Multimedia Tools and Applications, 2014, Springer
(The final publication is available at http://link.springer.com)

Keyframe Extraction in Endoscopic Video

Klaus Schoeffmann · Manfred Del Fabro · Tibor Szkaliczki · Laszlo Böszörmenyi · Jörg Keckstein

The final publication is available at Springer via http://dx.doi.org/10.1007/s11042-014-2224-7

Abstract In medical endoscopy more and more surgeons archive the recorded video streams in a long-term storage. One reason for this development, which is enforced by law in some countries, is to have evidence in case of lawsuits from patients. Another more practical reason is to allow later inspection of previous procedures and also to use parts of such videos for research and for training. However, due to the dramatic amount of video data recorded in a hospital on a daily basis, it is very important to have good preview images for these videos in order to allow for quick filtering of undesired content and for easier browsing through such a video archive. Unfortunately, common shot detection and keyframe extraction methods cannot be used for that video data, because these videos contain unedited and highly similar content, especially in terms of color and texture, and no shot boundaries at all. We propose a new keyframe extraction approach for this special video domain and show that our method is significantly better than a previously proposed approach.

Keywords keyframe extraction, video segmentation, endoscopy, medical imaging

1 Introduction

Medical endoscopy is a minimally invasive approach for diagnostic and therapeutic interventions in human body regions (e.g., abdomen, colon, joints). The operating endoscopist is guided by a video signal generated by a tiny camera that is inserted into hollow organs or cavities via a natural or artificial orifice. Nowadays, the endoscope typically generates an HD video signal that is displayed on one or more large screens in the operating room and thus allowing the whole operation team

Klaus Schoeffmann · Manfred Del Fabro · Laszlo Böszörmenyi
Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria

Tibor Szkaliczki
Institute for Computer Science and Control, Hungarian Academy of Sciences, MTA SZTAKI, Budapest, Hungary

Jörg Keckstein
Endometriosis Centre Stufe III, LKH Villach, Austria

to inspect the area of the surgery/procedure inside the patient and to watch the work of the operating surgeon.

Over the last years surgeons adopted to archive videos recorded during medical endoscopy in a long-term storage for different reasons. First, these *endoscopic videos* can be used as supporting material for explanations to the patients. Here, surgeons select important segments of the recordings and show them to the patients after the surgery in order to clarify what has been done. Next, the recordings are a good source of information to be used for training of young surgeons and for meetings where new operation methods are discussed. Additionally, recording endoscopic videos allows the surgeon for retrospective and comparative analysis of procedures and surgeries of different patients. Also, the recorded video is a much better source of information for follow-up surgeries than a textual surgery report, because it shows exactly the same images the surgeon has seen during the intervention. Finally, in some countries (e.g., the Netherlands) the long-time archival of videos from endoscopic interventions is enforced by law.

Unfortunately, endoscopic video data are typically stored without content-based information such as textual annotations that would allow for automatic search in the video data. Only some basic meta-data are available, like the assignment of the video to the corresponding patient, the recording date, and often a one-word textual classification of the entire video, which can be several hours long. Therefore, content-based indexing methods need to be performed in order to facilitate search in the endoscopic video archive and to enable automatic search in the scenarios described in the paragraph above. An important special problem is the selection of representative and informative keyframes, which we discuss later in detail.

Videos of medical endoscopy have very specific characteristics that make content analysis a challenging task (see also in our previous work [17] that covers summarization of arthroscopic videos). First of all, the videos are stored as recorded during the intervention, without any editing. This means that the videos contain highly similar content and no shot boundaries. Depending on the preference of the operating endoscopist – as well as the law regulations in the corresponding country – either many small recordings or one full-length recording of the intervention is saved in the long-term storage archive. In particular in the latter situation, the video typically contains a lot of unimportant content like small segments where nothing important happens (e.g., during exchange of instruments) or segments that show recordings from the endoscope lying outside the patient (because the recording often starts too early or stops too late, for example; see [21] for an algorithm identifying such, so-called *out-of-patient* scenes). The video data recorded inside the patient contains visually highly similar content, typically with slow or minor motion. Endoscopic video usually contain a lot of blurred and noisy frames due to two reasons. First, the endoscopist moves the endoscope with his hands, which results in unstable images and motion blur because the lens of the endoscope uses a high zoom factor. Also, the area currently not in focus cannot be seen clearly because of the short and fixed focus of the camera in the endoscope. Secondly, many frames suffer from noise that is caused by the flushing liquid, draining blood, or draining tissue cut by the endoscopist. In addition to that, depending on the current configuration of the endoscope – which can be changed by the endoscopist during the intervention – the frame either shows a *full image with zoomed content* or a *centered circular content area* only, which is surrounded by a large

black border as shown in Figure 1. The latter case is particularly challenging for content-based indexing methods because the centered content circle is not stable but can slightly move within the frame and scale over time, or even disappear as given in Figure 2. The easiest way for content-based indexing methods would be to consider the entire frame for content analysis. However, this gives a distorted result due to the usually available outer border area with black content (which is often not purely black but suffers from noise due to optical characteristics – like in Figure 1b – and is, hence, not easy to filter, as described in [20]). A simple way to solve this problem would be to use a smaller centered area (e.g., a rectangle) for content analysis to ignore the outer black area. This is exemplified in Figure 3, where RGB histograms of different sampling strategies are contrasted. In the first case the color histogram was extracted from the entire frame, while in the second case the color histogram was computed for a static centered rectangle of the frame to avoid distortion by border pixels. As can be seen from the figure, the resulting RGB histograms are quite different. However, this solution does not work well either because the position and size of the content area can significantly change over time in a video (see Figure 2). More importantly, for frames with zoomed content that have no black border area, like shown in the last example of Figure 2, content analysis limited to a centered rectangle could miss important information from the outer region of the frame. Optimally, content analysis should be based exactly on the actual content area in each frame. Therefore, as a very first step a content circle detection needs to be performed. We have already proposed an efficient method for that purpose in [19], and used it also for this work.

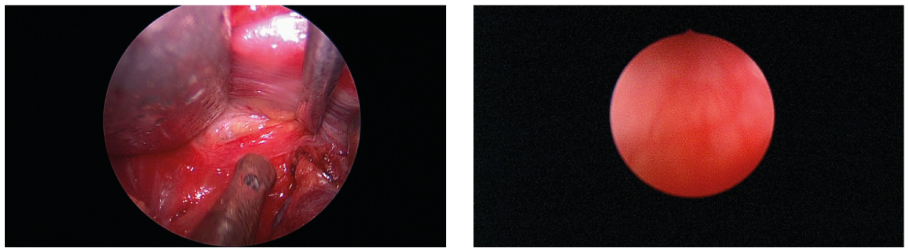

Fig. 1 Example frames taken from endoscopic videos showing the typical black border. (a) A frame with rather good visual quality and pure black border. (b) Another frame with blurry content and a large and noisy border area.

As already mentioned above, common shot detection methods like those proposed in [26] cannot be used with endoscopic video due to the very specific characteristics of the video content. For the same reason usual keyframe extraction methods – such as medium frame of shots – do not work either. Therefore, special segmentation methods and new methods for keyframe extraction are needed.

