is designed using a program written in C/C++ language. Each subject received the following instructions. On the one hand, the summary must be as short as possible and preserve the whole content. On the other hand, the summary must be as neutral as possible. It is thus the subject who distinguishes by himself the degree of acceptance of the summary. The creation of a video summary proceeds in three stages.

1st stage: viewing of the video

In the first stage, the subject viewed the whole video (frames and sound) then he had to give an oral summary in order to make sure that the video content was understood. He viewed the video a second time.

2nd stage: annotation of the video extracts

In the second stage, the video was viewed in the form of extracts presented in chronological order in the top left-hand corner of the screen (see Figure 8). Subject was asked to indicate the degree of importance of each extract. The extracts corresponded to successive shots of the video. They were presented to the subject as extracts and no information was given about the shots. Once the extract had been viewed, the subject specified the degree of importance by indicating if, according to him, this extract was "very important," "important", or "not important" for the summary of the video. The subject clicked on the corresponding notation in the top right-hand corner of the screen. Then, the subject was asked to choose frames to summarize the extract. In the bottom right-hand corner, the frames were presented according to a regular sampling (one frame out of ten). The subject had to select the frames which seemed to be the most representative

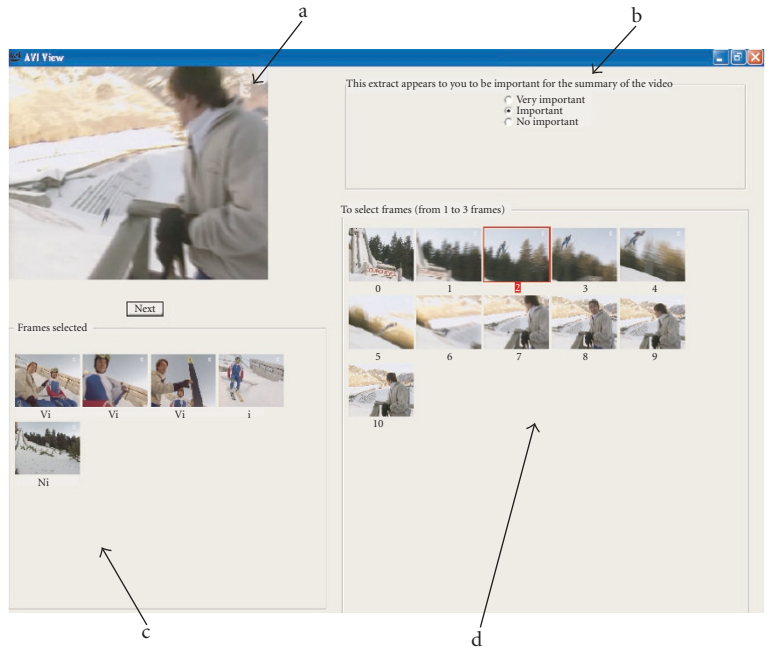


FIGURE 8: Second stage of the reference summary creation for the "documentary" video. The subject had to indicate the degree of importance of the extract in zone b. Then in zone d, he had to select the frames which seemed relevant to him for the summary of the extract presented in zone a. As the frames were displayed with a spatial undersampling by four, the subject could see them with a normal resolution by placing the mouse on a frame of zone d in order for it to appear in zone a. In zone c, the frames already selected from the preceding extracts were displayed to keep a record of the selection.

of the shot (from at least one to three) bearing in mind that the selection had to be as concise as possible and represent the entirety of the content. The maximum number three was selected by preliminary tests. During this stage, when subjects were allowed to choose five keyframes, the majority of them chose fewer than three keyframes per shot, except for some of them who systematically chose five frames to describe even very short shots. Once the subject had finished his annotation for a given extract, he validated it and the results were displayed in the bottom left-hand corner of the screen to keep a record of the annotations already given.

The second stage is illustrated in Figure 8 ("Documentary" video). The subject indicated here if the extract was important for the summary of the video. He also selected one frame (frame $n°2$) to summarize this extract. The annotation of the previous extracts is displayed in the bottom left-hand corner where 5 frames were selected.

