Multimed Tools Appl
DOI 10.1007/s11042-009-0353-1

A novel tool for summarization of arthroscopic videos

**Mathias Lux · Oge Marques · Klaus Schöffmann ·
Laszlo Böszörmenyi · Georg Lajtai**

Abstract Arthroscopic surgery is a minimally invasive procedure that uses a small camera to generate video streams, which are recorded and subsequently archived. In this paper we present a video summarization tool and demonstrate how it can be successfully used in the domain of arthroscopic videos. The proposed tool generates a keyframe-based summary, which clusters visually similar frames based on user-selected visual features and appropriate dissimilarity metrics. We discuss how this tool can be used for arthroscopic videos, taking advantage of several domain-specific aspects, without losing its ability to work on general-purpose videos. Experimental results confirm the feasibility of the proposed approach and encourage extending it to other application domains.

Keywords Multimedia tools · Video summarization · Arthroscopic videos

1 Introduction

Digital imaging has provided surgeons with new, powerful tools that enable a vast number of applications. In the domain of arthroscopy, a small camera—built as part of a device

M. Lux (✉) · K. Schöffmann · L. Böszörmenyi
Institute for Information Technology, Klagenfurt University, Universitätsstrasse 65–67, 9020 Klagenfurt, Austria
e-mail: mlux@itec.uni-klu.ac.at

K. Schöffmann
e-mail: ks@itec.uni-klu.ac.at

L. Böszörmenyi
e-mail: laszlo@itec.uni-klu.ac.at

O. Marques
Department of Computer Science and Engineering, Florida Atlantic University, 777 Glades Road,
Boca Raton, FL 33431, USA
e-mail: omarques@fau.edu

G. Lajtai
Private Clinic Althofen, Moorweg 30, 9330 Althofen, Austria
e-mail: laj@shoulder.org

Published online: 18 September 2009

called *arthroscope*—can be used to serve as the "eyes of the surgeon" as they diagnose or operate on a patient. The resulting video stream can be recorded for several purposes, such as: inspection and in-depth diagnosis, comparison of different diagnoses/surgeries, explanations to the patients, and training of surgeons.

Arthroscopic videos have certain distinguishing characteristics and requirements, which bear implications for video analysis and summarization. For example, the video coding method and visual quality of video streams are constrained by restrictions on the sensor, the sterile environment of the operating room and the requirement that surgical instruments need to be error-free during the course of a surgery. The way by which the video is acquired, encoded, and stored—using an expensive proprietary equipment—cannot be changed. Additionally, the recorded videos are typically short (60–70 s in average), as the surgeon only records the most important parts of the surgery. Despite the fact that only a small set of videos per arthroscopy is recorded for archival purposes, the total number of stored videos quickly builds up, as a result of the large number of surgeries performed by a surgeon. Consequently, the video archives of arthroscopists grow to a size for which multimedia information management is crucial.

A first step for multimedia information management is to provide *video summaries* (also called *video abstracts*) for video streams. While in general both types of video summaries known in literature could be used for that purpose, namely dynamic summaries (i.e. video skims) and static summaries, the latter ones are usually suited better for that domain. The reason is that static summaries can be easily attached to a surgery report or a medical history of a patient, and they also can be easily stored on a chip card, commonly used in the medical domain. These static summaries may consist of a set of keyframes, whose goal is to provide insight on the video without the need to watch it in its entirety. Ideally, these summaries should also exclude frames that are not appropriate, e.g., frames that do not show surgical instruments, frames that do not show the areas of interest, or frames that have too much motion, resulting in a blurred still image.

In this contribution we present a novel tool for video summarization of arthroscopic videos that takes advantage of several domain-specific heuristics and yet is flexible enough to be used in other domains.

The main features of the tool are:

- Summarization is done by clustering similar keyframes and selecting a representative keyframe from the largest clusters.
- Visualization of the video summary is based on representative keyframes from the largest clusters (one keyframe per cluster), with graphical indication of where the frames belonging to that cluster occur in time.
- Shot detection techniques are not needed, since temporal dependency is not a critical requirement for meaningful summarization.
- Visually similar keyframes are compared based on low-level features and dissimilarity metrics that can be selected by the user, if desired.

The paper is structured as follows. After reviewing relevant related work in video summarization (Section 2), we describe the domain of arthroscopy and arthroscopic videos and summarize the main characteristics of the domain that are relevant from a multimedia research standpoint (Section 3). Section 4 describes our approach for keyframe selection for video summaries in this specific domain, whose goal is to find keyframes that describe the video in an optimal way. Section 5 presents a qualitative evaluation of the tool for the specialized domain of arthroscopic video and compares the results against an evaluation of the same keyframe selection approach in another, non-technical, domain (short animation videos). Finally (Section 6), we identify research challenges and opportunities in arthroscopic videos and point out steps for further work.

2 Related work

The use of imaging and video equipment—and associated software—by arthroscopic surgeons has been growing over the past years. Arthroscopy specialists are being educated on—and encouraged to use—image editing (e.g., crop, rotate), enhancement (e.g., brightness, contrast, and color adjustments), annotation, and retrieval tools and techniques [9, 18], but there has not been—to our knowledge—a systematic study of video summarization tools and needs for arthroscopic videos. It appears to us that doctors and surgeons are being taught to use existing tools, understand and maximize their potential, and be creative about their use, without taking the extra step of helping design a tool that could be customized to serve their specific needs and overcome some of the limitations they may have faced while using their current tools. One first step towards the creation of such a tool is the development of video summarization methods especially suited for the very uncommon domain of arthroscopic videos. While this specialized domain has not yet been covered in the literature of video summarization, several general approaches for video summarization have been proposed during the past 15 years. The remainder of this section gives a brief overview of existing methods and roughly explains how they work.

Money and Agius [16], who recently presented a survey on video summarization, classified existing methods into *internal*, *external*, and *hybrid* ones. While internal methods perform analysis directly on the video stream, external methods use information not directly contained in the video stream (e.g., an MPEG-7 manual annotation), and hybrid methods use a combination of both. In the following we concentrate on internal methods which are by far the most common ones.

Video summarization methods can be first classified by the features they use for analysis. In general, summarization methods use either image features, audio features, textual features, or a combination of multiple features (these are called *multi-modal* methods). Video summarization can be further classified into *domain-specific* and *non-domain specific*. Moreover, a classification can be made based on the presentation of the summary. While *static methods* use representative keyframes (e.g., in a storyboard visualization), *dynamic methods* use video skims (e.g., a slide-show of keyframes or an extracted segment). A static presentation has the advantage that a user can more quickly watch the entire summary, while a dynamic presentation may allow a more comprehendible summary not only because usually audio playback is also available. In addition, *interactive video summarization* methods allow a user to selectively see parts of the summary according to a query.

Truong and Venkatesh [21] also considered keyframes and video skims as the two basic forms of video summaries. However, they conclude that the optimal visualization of summarized content remains an open question. The lack of a consistent evaluation framework is complained, leading to proprietary evaluation methods in the different visualization approaches. Starting in 2003 the TRECVID evaluation meetings [17] (which were already part of TREC [20] in 2001 and 2002) focus on content-based retrieval from digital video via open, metrics-based evaluation.

Ciocca and Schettini [4] use shot transitions (e.g. cut, fade-in, fade-out, etc.) to distinguish between *informative shots* and *uninformative shots*. They concentrate on informative shots only and use a frame difference measure based on color histogram, edge direction histogram, and a wavelet statistic to find the keyframe of a shot. The concept of *mutual information* (MI), representing the information changes between consecutive frames of a video sequence, is used by Cernekova et al. [1] to perform shot-boundary detection and to select keyframes from shots for the purpose of summarization. More precisely, they use a probabilistic model of gray-level changes of a pixel in adjacent frames to compute the MI.

Xu et al. [22] used several audio features with a SVM (Support Vector Machine) classifier to find keywords according to a particular audio event in soccer videos (e.g. whistling, commentator speech, and audience sound).

Matos and Pereira [15] used multimodal analysis to create personalized video summaries based on MPEG-7. Their system models the *users arousal*, similar to the model already proposed by Hanjalic and Xu [6], based on three features: (1) *motion intensity*, (2) *density of shot cuts* and (3) *sound energy*. According to the arousal curve, frames are classified into *top highlights*, *key points*, and *extended summary*. The *HierarchicalSummary* descriptor of MPEG-7 [10, 14] is used to store information about segments of these three classes. The MPEG-7 file is further used as an input to the summary generation process, where a user can partially control the generation of the summary by specifying which segments should be contained or how long the generated summary should be.

