Multimed Tools Appl (2018) 77:16813–16832
https://doi.org/10.1007/s11042-017-5252-2

Video retrieval in laparoscopic video recordings with dynamic content descriptors

Klaus Schoeffmann[1] · Heinrich Husslein[2] · Sabrina Kletz[1] · Stefan Petscharnig[1] · Bernd Muenzer[1] · Christian Beecks[3]

Received: 15 October 2016 / Revised: 3 August 2017 / Accepted: 21 September 2017 /
Published online: 3 November 2017

Abstract In the domain of gynecologic surgery an increasing number of surgeries are performed in a minimally invasive manner. These laparoscopic surgeries require specific psychomotor skills of the operating surgeon, which are difficult to learn and teach. This is the reason why an increasing number of surgeons promote checking video recordings of laparoscopic surgeries for the occurrence of technical errors with surgical actions. This manual surgical quality assessment (SQA) process, however, is very cumbersome and time-consuming when carried out without any support from content-based video retrieval. In this work we propose a video content descriptor called MIDD (Motion Intensity and Direction

✉ Klaus Schoeffmann
ks@itec.aau.at

Heinrich Husslein
heinrich@husslein.at

Sabrina Kletz
sabrina@itec.aau.at

Stefan Petscharnig
spetsch@itec.aau.at

Bernd Muenzer
bernd@itec.aau.at

Christian Beecks
christian.beecks@fit.fraunhofer.de

[1] Institute of Information Technology, Alpen-Adria-Universität Klagenfurt, Universitaetsstrasse 65-67, 9020 Klagenfurt, Austria

[2] Universitaetsklinik fuer Frauenheilkunde, Medical University of Vienna, Waehringer Guertel 18-20, 1097 Wien, Austria

[3] Institute for Applied Information Technology FIT, Fraunhofer, Schloss Birlinghoven, 53754 Sankt Augustin, Germany

Descriptor) that can be effectively used to find similar segments in a laparoscopic video database and thereby help surgeons to more quickly inspect other instances of a given error scene. We evaluate the retrieval performance of MIDD with surgical actions from gynecologic surgery in direct comparison to several other dynamic content descriptors. We show that the MIDD descriptor significantly outperforms the state-of-the-art in terms of retrieval performance as well as in terms of runtime performance. Additionally, we release the manually created video dataset of 16 classes of surgical actions from medical laparoscopy to the public, for further evaluations.

Keywords Video retrieval · Video descriptor · Surgical quality assessment · Laparoscopic video · Medical multimedia · Similarity search

1 Introduction

In the medical domain, many surgeries are performed minimally invasively, with an approach called *medical endoscopy*, also known as '*keyhole-surgery*'. There are several special areas of medical endoscopy, the most frequent ones are *arthroscopy* (operations performed on joints), *colonoscopy* (procedures in the colon), and *laparoscopy* (operations in the abdomen). In the particular field of gynecologic laparoscopy procedures are performed in the area of the female reproductive systems. Typically, when laparoscopy is performed, three to four orifices to the human body are created, where one is used for the endoscope, and the others are used for operation instruments. The endoscope is equipped with a light source, some fiber optics, and a high-resolution video camera (e.g., Full HD or 4K), whose images are transmitted to a large display in the operation room. The images on this display are then used by the operating surgeon to control the endoscope and supervise actions performed with the operation instruments.

Nowadays, surgeons commonly record the real-time images of the endoscope and store them as digital videos in a long-term archive. The reasons behind are manifold – among some other motivations the videos are used for: (1) teaching purposes; e.g., to train surgery techniques with inexperienced surgeons, (2) post-operative explanations to the patients, (3) detailed visual information for follow-up surgeries, and (4) evidence in case of lawsuits from patients [34, 43]. Another, more recently emerged purpose for usage of the recorded endoscopic video footage is *surgical quality assessment* (SQA) [10, 22]. Through the recorded video the surgery can be revisited, the surgeon's level of skill can be assessed, and the video can be thoroughly checked for the occurrence of technical errors, which is especially important in medical endoscopy. In comparison to traditional open surgery, endoscopic surgery requires a unique set of skills to adapt to the challenges of 2D to 3D orientation, the fulcrum effect of the abdominal wall, tissue handling with decreased tactile sensation and amplification of tremor [21].

Figure 1 shows a technical error scene, which could happen during the initial access to the abdomen (the so-called *Abdominal Access* action). Such an event is called *too much use of force/distance* and is a very common surgical error in general laparoscopy, because it is hard for surgeons to determine the distance of instruments within the body while looking only at the monitors in the operation room. In the second-last frame of the sequence, we can see that the surgeon used too much force and the laparoscopic trocar, which has a sharp triangular point at the end, is not visible for a short amount of time. This situation could lead to inadvertent injuries as well as extensive complications. Therefore, surgeons are advised to revisit their video recordings of laparoscopic surgeries after the intervention and look for

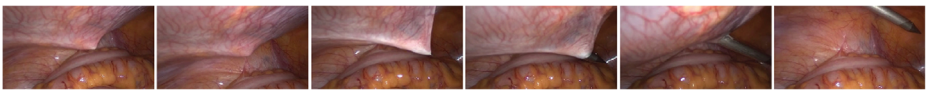

Fig. 1 Example of abdominal access with *inadequate use of too much force/distance*

surgical errors in order to increase the subjective awareness and thereby improve patient safety [1, 12, 22].

However, doing so by manual inspection without any support from content-based retrieval methods is very tedious. Therefore, our long-term goal is to support surgeons at SQA by providing automatic suggestions of segments to inspect. The target scenario we are focusing on is a use case where the surgeon has already found a relevant segment.[1] In the case of surgical quality assessment this segment would contain a surgical error and the surgeon would like to retrieve other similar segments from the video archive by performing automatic content-based retrieval, i.e., similarity search for video segments. For this particular use case we want to investigate whether dynamic content descriptors work better or worse than static content descriptors (which are known to work well in the medical domain).

More precisely, in this work we provide the following contributions. First of all, we provide a public video dataset showing short surgical actions that are very common in gynecologic laparoscopy. The contained surgical actions are subject of surgical errors and therefore constitute a good test data within the context of SQA. Next, we propose a novel dynamic video content descriptor (MIDD) that clearly outperforms other recently proposed dynamic descriptors in both terms of retrieval performance and runtime complexity. Furthermore, we investigate whether a dynamic extension of the Feature Signatures descriptor achieve equal or better performance than static Feature Signatures, which has shown good performance for content-based retrieval in the medical domain [8]. Finally, to the best of our knowledge we are the first to investigate content similarity search for video segments in the domain of gynecologic laparoscopy.

The paper is organized as follows. In Section 2 we describe related work in the field of surgical quality assessment, as well as some recent works focusing on dynamic content descriptors. Section 3 describes the proposed dynamic content descriptor MIDD (Motion Intensity and Direction Descriptor). Section 4 introduces the dataset we use for our evaluations, before detailed evaluation results are presented (in terms of MAP and runtime performance). Finally, Section 5 concludes the findings of our work.

2 Related work

Over the last two decades several works were published that focus on medical image and video analysis and processing. These works include (i) pre-processing of images such as image enhancement [14, 41] and content filtering [2, 36], (ii) real-time support at procedure time such as diagnostic decision support and computer-integrated surgery [44, 45], as well as (iii) post-procedural applications such as quality/skills assessment [31, 51] and content-based retrieval [47, 48]. A broad overview of such works is provided in an extensive survey by Muenzer et al. [35].