Two remarks can be made about this stage. The first concerns the limited number of levels of importance. Only three levels of importance are proposed: "very important," "important", or "no important." A scale with more levels would have made the task more complex and perhaps disconcerting for the subject because of the difficulty of making the difference between levels. The second is about the sampling of the frames of the extract. We chose the sampling of one frame out of ten to avoid displaying the complete shot on the screen, which would render the task of keyframe selection difficult and fastidious. Because of temporal redundancy of the frames, it seemed advisable to carry out this sampling and thus 5 frames displayed on the screen correspond to 2 seconds of the video.

3rd stage: confirmation of the annotations and construction of a short summary

In the third stage, once all the extracts had been annotated, the complete summary was displayed on the screen. The aim is to provide a global view of the summary and to allow the user to modify it and to validate it. Each extract was represented by the chosen frames and the degree of importance was indicated in the lower part of each frame. The subject was asked to modify, if he wished, the degree of importance of the extracts, then to remove the frames which appeared redundant and finally to select only a limited number of frames. The purpose of this stage is to provide a hierarchical summary with a fine level on a shot scale and a coarser level on the scale of the video.

In order to understand the experiment, a training phase is carried out with a test video with 5 shots and 477 frames.

3.2. Construction of a reference summary

The difficulty consists in creating a reference summary from the summaries created by various subjects. On the assumption that the summaries of subjects have a semantic significance, an "optimal" summary has to be built which takes into account these various summaries. Nevertheless, the differences between summaries are not measured by applying

a distance between the frame descriptors since the gap between low-level descriptors and semantic content has not yet been bridged. The process is based on elementary considerations to create the optimal summary. We develop two methods to create a reference summary, one designed for each shot called "fine summary" and the other created from comparison between shots called "short summary." As the summary method from camera motions provides the keyframes for each shot, we only present the fine summary in this paper.

The construction of summary on the shot level is carried out only from the annotations of stage 2 . As already mentioned above, each extract viewed corresponds to a shot, and only the frames chosen by the subjects will be examined and not the degrees of importance of the shots. As the possible number of frames selected varies from one subject to another, the optimal number of keyframes must be given to represent an extract. The arithmetic mean could be used to determine the optimal number. Nevertheless, as the mean is influenced by a typical data, the median is privileged because of its robustness.

Once the number of keyframes has been found, it is necessary to determine how the frames chosen by the various subjects are distributed on a given level. Nevertheless, the temporal distribution of the frames is not enough, since it is not possible to take into account the temporal neighbourhood of frames. As frames were sampled one out of ten, two neighbouring frames can be selected by various subjects and can have the same content. Moreover, it is also necessary to differentiate the subjects who selected a few frames from those who selected many. According to the number of frames chosen by a subject for a given shot, a weight is given to each frame. If only one frame is selected for a given shot, the weight associated with the frame is worth three, whereas if three frames are chosen, the weight of each frame is equal to one. This strategy ensures an average weight by shot which is equal for each subject. This remains coherent with the fact that if a subject chose many frames, they would have a weak weight and inversely.

In order to take into account the neighborhood of the selected frame, a Gaussian, centered on the frame and with a standard deviation σ, is positioned according to a temporal axis. The magnitude of Gaussian is according to the weight given above. If the subject chose, for example, only one frame to represent the shot, then only one Gaussian was placed on the temporal axis with a magnitude of three. The standard deviation is an important parameter for the creation of the reference summary. The greater this parameter is, the more frames selected by the different subjects will be combined. Figure 9 shows how the weight of the close frames varies according to the parameter σ. As the frames to be chosen were displayed according to a regular sampling, the weight of the close frame depends directly on this parameter and is located at index 10. For example, if $\sigma = 20$ then the weight of the close frame is worth 0.88.