In this paper, we propose a novel keyframe extraction method for selection of representative frames in videos of medical endoscopy. Such keyframes are an important source for surgeons, who need to filter relevant content when searching for specific segments in endoscopic videos. They can act as preview images for interactive search in endoscopic video and they can also serve as input for video

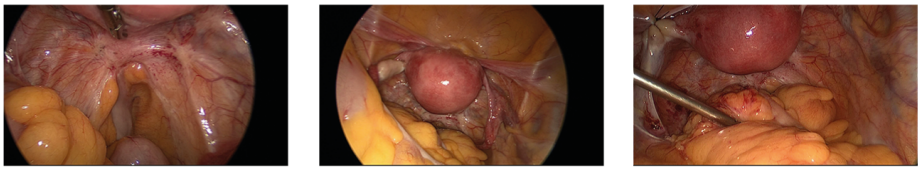

Fig. 2 These are example frames from the same video. As can be seen from the first two examples (left and center), the circular content region is not stable but can move over time. Moreover, the surgeon can enable *zoom mode*, to see a full frame without borders (right).

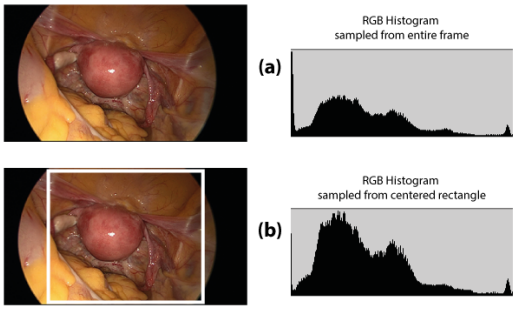

Fig. 3 a) Resulting RGB histogram when sampled from the entire frame. b) Resulting RGB histogram when sampled from a centered rectangle.

summarization methods (e.g., the one we proposed earlier in [17]). They can also be used to improve the quality of printed documentation. More importantly, the proposed method can be used as a temporal segmentation approach of endoscopic video into segments of significantly different content and therefore act as a kind of shot detection method for these videos. Shots are often used as basis for high-level semantic video segmentation, i.e. scene detection [7]. In our previous work we have already proposed a low-level segmentation method [22] for endoscopic videos. However, that work, which is based on motion tracking, targets fine-grained segmentation with low sensibility in order to produce many short segments that can be used as input for a fine-grained relevance filtering [21] or further content analysis (e.g., similarity search in retrieval tools).

In contrast, the keyframe extraction approach proposed in this work is designed for coarse-grained segmentation in order to select only a few representative frames for each video, with *low redundancy* and *good quality* (i.e., no similar frames and no blurred frames). The ultimate goal for such a keyframe extraction approach is to provide a small representative set of frames for each video in order to allow a user (i) to quickly decide whether the underlying video is interesting for him/her when browsing through the archive (e.g., does any of the preview images show an important scene of a surgery technique with clear view/no blur?) and (ii) to jump to a specific segment in the underlying recording of the endoscopic intervention; for example, to show a segment to the patient.

Our method is strongly adapted to both the characteristics of endoscopic video content and the requirements for a keyframe extraction method to be used for a quickly increasing video archive, as existent in a hospital with many surgeons. It uses the circle detection method proposed in [19] to exclude the potential black

border and limit content analysis to the real content only (cf. Figure 1). Keyframes are only selected if their content is not too blurry, such that structures (e.g., blood vessels) are still clearly visible according to a *Difference-of-Gaussians* (DoG) check and empirical tests. In order to find candidate keyframes we compare local features of frames, extracted by the ORB [25] keypoint descriptor. ORB (*Oriented FAST and Rotated BRIEF*) is a recently proposed binary keypoint extractor that has similar detection performance as SIFT [16] (and better than SURF [2]) being much faster at the same time. ORB is not scale invariant but scale invariance is not required when comparing frames from a single video in medical endoscopy. ORB is particularly well suited for our purpose, because it is very robust to image noise and allows for real-time processing, which is an important advantage when considering the situations of hospitals, where dozens of endoscopic surgeons work in parallel and record many hours of video content to be processed each day. Using ORB descriptors we match temporally close frames in order to determine frames of significant content change, which are considered as candidate keyframes. These candidate keyframes are further inspected and blurry frames as well as frames that are similar to already selected keyframes are filtered out in order to keep the number of representative frames as low as possible.

We present results from a qualitative evaluation performed with 1 surgeon, 5 endoscopic video experts that witnessed the surgery, and 9 general video experts. These 15 people were asked to rate 4 different sets of keyframes for randomly selected segments of a 160 minutes recording of one single endoscopic surgery. Our results show that the proposed keyframe extraction approach is significantly better in terms of user rating than a keyframe extraction approach proposed by Mendi et al. [18], which is the only keyframe extraction method proposed so far for endoscopic video, to best of our knowledge. Moreover, the results also show that our method performs better than two very simple baseline approaches that need no content analysis: *random selection* and *uniform selection*, where a specified number of keyframes is extracted from random positions or from temporally equidistant positions in the video. It is obvious that these two approaches will typically not result in satisfying results, but we used them for double-checking the performance of our proposed approach.

2 Related Work

The easiest procedure for keyframe extraction is uniform sampling, where keyframes are selected uniformly distributed from the content of a video. On the one hand this method is fast and efficient in terms of run-time, but has a significant drawback on the other hand: for certain short but semantically important segments no keyframes may be extracted, while for long segments with identical content too many keyframes are selected. Because of these reasons more sophisticated keyframe detection methods were proposed.

Truong and Venkatesh [29] present a comprehensive survey of video abstraction methods. They review keyframe extraction techniques under five different aspects: (i) the size of the keyframe set, (ii) the temporal unit a keyframe presents, (iii) the representation scope of individual keyframes, (iv) the underlying computational mechanisms, and (v) the visualization method. In this paper, we focus on a specific

method for keyframe extraction from videos of endoscopic surgeries, thus we also focus on the underlying computational mechanisms of related approaches.

Many keyframe extraction methods proceed by sequentially comparing frames. Whenever a significant content change between a frame and its preceding frame(s) can be observed that frame is selected as candidate keyframe. Several algorithms relying on this method have been proposed in the literature [13,31,34] and also the approach presented in this paper relies on this principle.

One disadvantage of this method is that the selected keyframes may not be well distributed temporally over the content of the whole video (i.e. they provide bad *coverage*). Therefore, algorithms were introduced where the video is first divided into a fixed number of segments and then for each segment a keyframe is extracted [8,14,28]. A good coverage of the content of a video can also be achieved with so-called *frame coverage* methods. Keyframes are selected in such a way that they represent as much other frames of the video as possible [4][5]. The coverage is computed based on visual similarity of the content.

Clustering-based approaches cluster frames of a video based on visual similarity and at the end those frames are selected that best represent each cluster [10, 33]. Similar to clustering are correlation-based methods, which extract keyframes in such a way that a minimum correlation among the members of the resulting keyframe set is achieved [9,15]. This approach is often used in combination with other methods. The keyframe extraction method presented in this paper also applies an additional step after the initial extraction, where similar keyframes are removed.

Lux et al. [17] proposed to use global image features with a k-means clustering approach to create static image summaries of arthroscopic videos. More specifically, color and edge/texture features of all frames of a video are used to create a predefined number of clusters (e.g., 3 or 5). From each cluster a representative frame is taken to create a composed summary image that describes the content of the video.