3 Arthroscopy: a short overview

Arthroscopy is a technique that enables a surgeon to examine directly into a joint, by inserting a specially designed device (called an *arthroscope*) into the joint through a small incision. The procedure of arthroscopy can be used for diagnosis or actual repair of the joint; the latter is called *arthroscopic surgery*. Arthroscopic surgery is—compared to open surgery—minimally invasive. The arthroscope and the surgical instruments are inserted through small incisions. Medical instruments and probes are used to manipulate bone and tissue and the arthroscope acts as “the eye of the surgeon”, allowing them to assess the area, observe the effect and position of the instruments, and adjust them accordingly.

Arthroscopy is typically used to perform surgery on joints, especially knees, shoulders, hands and feet. For each type of arthroscopy only a limited number of different surgical tasks are commonly performed. For shoulder arthroscopy, which is our use case for evaluation in Section 5.2, typical surgical procedures are[1]:

1. Diagnosis: The arthroscope is used to examine the joint.
2. Removal of bone and tissue: There are standardized common procedures where bone and tissue are removed using an electro scalpel, a drill or a burr.
3. Fixation of the labrum in the joint: Grasps, anchors, osteoms and fibers are used to fix the labrum.

The arthroscope itself (Fig. 1) is a tube containing a set of optic fibers. At the head end, a digital camera and a light source are attached. At the tail end (the one inserted into the patient’s body), rod lenses magnify the image and bend the light (typically by 30°) to one side. The figure also shows the inlet and outlet for an irrigation fluid that flows under pressure through the patient’s body. This constant flow of irrigation fluid serves two main purposes: (i) it creates an artificial cavity inside the patient’s body; and (ii) it cleans out removed tissue and blood. The view inside the patient’s body is illustrated by the dashed lines and is bent 30° from the main axis of the arthroscope tube. This allows the surgeon to alter (expand) the field view by rotating the arthroscope (not the camera). The camera

[1] For an in-depth explanation of surgical techniques and instruments employed in arthroscopy refer to [11]

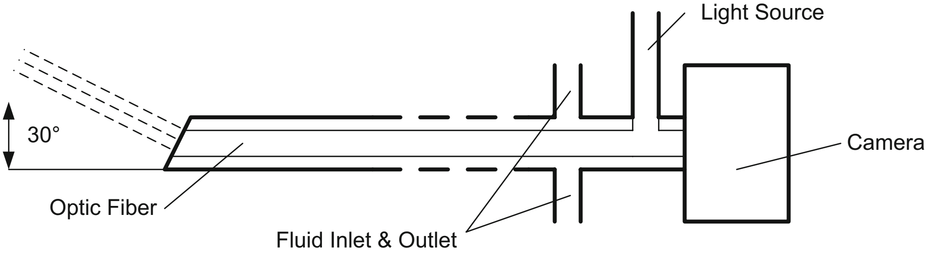


Fig. 1 Simplified schematic view of an arthroscope. The tail end, which is inserted in the patient's body, and the direction of vision are shown on the left-hand side. The head end with the camera is illustrated on the right-hand side

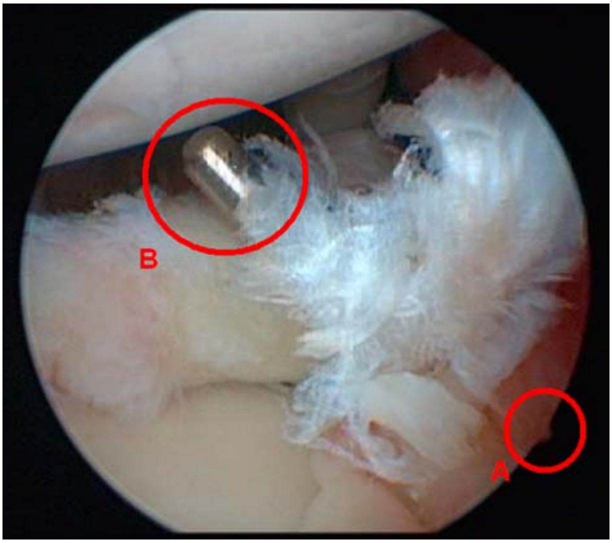
attached to the arthroscope is connected to a screen, where the surgeon can observe the advance of the surgery inside the patient's body.

3.1 Arthroscopic videos: visual and semantic properties

Due to the setup of the arthroscope's camera, the resulting video frames are of circular shape—with part of the circle cut out at the top and the bottom—against a dark background. Figure 2 shows a sample frame from an arthroscope's camera video stream, which can be used to highlight two characteristic aspects of this type of video:

1. The presence of a small spike in the right lower corner of the figure, marked by circle A. This spike indicates the direction of the bend in the vision through the arthroscope (the above mentioned 30° angle).
2. The presence of blown-out highlights in the video frame, which cause clipping of the video signal. Clipping occurs either when the light is reflected by an object in the scene (e.g. instrument, cartilage), or the exit of the light is too close to the subject of illumination. Circle B marks an example of clipping.

Fig. 2 Still frame from an arthroscopic surgery video stream. The fiber is inserted through the instrument marked with *circle B*. The spike in *circle A* indicates the direction of the bend in the vision through the arthroscope



Both aspects have implications from a video analysis viewpoint: spikes can be used as markers (or reference points), in a way that resembles the use of control points in image registration; clipping removes the color information from a portion of the image, resulting in an artificially white area, which is disadvantageous especially for color-based analysis and matching.

A beneficial characteristic of arthroscopic videos, from a video analysis perspective, is the limited number of possible camera movements and associated viewport changes. Due to the nature of the arthroscope, pan and zoom are not possible. Insertion of the arthroscope is only possible through standardized portals to prevent soft tissue or bone damage. Consequently, there are only two major situations where the viewport is changed significantly:

1. *Rotation of the light cable*: Because of the bend view, rotation of the light cable moves the center of the viewport in a circle around the main axis of the arthroscope tube. Note that neither the arthroscope nor the camera are rotated, just the optic fiber in the arthroscope tube.
2. *Change of portals*: If the visual information at the current point of insertion is not sufficient, the surgeon can insert the arthroscope through another portal. However, in this case the recording of the video stream is stopped and started again after adjusting the arthroscope and therefore a change of portals will not occur within a single video file.

The semantic content of an arthroscopic video is clearly defined by the context and the body part where the procedure is being performed: only a limited number of bones, ligaments, cartilages and muscles can be visualized within the area captured by the camera. The type, quality, and amount of visual information should be appropriate to allow the surgeon to diagnose on the associated pathologic structures. Therefore, an ontology of pathologic structures of the area of interest can be created.

The set of medical instruments used in arthroscopic surgeries is standardized, as well as the circumstances and processes in which they are employed. Medical instruments include probes, hooks, scalpels, drills, stitching devices, grasps and burrs. For a certain type of procedure, the set of possible occurring objects and environments in the video can be known and stored as *domain knowledge* for video analysis purposes.

The process of an arthroscopy is usually recorded in parts. The surgeon decides which part should be archived for documentation and diagnostic purposes. Typically the recorded videos are rather short (see Section 5.2 for statistics on our test data set).

The arthroscopic surgeon has a clear definition of what constitutes a video stream (or its representative keyframes) of good visual quality: the video should provide a clear, sharp, and unobstructed view of the scene of action. Highlight clipping is not an issue for image quality for the surgeon as it typically occurs within a specific object such as the medical instrument in Fig. 2. Relevant factors that impact image quality from a surgeon's point of view include:

- Loose and floating tissue, which might cover parts of the subject of the pathologic structures.
- Blood droplets floating in the cavity before they get sucked out with the irrigation fluid, which might also cover structures of interest and add reddish color to the scene.
- Fast movement of the arthroscope, which results in motion blur in the video.
- Small moves of the surgeon's hands, which result in significant changes of the view, because: (i) the surgeon holds the arthroscope at the head end and therefore virtually steers the arthroscope's tip with a long stick; and (ii) because of the magnification of the scene.