After accumulation of the answers, we obtain the temporal distribution of selected frames. Figure 10 shows the results for the "documentary" sequence. We can note for example that the first shot is very long and has many local maxima

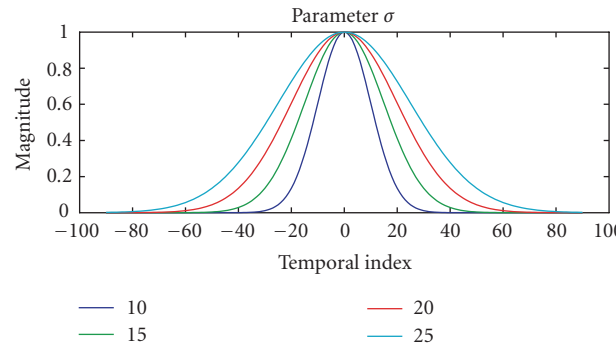


Figure 9: Parameter σ according to the frame chosen by the subject. The Gaussian is positioned on the selected frame. For example, if the parameter $\sigma = 10$, then the close frame (on the left or on the right) has a weight of 0.6 and the following frame has a weight of 0.13, since the frames are displayed according to a regular sampling (all ten).

whereas the second shot has one maximum. The maxima symbolize the locations where the frames must be selected to summarize the video, since these locations are chosen by the subjects. We obtain the maxima by calculating the first derivative and by finding the changes of sign. They are sorted by decreasing order. The close local maxima are combined to avoid the presence of local maxima on a window lower than 2 seconds (or 50 frames). Moreover, all local maxima whose magnitude is lower than 20% of the global maximum are removed.

Finally, for each shot, we retained only the n first local maxima sorted by descending order according to the optimal number of frames required. They correspond to the keyframes selected to summarize the shot and thus the video. The chosen parameter σ is explained with the description of our results.

3.3. Comparison between the automatic summary and the reference summary

The comparison between the reference summary and the automatic summary obtained by an algorithm, called candidate summary, is a delicate task since it requires the comparison of frames. The process of comparison between the reference summary and the candidate summary for the shots is carried out in 4 stages. Figure 11 illustrates the comparison of the summaries for each shot. We can note in this example that the reference summary has 3 keyframes whereas the candidate summary has 4.

The first stage consists in determining the frames of the reference summary with which each frame of the candidate summary could be associated. Each candidate frame is thus associated if possible with two frames of the reference summary, which are temporally the closest frames in the same shot. For example, frame B of the candidate summary is associated with frames 1 and 2 of the reference summary (see Figure 11(a)). On the other hand, frame A is only associated with frame 1, because it is the first frame of the shot.

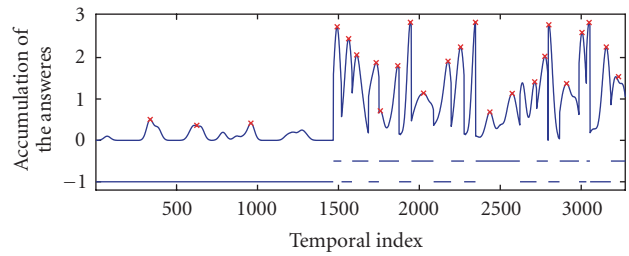


Figure 10: Distribution of keyframe selection on the "documentary" video standardized by the number of subjects (horizontal axis corresponds to the frame number). The maxima on this curve gives the selection of keyframes. The crosses on the curve are the frames chosen to summarize the video. The curve at the bottom corresponds to the staircase function between -0.5 and -1 that locates the changes of shot. In this example, the parameter σ is fixed at 20.