Most of the above mentioned content-based methods were already proposed over a decade ago. They all have in common that the similarity matching is based on low-level global visual features (color, motion, etc.) and form a profound base for keyframe extraction. In recent years keyframe extraction approaches have focused on a higher semantic level. The idea is to extract frames that are semantically important, e.g., characters in movies [30] or frames that show certain concepts [27][32]. *Guan* et al. [11] rely on low-level features for the keyframe extraction, but instead of global features they use local keypoints.

Mendi et al. [18] present a simple keyframe extraction algorithm for endoscopic videos that compares adjacent frames based on HSV color histogram similarity to detect shots. From each shot keyframes are extracted. The algorithm is described in section 4. It was implemented for this work in order to have a reference system for comparison with the keyframe extraction algorithm introduced in this paper.

3 Proposed Approach

3.1 Requirements

In order to provide a meaningful set of keyframes for videos recorded during endoscopic medical interventions, basically three requirements should be met. First, the set of keyframes should be rather small yet informative. This means that keyframes, which also act as shortcuts to specific positions in the video, should be stored only for important moments in the video. Secondly, the set of keyframes should summarize the content of the video in a compact way, such that it can be used for preview purpose as well. These two requirements make it also clear that the set of keyframes should contain low redundancies, i.e., few frames showing highly similar content. Finally, the third requirement is to extract only those keyframes that show clearly visible content, which is an important and challenging goal since the content in endoscopic videos regularly suffers from lens or motion blur as well as noise (see Section 1).

Optimally, in the first requirement the locations for informative keyframes would be those positions in the video where a content change happens that is relevant to the medical domain expert. However, even with state-of-the-art content analysis methods this remains a very challenging goal, for two reasons: (i) it is hard to automatically determine semantic relevance of content due to shortcomings of content-based analysis (e.g., *semantic gap* [6]), and (ii) the definition of relevance could change from one medical expert to another and even over time. Therefore, in this work we aim on the first step of providing a general and currently feasible solution for keyframe extraction from endoscopic video, by assuming every significant content change as an important one (see below in more detail).

Previous work on keyframe extraction, however, is not able to meet the two requirements defined above, because endoscopic video content has very special characteristics (see Section 1 in detail): (i) a high amount of content in the video suffers from stark content blur, (ii) highly similar content might repeat over time in one recording, and (iii) only a small part of the frame might show the actual content (cf. Figure 1b). Existing methods (see Section 2) do not consider these special properties. Also, many of them use color-based analysis, which is not optimal for endoscopic videos because most of the frames have a similar color composition.

3.2 Overview

The basic idea of our keyframe extraction approach is highly influenced by the characteristics of endoscopic video (e.g., a recording contains visually highly similar content and a lot of blurry frames) and practical requirements (which is the need for representative keyframes in good visual quality with a minor number of near duplicates): select only those keyframes that conform to the following conditions:

1. the frame differs significantly from neighboring frames in the video,
2. the frame is significantly different from already selected keyframes, and
3. the frame is not too blurry.

The first condition should avoid selection of too many similar frames from small temporal regions in the video while the second condition ensures that the overall

set of keyframes does not contain frames with highly similar or redundant content. As endoscopic videos capture whole operations and procedures they can be very long in duration and contain highly similar frames reoccurring throughout the whole video. Therefore, these two requirements ensure that the selected keyframes show different situations of the video and hence give a good overview of the whole intervention. Finally, the last condition ensures that only frames with clearly visible content are selected, because endoscopic video typically contains a high amount of blurry frames [17,21].

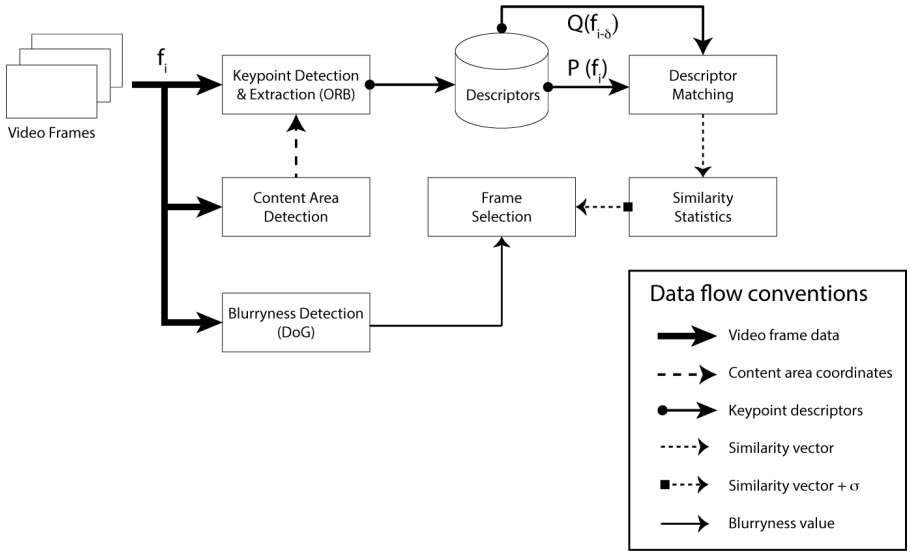

Fig. 4 General scheme of the proposed keyframe extraction approach.

Figure 4 shows the general scheme of our keyframe extraction approach, which is described in detail in the following.

To determine the similarity of neighboring frames we use salient points (i.e., keypoints) in the current frame f_i and compare them to keypoints of a past frame $f_{i-\delta}$, whereas δ is the number of frames in a predefined small duration, set to 1 second in our evaluations. The distance δ is necessary because adjacent frames would be too similar due to the rather slow motion in endoscopic video.

Instead of considering the entire frame, we limit content analysis to the actual content area (i.e., filter out the black border), with the help of the method proposed in [19]. This method first determines if a frame shows zoomed content as full-frame or a circular content area only. In both ways the method returns coordinates of the rectangular or circular content area, which is subsequently used by our method for content analysis. For frames with circular content this avoids detection of keypoints outside the content area (Figure 5a) or along the high-contrast region around the content circle (Figure 5b), which would typically occur when we simply apply keypoint detection to the whole frame.

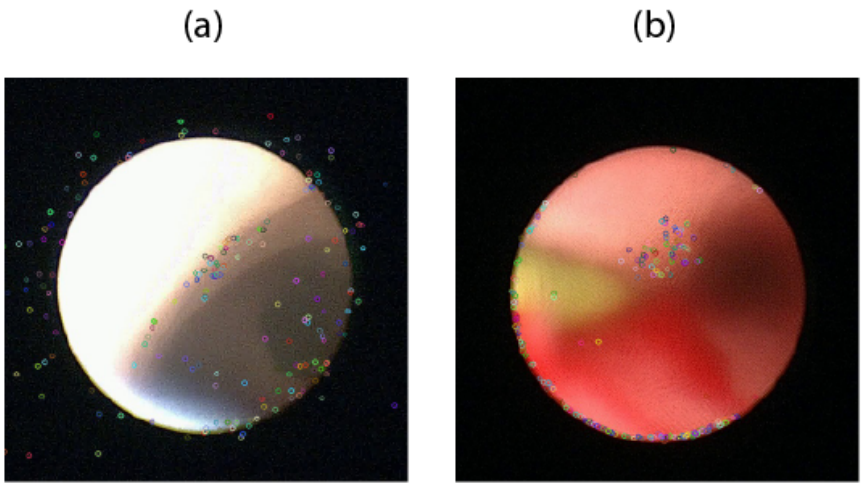

Fig. 5 (a) Without content area detection [19] many keypoints would be detected in the noisy dark region outside the content area. **(b)** Due to the typical circular content excerpt of endoscopic videos detected keypoints are often clustered around the content circle. To avoid this problem we limit content analysis to the inner part of the content circle [19].