Arthroscopic surgery instruments and devices (cameras, monitors, etc.) are of high quality and accuracy, and manufactured to comply with strict medical and surgical standards. Most currently used arthroscopy cameras generate standard definition (PAL or NTSC) color video streams. It is expected that the use of high definition, three chip cameras, which provide a clearer picture with richer contrast, will become increasingly more popular.

In summary, the domain of arthroscopy videos has very specific characteristics, which are relevant from a multimedia research point of view, such as:

- The domain is very well defined, with a limited number of environments (shoulders, knees, hands and feet), a small number of surgical techniques, and a limited set of objects—such as instruments and cartilage—that may be present (and visible) in the scene.
- The actual image captured by the camera attached to the arthroscope has a circular shape.
- The recorded videos are usually of short duration.
- There are no cuts or transitions in arthroscopy videos. All captured arthroscopy videos are raw data and are not edited.
- The set of possible camera movements is very small: just rotation is possible.
- The viewport is well defined by the portals through which the arthroscope is inserted.
- A small spike at the border of the circular image shows the direction of view of the arthroscope. Therefore the relative direction of the camera can be found from the video stream.
- Image quality semantics (from the surgeon's perspective) are well defined in the domain. A video is considered to be of high quality if it provides a sharp and clear picture with unobstructed view on the subject of operation.
- Moreover, the main reasons for quality reduction (from the surgeon's perspective) are well defined: Quality is mainly reduced by (i) floating tissue overlapping the scene of action, (ii) blood droplets floating in the cavity, (iii) image blurring due to fast camera rotation and (iv) too close capture of a subject such as a bone or a medical instrument.
- Due to the optic fiber in the arthroscope as well as the tiny lens at the tail end, the visual quality of the image is low in terms of contrast and lighting.
- The lighting setup of the arthroscope and the color and reflection of bone and medical instruments lead to highlight clipping and therefore loss of visual information.
- All the content captured by the camera is relevant to the surgical process as the arthroscope is the "eye of the surgeon". Therefore there are no irrelevant parts within a recorded arthroscopy, just less relevant sequences (e.g., noisy or blurry sequences).

4 Summarization of arthroscopy videos

A specialized surgeon performs, in average, 300–500 arthroscopies per year. For each arthroscopy, 3 to 8 videos are captured and recorded, resulting in a minimum of 900 and a maximum amount of 4,000 videos recorded per year. The videos are archived in a way that is interlinked with the patients' records and the report of the surgery during which the video was captured. Consequently, in a multimedia management system, queries on patients' and surgeries' metadata can be answered easily. However, as mentioned in Section 3, the visual quality of a video sequence cannot be guaranteed. Therefore, searching for videos showing specific features of a joint, an extraordinary anatomy, or videos with a high value for training and education is typically achieved through time-consuming browsing tasks, based on trial and error. Furthermore, surgeries are summarized by using still images for: (i) a formal report for the records and (ii) explanations to the patients. These still images show critical parts of the surgery and must be of high visual quality. While these images can be manually extracted from the video stream, the task of browsing the video and selecting keyframes is cumbersome and tedious.

In this paper we present a tool for keyframe selection and video summarization that is particularly well-suited for arthroscopic videos. The proposed video summarization tool takes advantage of several domain-specific characteristics, as follows:

- Arthroscopy videos have no shots. Therefore, since shot detection would not enhance the quality of results, shot detection algorithms are not necessary.
- In arthroscopy, frame similarity is more relevant than temporal order, as the surgical process consists of recurring events, like scrubbing, sawing and cutting, etc. Consequently, our approach highlights relevant keyframes and displays them to the user in a friendly way, which preserves temporal information without assigning excessive importance to the temporal arrangement of the frames.
- Blurred frames and those with visual noise are considered *irrelevant*, but they represent a reasonable portion of the videos spread over different points in the temporal domain. This motivates the idea of a *junk cluster*, where all these irrelevant frames are grouped, regardless of the time in which they appear in the video stream.
- There are recurring events with semantic meaning, e.g., “instrument is visible”. Our summarization tool allows the expert to visualize the most important keyframes and which part of the process they describe, therefore providing an implicit indicator of their meaning.
- The video has a circular shape. We use part of the unnecessary black portions of the rectangular frame for additional visualization hints.
- Camera movement is sparse in most arthroscopic videos, so the surroundings of the area, where the surgical process takes place stays mostly the same. This led to the design decision of presenting only one keyframe in its full size, where the scene is shown in detail.

Our video summarization approach consists of the following steps:

1. Extraction of global features of frames
2. Clustering of frames
3. Composition of the summary image

In the subsequent sections each of the steps is described in detail.

4.1 Feature extraction

For the sake of keyframe selection, an uncompressed input video is interpreted as a sequence of still images. In our approach, a low-level feature vector has to be extracted from each of the images (frames). The algorithms for low-level feature extraction employed for arthroscopic videos were originally made available in an open source Java-based CBIR framework, LIRe [12]. Additional feature extraction methods can be easily integrated by implementing a simple Java interface. In our summarization approach, we employed five different combinations of features and dissimilarity functions:

1. 64-bin RGB color histograms with L1 distance.
2. Tamura global texture features [19].
3. Color and edge directivity descriptor, CEDD, with the Tanimoto coefficient [2].
4. Fuzzy color and texture histogram, FCTH, with the Tanimoto coefficient [3].
5. Auto color correlograms [7] with L1 distance.

The RGB color histogram captures color information only, without taking texture or color distribution into account. The Tamura features are common global texture features, which do not take color into account. CEDD and FCTH are features that combine color and texture. They differ in the fuzzy color quantization scheme and the granularity of texture they capture. FCTH, in particular, is more sensitive to small changes in color and fine textures. Auto color correlograms are histograms of color correlations and capture how often a color occurs in the neighborhood of itself, e.g., how often red pixels are surrounded by other red pixels. Therefore the feature also is a combination of texture and color, whereas not only the amount but also the distribution of colors is captured.

The currently implemented descriptors include traditional feature extraction methods (1, 2 and 5 in the above list) as well as recently proposed ones (3 and 4). Most importantly, however, is the fact that the user of our tool can select the feature (and corresponding dissimilarity metric) that best suits their needs. To the best of our knowledge, this is a novel feature for a video summarization tool.

4.2 Clustering

Since different feature-dissimilarity combinations yield different results for keyframe selection, the most appropriate combination should be selected. Our approach was motivated by the idea that the arthroscopist could assess the appropriateness and quality of video summaries with different combinations of low-level features and dissimilarity measures. To assess different combinations of low-level features and dissimilarities, a clustering algorithm for keyframe selection that works satisfactorily in many different feature spaces is needed.

Under the assumption that for each frame a low-level feature vector exists and that we can compute a pair-wise distance between frames based on the feature vector and a dissimilarity function, we employ a clustering algorithm to assign each frame to one of n clusters, where n is the number of keyframes we want to select. The choice of a clustering algorithm is limited to those that rely on a distance (dissimilarity) measure without imposing additional requirements on the feature space. We chose the k-medoid clustering algorithm, which is a very common partitioning clustering algorithm similar to k-means [8]. The k-medoid approach is applicable to keyframe selection as it has been shown for instance in [5].

The k-medoid algorithm has two main advantages for our application:

1. The cluster centre is always represented by a real data point and not an “artificial cluster centre” (which is the case for instance with the k-means algorithm)
2. The clustering only depends on the result of the dissimilarity function applied to the image feature vectors, and not the feature itself or the feature space.

The resulting n clusters group frames that are visually similar according to the chosen image feature. The clusters’ medoids M_1, M_2, ... M_n minimize the distance to all elements of a cluster. Therefore we interpret the medoids as most descriptive and representative elements for the respective groups. Furthermore, to allow a ranking of chosen keyframes relative to their ability to describe the content of the video, we introduce a relevance function for medoids M_k. The relevance $r(M_k)$ of the medoids M_k depends on the number of frames in cluster C_k.

$$r(M_k) = |C_k|$$

Consequently, the bigger a cluster, the more keyframes are in it and more of the video’s duration is covered by the cluster. Therefore, the medoid of the biggest cluster summarizes

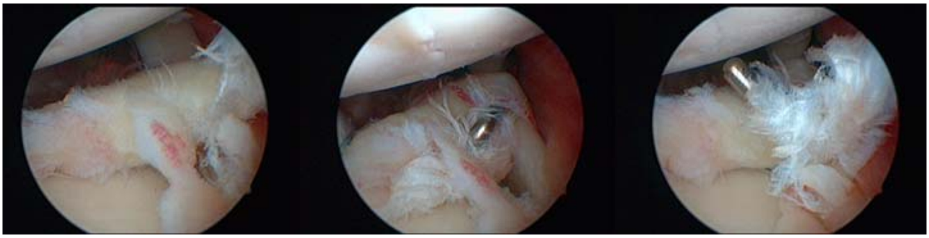


Fig. 3 List based visualization of keyframes of an arthroscopy with three keyframes. The keyframes are ordered by the proposed relevance function, i.e. the size of the respective clusters. Leftmost image represents the largest cluster and therefore the largest part of the video

the largest part of the video. Also, the medoid of the smallest clusters summarizes the smallest part of the video. The relevance function $r(M_k)$ utilizes this observation and scores the representative of the biggest cluster highest and the one of smallest cluster lowest.