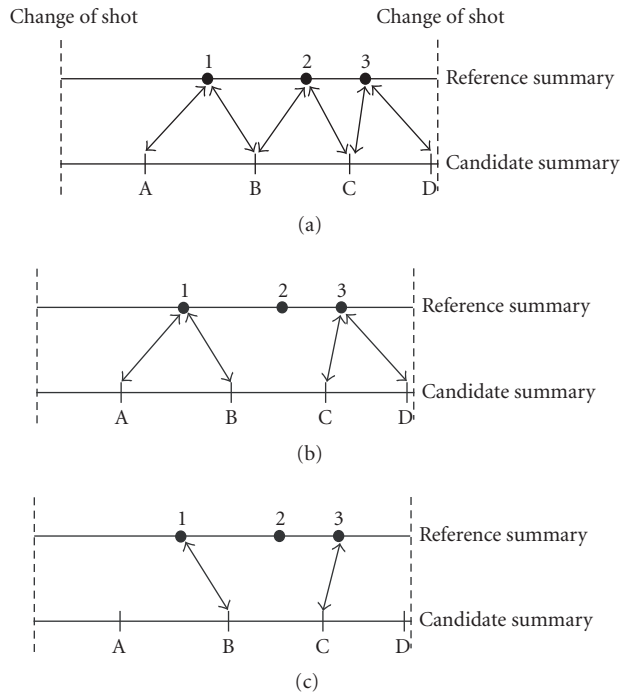


Figure 11: Illustration of the comparison for each shot between the reference summary and the candidate summary. The reference summary has 3 frames (from 1 to 3) whereas the candidate summary presents 4 frames (of A with D). (a), (b), and (c) represent the first three stages of the comparison.

The second stage consists in determining the most similar frame to the frame of the candidate summary among the two potential frames of the reference summary. For example, frame B which can be associated with either frame 1 or 2 is finally associated with frame 1 (see Figure 11(b)) because it is assumed to be closer in terms of content. This requires the representation of frames by a descriptor and the definition of a distance between two frames. Nevertheless, it is difficult to compare the content of two frames. However, as the frames belong to the same shot, there is a temporal continuity between the frames and the comparison between the frames can be carried out by comparing their color histograms. Indeed, two similar histograms will have the same content since the frames are temporally continuous. Inside the same shot, the probability that two similar histograms correspond to different frame contents is very low. The descriptor used here is a global color histogram obtained in color space YCbCr and the distance between histograms is obtained by the L1 norm. We chose not to present a color histogram, as it is not essential to understand the method. However, a detailed description can be found in [19].

The third stage deals with the case where several frames of the candidate summary are associated with the same frame of the reference summary. For example, frames A and B are associated with the same frame 1 (see Figure 11(b)), and finally, only frame B is associated with frame 1 (see Figure 11(c)) since the distance between frames 1 and B is assumed to be weaker.

Lastly, the fourth stage consists in preserving only the clustering where the distances are lower than a threshold δ_s. The frames which were gathered can have great distances. Thresholding makes it possible to preserve only the frames gathered with similar content. The parameter δ_s is fundamental and will be largely studied in the presentation of the results.

The comparison between the reference summary and the candidate summary leads to the number of frames gathered. The standard measures Precision (P), Recall (R), and F_1 (F_1 is a harmonic mean between Recall and Precision) can then be used to evaluate the candidate summary.

3.4. Evaluation of automatic summary

As the summary method from camera motion provides a shot-level summary, we only study the evaluation method on the shot level. Five methods of creating summaries are tested: four are elementary summarization methods and one is our summarization method. For the first method, a number of keyframes is chosen randomly (between 1 and 3) for each shot, then the keyframes are chosen randomly (random summary). For the second method, keyframes are chosen randomly in each shot, but the number of keyframes is defined by the reference summary (semirandom summary). For the third method, only one keyframe is selected in the middle of each shot (center summary). For the fourth method, keyframes are selected with a regular sampling rate as a function of the shot length (one keyframe per 200 frames) (regular sampling summary). Finally, the last one is the one that we proposed using camera motion (camera motion-based summary).

It is important to note that the third method is classically used in the literature. The second one is, in practice, unfeasible. In fact the reference summary is not known, so the number of keyframes to be selected in each shot is unknown. This method might offer good candidate summaries, because they have the same number of keyframes as the reference one.

Table 1 recapitulates the evaluation of the five video summarization methods. As we can see, the method that we

Table 1: Results of the four summarization methods for the three videos. The threshold δ_s of clustering between two frames is fixed at 0.3 and the parameter σ is 20 (R: Recall, P: Precision, F_1). $n°1$: random summary, $n°2$: semirandom summary, $n°3$: summary by selecting the frame in the center of each shot, $n°4$ summary based on a regular sampling, and $n°5$ summary based on camera motion.