3.3 ORB Keypoint Matching

For the keypoint detection and extraction we use ORB [25] (Oriented FAST and Rotated BRIEF) because it is a promising alternative to SIFT [16] and SURF [2], with high run-time performance and robustness to image noise, which occurs frequently in endoscopic video.

ORB uses an improved version of FAST [24] to detect keypoints in an image. To find keypoints FAST uses a simple intensity threshold to compare the center pixel to the circular ring around it. ORB uses a circular radius of 9 (FAST-9) for that purpose but additionally considers the cornerness of potential keypoints with a Harris corner filter [12] to avoid too many detections along edges. Moreover, ORB uses a scale pyramid of the image and considers the orientation of the keypoint [23], i.e., of the image patch around it. For the keypoint description ORB uses an extended variant of BRIEF [3], where each keypoint is described through a 256-tupel vector that holds results of simple binary intensity tests on pixels in the image patch around the keypoint. With ORB the intensity tests are based on a Gaussian distribution of the image patch around the center/keypoint and performed on a smoothed image. Additionally, in difference to BRIEF the image patch is steered according to its pixel orientation to allow for orientation invariance.

The authors of [25] could show that the ORB keypoint descriptor is invariant to image rotation and much more robust to image noise than SIFT, SURF, and BRIEF. Moreover, their evaluation shows that keypoint description with ORB is about 15 times faster than with SURF (and more than 300 times faster than with SIFT). These characteristics – fast and efficient detection of relevant keypoints as well as fast extraction and matching of binary descriptors – render ORB a well-

suited keypoint processing method for the domain of endoscopic video, where high run-time performance is crucial due to the quickly increasing size of video archives.

For this work we used the ORB implementation of OpenCV[1] (version 2.4.2), which was originally contributed by the authors of [25]. We use default settings, which means that 500 keypoints are detected for each frame (our videos use PAL resolution with 720x576 pixels) using a scale pyramid with of 8 levels and a pyramid decimation level (i.e., level scale factor) of 1.2.

3.4 Keyframe Extraction Method

Let $F = \{f_i : i = 1, ..., n\}$ be the set of frames of a video where n denotes the number of frames in the entire video. Let P_i denote the set of ORB features of frame f_i. $d(p, q)$ denotes the Hamming distance of two ORB features p and q. $q(p, Q)$ denotes the element of set Q of ORB features that is the closest to ORB feature p: $d(p, q(p, Q)) = \min_{q \in Q} d(p, q)$.

Let P'_i denote the set containing those elements from P_i which match best with elements in $P_{i-\delta}$. Let us introduce a weighting factor β giving the ratio of the number of the elements in P'_i and P_i: $|P'_i| = \beta \cdot |P_i|$. Currently, the best 25 % matches are used: $\beta = 0.25$. P'_i is formed by the ORB features of P_i whose distances to the elements of $P_{i-\delta}$ are the smallest: $d(p_j, q(p_j, P_{i-\delta})) \leq d(p_k, q(p_k, P_{i-\delta}))$ for all $p_j \in P'_i$ and $p_k \in P_i \setminus P'_i$.

In order to find the best matching keypoint $q(p, P_{i-\delta})$ in frame $f_{i-\delta}$ for a keypoint p in frame f_i we use a brute force matching algorithm. The distance D between two frames is computed as the average distance between all best matching keypoints of the two frames. Equation 1 gives the distance between frames f_i and $f_{i-\delta}$:

$$D_i = \frac{1}{|P'_i|} \sum_{p \in P'_i} d(p, q(p, P_{i-\delta})) \tag{1}$$

Figure 6 shows an example of matching two frames with highly similar content. We can see that for all keypoints perfect matches can be found, resulting in a low distance between the two frames. In contrast, Figure 7 shows another example of matching two frames with moderately different content. Because for several keypoints in frame $f_{i-\delta}$ the best matching points in frame f_i are not directly related, the distance D_i between these frames is much larger than than for the example shown in Figure 6.

Instead of using a fixed threshold for detecting frames with a significant difference to previous neighboring frames, we use an *adaptive threshold* that is based on the standard deviation (σ) of the frame distance computed over a sliding window over S sampled frames to the past of the current frame i:

$$\sigma_i = \sqrt{\frac{1}{S} \sum_{s=1}^{S} (D_{i-s} - \mu_i)^2}, \text{where } \mu_i = \frac{1}{S} \sum_{s=1}^{S} D_{i-s} \tag{2}$$

[1] http://www.opencv.org/

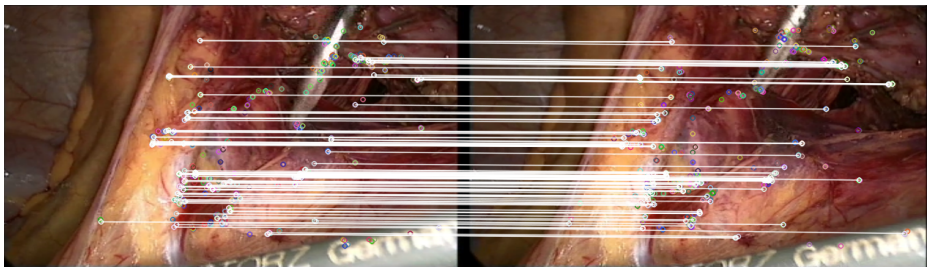

Fig. 6 ORB keypoint matching example for two frames with *highly similar content*. We can see that for all keypoints a very good match has been found.

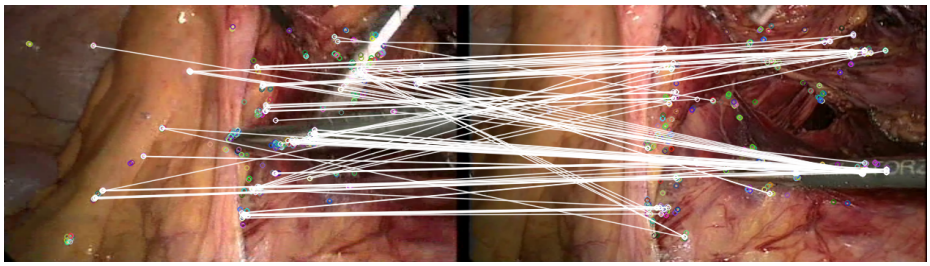

Fig. 7 ORB keypoint matching example for two frames with *moderately different content*. We can see that for many keypoints no good match can be found.

Only if the current frame distance significantly differs from the average frame distance μ_i observed in the sliding window, a keyframe is selected (we use $k = 4$):

$$D_i \leq \mu_i - k\sigma_i \text{ or } \mu_i + k\sigma_i \leq D_i \tag{3}$$

This low sensitivity should ensure that only frames containing completely different keypoints than frames in the past are selected as candidate keyframes. Figure 8 exemplifies the keyframe extraction approach based on significant changes of the the frame distance.