4.3 Summary image composition

The final step to complete a video summary is the visualization of the medoids. Proposed methods range from simple keyframe list to complex arrangement with thumbnails of varying size. A simple storyboard summary, for instance, presents all found keyframes from left to right in sequence as shown in Fig. 3. The ordering of selected keyframes is also an issue: Which one is the most relevant or most descriptive and should be presented first, biggest or in a prominent position? While time of occurrence provides an ordinal scale, in the domain of arthroscopy videos temporal order is not as important (as stated earlier in this section).

In our approach we used the relevance function presented in Section 4.2. The relevance function ensures that the keyframe representing the largest cluster is ranked first. Moreover, we chose to visualize a single top ranked image in full size while the lower ranked images are visualized at a quarter of their original size (half width and half height). An example of such a visualization with five keyframes can be found in Fig. 4.

The visualization scheme shown in Fig. 4 benefits from its reduced size, compared to a storyboard visualization: for the visualization of five keyframes only the size of two

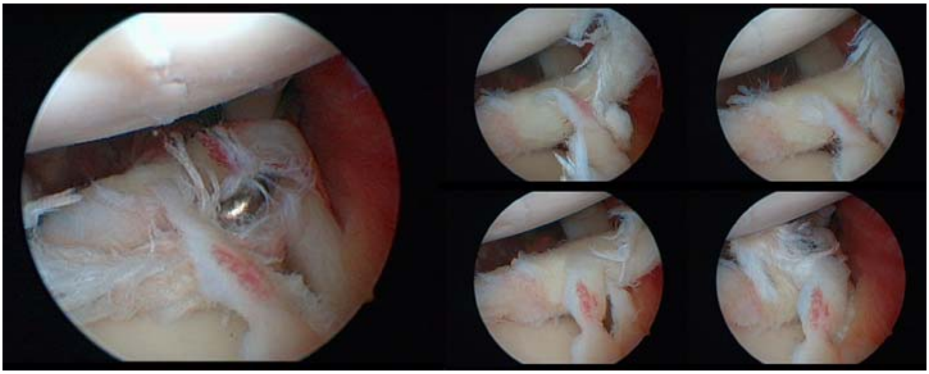


Fig. 4 Summary of an arthroscopy with five keyframes. The *top ranked* keyframe is presented in full size, while *lower ranked* keyframes are visualized at a quarter of their size

keyframes is needed. Moreover, due to the full size visualization of the top ranked keyframe, a lot of detailed visual information about keyframes that make up the largest cluster is still available. Despite of their reduced resolution and size, the remaining keyframes still carry a significant amount of information for a video summary. Especially in the domain of arthroscopy videos, where camera movement is sparse, the smaller keyframes show scenes that might be very similar to the one shown in the full size keyframe, as it can be seen in Fig. 4.

Since each selected keyframe is the medoid of a cluster of frames, for each keyframe the distribution of the underlying cluster can be visualized in addition to the keyframe's content. Due to the fact that arthroscopy videos are circular, but the video itself is stored in a rectangular resolution, in each keyframe a lot of unused, black pixels are available. In our approach we utilize this available additional space by visualizing the distribution of cluster members relative to the timeline of the video.

For the visualization we assume that the timeline of the video is represented by the width of the keyframe. For each member of the cluster a green pixel column is painted on a position relative to the start and the end of the video. The example in Fig. 5 shows a keyframe as well as a magnified view of the cluster distribution visualization. From the visualization one can infer that the cluster members can be found exclusively in the first third of the video. Figure 6 shows a video summary featuring the cluster distribution visualization. The most relevant keyframe summarizes mainly the second half of the video, whereas the keyframe in the top right corner (the same as in Fig. 5) summarizes the first third of the video.

Even though the temporal order of frames is not relevant for clustering, our tool still allows users to visualize the temporal dependency of each cluster, by inspecting the green line. For example, in Fig. 6, in the biggest keyframe, representing the biggest cluster, no medical instrument can be seen. In arthroscopy instruments often move in and out of the picture very quickly due to the high lens magnification and the instruments' small size. From the green line the surgeon can infer *where* in the video no instrument is visible and relate that to the timing of the events. Also, a difference between the second and the third cluster in Fig. 6 is the amount of loose and floating tissue (colored white due to proximity

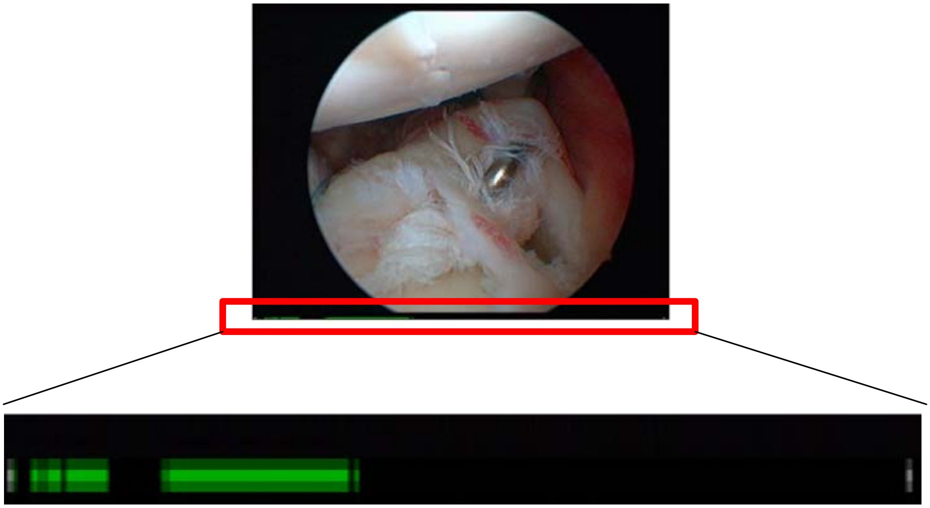


Fig. 5 Cluster distribution visualization within the keyframe

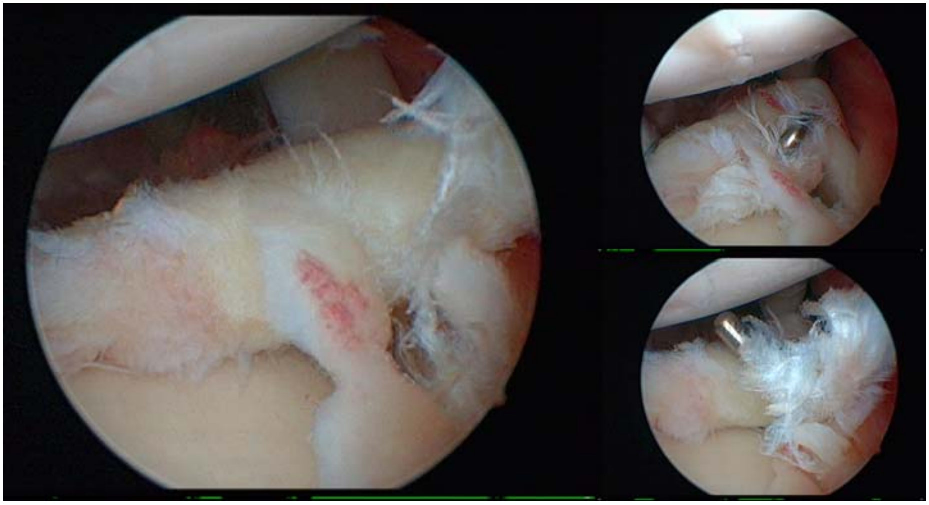


Fig. 6 Summary of an arthroscopy video with three keyframes and the cluster distribution visualization

to the lighting source). The green line helps identifying when loose tissue is floating in the cavity and the timing of events leading to the floating tissue, e.g., a cut with a scalpel.