Summary	Documentary			TV news			Series		
	R	P	F_1	R	P	F_1	R	P	F_1
$n°1$	62 (15/24)	40 (15/37)	49.1	83 (46/55)	50 (46/91)	63.0	80 (24/30)	40 (24/59)	53.9
$n°2$	54 (13/24)	54 (13/24)	54.1	72 (40/55)	72 (40/55)	72.7	76 (23/30)	76 (23/30)	76.6
$n°3$	50 (12/24)	60 (12/20)	54.5	63 (35/55)	83 (35/42)	72.1	73 (22/30)	78 (22/28)	75.8
$n°4$	62 (15/24)	54 (15/28)	57.7	69 (38/55)	70 (38/54)	69.7	73 (22/30)	73 (22/30)	73.3
$n°5$	79 (19/24)	55 (19/34)	**65.5**	80 (44/55)	77 (44/57)	**78.5**	86 (26/30)	72 (26/36)	**78.7**

propose according to the succession and the magnitude of motions provides the best results (in term of F_1) for the three videos. For the "series" video, methods $n°2$, $n°3$, and $n°4$ present close results compared to the method according to the magnitude and the succession of motions. This confirms that the methods which select only one frame by shot (either a frame in the middle of the shot or at a random location in the shot) are relatively effective when the shots are of short duration. The "series" video contains 16 shots out of 28 of less than 3 seconds whereas the "documentary" and "TV news" video have, respectively, 8 shots out of 20 and 9 shots out of 42 of less than 3 seconds. It is indeed natural to select only one frame for these shots. However, the results for the three videos confirm the interest of using camera motion to select frames. The longer the shots are, the more likely the contents are to change and thus the more effective the method is.

However, the comparison method of summaries requires various parameters to be fixed which can influence the results. In the method of reference summary construction, the parameter studied is the standard deviation of Gaussian σ around the frame chosen by a subject. Indeed, if the parameter σ selected is low, then the close frames selected by the subjects cannot be combined. In the same way, if the parameter σ selected is large, then the frames will be gathered easily. Thus, the number of local maxima inside a shot depends on this parameter σ. Figure 12 illustrates the results of the summarization method with the keyframe selection in the center of the shot, and the method using succession and magnitude of motions according to parameter σ. Moreover, the results of the two methods presented remain relatively stable according to parameter σ. We can also note that the number of keyframes of the reference summary for the three videos does not decrease greatly with the increase of parameter σ. Thus, we can conclude that this parameter σ does not call into question the performance of the methods. Thereafter, this parameter σ will be fixed at 20.

Lastly, with regard to the comparison between the reference summary and the candidate summary, although the description of the frames is carried out by color histogram, clustering between frames is preserved only if the distances are lower than the threshold δ_s. However, this threshold plays an important role in the results. Indeed, if the threshold selected is rather low, then the frames will be gathered with difficulty, whereas if the threshold is too large, the dissimilar frames can be matched together. Figure 13 illustrates the results of various methods according to threshold δ_s. As expected, the more the threshold increases, the more the performances increase (up to a certain value). Nevertheless, whatever the threshold selected, the method according to the succession and the magnitude of motions presents the best results for the "documentary" and "TV news" videos. With regard to the "series" video, the most competitive method is that based on the magnitude and the succession of motions for thresholds 0.1, 0.2, 0.3, and 0.4. On the other hand, for thresholds 0.5 and 0.6, the summarization method with the frame in the center of the shot is more competitive. Generally, the performances obtained for thresholds 0.5 and 0.6 are fairly similar for the same video. That means that parameter δ_s is too high and that dissimilar frames can be gathered. Parameter δ_s should be selected inferior to 0.5 because the slope is nonnull.

4. CONCLUSION

In this paper, we have presented an original video summarization method from camera motion. It consists in selecting keyframes according to rules defined on the succession and the magnitude of camera motions. The rules we used are "natural" and aim to avoid temporal redundancy between frames and at the same time to keep the whole content of the video. The camera motion brings "high-level" information; in fact the camera motion is desired by the film maker and contains some cues about the action or an important location in a scene. The keyframe selection is directly based on the camera motion (succession and magnitude) and offers the advantage of not calculating differences between frames as it was done in other research.

A new evaluation method was also proposed to compare the different summaries created. A psychophysical experiment was set up to make it possible for a subject to create manually a summary for a given video. Twelve subjects summarized three different videos (duration from 1.5 to 5 minutes). A protocol was designed to combine these twelve summaries into a unique one for each video. This reference summary provided us with the "ideal" or "true" summary.