The candidate keyframes are further compared to already selected keyframes, using the metric described in Equation 1 (we use an empirically selected threshold for D_i) to ignore similar keyframes detected in a later portion of the video. Finally, to also ignore frames with too blurry content we compute the *Difference-of-Gaussians* for every candidate keyframe and select only those keyframes having a large difference ($DoG > 0.75$), which should ensure good visual quality (we used $K \sim 1.6$):

$$DoG = f_i * \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} - f_i * \frac{1}{2\pi K\sigma^2} e^{-\frac{x^2+y^2}{2K\sigma^2}} \tag{4}$$

4 Keyframe Extraction Methods Used for Comparison

In the literature we could find only one approach for keyframe extraction in endoscopic video, proposed by Mendi et al. [18]. Therefore, we have implemented this existing keyframe extraction algorithm and used it for a performance comparison

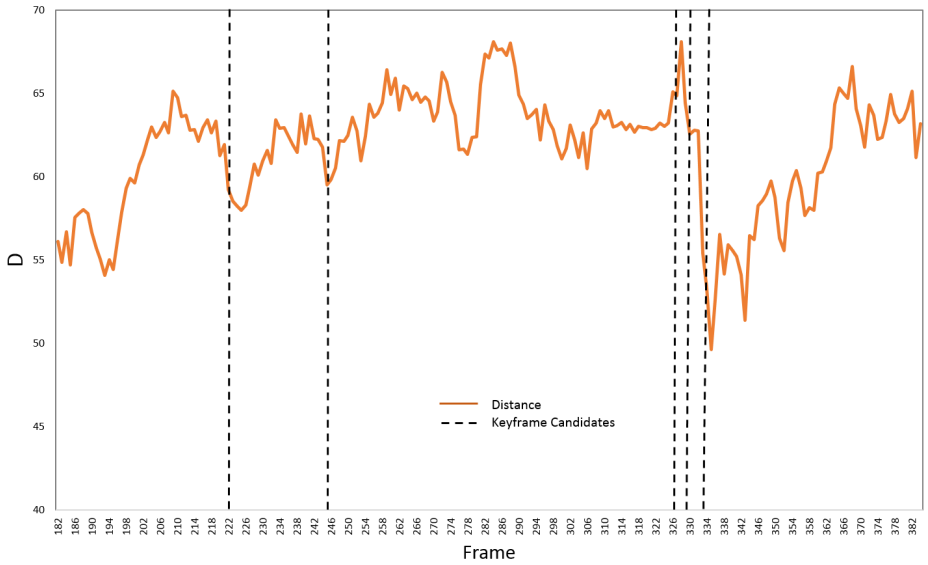

Fig. 8 Keyframe candidates are detected when the distance D_i between neighboring frames (see Eq. 1) significantly changes (i.e., the absolute value of its difference from the mean distance observed within a sliding window of size S frames is greater than 4 times standard deviation σ_i of the distance within the same window)

with our approach. The algorithm is described in detail in [18]; in the following we quickly summarize how it works.

4.1 HSV Color Histograms

First, the video stream is divided into shots. Let $h_i(b)$ denote the b^{th} bin of the HSV color histogram of frame f_i. A shot boundary is detected where the difference d_2 between two consecutive frames exceeds a certain threshold. As no threshold is stated by the authors of the original paper [18] and the authors did not respond to email requests, we used an own threshold determined by empirical observations.

$$d_2 = \sum_{b=1}^{B} |h_k(b) - h_{k-1}(b)| \tag{5}$$

B is the number of bins. Color quantization is performed using 256 colors (16 levels for hue, 4 levels for saturation, and 4 levels for value). For each shot a k-means clustering algorithm is applied to cluster the frames and to select k keyframes. This clustering process is not described in detail in the original paper [18] and thus we decided to select exactly one frame of each shot as keyframe. Furthermore, to avoid extracting shots that consist only of a few frames, we introduced a minimum shot length of 25 frames.

4.2 Uniform and Random Sampling as Baseline

We also compare both the existing keyframe extraction approach described above and our newly proposed keyframe extraction method to two very simple non-content based keyframe extraction methods: (1) a uniform frame sampling approach and (2) a random frame sampling approach. This should allow for assessing the real benefit of content-based approaches.

The advantage of uniform sampling is that it is quick and easy to implement because it requires no analysis of the content of the frames, and it provides a good coverage of the entire video. It has been used successfully for several applications, e.g., to summarize the content of a video [1]. It works by uniformly sampling frames from the video at equidistant positions. Let $u(F, m)$ denote a temporally uniform selection of m frames from F. Every ith frame from F is selected as keyframe until m frames are extracted, where $i = |F|/(m+1)$. The drawback of uniform sampling is the fact that it is hard to determine how many frames should be sampled for a video. Moreover, it could happen that only blurry frames are sampled because the content of the frames is not inspected.

At random sampling, $r(F, m)$ denotes a random selection of m frames from F. Similarly, random sampling of frames is quick and easy to implement but has the same drawbacks as uniform sampling. Additionally, at every run we would get a different selection of frames for the same video.

In our evaluations we set m to the number of frames detected by the proposed ORB-based method for each tested video respectively (for both the uniform sampling and random sampling approach). We would like to note that this might have positively skewed the evaluation results for uniform sampling and random sampling in our evaluation. In practice it will be hard to select a good value for m, which works for videos of different length and different content characteristics. In other words, if the proposed ORB-based keyframe extraction method selected only two keyframes for a 294 seconds long video clip, because most of the content was blurry, we also sampled only two frames ($m = 2$) with the uniform sampling and random sampling approaches. Most likely, however, in practice one would rather use a much higher value for an almost five minutes long video clip (that for a mostly blurry video would produce several irrelevant keyframes).

5 Experimental Results

The proposed keyframe extraction approach has been evaluated together with the HSV-based approach [18] as well as the baseline approaches uniform sampling and random sampling. For the evaluation we used 10 segments extracted from a recording of an endoscopic surgery that lasted over 3 hours. The duration of the extracted clips range from 102 seconds to 294 seconds (average 183 seconds). For each of the 10 videos we generated the 4 different sets of keyframes in random order and asked 15 users to rate the results for their *appropriateness as representative preview images* with a Likert-scale ranging from -3 (very poor) over 0 (neutral) to 3 (very good).

Out of the 15 users in our study, one participant was the surgeon who performed the surgery the video were recorded from, so he could perfectly assess the

appropriateness from an expert's perspective. The remaining 14 participants (5 female, 9 male; mean age 33.93, s.d. 12.35) were video experts, most of them (12 out of 14) are researchers working in the field of multimedia, 5 out of 14 are working in endoscopic video analysis. These 5 participants were also eye witnesses of the 3-h surgery the video data has been taken from. For every task the user started with watching the corresponding video clip and then performed a Likert-scale rating for each of the 4 keyframe sets that were listed in random order using a latin-square principle.