[1] Please note that the second author of this paper is a surgeon that is specialized in the field of surgical quality assessment in gynecologic laparoscopy.

While many works have been proposed for the first two categories mentioned above, content similarity search for supporting surgical quality assessment has been addressed only sparsely in the literature so far. De facto standards for objective assessment methods are just starting to establish [10, 22], and related research in this medical field is a niche area. However, it has been shown in recent work that manual post-hoc analysis of video recordings of laparoscopic surgeries can adequately categorize surgeons according to skill and training level. Additional error analysis allows the detection of the surgeon's specific strengths and respective weaknesses. The application of this kind of post-hoc video analysis of surgical videos through comprehensive surgical coaching [11] significantly improves the performance of trained surgeons [9, 16, 32].

In the context of surgical action classification, the authors in [33] provide a preliminary work for automatic instrument recognition and tracking, especially without the need to modify instruments to support automatic recognition. Their approach is based on particle filtering for the tracking of instruments during surgical actions. The idea is to use RGB histograms and Bayes' rule to distinguish between instrument or non-instrument pixels. Otherwise for instrument classification, Primus et al. [39] uses several keypoint detections methods as well as Support Vector Machines (SVM) with the Bag-of-visual-Words (BoW) approach for segmentation of video content. In the field of video summarization of laparoscopic surgeries, Ionescu et al. [23] use temporal visual changes to create an automatic video highlight detection. In preliminary studies, they found that keypoint scenes in such videos have no significant motion, whereas the camera motion is very scattered and discontinuous in other parts. With the comparison of adjacent color histograms and thresholds for significant motion changes, they are able to detect such keypoint moments in laparoscopic surgeries. Content classification has also been addressed recently by Petscharnig and Schoeffmann [37, 38], who evaluate well-known convolutional neural network architectures for the purpose of semantic segment annotation.

Apart from the special content of medical endoscopy, a few works can be found in the literature that focus on dynamic content descriptors. One approach utilizes the analysis of spatial and temporal information by building a space-time volume. More specifically, DeMenthon and Doermann [17] introduce a spatio-temporal descriptor that exploits the location, color, and dynamics of independently moving regions for a small number of consecutive frames. In their proposed approach, temporally ordered frames are stacked together to capture stable color-motion information of each pixel with respect to time. The regions are then produced by a Hierarchical Mean Shift clustering approach to summarize individual motion patterns. The authors use this feature to detect patterns of motion, such as actions in surveillance videos. Since their approach is focused on centralized and distinctive actions, it is not applicable to the medical domain where the content is highly self-redundant and contains both global camera and local object motion.

Further approaches for dynamic action recognition extends the notion of interest points into the spatio-temporal domain. Laptev [29] introduces a temporal extension of the Harris-Laplace detector. Spatio-temporal interest points are often used by numerous motion-based histograms (e.g., Histogram of Oriented Gradients, Histogram of Optical Flow, Motion Boundary Histograms) to represent motion information in a compact way. These methods are hard to use in the medical domain since already the very first step (corner detection) is difficult due to the very special video content, as pointed out by Schoeffmann et al. [43] and Primus et al. [39]. Therefore, it is unsure how well spatio-temporal keypoints perform for content retrieval in the medical domain (this should be investigated in future work). On the other side Duta et al. [20] presented recently a descriptor that captures motion information

by using fast temporal derivation, instead of optical flow. However, evaluations in these works were usually limited to a single action recognition dataset.

3 Motion intensity and direction descriptor (MIDD)

In this section we introduce the proposed method to describe the motion of video segments in videos from medical laparoscopy. Since these medical videos typically contain similar color and similar instruments as well as similar anatomy, we argue that motion is a very important and discriminative feature in this domain.

The proposed *Motion Intensity and Direction Descriptor (MIDD)* builds on our previous work [42], where motion vectors were extracted from every frame of an MPEG compressed video and classified into different directions. These motion vector classifications from all frames of a segment were used to detect similar segments in sports videos. The work of Droueche et al. [19] combined this idea with dynamic time warping (DTW) in order to match segments of different lengths.

In the current work, however, we propose a refined descriptor that is easier to compute, more flexible (in terms of different segment lengths), and has better performance. Our descriptor consists of a normalized motion direction histogram with assigned normalized motion intensity values and is created in pixel domain for the entire segment instead of a per-frame basis (see Fig. 2).

Let $F = f_i : i = 1, ..., n$ denote the frames of a video segment of size $W \times H$ and P_i denote the set of densely sampled feature points in frame f_i. First, the Lukas-Kanade [13] method is used to compute the optical flow for P_i from frame f_i to frame f_{i+1} ($\forall i < n$). This way, for each feature point $p \in P_i$ a motion vector $\mu_p = (x, y)$ is computed. For this motion vector the motion direction D_p is determined:

$$D_p = \begin{cases} \arccos \frac{x}{|\mu_p|} & \text{if } y \geq 0 \\ 2\pi - \arccos \frac{x}{|\mu_p|} & \text{if } y < 0 \end{cases} \tag{1}$$

In order to harmonize motion directions in small areas of a frame, we use an *averaging filter* for motion vectors of feature points in close proximity. For these predicted feature points (within a distance of $\delta = 4$ pixels), we average their x and y components.

Next, based on D_p each motion vector μ_p is classified into $k = 12$ equidistant motion directions (each accounting for $\frac{2\pi}{k}$ degrees) and point p is assigned to the corresponding bin $b \in 1...k$ in a motion direction histogram, where each bin simply counts the number of assigned points. One additional bin $(k + 1)$ is used in the motion direction histogram to count the number of feature points without motion.

$$b = \left(D_p \frac{k}{2\pi} \bmod k \right) \tag{2}$$

However, we do not only want to include the direction of motion in our final descriptor, but also the intensity of each direction to make the descriptor more distinctive. Also, we want to exclude points with no motion in our direction classification. Therefore, we also compute motion intensity I_p for each point in a frame:

$$I_p = \sqrt{x^2 + y^2} = |\mu_p| \tag{3}$$

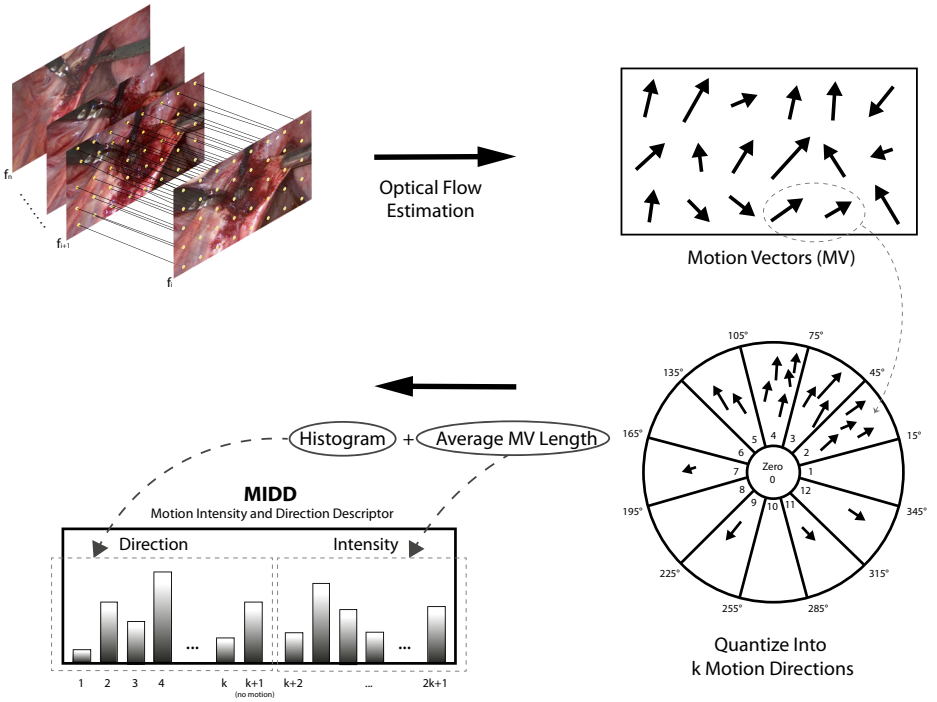

Fig. 2 For the MIDD descriptor we start with dense sampling of feature points from adjacent frames, for which we compute the optical flow. The resulting motion vectors are assigned to $k = 12$ equidistant directions and counted in a motion direction histogram. An additional bin $(k+1)$ is used to count feature points without motion, i.e., having a motion vector length of zero. The average motion vector length is computed for every bin and added to the second part of the descriptor. Both parts are normalized

Finally, for each bin b in the motion direction histogram the average motion intensity I_b of all assigned points is computed and added as second part of the descriptor (with k values, one for each motion direction). This way we end up with a descriptor that has $2k+1$ dimensions for each frame.

The descriptors of all frames are averaged over the whole video segment. Additionally, the descriptor is normalized in the following way: the first part of the descriptor – the *Motion Direction* – is normalized by the number of feature points $|P_i|$ in frame i, whereas the second part – the *Motion Intensity* – is normalized by the maximum of W and H.

4 Experimental evaluation

As mentioned in the introduction, we focus on a scenario where the surgeon has already found a relevant segment that shows some surgical action with a technical error and wants to find other instances by content-based similarity search. In this *query-by-example* scenario one segment is used as input query to retrieve a ranked list of results, where an optimal result would return other similar instances in the top part of the list. To decide which of the segments are regarded as similar (i.e., as correct results), we use a small number of pre-defined classes and regard membership to the same class as similarity criterion.

Table 1 Surgical actions in the manually annotated *SurgicalActions160* dataset (publicly available; see link above)

Surgical action	Segments	Description
Abdominal access	10	Initial access to the abdomen (puncture)
Injection	10	Injection of anesthetization liquid
Cutting	10	Cutting tissue with scissors
Blunt dissection	10	Dissection of tissue with blunt instruments
Dissection (thermal)	10	Thermal dissection of tissue with an electrosurgical instrument
Irrigation	10	Cleaning of the operation area with the *Suction* and *Irrigation Tube* instrument
Coagulation	10	Coagulation of tissue with the *Coagulation Forceps* instrument
Suction	10	Cleaning of the operation area with the *Suction and Irrigation Tube* instrument
Needle positioning	10	Bringing the needle into right position and orientation
Needle puncture	10	Puncturing with the suturing needle
Knot pushing	10	Pushing an externally tied knot to the suturing area with the *Knot Pusher* instrument
Knotting	10	Tying a knot during suturing (inside of the patient)
Thread cut	10	Cutting a thread after suturing
Sling-In	10	Insertion of the *Dissection Sling* instrument
Endobag-In	10	Insertion of the *Endobag* tool
Endobag-Out	10	Removal of the *Endobag* tool

Each video segment is about 5 seconds long, except for *Abdominal Access*, which contains a bit shorter segments

Consequently, we evaluate our results with the *Mean Average Precision* (MAP) metric and plot the Recall/Precision curves for all tested classes. Additionally, since we are interested in runtime performance, we further evaluate the processing performance to create the descriptors (for all videos) and the average time to retrieve the results for one query. All experiments were performed on a Mac Pro (Late 2013) desktop computer with a 3.5 GHz 6-Core Intel Xeon E5 CPU, 32 GB DDR3 RAM running at 1866 MHz, and a PCIe-based flash storage.

4.1 Dataset

In this work we use a manually created dataset representing typical surgical actions in gynecologic laparoscopy, which we make available to the public with this paper.[2] The entire dataset consists of 16 different classes (see Table 1), where each class is represented by exactly 10 examples. The 160 video segments were extracted from 59 different recordings and have a resolution of 427×240 pixels, are encoded with H.264/AVC, and are very short in terms of duration (min: 51 frames, max: 126 frames, avg: 119.8 frames). In total, the dataset consists of 19181 frames. Figure 3 shows example images of the 16 classes. As visible in the figure, the content of the different classes is highly similar in terms of color and texture, and therefore hard to distinguish for medical non-experts (e.g., researchers in the field of multimedia). In fact, the semantics of different content classes can only be completely understood by having medical expert knowledge. Therefore, automatic

[2] http://www-itec.aau.at/ftp/datasets/SurgicalActions160

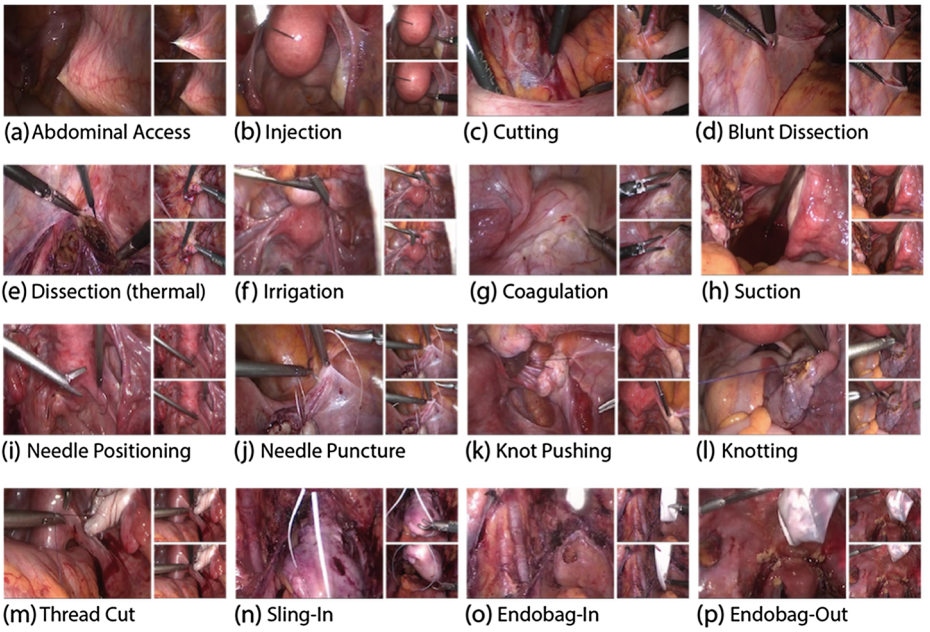

Fig. 3 Overview of the 16 different classes of surgical actions in the *SurgicalActions160* dataset (released together with this paper)

content-based similarity search is a very challenging task in this domain and does not work well with common static content descriptors [8].