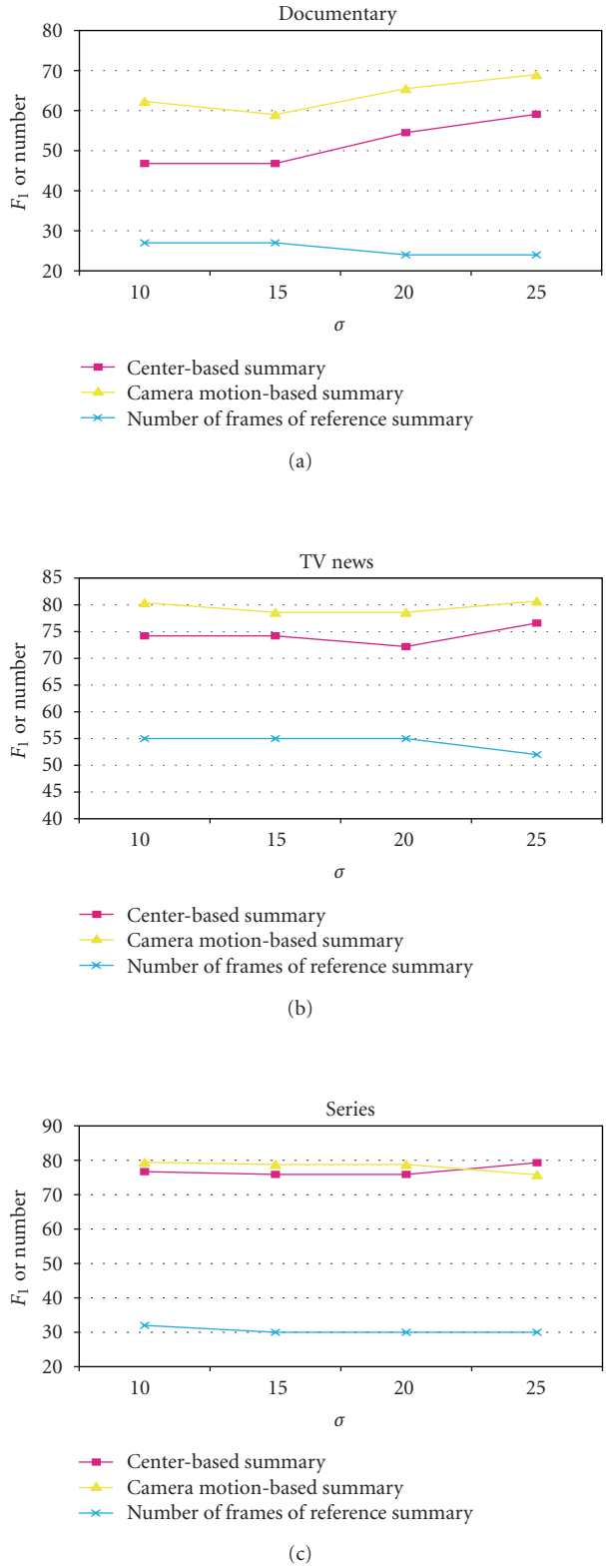


FIGURE 12: F_1 as a function of the parameter σ for two summarization methods (summaries by selecting the center of each shot and based on camera motion) for three videos. The threshold σ_s is fixed at 0.3. The third curve, at the bottom of each figure, corresponds to the number of keyframes for the reference summary as a function of the parameter σ.