In all presented examples a number of *n* keyframes have been extracted from an arthroscopy video and all of the *n* frames were presented in the visualization. However, we know from the domain of arthroscopy videos, that there are several frames that are less relevant due to a blurred image, floating tissue and blood. Motivated by a series of experiments (and later on supported by the evaluation), and the observation that the medoid frame of the smallest cluster very often featured rather low quality, we further introduced a second visualization scheme: Extract *n* keyframes based on *n* clusters, but just present *n*-1 of the keyframes in the visualization. Under the assumption that (i) there were more images of high quality than images of low quality within a video and (ii) images with reduced quality (e.g., blurred, reddish ones) are grouped in one cluster, the least relevant (smallest) cluster is omitted from the presentation in this approach. In this context we could speak of a “junk cluster”, where all images of low quality are dumped into. While this is heavily depending on the employed global feature, we found that this approach yields surprisingly good results.

Figure 7 shows a summary of four clusters, where only the medoids of the three biggest clusters are presented. The fourth one, which can be seen in Fig. 8 has been omitted. The keyframe shown in Fig. 8 has a low visual quality as loose tissue covers instrument and pathology. Also, due to lighting issue the contrast is rather low, so the tip of the instrument can barely be seen.

5 Evaluation

The video summarization tool presented in this paper has been evaluated before in a general domain [13]. In this section, we expand upon the previous evaluation by presenting an extended version of the study with 17 participants (instead of seven) in Section 5.1. We also present results of a qualitative interview with a specialized arthroscopic surgeon, after having used the tool for several weeks, in Section 5.2. The section concludes with a comparison between the two studies.

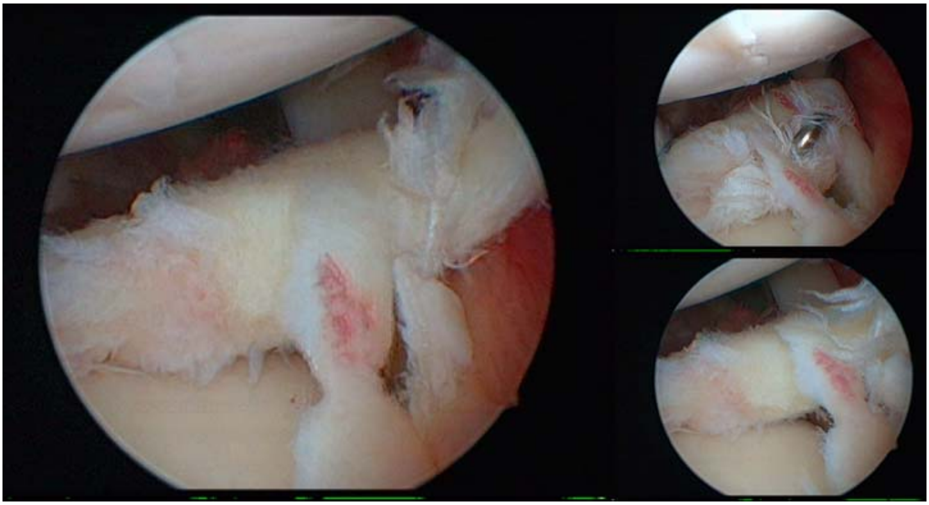


Fig. 7 Summary with three keyframes where the fourth keyframe (corresponding to the smallest cluster) has been omitted

5.1 Exploratory user study

5.1.1 Method, participants, and materials

We surveyed 17 users on three different short animation videos. The videos were taken from YouTube[2] to select a domain common to the participants. By selecting YouTube videos as domain we ensured that users understood the concept of a video summary. For the participants, the concept of a video summary was explained as a possible replacement for the video thumbnail. The videos were selected from the overall most viewed animations (Table 1).

Two parameters have been varied for the study: number of clusters (n) and feature/dissimilarity metrics combinations. The case where $n=1$ has been omitted due to its triviality and the case where $n=2$ has been omitted due to disappointing results in a first exploratory investigation. Based on the selected visualization metaphor we wanted to study if users preferred three still images (one big and 2 small) or five still images (one big 4 small). Also we wanted to find out whether a visualization with three still images should be generated based on three or four clusters, i.e., whether a junk cluster makes sense for this domain or not. We investigated three possibilities:

- $n=3$ with a visualization displaying all three medoids;
- $n=4$ displaying only the three most relevant medoids; and
- $n=5$ displaying all five medoids.

Note that the selected visualization metaphor features an odd number of images, so we did not test with four clusters showing all four keyframes. Furthermore, for each n and video under consideration we created five different video summaries with different feature and dissimilarity combinations as mentioned in Section 4.1. This results in a set of 15 video summaries to assess per video.

[2] URI: http://www.youtube.com

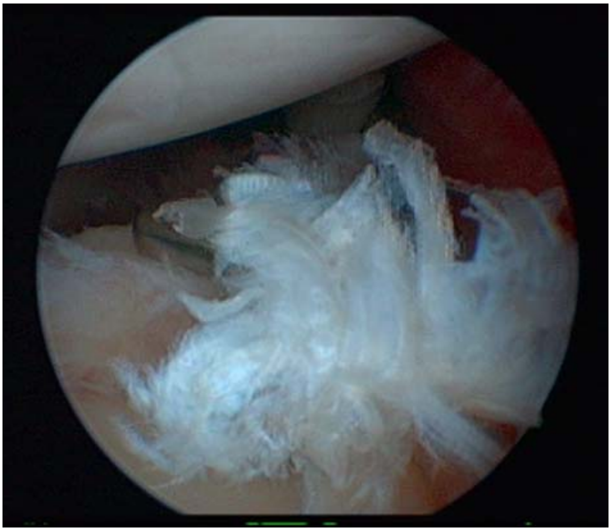


Fig. 8 Omitted keyframe of the summary in Fig. 7. This keyframe has a low visual quality as loose tissue covers instrument and pathology. It is included in the *junk cluster* and not available for viewing in Fig. 7

The participants were experienced computer users, who use YouTube on a regular basis (at least once a week) and the computer on a daily basis. The survey group consisted of three female and 11 male participants, with ages ranging from 15 to 30 years old. For each participant the survey took place in a single session, where only the participant and the moderator (the same for each test) were present. For each video the moderator showed the actual video first. Then three groups of summaries were presented: (i) the group of summaries generated with $n=3$, (ii) the group of summaries generated with $n=4$, and (iii) the group of summaries with $n=5$. Each of the groups consisted of five different summaries generated based on the five low-level features described in Section 4.1. The participant was asked to choose the best summary out of each group and to rank the three chosen summaries according to their descriptiveness for the video. In addition to selection and ranking, the moderator further asked the participant *why* the specific summary was chosen and *which criteria* were used to assess the ranking.

5.1.2 Results

Out of the 1,533 chosen images (three images per video with three videos per participant) there was no clear winner in terms of low-level features, although one of the features (namely, color histogram) has been chosen the most times in absolute terms, as it can be seen in Fig. 9. The visualization based on the color histogram feature has been chosen 49 times as most appropriate video summary followed by the auto color correlogram (ACC, 29 times), the fuzzy color and texture histogram (FCTH, 28 times), the color and edge

Table 1 Videos employed for exploratory study

Title	Length	Views[a]
Hippo bathing	30 s	2,168,000
The Room—Vancouver Film School (VFS)	194 s	871,000
Dinosaurs vault	49 s	645,000

[a] Approximate numbers, as of June 2009

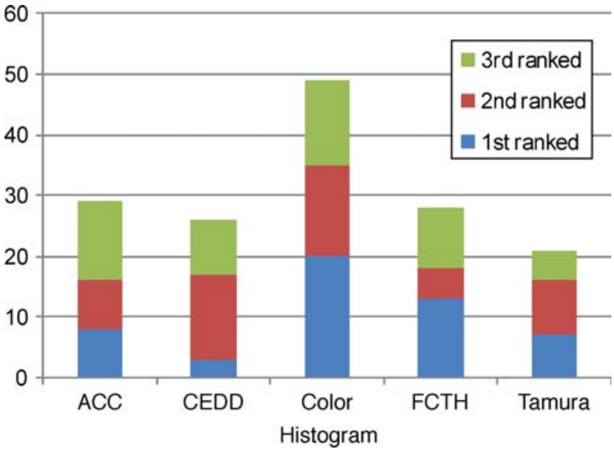


Fig. 9 Low-level features used for keyframe selection and a visualization of how often they have been selected at a specific rank

directivity descriptor (CEDD, 26 times) and the Tamura global texture descriptor (21 times). Table 2 shows how often participants have picked a specific feature for different values of n.