4.2 Comparison to static content descriptors

First of all, we compare the performance of the dynamic *Motion Intensity and Direction Descriptor* (MIDD) to the following three static image content descriptors. These descriptors are created on a frame basis rather than on a segment basis. More specifically, they are extracted from the centered frame of each video segment.

- CNN Features (CNN_A) extracted from AlexNet [27]
- CNN Features (CNN_G) extracted from GoogLeNet [46]
- Feature Signatures (FS) [7, 8]

The first two descriptors are the so-called *CNN Features*, also known as *neural codes*. These are activation weights of the last fully-connected layer in a deep convolutional neural network. Evaluations of image retrieval tasks with these features have shown good performance [3, 18]. We use the two widely known network architectures AlexNet [27] and GoogLeNet [46] from the Caffe model zoo [25], which were trained on ILSVRC 2012 with 1000 classes from ImageNet, as described in [28]. The AlexNet architecture uses 4096 weights in the last fully-connected layer (layer fc7), while the GoogLeNet architecture uses only 1024 weights (layer pool5/7×7_s1). For similarity search with CNN Features (i.e. comparing feature vectors) we simply use Manhattan distance (L1 norm), which produces slightly better results than Euclidian distance. While these networks were trained with common videos instead of medical videos, their performance represents an out-of-the-box baseline that one could easily achieve without specific adaptation.

The third descriptor is the *Feature Signatures* descriptor, which has shown good performance for image retrieval (i.e. similarity search) in images extracted from medical videos [8]. Feature Signatures are used to describe the visual content of images individually. They are local features and commonly known in the literature as adaptive-binning histograms.

In the following, we describe Feature Signatures in more detail, since in the next subsection we also describe a dynamic extension of them, specifically developed for the evaluation of this work. Let us assume an endoscopic image is described by means of features $f_1, \ldots, f_d \in \mathbb{F}$ in a numerical feature space $(\mathbb{F}, \delta)$ such as the d-dimensional Euclidean space $(\mathbb{R}^d, L_2)$ [4]. The distance function δ is used to determine the (dis)similarity between features. The signatures of an image are extracted by clustering local features with regard to their position. To this end, each feature f is assigned a real-valued weight indicating its contribution to the corresponding endoscopic image. In [4], Feature Signatures are defined as follows.

Definition 1 (Feature Signatures) Let $(\mathbb{F}, \delta)$ be a feature space. A feature signature X is defined as

$$X : \mathbb{F} \to \mathbb{R} \text{ subject to} |R_X| < \infty,$$

where the representatives $R_X = \{f \in \mathbb{F} | X(f) \neq 0\} \subseteq \mathbb{F}$ are determined by cluster centroids and their weights $X(f)$.

According to this definition and the definition of feature representations in [4], a feature signature X is restricted to a finite number of representatives $R_X \subset \mathbb{R}^d$. Each representative (i.e., feature vector) has a weight unequal to zero $w_X : R_X \to \mathbb{R}^{\geq 0}$ and a set of representatives is computed for each endoscopic image individually. The computation is frequently carried out by applying a clustering algorithm such as the *k-means* algorithm to the extracted features of an endoscopic image and taking the cluster centroids as characteristic features. The weights can be determined by the relative frequencies, i.e. the number of assigned features of the cluster centroids.

Figure 4 depicts four laparoscopic images together with their Feature Signatures over a multi-dimensional feature space comprising position, color, and texture information. For each frame, a fixed number of points are chosen and described by a seven-dimensional feature vector $(x, y, l, a, b, c, e) \in \mathbb{F} = \mathbb{R}^7$. The vector contains information about x- and y-position, Lab-color information, as well as contrast and entropy. These vectors are then aggregated using k-means clustering. Following the work of Beecks et al. [8], the characteristic features are visualized by colored circles with diameters indicating their weights.

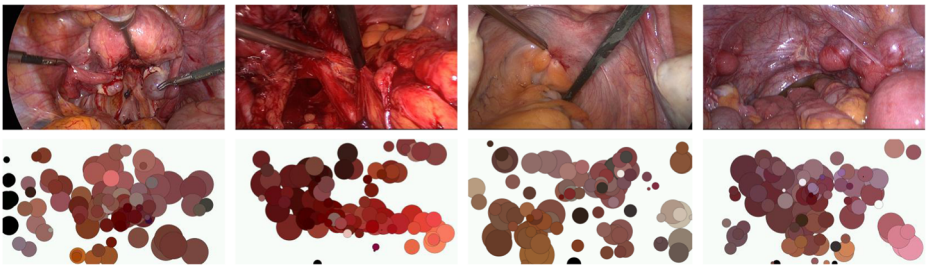

Fig. 4 Four example endoscopic images and the corresponding Feature Signatures. Images are taken from the work of Beecks et al. [8]

This example shows that Feature Signatures are able to visually approximate the content of endoscopic images by utilizing individual characteristics.

The (dis)similarity between two Feature Signatures is often determined in a distance-based manner by means of signature-based distance functions [6], such as the Earth Mover's Distance [40], the Signature Quadratic Form Distance (SQFD), or the Signature Matching Distance (SMD) [5]. In particular the latter is used as an asymmetric variant for the purpose of linking endoscopic images with video segments [4].

For our evaluations we use the following settings for Feature Signatures. We sample 8000 pre-computed random sample points (for keyframes having a resolution of 427×240 pixels). The feature vectors of the sample points are then clustered into 90 clusters. As similarity measure we use SMD.

4.3 Comparison to dynamic content descriptors

We further compare the performance of the *Motion Intensity and Direction Descriptor* (MIDD) to several other dynamic video content descriptors:

- Histogram of Oriented Gradients (HOG) [15, 30]
- Histogram of Optical Flow (HOF) [30]
- Histogram of Motion Gradients (HMG) [20]
- Dynamic Feature Signatures (DFS)

The *Histogram of Oriented Gradients* (HOG) was originally proposed by Laptev et al. [30] and designed to effectively represent human actions in videos. The authors use a spatio-temporal Bag-of-Features (BoF) approach to encode video segments and perform content classification with Support Vector Machines (SVM). In this work, however, we use them for similarity search, i.e., retrieval purpose, and build on the extension proposed by Uijlings et al. [49] that encodes HOG descriptors using Vectors of Locally Aggregated Descriptors (VLAD) [24] as well as Fisher Vectors (FV) [26]. Furthermore, we also evaluate with Histogram of Optical Flow (HOF) [30], and with the Histogram of Motion Gradients (HMG) descriptor, recently proposed by Duta et al. [20], which showed superior performance than HOG and HOF for human action recognition.

We use the same settings as proposed in the work of Duta et al.: the codebook for VLAD and FV is trained with 500000 training examples, which were extracted from additional training segments selected from the 16 classes of the dataset (these training segments are not contained in the actual dataset and are not contained in the test set). For encoding VLAD we use 512 clusters, for encoding Fisher Vectors we use 256 vectors; both encodings are created with the VLFeat library [50].

Finally, we compare our proposed MIDD descriptor to a dynamic extension of Feature Signatures, which we explicitly created for this work – this also allows us to investigate whether a dynamic representation of Feature Signatures works better than the static counterpart described above.