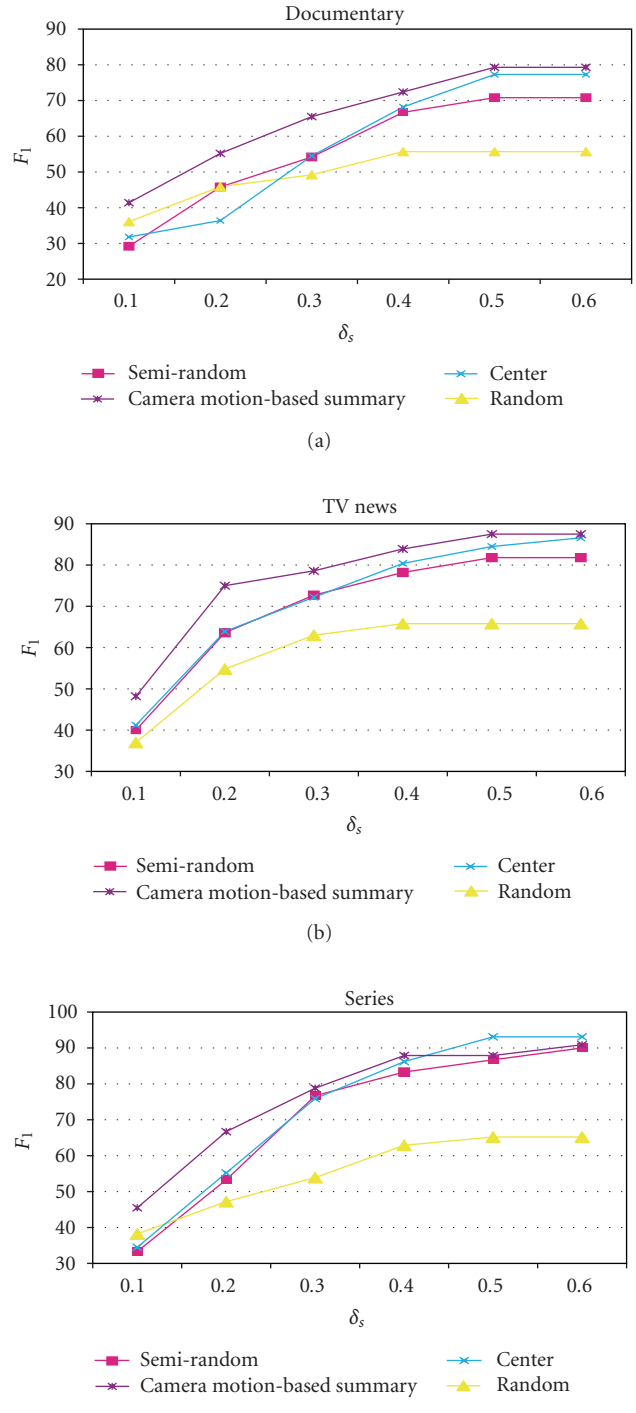


FIGURE 13: F_1 as a function of the parameter σ_s for four summarization methods and for the three videos. The parameter σ is fixed at 20.

Finally, we proposed an automatic comparison between this reference summary and the summary built by our method. This method can also be used to compare different kind of summaries, with different lengths.

One of the future lines of investigation would be to create what we previously called a hierarchical summary. This

hierarchical summary would be based on our camera motion-based summary (per shot) and would include some criteria to measure the relative importance of each shot. This new criteria would be for example the magnitude of motion in a segment or for static segment, the relative "interest" of the segment. The relative interest can be described by a biological model of saliency. A "degree of importance" could be linked to each shot and the keyframes of the shot (selected by the camera motion) would be weighted with this index of "importance." A hierarchical summary can be easily evaluated with our subjective evaluation method. In fact, with this method, we already have access to the "important" information for each shot.

ACKNOWLEDGMENTS

The authors would like to thank C. Marendaz and D. Alleyson (Laboratoire de Psychologie et NeuroCognition, Grenoble, France) for having welcomed and helped them with the experiments. They also would like to thank S. Marat for her help in testing some summarization methods.

REFERENCES

[1] S. Kopf, T. Haenselmann, D. Farin, and W. Effelsberg, "Automatic generation of video summaries for historical films," in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME '04)*, vol. 3, pp. 2067–2070, Taipei, Taiwan, June 2004.

[2] Y.-F. Ma and H.-J. Zhang, "Video snapshot: a bird view of video sequence," in *Proceedings of the 11th International Multimedia Modelling Conference (MMM '05)*, pp. 94–101, Melbourne, Australia, January 2005.

[3] X. Zhu, A. K. Elmagarmid, X. Xue, L. Wu, and A. C. Catlin, "InsightVideo: toward hierarchical video content organization for efficient browsing, summarization and retrieval," *IEEE Transactions on Multimedia*, vol. 7, no. 4, pp. 648–666, 2005.