From Table 2 one can see that the type of chosen features heavily depends on the chosen n. An example is the CEDD feature, which performs well on $n=3$ but has only been chosen once for $n=4$. Table 3 however also indicates that the preference for low-level features also changes with different videos. CEDD was mostly selected for the *Dinosaurs* video while FCTH was mainly used for the other two. The same aspect is shown in Fig. 10, where the colors in the bar graph indicate the different videos.

When asked to rank the three selected video summaries, the users ranked first the $n=5$ video summary (28 times), followed by the $n=4$ video summary (15 times) and the $n=3$ video summary (eight times). Most users voted for the 5-cluster-based summary because more of the video was captured in the more extensive summary (five frames compared to three in the other two approaches). According to the feedback of the users, assessment was based the *appropriateness* of presented frames, i.e. how well keyframes represent the point of the video, and the *coverage* of the summary, i.e. how much of the video is represented or how much of the story can be deduced from the keyframes.

5.1.3 Conclusions

The results of the study indicate that—not surprisingly—some feature extraction methods perform better than others for this video summarization approach. Results presented in Table 3 and Fig. 10 indicate that the performance of a chosen feature depends on the

Table 2 Selected features for different values of n

	$n=3$	$n=4$	$n=5$
ACC	14	9	6
CEDD	17	1	9
Color Histogram	11	18	18
FCTH	1	18	10
Tamura	8	5	8

Table 3 Selected features for specific videos

	Hippo	Dino	Vfs
ACC	9	9	11
CEDD	8	13	6
Color Histogram	7	23	17
FCTH	18	2	9
Tamura	9	4	8

selected video. Auto color correlograms worked well for one of the videos, while color histograms worked fine for the other two. The same was true for FCTH and CEDD: while the former works fine for two of the videos, the latter only yields a good result for the remaining third video. Therefore, varying the selection of low-level feature and the corresponding metric changes the quality of results, which qualifies the research question: "which feature and metric combination performs best for a specific domain".

Tested users generally preferred the visualization with five keyframes. In interviews, they justified their preference by stating that the larger the number of presented keyframes, the better the quality (or greater the completeness) of the summary. This leads to an interesting question for future investigations: Is there an optimal number of frames to be displayed within a video summary which is enough to cover the content of the video but still not too many to be investigated by the user in a short time conveniently? An interesting observation is that users in the study preferred visualizations based on $n=4$ keyframes, i.e. the approach with the junk cluster, over $n=3$ keyframes. This hints towards the existence of a junk cluster also in other domains. In our experiments we found that the smallest cluster often contained intro and end credits of a video, which—in many cases—are not relevant for a summary.

5.2 Evaluation of the summarization approach for arthroscopic videos

In this section we present results of a qualitative evaluation of the proposed tool by an arthroscopic surgeon expert.

5.2.1 Method, participants, and materials

For development and testing we obtained a test data set of 377 arthroscopic videos recorded during one year. The videos are stored as MPEG-2 files with a bitrate of 7 Mbps, 25 frames

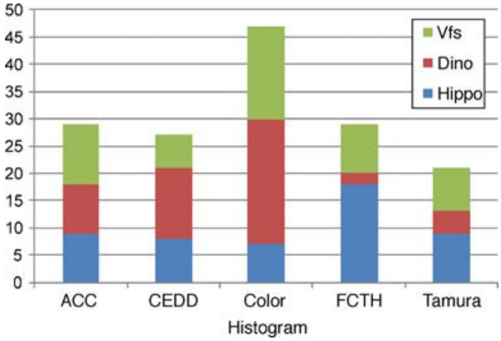


Fig. 10 Low-level features used for keyframe selection and a visualization of how often they have been selected per video

per second, and a resolution of 720×544 pixels. Overall, the videos require 29 GB storage space and have a cumulative duration of 9 h, 30 min and 55 s. Video sequences have a maximum duration of 5 min and a median duration of 1 min and 8 s. The median number of videos created per surgery is five with a maximum of 15. The length histogram of our test set is shown in Fig. 11. All videos show shoulder arthroscopies.

From this data set we randomly selected eight videos for the test. For each of the videos 15 summaries were created, just as we did earlier with the general animation videos. For each of the five chosen descriptors a summary with $n=3$, one with $n=4$ and one with $n=5$ was prepared.

For evaluation, a surgeon specialized in arthroscopy was interviewed. For each video we followed the following interview structure:

1. Participant watches the video
2. Summaries with $n=3$ are presented and the participant selects the top candidate.
3. The same for summaries with $n=4$.
4. The same for summaries with $n=5$.
5. Participant ranks the three candidates and thereby selects the best summary.

5.2.2 Results

The ranking results for different n are shown in Table 4. The visualization of five keyframes has been ranked first four times and ranked on the second place three times. For $n=3$ and $n=4$ the visualization were ranked first two times each and ranked on the second place two times and three times respectively. Figure 12 shows a histogram of the features selected by the surgeon.

In the interview the surgeon noted that all presented summaries—with one exception—describe the corresponding video very well. It was also pointed out that each of the chosen

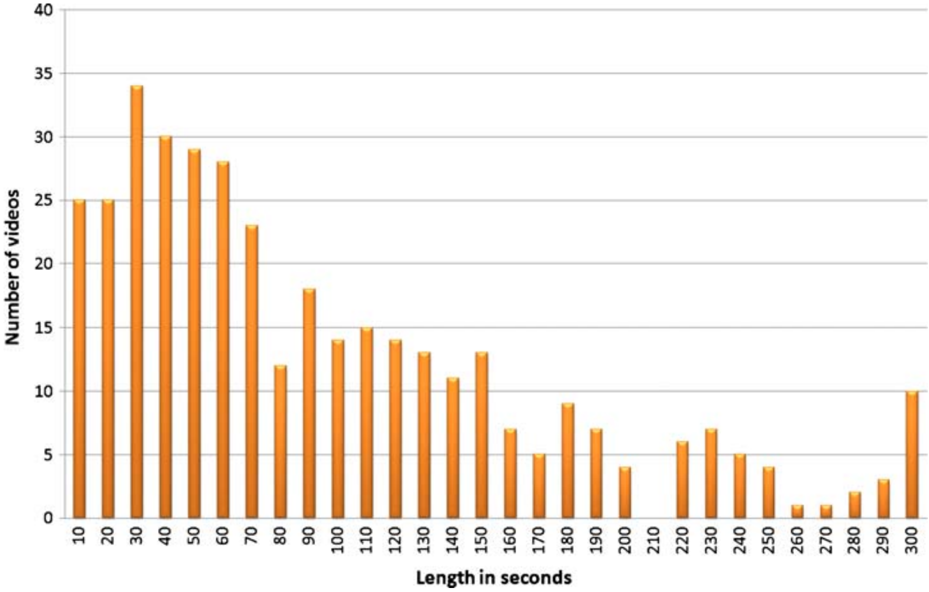


Fig. 11 Length histogram of the test data set

Table 4 Results of the selection in terms of *n*

	ranked 1st	ranked 2nd	ranked 3rd
$n=3$	2	2	4
$n=4$	2	3	3
$n=5$	4	3	1

keyframes has high quality, shows significant parts of the arthroscopy and is therefore usable for reports and illustrations. Criteria for ranking reported by the arthroscopy expert were (i) to what amount the context of the surgical process was captured and (ii) how relevant the biggest image (the one ranked highest by the relevance function) was. While for seven videos all or the major part of the summaries were satisfactory for the surgeon, there was one exception: one of the tested eight videos was a diagnostic arthroscopy, where the view moved a lot and can be described best as a round trip through the whole joint. The surgeon pointed out that on all summaries relevant parts for the diagnostic task were not visible. In this special case of diagnostic arthroscopy typically several different videos are captured, whereas one of them is such a round trip. A special characteristic of these videos is that they do not contain instruments. In our test data set ~5% of the video are such round trips. Figure 13 shows a video summary of such a round trip video. For each of the keyframes the spike indicating the direction of view is in another position. This is a clear indicator for such a round trip.