Following the conventional static approach, the Feature Signatures are extracted and aggregated for each frame individually. To describe their dynamic changes, the features $f \in \mathbb{F}$ of the resulted representatives $R_X \subseteq \mathbb{F} = \mathbb{R}^6$ of each feature signature X are extended by two additional dimensions. These dimensions represent the start and end positions of the displaced cluster centroids within a sequence of Feature Signatures. Therefore each feature f contains additional information of its spatial movement and is described as follows:

$$(x_{start}, y_{start}, l, a, b, c, e, x_{end}, y_{end}) \in \widetilde{\mathbb{F}} = \mathbb{R}^9,$$

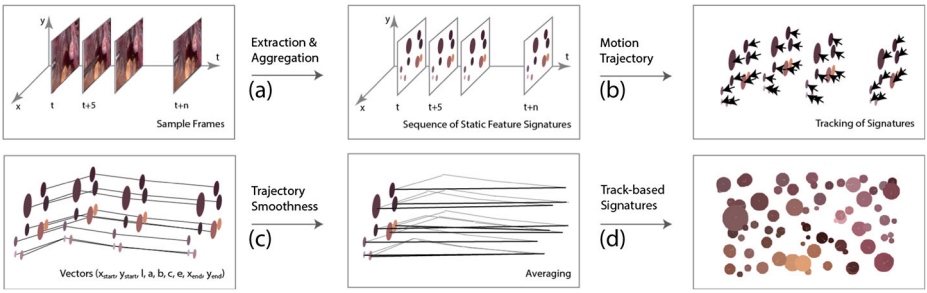

Fig. 5 To create Dynamic Feature Signatures (DFS), static Feature Signatures are extracted from several frames of the segment. Their clusters (i.e., cluster centroids) are tracked over the video segment in order to compute a motion vector for each cluster, which is stored in addition to position, color and texture

where $\widetilde{\mathbb{F}}$ represents the extended feature space comprising nine-dimensional feature vectors. Figure 5 illustrates this approach to extract Dynamic Feature Signatures (DFS). In the first step (Fig. 5a), static Feature Signatures are extracted for a set of frames. For tracking of clusters we use Feature Signatures of the extracted subset of frames (Fig. 5b) and calculate their displacement to the previous frame. The spatial movement is found by the nearest neighbor search within two frames and the feature vectors are extended by the end position of it. As can be seen, instead of storing each track separately the average value of all tracks is calculated (Fig. 5c). This way, each video segment is modeled via a dynamic feature signature $\widetilde{X}$ (Fig. 5d). For the Dynamic Feature Signatures we use the same settings for each sample frame as for the static Feature Signatures (described above). As frame sampling rate we use 50 frames per segment. The similarity of DFS is also computed with SMD.

4.4 Retrieval performance

For evaluation of the retrieval performance of the different descriptors, we employ a typical query-by-example approach, where each video segment – or representing keyframe, in case of the static descriptors – is used as input for a similarity search query. The retrieved result list (containing all 160 segments of the dataset; in optimal case the query segment at first place) is ranked by distance. For MIDD, CNN_A, CNN_G, HOG, HOF, and HMG we evaluate with Manhattan distance (L1 norm) and Euclidian distance (L2 norm). Since Feature Signatures have varying dimensionality they need an own distance measure (see [6]). We employ the Signature Matching Distance (SMD) for that purpose, using the L1 norm as ground distance.

Table 2 presents the performance of the proposed MIDD descriptor in direct comparison to the static content descriptors. The performance is measured in MAP, averaged over all

Table 2 Retrieval performance of the proposed dynamic content descriptor (MIDD) and the static content descriptors

Descriptor	MIDD (proposed)	CNN_A	CNN_G	FS
MAP (L1)	29.47%	21.19%	25.64%	23.25%
MAP (L2)	28.94%	21.62%	25.03%	23.10%

Values are given in *MAP*, averaged over all 160 queries of the 16 classes

Table 3 Retrieval performance of the other dynamic content descriptors

Descriptor	HOG_{VLAD}	HOG_{FV}	HOF_{VLAD}	HOF_{FV}	HMG_{VLAD}	HMG_{FV}	DFS
MAP (L1)	22.96%	23.73%	21.73%	21.84%	23.14%	24.11%	23.56%
MAP (L2)	25.17%	21.63%	25.31%	22.32%	25.79%	23.84%	23.50%

Values are given in *MAP*, averaged over all 160 queries of the 16 classes

16 classes of the dataset. We can see that – despite the fact that the performance is low in general (which reflects the challenging content of the medical video data) – the proposed MIDD descriptor clearly outperforms static Feature Signatures [7] as well as the CNN features, with an average MAP value of nearly 30%. Furthermore, it is also obvious that static Feature Signatures provide better performance than CNN features extracted with AlexNet, but worse performance than those CNN features extracted with the GoogLeNet architecture.

Table 3 contains the performance of the Histogram of Gradients (HOG) [29], Histogram of Optical Flow (HOF) [30], Histogram of Motion Gradients (HMG) [20], and the Dynamic Feature Signatures (DFS). Similar to the findings of [20], HMG achieves better results than HOG and HOF with both type of encodings – VLAD and Fisher Vectors. Interestingly, when using L2 as distance, the VLAD encodings produce better results than the FV encodings for all three descriptors (the vice-versa is true when using L1). Furthermore, we can see that the DFS achieve similar performance as HOG and HOF, slightly better results than their static counterparts (compare Table 2), but is beaten by HMG. When considering the Recall/Precision curve of selected descriptors (see Fig. 6) we can see that HMG (using

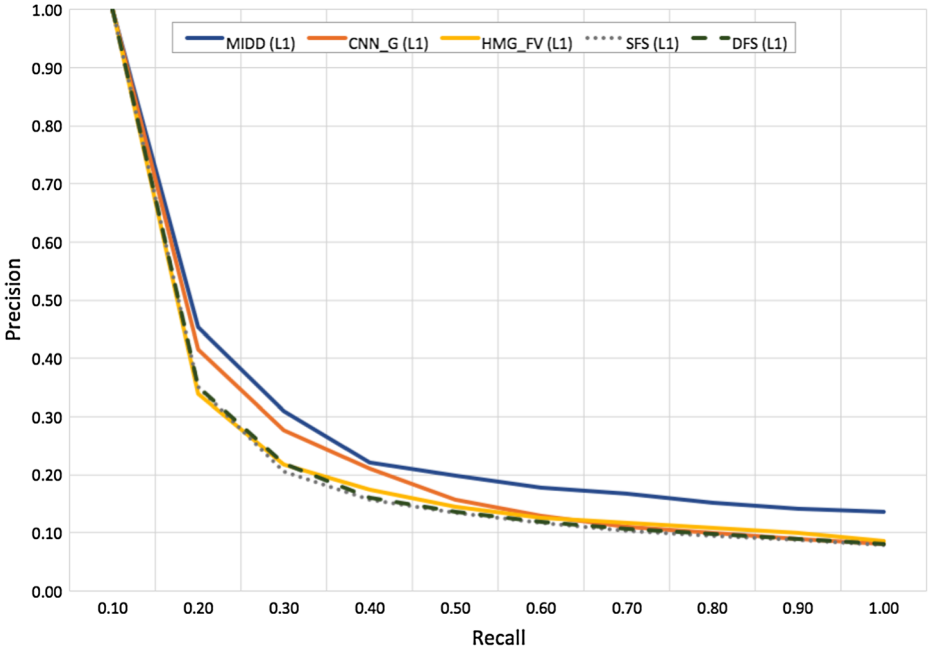

Fig. 6 Recall/Precision curve for MIDD, CNN_G, HMG_{FV}, SFS, and DFS, evaluated over all 160 queries (of all 16 classes)