Figure 9 shows the sets of keyframes created by the 4 different approaches for an example clip in the test set (duration: 294 seconds) with high amount of irrelevant content (blurry frames and out-of-patient frames, which are typically also kind of blurry). As clearly shown in the figure, our proposed algorithm extracts two visually different keyframes that show no blurry content. The uniform sampling approach and the random sampling approach also extract two frames as these algorithms were configured to use exactly the same number of keyframes as the proposed ORB-based algorithm, in order to enable a fair comparison. However, as can be seen from Figure 9 the uniform sampling approach returns two highly similar frames because accidentally the corresponding video starts with almost the same scene as it ends with in the second half. The random sampling approach returns an even worse set of two frames of which the first one shows the situation where the endoscope is removed from the joint of the patient. Much worse, the HSV-based algorithm shows even a lot of frames from an out-of-patient segment in the video, which is completely irrelevant to the surgeon. In addition, most of the frames from inside the patient are rather blurry, posing a great challenge for the HSV approach with this example clip. The reason for the low performance of the HSV approach in our evaluation is the fact that every time the content changes in terms of colors, it will detect a new keyframe, though mostly by mistake. However, we warrant caution on the fact that the threshold for keyframe detection with the HSV algorithm is not available in the paper and, therefore, was selected by ourselves through empirical investigations, as described in Section 4.1.

In Figure 10 we can see the average rating of the medical expert for the 10 video clips of our study. Obviously, the HSV-based keyframes selection [18] performed worst with an average Likert rating of -2.8 (out of -3...3), which means the expert considers the extracted keyframes as *very poor*. As could be expected, also the random sampling approach performed not very well for the surgeon with an average rating of -0.7 (*rather poor*). The uniform sampling method achieved a rating close to neutral (-0.2), whereas the proposed keyframe extraction method achieved a slightly better average rating of 0.2, which means the medical expert considered those keyframes as best but still improvable.

We asked the surgeon about details of his rating in an interview after the evaluation. It turned out that the reason for the rather low rating of the results generated by the proposed approach has semantic reasons. More specifically, to the experts understanding an optimal keyframe extraction approach would consider not only clearly visible keyframes with diverse content but also consider semantics like: (i) how many instruments are visible in the keyframe, (ii) do instruments cross each other if several ones are visible (which would not be a good representative frame), (iii) does it clearly show the anatomy, and (iv) does it show the pathology and how it was repaired? These are semantic events in endoscopic videos that might be relevant to one surgeon but not necessarily to another one; furthermore, these

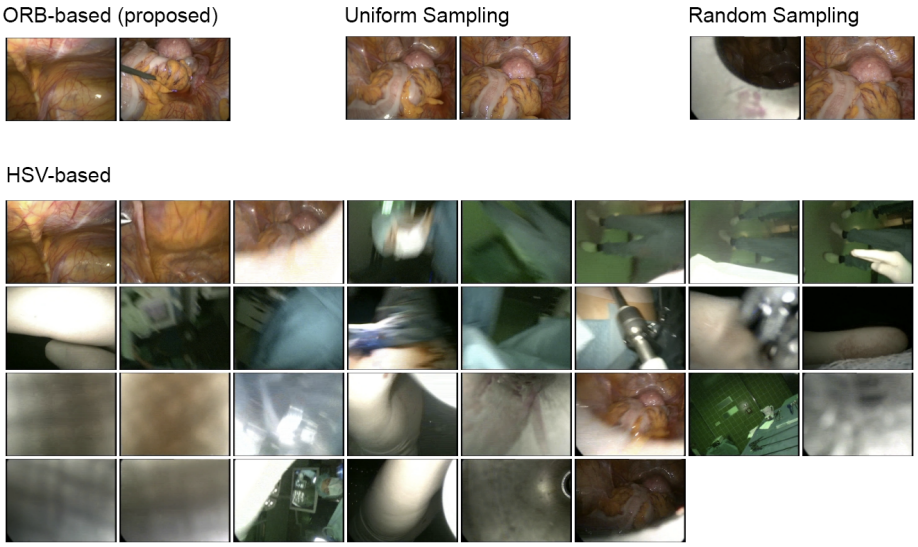

Fig. 9 Example results obtained by all four approaches for a video used in our study with a duration of 222 seconds and many blurry frames as well as several out-of-patient frames. The proposed ORB-based approach returns two clearly visible frames that show different situations of the surgery while uniform sampling shows two similar frames and random sampling shows only one relevant frame. The HSV-based approach returns a lot of blurry and out-of-patient frames with minor relevance to the actual surgery.

semantic needs might also change over time. As already mentioned in Section 3.1, the goal for this work was to provide a generally meaningful solution for the keyframe extraction problem in endoscopic video, instead of an optimal solution that is hard to achieve with current content-analysis methods.

However, we would like to note that the proposed approach results in significantly better results than the one approach proposed in the literature so far [18]. Out of the 10 video clips, for all but 2 the surgeon ranked the proposed ORB-based approach as the best one.

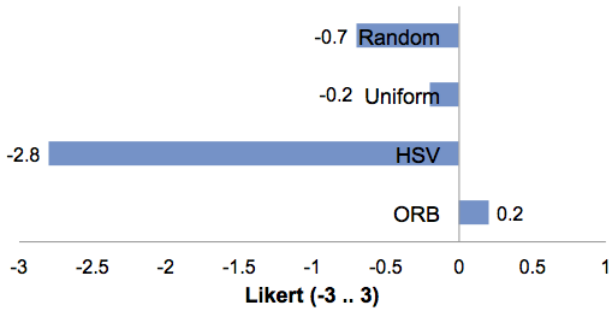

Fig. 10 Average rating of the surgeon/medical expert.

Figure 11 shows the average rating of the non-medical experts (but video experts, as described above). In general these medical novices gave a better rating for all keyframe extraction approaches but with the same final ranking as performed by the surgeon. The HSV-based method performed worst with an average rating of -1.87 (*poor*), next was random sampling with an average rating of 0.36 followed by uniform sampling (0.56) and the proposed ORB-based approach with 0.91 (*rather good*).

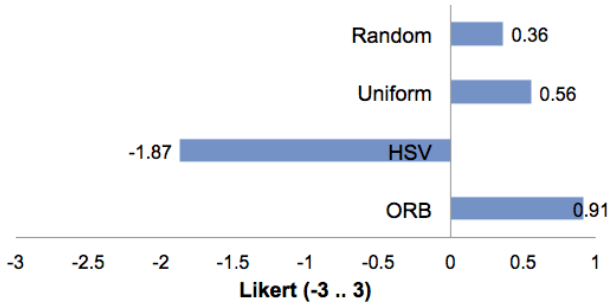

Fig. 11 Average rating of the non-medical experts.

We performed statistical analysis on the rating collected from the 14 non-medical experts for the 10 videos in our study. A non-parametric Friedman test for all approaches showed that the keyframe extraction methods show significant differences ($\chi^2(3,N=14)=30.942$, $p=0.000$). Bonferroni adjusted post-hoc Wilcoxon signed-rank tests showed that participants rated the keyframes extracted by the ORB-based approach significantly better than keyframes extracted by (i) the HSV-approach ($p<0.001$), (ii) the uniform sampling approach ($p<0.007$) and (iii) the random sampling approach ($p<0.004$). In other words, the keyframes generated by the proposed approach were clearly better than all other evaluated approaches. The statistical analysis also revealed that the HSV-based approach was rated significantly worse than the uniform sampling approach ($p<0.001$) and the random sampling approach ($p<0.001$).