5.2.3 Conclusions

Color histogram and CEDD performed best in a cumulative view, while FCTH was chosen as first ranked most often (blue/lower part of the bars in the graph). Tamura was selected least often. The visualization with $n=4$ clusters has been ranked higher than the one with $n=3$ clusters five out of eight times. While this trend does not prove the theory of a junk cluster, it still supports the idea that by dropping the smallest cluster no critical information is removed.

More importantly, the surgeon pointed out that the selected keyframes and summaries presented by the tool are of high quality and can be used in summaries and reports, such as the ones prepared by surgeons on a regular basis. The small amount of round trip videos where this approach is potentially not applicable can be easily treated differently, e.g., by exploiting metadata available on the patients records, namely the type of surgery performed. Based on this evaluation, we conclude the overall usefulness of the tool for this domain.

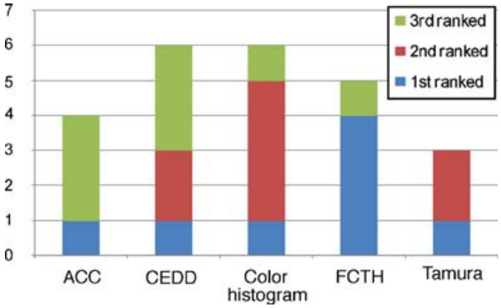


Fig. 12 Low-level features used for keyframe selection and a visualization of how often they have been selected at a specific rank

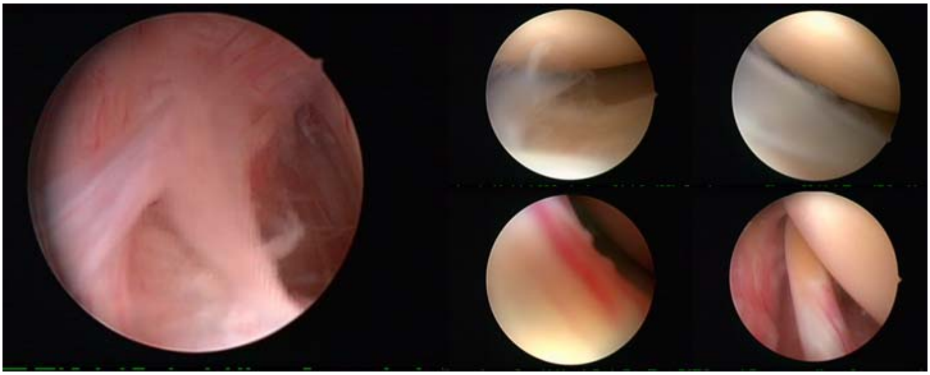


Fig. 13 Summary of a video featuring a camera round trip for diagnostic arthroscopy. Note that the spike indicating the direction of view is at a different position in each of the frames

5.3 Comparative analysis

In this section we present a comparative analysis of the results obtained in Section 5.1 and the ones in Section 5.2, whose goal is to outline the correlation between them. More specifically, since the first batch of results was obtained with a number of participants that cannot be matched in any experiment involving domain experts (since their number and time availability are limited), this analysis aims at showing that, if the results from both evaluations are in agreement, there is no reason to believe that the surgeon's opinion is biasing the evaluation of the tool for the proposed specialized domain in any significant way. The results are summarized in Table 5.

As Table 5 indicates, the results of the interview (Section 5.2) are highly correlated with the results of more extensive tests on general domain (Section 5.1). Color histogram seems to be a robust feature for both cases as it has been chosen often in both evaluations. Also, Tamura has been chosen fewest times in both evaluations. The visualization with five keyframes has been ranked first most often also in both evaluations.

However, there is a difference in the qualitative results. For the surgeon all presented summaries—with one exception—describe the according video very well. The users in the

Table 5 Comparing the results of both investigations

Aspect	YouTube	Arthroscopy
Number of keyframes presented	5 keyframes preferred over 3	5 keyframes preferred over 3
Indication of a junk cluster	Visualizations with $n=4$ preferred over $n=3$	Visualizations with $n=4$ preferred over $n=3$
Selection of low level feature	Color histogram was selected most often, Tamura has been selected least often	Color histogram and CEDD were selected most often, Tamura has been selected least often
Interview on quality of summaries	Better than a single keyframe, but there is room for improvement in terms of coverage and accuracy, especially for the longest video	Quality and tool are significant improvement (in terms of time & quality) over the current process, but summaries are not capable of capturing round trip videos very well

exploratory study gave the feedback that each of the summaries provides added value compared to the commonly used single keyframe. However many of the tested summaries did not present the semantic content of the videos completely. This is especially true for the longest of the tested videos in the exploratory study. Participants pointed out that even five keyframes were too few to provide a reasonable coverage of the semantics of the long video.

6 Conclusions and future work

In this paper we have provided an introduction to the domain of arthroscopic videos and outlined some of its characteristics that are relevant for multimedia research, particularly video summarization. Based on our analysis of that domain we introduced a novel tool for video summarization based on keyframe selection.

The video summarization prototype was evaluated qualitatively by a domain expert who judged the results to be so good that the tool will be incorporated into their daily arthroscopic surgery routine. We have also presented an exploratory study involving 17 subjects on videos from outside the domain of arthroscopic videos. A comparative analysis of the two studies confirmed that the results were consistent and hint towards a best performing value for the number of keyframes presented in the summary, as well as a best performing feature/dissimilarity combination for specific domains.

From a computational standpoint, since the videos are relatively short (see Section 5), and the presented approach employs the k-medoid clustering algorithm—which is very fast for small sets objects to be clustered—, our tool performs efficiently in terms of runtime and memory consumption. The prototypical implementation allowed for analysis of a video with a speed of 20–30 frames per second on a common home computer (Core 2 duo CPU with 2.5 GHz, 2 GB RAM and Windows XP). Due to the good results reported in this evaluation, we postulate that a more complex competitive approach has to provide significantly better results than the ones obtained with our tool to legitimate the additional runtime, storage and development complexity that it may require.

The approach presented in this paper utilizes several facts from the domain of arthroscopic videos (see Section 4), among them: (i) the absence of shot detection algorithms, since there are no shots in arthroscopy videos; (ii) the fact that frame similarity is more important than temporal ordering of the frames; (iii) the use of black pixels outside the circular shape of the captured video to present temporal information about when the frames in a cluster (represented by its keyframe) appear in the video sequence; and (iv) the decision to present only one keyframe in full size, motivated by the fact that camera movement is sparse in most arthroscopic videos.

Our approach also successfully employed the idea of a *junk cluster*, where images of low quality (e.g., blurry frames in arthroscopic videos) or relatively little semantic meaning (e.g., scrolling credits at the end of a general video) are dumped.

While we did exploit some of the arthroscopy domain characteristics in the development of the presented tool, our approach is modular and configurable and can be easily extended with other techniques for feature extraction, clustering, and visualization.

Future work will proceed in two different directions, one specific to arthroscopic videos, the other more general and applicable to short videos from any domain. Within the domain of arthroscopic videos, we shall focus on taking more advantage of the domain heuristics to extract relevant semantic information that may lead to better summaries, e.g., a semantic storyboard of the key steps in an arthroscopy surgical procedure. Possible directions include: (i) adding a block to detect motion blur (and sort out those frames) before the actual summarization step; (ii) adapting the global image features to the circular shape of the arthroscopic videos; (iii) taking the direction of view of the arthroscope into account, e.g., to use rotation information, indicated by the movement of the spike (see Section 3), to identify round trip videos of diagnostic arthroscopies and treat them differently; and (iv) employing computer vision techniques (e.g., for recognition of instruments and pathologies, identification of steps in the surgical procedure) to enhance the quality of the summaries. In a more general sense, since there exist several other domains with similar characteristics—namely those where endoscopes and arthroscopes are involved, e.g., surgeries in human and veterinary medicine, inspection of complicated and expensive machinery—we plan to adapt the tool to some of those domains and evaluate the resulting performance.