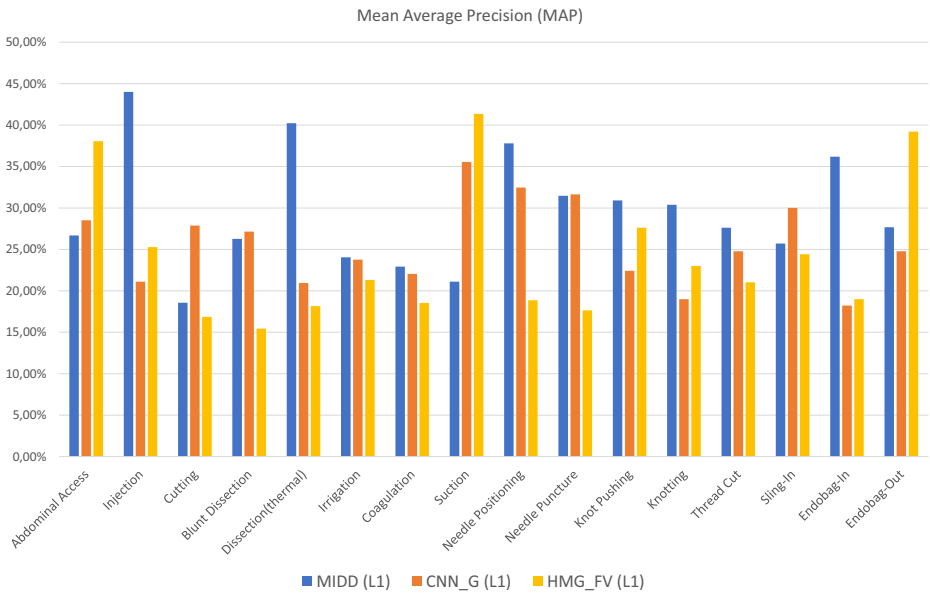

Fig. 7 Comparison of best performing descriptors for each class individually

L1 norm with Fisher Vector encoding, as proposed by [20]) is only marginally better than DFS. However, all dynamic content descriptors in Table 3 are clearly outperformed by the proposed MIDD descriptor, which achieves better performance over the whole Recall range (see Fig. 6).

In order to further investigate the retrieval performance, we evaluate the achieved MAP values of each class for the three best performing descriptors (see Fig. 7): MIDD (L1), CNN_G (L1), and HMG_{FV} (L1) – the latter setting (FV with L1) was also found optimal for HMG in [20].

As shown in the figure, HMG can beat MIDD for only three classes – *Abdominal Access*, *Suction*, and *Endobag-Out*. For all other 13 classes, MIDD clearly outperforms HMG, with nearly double performance for the classes *Injection*, *Dissection (thermal)*, and *Endobag-In*.

The significantly better performance of CNN_G (over both other dynamic descriptors) for the two classes *Cutting* and *Sling-In* is quite remarkable when keeping in mind that the information is extracted from a single keyframe of the segment. However, we hypothesize that

Table 4 Runtime performance of the proposed dynamic content descriptor (MIDD) and the static content descriptors

Descriptor	MIDD (proposed)	CNN_A	CNN_G	FS
Feature vector size	25	4096	1024	630
Extraction performance (frames/s)	103	12*	8*	3
Retrieval time (ms/query)	2	97	29	8

*...Please note that extraction time for the CNN features cannot be directly compared, since they were extracted on a similarly powerful computer, but with GPU support, using an NVIDIA GeForce GTX 1080 Ti graphics card

Table 5 Runtime performance of the other dynamic content descriptors (HOG and HOF omitted, but with almost identical values to HMG)

Descriptor	$HMGV_{LAD}$	HMG_{FV}	DFS
Feature vector size	73728	73728	810
Extraction performance (frames/s)	10	11	1
Retrieval time (ms/query)	1700	1788	8

the reason for this is the rather unique content – namely the distinctive *Scissors* instrument in the class *Cutting*, and the very unique *Dissection Sling* instrument in class *Sling-In* (also compare with Fig. 3). This assumption is further reinforced when considering the rather low performance of CNN_G for the classes *Injection*, *Dissection (thermal)*, and *Knotting*, which lack of distinctive instruments (or have varying usage of different instruments, like with *Dissection (thermal)*) but contain distinctive motion. MIDD performs much better in these motion-based classes, hence we can conclude that the proposed dynamic content descriptor is very well suited for content-based retrieval in videos of laparoscopic surgery.

4.5 Runtime performance

We further compare the complexity of the descriptors in terms of (i) feature vector length (i.e., dimensions) per segment, (ii) extraction performance in frames per second, and (iii) average retrieval time for one example query in milliseconds. The corresponding values for the proposed MIDD descriptor, the CNN features, and the static Feature Signatures are given in Table 4, while the values for HMG and DFS are presented in Table 5 (the values of HOG and HOF are almost identical to HMG, but are omitted due to space limitations).

First of all, it can be seen that the MIDD uses the smallest feature vector size, namely only 25 floating point values, whereas HMG (encoded with VLAD using 512 clusters, or with FV using 256 clusters) results in 73728 floating point values. This extreme difference has direct impact on the retrieval performance, where MIDD requires only 2 ms to retrieve all 160 results for a query (i.e., perform 160 comparison operations), HMG with FV encodings requires 1788 ms (nearly 900 times slower).

Also the feature extraction performance of MIDD is remarkably high: it can process 103 frames per second on our evaluation system, while HMG can only process 11 frames per second. It has to be noted though that MIDD is implemented in C++ with OpenCV (as is FS and DFS), while for HMG/HOG/HOF we used the MATLAB implementation provided by Duta et al. [20] together with the VLFeat library [50]. The feature extraction performance of CNN_A and CNN_G cannot be directly compared, since we extracted them with the Caffe framework [25] on a similarly powerful computer, but with GPU support (see caption of Table 4).

5 Conclusions

In this work we evaluated video content descriptors in terms of retrieval performance for video recordings from gynecologic laparoscopy. For that purpose, we have manually created a dataset consisting of example segments of 16 different surgical action classes, which are very typical in medical laparoscopy. We are releasing the dataset together with this paper in order to allow other researchers to compare to our results. We have proposed a novel video content descriptor called MIDD (*Motion Intensity and Direction Descriptor*), which

outperforms other dynamic content descriptors recently proposed in the literature for the specific domain of medical laparoscopy. We have shown that MIDD achieves not only better retrieval performance, but also is much faster to extract and to compare than alternative descriptors, such as the *Histogram of Motion Gradients* (HMG) [20]. We have also shown that our dynamic extension of *Feature Signatures* (DFS) can achieve slightly better retrieval performance than static FS– which are known to work well for content-based retrieval in medical videos [8] – but cannot outperform HMG for our dataset. To the best of our knowledge, we are the first to focus on content similarity search in video archives of gynecologic laparoscopy, with the ultimate goal to support the surgical quality assessment process [11, 22]. This work is only a first step in this domain, which would also benefit from semantic content classification approaches, which we want to investigate in the near future.