From the results obtained through our user study we find that the proposed approach generates meaningful keyframes for the selected endoscopic video data set. Our evaluation is based on a recording of one surgery only. However, we have performed experiments with several other endoscopic videos taken from different interventions and different surgeons, for example the videos used for our studies in [21,19,22]. From these experiments we can confirm that the ORB-based approach generates meaningful sets of keyframes (i.e., a small set of clearly visible frames that show different situations with non-redundant content in the corresponding video). A real large-scale evaluation through user studies with experts for all these videos could only be performed by requesting all the corresponding surgeons to rate the generated sets of keyframes, which is unfortunately hard to achieve and very time-consuming in practice due to the low availability of surgeons. However, we are currently setting up a cooperation with a larger group of surgeons for our future extension of the keyframe extraction approach, where we plan to integrate

expert knowledge in order to recognize semantics. This future cooperation will also allow for user studies with a larger group of surgeons/medical experts.

It can be expected that the rather good performance of uniform sampling and random sampling in our study (which were rated close to ORB-based keyframe extraction, although still significantly worse) will not hold in practice for the following reasons. First, the actual content of randomly and uniformly sampled frames will highly depend on the underlying content. For example, in worst case it can happen that both approaches extract blurry and irrelevant frames only. Secondly, in practice it is hard to determine how many frames should be sampled. In our study we simply used the same number as generated by the ORB-based approach. In Figure 12 we can see uniform sampling (left in the figure) and random sampling (right) for the same video clip Figure 9 is based on. Sampling one random frame results in a completely irrelevant frame but also sampling 8 random frames does not show many diverse frames (frame 1 and 2 are highly similar, frame 3-6 are quite similar). The uniform sampling approach with 8 frames also results in two irrelevant frames and at least two highly similar frames. Sampling 8 frames with the proposed ORB-based approach (by using a much lower threshold) would still result in diverse frames with no blurry content.

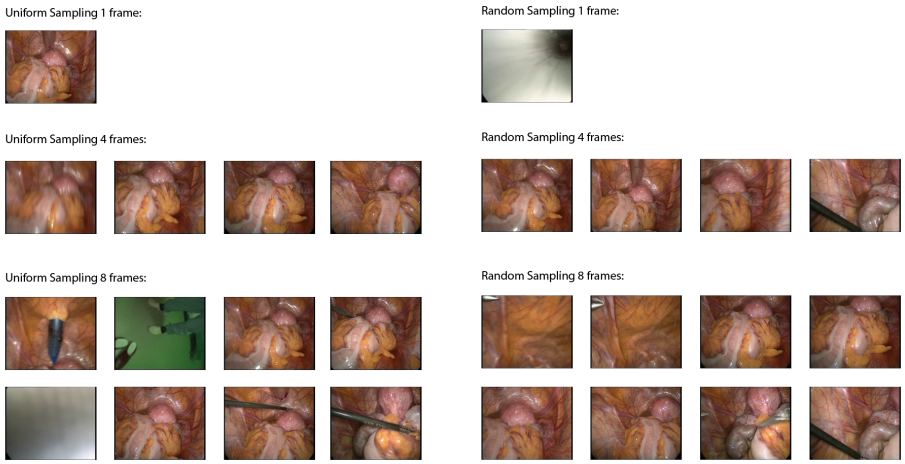

Fig. 12 Random (left) and uniform (right) sampling with different number of frames for the same video used in Figure 9.

Table 1 Run-time of all tested approaches for a 45-min. video.

	Run-time in seconds	Frames-per-Sec
Random sampling	99	27.27
Uniform sampling	97	27.84
HSV	392	6.89
ORB	546	4.95

We have also investigated the run-time of the keyframe selected approaches evaluated in this paper for a video with PAL resolution (720×576 pixels) and 67500 frames. The test has been performed on an Intel(R) Xeon(R) CPU E5-2620 running at 2.0 GHz. As given in Table 1, the proposed approach is, of course, slower than the other approaches due to much more complex content analysis. However, it is able to perform keyframe extraction 5 times faster than real-time, which makes it well-suited for video processing in large and even huge endoscopic video archives as getting usual nowadays.

6 Conclusions

In this paper we have investigated the challenging problem of keyframe extraction in video recorded by medical endoscopy. In a comparative evaluation through a user study with a surgeon and several multimedia experts (a few of them are experts in medical video analysis) we have shown that simple keyframe extraction methods like uniform sampling and random sampling do not work well for the purpose of providing a good preview. We have also shown that color-based methods like the one proposed by Mendi et al. [18] do not work for the challenging content of endoscopic video. The reason for the bad result of the HSV method is twofold. First, endoscopic video contains many highly similar frames with low color variance. Secondly, the method does not account for blurry content and fails for out-of-patient segments, which typically show shaky content with high motion blur (cf. Figure 9). However, we need to mention that the threshold for keyframe detection is not available in [18] and, therefore, was selected by ourselves through empirical investigations. The relatively good performance of uniform sampling and random sampling is somehow influenced through our ORB-based approach, i.e., through the test setup: the number of frames to sample has been chosen based on the number of frames found by the ORB-based algorithm. Therefore, for scenes like to one of the example given in Figure 9 the performance will highly depend on the number of frames to extract, which is hard to determine in practice.

The proposed keyframe extraction approach is designed to select only frames that have visually clear content and high entropy (i.e., visually different frames rather than several similar ones). The statistical analysis of the rating data obtained through our user study has shown that multimedia experts consider our method as significantly better suited for summarizing endoscopic videos than the other tested ones. Furthermore, the surgeon who performed the procedures and recorded the video content considers our proposed algorithm as the best one of the tested ones. Nevertheless, the rating of the medical expert shows clearly that improvements are still required to provide an optimal preview or overview of an endoscopic video. An additional interview with the surgeon revealed that optimal results of a keyframe extraction method for endoscopic video would include only those frames that show semantically important positions of the surgery. This would be keyframes clearly showing the anatomy, the pathology, how it was repaired, as well as keyframes showing a clear view of the operation scene during the surgery (e.g., without crossed instruments etc.).

However, it is not easy to extract these semantics with current state-of-the-art content analysis methods, a problem that is known as the *semantic gap* [6]. Moreover, as mentioned in Section 3.1 these are specific requirements of one domain

expert that might not necessarily hold for other domain experts and could further change over time. Therefore, in our future work we will focus on the integration of knowledge from domain experts into our keyframe extraction approach to further improve its performance. Furthermore, we are also looking into adapting such a method over time by considering relevance feedback.

Acknowledgements This work was supported by Lakeside Labs GmbH, Klagenfurt, Austria, funding from the European Regional Development Fund and the Carinthian Economic Promotion Fund (KWF) under grant KWF-20214 22573 33955 and partially by the Hungarian National Development Agency under grant HUMAN_MB08-1-2011-0010.