References

1. Cerneková Z, Pitas I, Nikou C (2006) Information theory-based shot cut/fade detection and video summarization. IEEE Trans Circuits Syst Video Technol 16(1):82–91
2. Chatzichristofis SA, Boutalis YS (2008) CEDD: Color and Edge Directivity Descriptor. A compact descriptor for image indexing and retrieval. In: Gasteratos A, Vincze M, Tsotsos JK (eds) Proceedings of the 6th International Conference on Computer Vision Systems, ICVS 2008, Springer, Santorini, Greece, pp 312–322
3. Chatzichristofis SA, Boutalis YS (2008) FCTH: Fuzzy Color and Texture Histogram. A low level feature for accurate image retrieval. In: Proceedings of the 9th International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS 2008, IEEE, Klagenfurt, Austria, pp 191–196
4. Ciocca G, Schettini S (2006) An innovative algorithm for key frame extraction in video summarization. J Real-Time Image Proc 1(1):69–88
5. Hadi Y, Essannouni F, Thami ROH (2006) Video summarization by k-medoid clustering. In: SAC '06: Proceedings of the 2006 ACM symposium on applied computing, ACM, New York, NY, USA, pp 1400–1401
6. Hanjalic A, Xu LQ (2005) Affective video content representation and modeling. IEEE Trans Multimedia 7(1):143–154
7. Huang J, Kumar SR, Mitra M, Zhu W-J, Zabih R (1997) Image indexing using color correlograms. In: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition, CVPR '97, IEEE, San Juan, Puerto Rico, pp 762–768
8. Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 31(3):264–323
9. Johnson D (2002) Basic science in digital imaging. Arthroscopy: The Journal of Arthroscopic and Related Surgery 18(6):648–653
10. Kosch H (2004) Distributed multimedia database technologies supported by MPEG-7 and MPEG-21, CRC, Boca Raton, Florida, USA
11. Lajtai G, Applegate G, Snyder SJ, Aitzetmuller G, Gerber CS (eds) (2003) Shoulder arthroscopy and MRI techniques. Springer, Berlin
12. Lux M, Chatzichristofis SA (2008) Lire: lucene image retrieval: an extensible java CBIR library. In: MM '08: Proceeding of the 16th ACM international conference on Multimedia, ACM, New York, NY, USA, pp 1085–1088
13. Lux M, Schöffmann K, Marques O, Böszörmenyi L (2009) A novel tool for quick video summarization using keyframe extraction techniques. In: Proceedings of the 9th Workshop on Multimedia Metadata (WMM'09), CEUR Workshop Proceedings, Vol. 441, Toulouse, France, March 19–20, 2009
14. Manjunath BS, Salembier P, Sikora T (2002) Introduction to MPEG-7: multimedia content description interface. Wiley, Chichester, West Sussex, UK
15. Matos N, Pereira F (2008) Using MPEG-7 for generic audiovisual content automatic summarization. In: Image analysis for multimedia interactive services, 2008. WIAMIS'08. Ninth International Workshop on, pp 41–45
16. Money AG, Agius H (2008) Video summarisation: a conceptual framework and survey of the state of the art. J Vis Commun Image Represent 19(2):121–143
17. NIST National Institute of Standards and Technology. Trec video retrieval evaluation. Online (last accessed on: 01/10/09): http://www-nlpir.nist.gov/projects/trecvid/
18. Pavlovich R, Vazquez-Vela G, Pardinas J, Bustos Villarreal J, Rico E, de la Mora Behar G (2002) Basic science in digital imaging. Arthroscopy: The Journal of Arthroscopic and Related Surgery 18(6):639–647

19. Tamura H, Mori S, Yamawaki T (1978) Textural features corresponding to visual perception. IEEE Trans Syst Man Cybern 8(6):460–472
20. Text REtrieval Conference (TREC). website. http://trec.nist.gov
21. Truong BT, Venkatesh S (2007) Video abstraction: a systematic review and classification. ACM Trans Multimed Comput Comm Appl (TOMCCAP) 3(1). doi:10.1145/1198302.1198305
22. Xu M, Maddage NC, Xu C, Kankanhalli M, Tian Q (2003) Creating audio keywords for event detection in soccer video. In: ICME '03: Proceedings of the 2003 International Conference on Multimedia and Expo, Vol. 1. IEEE Computer Society, Washington, DC, USA, pp 281–284, isbn 0-7803-7965-9

Mathias Lux is University Assistant at the Institute of Information Technology at the University of Klagenfurt in the distributed multimedia systems group. He received his master degree in Mathematics in 2004 and his Ph.D. in Telematics in 2006, both with distinction from Graz University of Technology. His Ph.D. thesis focused on semantics in multimedia metadata, especially MPEG-7. Having worked as researcher and project manager in the Know-Center, the Competence Center for Knowledge Based Applications in Graz, he has experience in the management and organization of research projects.

Mathias Lux has an extensive research record in the field of multimedia metadata. His experience encompasses research on annotation, metadata based multimedia retrieval, and social software and user intentions. He published numerous papers at peer reviewed conferences and some peer reviewed journals. Lately, he published a book on multimedia semantics as editor. His current research focus is on intentional metadata and social aspects of multimedia annotation and retrieval. He is a founding member of the Multimedia Metadata Community (http://www.multimedia-metadata.info), which has organized several successful scientific events.

Dr. Oge Marques is Associate Professor in the Department of Computer and Electrical Engineering and Computer Science at Florida Atlantic University in Boca Raton, Florida. He received his Ph.D. in Computer

Engineering from Florida Atlantic University in 2001, his Masters in Electronics Engineering from Philips International Institute (Eindhoven, NL) in 1989 and his Bachelor's Degree in Electrical Engineering from UTFPR (Curitiba, Brazil) in 1987. His research interests have been focused on: image processing, analysis, annotation, search, and retrieval; human and computer vision; and video processing and analysis. He has published three books, several book chapters, and more than 40 refereed journal and conference papers in these fields. He is a member of ACM, IEEE, and the honor societies of Phi Kappa Phi and Upsilon Pi Epsilon.

Dr. Klaus Schöffmann is Assistant Professor at the Institute of Information Technology at the University of Klagenfurt in the Distributed Multimedia Systems group. He received his Ph.D. (Dr.techn.) degree in June 2009 and his M.Sc. (Dipl.-Ing.) degree in informatics in March 2005, both from the University of Klagenfurt and both with distinction. His master thesis focused on the design and implementation of a video session migration system. In his Ph.D. thesis he investigated possibilities to combine video browsing, video retrieval, and video summarization in order to allow immediate video exploration. In his current research he further pursues that field of research with more focus on collaborative video exploration. He is the author of numerous peer-reviewed publications on video browsing and video content analysis.

Dr. Laszlo Böszörmenyi is a full professor and the head of the institute of Information Technology at the Klagenfurt University, Austria. He is a member of the ACM, IEEE and OCG and deputy head of the Austrian delegation to the Moving Picture Experts Group (MPEG, the ISO/IEC JTC1/SC29 WG11). Furthermore he is a founding member of the European Chapter of the SIGMM (Special Interest Group on Multimedia). In his research, he is currently focusing on Distributed Multimedia Systems, with special emphasis on adaptation, video delivery infrastructures, advanced video coding and multimedia languages. He is the author of several books and regularly publishes in refereed international journals and conference proceedings. In addition he organized several international conferences and workshops.

Dr. Georg Lajtai is Medical Director at the Humanomed Center Althofen, Austria and shoulder specialist. He degreed at the Medical School University of Graz in 1992. After his Medical degree he finished the Trauma school of Wels with specialization in shoulder surgery. Furthermore, he was in a fellowship program in Los Angeles at the Southern California Orthopedic (Los Angeles) Institute with Steven S. Snyder and finished a fellowship program in Zurich with Prof. Christian Gerber at the University of Zurich. In 2004 he finished his habilitation in Graz and is Associate Professor for Traumatology. He has worked as a researcher at the University of Zurich, Wels as well as at the Southern California Orthopedic Institute. He published numerous papers at peer-reviewed conferences and in peer-reviewed journals. Moreover, he published several books on shoulders, e.g.: Shoulder Arthroscopy and MRI techniques (Springer). Over the last 10 years he organized the International Shoulder Course (ISC, www.shoulder.org) and he is an internationally well-known and consistently booked teacher in terms of shoulder surgery. Since 1999 he is a consultant in the shoulder development group of Karl Storz (Medical Company) where he has patents for various shoulder implants, instruments and techniques.