Acknowledgements Open access funding provided by University of Klagenfurt. This work was supported by Universität Klagenfurt and Lakeside Labs GmbH, Klagenfurt, Austria and funding from the European Regional Development Fund and the Carinthian Economic Promotion Fund (KWF) under grant KWF-20214 U. 3520/26336/38165.

References

1. Alba Mesa F, Sanchez Hurtado M, Sanchez Margallo F, Gomez Cabeza de Vaca V, Komorowski A (2015) Application of failure mode and effect analysis in laparoscopic colon surgery training. World J Surg 39(2):536–542
2. Atasoy S, Mateus D, Lallemand J, Meining A, Yang G-Z, Navab N (2010) Endoscopic video manifolds. Med Image Comput Comput Assist Interv 2010:437–445
3. Babenko A, Slesarev A, Chigorin A, Lempitsky V (2014) Neural codes for image retrieval. In: European conference on computer vision, pp 584–599. Springer
4. Beecks C (2013) Distance-based similarity models for content-based multimedia retrieval. RWTH Aachen University, PhD thesis
5. Beecks C, Kirchhoff S, Seidl T (2013) Signature matching distance for content-based image retrieval. In: ICMR, pp 41–48
6. Beecks C, Kirchhoff S, Seidl T (2014) On stability of signature-based similarity measures for content-based image retrieval. Multimed Tools Appl 71(1):349–362
7. Beecks C, Lokoč J, Seidl T, Skopal T (2011) Indexing the signature quadratic form distance for efficient content-based multimedia retrieval. In: Proceedings of the 1st ACM international conference on multimedia retrieval, p 24. ACM
8. Beecks C, Schoeffmann K, Lux M, Uysal MS, Seidl T (2015) Endoscopic video retrieval: A signature-based approach for linking endoscopic images with video segments. In: 2015 IEEE international symposium on multimedia (ISM), pp 33–38. IEEE
9. Bonrath E, Dedy N, Zevin B, Grantcharov T (2014) International consensus on safe techniques and error definitions in laparoscopic surgery. Surg Endosc 28(5):1535–1544
10. Bonrath EM, Zevin B, Dedy NJ, Grantcharov TP (2013) Error rating tool to identify and analyse technical errors and events in laparoscopic surgery. Br J Surg 100(8):1080–1088
11. Bonrath E, Dedy N, Gordon LE, Grantcharov T (2015) Comprehensive surgical coaching enhances surgical skill in the operating room: A randomized controlled trial. Ann Surg 262(2):205–212
12. Bonrath EM, Gordon LE, Grantcharov TP (2015) Characterising 'near miss' events in complex laparoscopic surgery through video analysis. BMJ quality & safety, pp bmjqs–2014
13. Bouguet J-Y (2001) Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm. Intel Corporation 5(1-10):4
14. Dahyot R, Vilariño F, Lacey G (2008) Improving the quality of color colonoscopy videos. EURASIP J Image Video Process 2008(1):1–7

15. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: IEEE computer society conference on computer vision and pattern recognition, 2005. CVPR 2005, volume 1, pp 886–893. IEEE
16. Dedy NJ, Fecso AB, Szasz P, Bonrath EM, Grantcharov T (2015) Implementation of an effective strategy for teaching nontechnical skills in the operating room: A single-blinded nonrandomized trial. Annals of surgery
17. DeMenthon D, Doermann D (2003) Video retrieval using spatio-temporal descriptors. In: Proceedings of the 11th ACM international conference on multimedia, MULTIMEDIA '03, pp 508–517, New York, NY, USA. ACM
18. Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, Tzeng E, Darrell T (2014) Decaf: A deep convolutional activation feature for generic visual recognition. In: ICML, pp 647–655
19. Droueche Z, Lamard M, Cazuguel G, Quellec G, Roux C, Cochener B (2012) Motion-based video retrieval with application to computer-assisted retinal surgery. In: 2012 annual international conference of the IEEE engineering in medicine and biology society, pp 4962–4965
20. Duta IC, Uijlings JRR, Nguyen TA, Aizawa K, Hauptmann AG, Ionescu B, Sebe N (2016) Histograms of Motion Gradients for Real-time Video Classification. In: 2016 14th international workshop on content-based multimedia indexing (CBMI). IEEE, pp 1–6
21. Fried GM, Gill H (2007) Surgery through the keyhole: a new view of an old art. McGill J Med MJM 10(2):140
22. Husslein H, Shirreff L, Shore EM, Lefebvre GG, Grantcharov TP (2015) The generic error rating tool: A novel approach to assessment of performance and surgical education in gynecologic laparoscopy. Journal of Surgical Education
23. Ionescu B, Vertan C, Florea L (2011) Automatic Abstraction of Laparoscopic Medical Footage Through Visual Activity Analysis. In: E-Health and Bioengineering Conference. IEEE, pp 1–4
24. Jegou H, Perronnin F, Douze M, Sánchez J, Perez P, Schmid C (2012) Aggregating local image descriptors into compact codes. IEEE Trans Pattern Anal Mach Intell 34(9):1704–1716
25. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: Convolutional Architecture for Fast Feature Embedding. In: Proceedings of the 22nd ACM international conference on multimedia, MM '14, pp 675–678, New York, NY, USA. ACM
26. Krapac J, Verbeek J, Jurie F (2011) Modeling spatial layout with fisher vectors for image categorization. In: 2011 IEEE international conference on computer vision (ICCV), pp 1487–1494. IEEE
27. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
28. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in Neural Information Processing Systems 25, pp 1097–1105. Curran Associates, Inc.
29. Laptev I (2005) On space-time interest points. Int J Comput Vis 64(2):107–123
30. Laptev I, Marszalek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. In: IEEE conference on computer vision pattern recognition, 2008. CVPR 2008, pp 1–8. IEEE
31. Lin HC, Shafran I, Yuh D, Hager GD (2006) Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions. Comput Aided Surg 11(5):220–230
32. Ma M (2013) The power of video recording: Taking quality to the next level. JAMA 309(15):1591–1592
33. McKenna S, Charif HN, Frank T (2005) Towards video understanding of laparoscopic surgery: Instrument tracking. Proc. of Image and Vision Computing, New Zealand
34. Münzer B, Schoeffmann K, Böszörmenyi L (2013) Relevance segmentation of laparoscopic videos. In: 2013 IEEE international symposium on multimedia (ISM), pp 84–91
35. Münzer B, Schoeffmann K, Böszörmenyi L (2017) Content-based processing and analysis of endoscopic images and videos: A survey. Multimedia Tools and Applications
36. Oh J, Hwang S, Lee J, Tavanapong W, Wong J, De Groen PC (2007) Informative frame classification for endoscopy video. Med Image Anal 11(2):110–127
37. Petscharnig S, Schoeffmann K (2017) Learning laparoscopic video shot classification for gynecological surgery. Multimed Tools Appl. https://doi.org/10.1007/s11042-017-4699-5
38. Petscharnig S, Schöffmann K (2017) Deep Learning for Shot Classification in Gynecologic Surgery Videos. Springer International Publishing, Cham, pp 702–713
39. Primus M, Schoeffmann K, Böszörmenyi L (2015) Instrument classification in laparoscopic videos. In: 2015 13th international workshop on content-based multimedia indexing (CBMI), pp 1–6
40. Rubner Y, Tomasi C, Guibas LJ (2000) The earth mover's distance as a metric for image retrieval. Int J Comput Vis 40(2):99–121
41. Saint-Pierre C-A, Boisvert J, Grimard G, Cheriet F (2011) Detection and correction of specular reflections for automatic surgical tool segmentation in thoracoscopic images. Mach Vis Appl 22(1):171–180