References

1. Bailer, W., Schöffmann, K., Ahlström, D., Weiss, W., del Fabro, M.: Interactive evaluation of video browsing tools. In: Proceedings of the Multimedia Modeling Conference, pp. 81–91 (2013). URL http://mmm2013.org/Video_browser_showdown.htm
2. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. In: Computer Vision–ECCV 2006, pp. 404–417. Springer (2006)
3. Calonder, M., Lepetit, V., Strecha, C., Fua, P.: Brief: Binary robust independent elementary features. In: K. Daniilidis, P. Maragos, N. Paragios (eds.) Computer Vision ECCV 2010, *Lecture Notes in Computer Science*, vol. 6314, pp. 778–792. Springer Berlin Heidelberg (2010). DOI 10.1007/978-3-642-15561-1_56. URL http://dx.doi.org/10.1007/978-3-642-15561-1_56
4. Chang, H.S., Sull, S., Lee, S.U.: Efficient video indexing scheme for content-based retrieval. Circuits and Systems for Video Technology, IEEE Transactions on **9**(8), 1269–1279 (1999)
5. Cooper, M., Foote, J.: Discriminative techniques for keyframe selection. In: Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on, pp. 4–pp. IEEE (2005)
6. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys (CSUR) **40**(2), 5 (2008)
7. Del Fabro, M., Böszörmenyi, L.: State-of-the-art and future challenges in video scene detection: a survey. Multimedia Systems **19**(5), 427–454 (2013). DOI 10.1007/s00530-013-0306-4. URL http://dx.doi.org/10.1007/s00530-013-0306-4
8. Divakaran, A., Radhakrishnan, R., Peker, K.A.: Motion activity-based extraction of keyframes from video shots. In: Image Processing. 2002. Proceedings. 2002 International Conference on, vol. 1, pp. I–932. IEEE (2002)
9. Doulamis, N.D., Doulamis, A.D., Avrithis, Y.S., Kollias, S.D.: Video content representation using optimal extraction of frames and scenes. In: Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on, vol. 1, pp. 875–879. IEEE (1998)
10. Gibson, D., Campbell, N., Thomas, B.: Visual abstraction of wildlife footage using gaussian mixture models and the minimum description length criterion. In: Pattern Recognition, 2002. Proceedings. 16th International Conference on, vol. 2, pp. 814–817. IEEE (2002)
11. Guan, G., Wang, Z., Lu, S., Deng, J., Feng, D.: Keypoint-based keyframe selection. Circuits and Systems for Video Technology, IEEE Transactions on **23**(4), 729–734 (2013). DOI 10.1109/TCSVT.2012.2214871
12. Harris, C., Stephens, M.: A combined corner and edge detector. In: Alvey vision conference, vol. 15, p. 50. Manchester, UK (1988)
13. Kim, C., Hwang, J.N.: Object-based video abstraction for video surveillance systems. Circuits and Systems for Video Technology, IEEE Transactions on **12**(12), 1128–1138 (2002). DOI 10.1109/TCSVT.2002.806813
14. Lee, H.C., Kim, S.D.: Rate-driven key frame selection using temporal variation of visual content. Electronics Letters **38**(5), 217–218 (2002)
15. Liu, T., Kender, J.R.: Optimization algorithms for the selection of key frame sequences of variable length. In: Proceedings of the 7th European Conference on Computer Vision-Part IV, ECCV '02, pp. 403–417. Springer-Verlag, London, UK, UK (2002). URL http://dl.acm.org/citation.cfm?id=645318.757467
16. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2), 91–110 (2004)

17. Lux, M., Marques, O., Schöffmann, K., Böszörmenyi, L., Lajtai, G.: A novel tool for summarization of arthroscopic videos. Multimedia Tools and Applications **46**(2-3), 521–544 (2010)
18. Mendi, E., Bayrak, C., Cecen, S., Ermisoglu, E.: Content-based management service for medical videos. Telemedicine and e-Health (2012)
19. Münzer, B., Schoeffmann, K., Böszörmenyi, L.: Detection of circular content area in endoscopic videos. In: Proceedings of the IEEE International Symposium on Computer-Based Medical Systems (CBMS13) (2013)
20. Münzer, B., Schoeffmann, K., Boszormenyi, L.: Improving encoding efficiency of endoscopic videos by using circle detection based border overlays. In: Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on, pp. 1–4 (2013). DOI 10.1109/ICMEW.2013.6618304
21. Münzer, B., Schoeffmann, K., Böszörmenyi, L.: Relevance segmentation of laparoscopic videos. In: Proceedings of 2013 IEEE International Symposium on Multimedia (ISM 2013), pp. –. Anaheim, California, USA (2013)
22. Primus, M.J., Schoeffmann, K., Bszrmenyi, L.: Segmentation of recorded endoscopic videos by detecting significant motion changes. In: Proceedings of the International Workshop on Content-Based Multimedia Indexing (CBMI 2013) (2013). To appear
23. Rosin, P.L.: Measuring corner properties. Computer Vision and Image Understanding **73**(2), 291–307 (1999)
24. Rosten, E., Drummond, T.: Machine learning for high-speed corner detection. In: Computer Vision–ECCV 2006, pp. 430–443. Springer (2006)
25. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. In: Computer Vision (ICCV), 2011 IEEE International Conference on, pp. 2564–2571 (2011). DOI 10.1109/ICCV.2011.6126544
26. Smeaton, A.F., Over, P., Doherty, A.R.: Video shot boundary detection: Seven years of trecvid activity. Computer Vision and Image Understanding **114**(4), 411–418 (2010)
27. Spyrou, E., Tolias, G., Mylonas, P., Avrithis, Y.: Concept detection and keyframe extraction using a visual thesaurus. Multimedia Tools and Applications **41**(3), 337–373 (2009)
28. Sun, X., Kankanhalli, M.S.: Video summarization using r-sequences. Real-Time Imaging **6**(6), 449–459 (2000). DOI 10.1006/rtim.1999.0197. URL `http://dx.doi.org/10.1006/rtim.1999.0197`
29. Truong, B.T., Venkatesh, S.: Video abstraction: A systematic review and classification. ACM Trans. Multimedia Comput. Commun. Appl. **3**(1), 3+ (2007)
30. Weng, C.Y., Chu, W.T., Wu, J.L.: Rolenet: movie analysis from the perspective of social networks. Trans. Multi. **11**(2), 256–271 (2009). DOI 10.1109/TMM.2008.2009684. URL `http://dx.doi.org/10.1109/TMM.2008.2009684`
31. Yeung, M.M., Liu, B.: Efficient matching and clustering of video shots. In: Proceedings of the 1995 International Conference on Image Processing (Vol. 1)-Volume 1 - Volume 1, ICIP '95, pp. 338–. IEEE Computer Society, Washington, DC, USA (1995). URL `http://dl.acm.org/citation.cfm?id=839282.841061`
32. Yong, S.P., Deng, J.D., Purvis, M.K.: Wildlife video key-frame extraction based on novelty detection in semantic context. Multimedia Tools Appl. **62**(2), 359–376 (2013). DOI 10.1007/s11042-011-0902-2. URL `http://dx.doi.org/10.1007/s11042-011-0902-2`
33. Yu, X.D., Wang, L., Tian, Q., Xue, P.: Multi-level video representation with application to keyframe extraction. In: Proceedings of the 10th International Multimedia Modelling Conference, MMM '04, pp. 117–. IEEE Computer Society, Washington, DC, USA (2004). URL `http://dl.acm.org/citation.cfm?id=968883.969425`
34. Zhang, X.D., Liu, T.Y., Lo, K.T., Feng, J.: Dynamic selection and effective compression of key frames for video abstraction. Pattern Recogn. Lett. **24**(9-10), 1523–1532 (2003). DOI 10.1016/S0167-8655(02)00391-4. URL `http://dx.doi.org/10.1016/S0167-8655(02)00391-4`