[4] M. Cherfaoui and C. Bertin, "Two-stage strategy for indexing and presenting video," in *Storage and Retrieval for Image and Video Databases II*, vol. 2185 of *Proceedings of SPIE*, pp. 174–184, San Jose, Calif, USA, February 1994.

[5] K. A. Peker and A. Divakaran, "An extended framework for adaptive playback-based video summarization," in *Internet Multimedia Management Systems IV*, vol. 5242 of *Proceedings of SPIE*, pp. 26–33, Orlando, Fla, USA, September 2003.

[6] A. Kaup, S. Treetasanatavorn, U. Rauschenbach, and J. Heuer, "Video analysis for universal multimedia messaging," in *Proceedings of the 5th IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI '02)*, pp. 211–215, Sante Fe, NM, USA, April 2002.

[7] S. V. Porter, M. Mirmehdi, and B. T. Thomas, "A shortest path representation for video summarisation," in *Proceedings of the 12th International Conference on Image Analysis and Processing (ICIAP '03)*, pp. 460–465, Mantova, Italy, September 2003.

[8] B. Fauvet, P. Bouthemy, P. Gros, and F. Spindler, "A geometrical key-frame selection method exploiting dominant motion estimation in video," in *Proceedings of the 3rd International Conference on Image and Video Retrieval (CIVR '04)*, pp. 419–427, Dublin, Ireland, July 2004.

[9] I. Yahiaoui, B. Mérialdo, and B. Huet, "Automatic video summarization," in *Multimedia Content-Based Indexing and Retrieval (MMCBIR '01)*, Rocquencourt, France, September 2001.

[10] G. Ciocca and R. Schettini, "Dynamic key-frame extraction for video summarization," in *Internet Imaging VI*, vol. 5670 of *Proceedings of SPIE*, pp. 137–142, San Jose, Calif, USA, January 2005.

[11] S. Corchs, G. Ciocca, and R. Schettini, "Video summarization using a neurodynamical model of visual attention," in *Proceedings of the 6th IEEE Workshop on Multimedia Signal Processing (MMSP '04)*, pp. 71–74, Siena, Italy, September-October 2004.

[12] A. M. Ferman and A. M. Tekalp, "Two-stage hierarchical video summary extraction to match low-level user browsing preferences," *IEEE Transactions on Multimedia*, vol. 5, no. 2, pp. 244–256, 2003.

[13] X. Shao, C. Xa, and M. S. Kankanhalli, "A new approch to automatic music video summarization," in *Proceedings of IEEE International Conference on Image Processing (ICIP '04)*, vol. 1, pp. 625–628, Singapore, October 2004.

[14] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li, "A user attention model for video summarization," in *Proceedings of the 10th ACM International Conference on Multimedia*, pp. 533–542, Juan-les-Pins, France, December 2002.

[15] S. Lu, M. R. Lyu, and I. King, "Video summarization by spatial-temporal graph optimization," in *Proceedings of International Symposium on Circuits and Systems (ISCAS '04)*, vol. 2, pp. 197–200, Vancouver, Canada, May 2004.

[16] C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang, "Automatic video summarization by graph modeling," in *Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV '03)*, vol. 1, pp. 104–109, Nice, France, October 2003.

[17] M. Guironnet, D. Pellerin, and M. Rombaut, "Camera motion classification based on transferable belief model," in *Proceedings of the 14th European Signal Processing Conference (EUSIPCO '06)*, Florence, Italy, September 2006.

[18] M. Huang, A. B. Mahajan, and D. DeMenthon, "Automatic performance evaluation for video summarization," Tech. Rep. LAMP-TR-114, CAR-TR-998,CS-TR-4605,UMIACS-TR-2004-47, University of Maryland, College Park, Md, USA, June 2004.

[19] M. Guironnet, D. Pellerin, and M. Rombaut, "Video classification based on low-level feature fusion model," in *Proceedings of the 13th European Signal Processing Conference (EUSIPCO '05)*, Antalya, Turkey, September 2005.