42. Schoeffmann K, Lux M, Taschwer M, Boeszoermenyi L (2009) Visualization of video motion in context of video browsing. In: 2009 IEEE International Conference on Multimedia and Expo, pp 658–661
43. Schoeffmann K, Del Fabro M, Szkaliczki T, Böszörmenyi L, Keckstein J (2015) Keyframe extraction in endoscopic video. Multimed Tools Appl 74(24):11187–11206. https://doi.org/10.1007/s11042-014-2224-7
44. Schulmann K, Hollerbach S, Kraus K, Willert J, Vogel T, Moslein G, Pox C, Reiser M, Reinacher-Schick A, Schmiegel W (2005) Feasibility and diagnostic utility of video capsule endoscopy for the detection of small bowel polyps in patients with hereditary polyposis syndromes. Am J Gastroenterol 100(1):27–37, 01
45. Summers RM, Johnson CD, Pusanik LM, Malley JD, Youssef AM, Reed JE (2001) Automated polyp detection at ct colonography: Feasibility assessment in a human population 1. Radiology 219(1):51–59
46. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
47. Twinanda AP, De Mathelin M, Padoy N (2014) Fisher kernel based task boundary retrieval in laparoscopic database with single video query. In: International conference on medical image computing and computer-assisted intervention, pp 409–416. Springer
48. Twinanda AP, Marescaux J, de Mathelin M, Padoy N (2015) Classification approach for automatic laparoscopic video database organization. Int J Comput Assist Radiol Surg 10(9):1449–1460
49. Uijlings J, Duta IC, Sangineto E, Sebe N (2015) Video classification with densely extracted hog/hof/mbh features: an evaluation of the accuracy/computational efficiency trade-off. Int J Multimed Inf Retr 4(1):33–44
50. Vedaldi A, Fulkerson B (2010) Vlfeat: An open and portable library of computer vision algorithms. In: Proceedings of the 18th ACM international conference on Multimedia, pp 1469–1472. ACM
51. Wang Y, Tavanapong W, Wong J, Oh J, De Groen PC (2013) Near real-time retroflexion detection in colonoscopy. IEEE J Bio Health Inf 17(1):143–152

Dr. Klaus Schoeffmann is associate professor in the distributed multimedia systems research group at the Institute of Information Technology (ITEC) at Klagenfurt University, Austria. He received his Ph.D. in 2009 and his habilitation (venia docendi) in 2015, both in computer science and from Klagenfurt University. His research focuses on human-computer interaction with multimedia data, multimedia content analysis, and multimedia systems (particularly in the mobile and medical domain). He has co-authored more than 80 publications on various topics in multimedia and he has co-organized international conferences, special sessions and workshops. He is co-founder of the Video Browser Showdown (VBS) – an international live evaluation competition of video search, and an editorial board member of the Multimedia Tools and Applications and Multimedia Systems journals (by Springer). Additionally, he is a steering committee member of the International Conference on MultiMedia Modelling (MMM), a member of the IEEE and the ACM, and a regular reviewer for international conferences and journals in the field of multimedia. Klaus Schoeffmann teaches various courses in computer science, including mobile app development (interactive multimedia applications), video retrieval, media technology, distributed systems, distributed multimedia systems, and operating systems.

Heinrich Husslein is an obstetrician/gynecologist who specializes in minimal invasive gynecological surgery. Dr. Husslein grew up in Vienna and spent one year during high School in Washington DC. He graduated from the Medical University of Vienna and completed his OB/GYN residency at the Klinikum am Wörthersee in Carinthia, Austria. He completed his 2-year clinical fellowship training in advanced minimal invasive gynaecological surgery at St. Michael's Hospital at the University of Toronto with Drs. Guylaine Lefebvre, Abheha Satkunaratnam, Sari Kives, Colleen Mcdermott, Deborah Robertson and Dana Soroka. Dr. Husslein?s research interests include surgical techniques, surgical education and objective skill assessement, endometriosis and urogynecology. He is a peer reviewer for multiple major obstetric and gynecologic journals Dr. Husslein is currently on the faculty of the Department of Obstetrics and Gynaecology of the Medical University of Vienna as Assistant Professor of Obstetrics and Gynaecology. His clinical focus is advanced laparoscopy and hysteroscopy as well as vaginal surgery for benign indications (eg, endometriosis, uterine fibroids, adnexal masses, uterovaginal prolapse, urinary incontinence, etc).

Sabrina Kletz is currently a Ph.D. candidate for a substantial research project on content-based video analysis of laparoscopic videos for semiautomatic surgical quality assessments. In 2009, she started her studies of computer science at Klagenfurt University and received her master's degree in 2017. In her master's thesis, she investigated the impact of temporal content features on the effectiveness of example-based video search in medical archives. In March 2016, she joined the Distributed Multimedia Systems group at the Institute of Information Technology, Klagenfurt University.

Stefan Petscharnig started his studies of computer science in 2010 at Klagenfurt University. He received his bachelor's degree in 2014 and his master's degree in 2015 from Klagenfurt University. In his master's thesis he investigated the impact of asynchronism on the Quality of Experience in social TV like scenarios using methods from Crowdsourcing, Human Computation, and Gamification. He was occupied as student research assisstant and after graduation as project assistant for the AdvUHD-DASH project in 2015. Since 2016, he works in the KISMET research project with a focus on the analysis of endoscopic video data using machine learning.

Bernd Muenzer is currently a postdoc researcher in the applied research project "EndoViP2". He received his Master's degrees in 2011 (Information Management, Mag.) and 2012 (Informatics, Dipl.-Ing.), and his Ph.D. degree (Dr. techn.) in September 2015 from Klagenfurt University, all with distinction. His research focus is on content-based analysis of videos from endoscopic surgeries in order to enable a comprehensive Video documentation and efficient post-procedural analysis of procedures. In his diploma theses he first surveyed the current situation of Video recording and usage in the medical domain of endoscopy and gave a comprehensive market overview of currently available systems. Based on that, he further investigated in detail how content-based video analysis could be incorporated in such systems in order to improve Management efficiency and generate an added value for physicians. In his Ph.D. thesis he investigated how the enormous data volume of endoscopic Video recordings can be substantially reduced by domain-specific Video compression that exploits domain-specific characteristics and also considers perceptual aspects of video quality. In EndoViP 2, he continues his research in this challenging field.

Dr. Christian Beecks worked for several years in the data management and data exploration group at RWTH Aachen University, Germany, before he moved to the Fraunhofer Institute for Applied Information Technology FIT, Germany. His research interests include efficient and adaptive multimedia data analysis, distance-based multimedia indexing and query processing, and real-time data management.