Surg Endosc (2018) 32:553–568
DOI 10.1007/s00464-017-5878-1

REVIEW

Video content analysis of surgical procedures

Constantinos Loukas[1]

Received: 14 February 2017 / Accepted: 7 September 2017 / Published online: 26 October 2017
© Springer Science+Business Media, LLC 2017

Abstract
Background In addition to its therapeutic benefits, minimally invasive surgery offers the potential for video recording of the operation. The videos may be archived and used later for reasons such as cognitive training, skills assessment, and workflow analysis. Methods from the major field of video content analysis and representation are increasingly applied in the surgical domain. In this paper, we review recent developments and analyze future directions in the field of content-based video analysis of surgical operations.
Methods The review was obtained from PubMed and Google Scholar search on combinations of the following keywords: 'surgery', 'video', 'phase', 'task', 'skills', 'event', 'shot', 'analysis', 'retrieval', 'detection', 'classification', and 'recognition'. The collected articles were categorized and reviewed based on the technical goal sought, type of surgery performed, and structure of the operation.
Results A total of 81 articles were included. The publication activity is constantly increasing; more than 50% of these articles were published in the last 3 years. Significant research has been performed for video task detection and retrieval in eye surgery. In endoscopic surgery, the research activity is more diverse: gesture/task classification, skills assessment, tool type recognition, shot/event detection and retrieval. Recent works employ deep neural networks for phase and tool recognition as well as shot detection.
Conclusions Content-based video analysis of surgical operations is a rapidly expanding field. Several future prospects for research exist including, inter alia, shot boundary detection, keyframe extraction, video summarization, pattern discovery, and video annotation. The development of publicly available benchmark datasets to evaluate and compare task-specific algorithms is essential.

Keywords Surgery · Content analysis · Video analysis · Video retrieval · Video annotation

It is well documented that minimally invasive surgery (MIS) offers important therapeutic benefits for the patient such as easier rehabilitation, shorter recovery and less pain. In addition, the endoscopic camera used for visualization of the patient's anatomy allows for effortless recording of video and images during the operation, an important feature for improving quality, efficiency, and safety of care. The recorded videos may be used for reasons such as analysis of the operational steps, review of the techniques employed, and evaluation of instrument usage. The stored digital media may also be employed for patient's briefing, or as educational material for cognitive training of junior surgeons. Another application is the assessment of surgeon's performance during the operation. Moreover, recording and archival of digital media from MIS operations is considered mandatory in some countries to provide evidence for lawsuits in case of malpractice [1]. Due to the sensitivity of the patient's personal data, guidelines for recording and archival of surgical videos are currently under development [2].

Taking into account the significance of surgical video recording, a fundamental issue is how these data could be managed and analyzed efficiently to help surgeons finding the visual information sought. Over the last few years, the videos uploaded on video-sharing websites (e.g., YouTube), or dedicated web-based resources (e.g., WebSurg),

✉ Constantinos Loukas
cloukas@med.uoa.gr

[1] Laboratory of Medical Physics, Medical School, National and Kapodistrian University of Athens, Mikras Asias 75 str., 11527 Athens, Greece

are constantly increasing. Moreover, hospitals are being equipped with mass storage servers for the archival of the videos recorded in the operating room (OR). However, software tools and techniques dedicated to the management of surgical video databases are still limited. Surgical content representation and indexing are mostly performed manually via keywords, tags, and short descriptions added into titles, metadata, etc. The inverse process of retrieval is performed in a similar manner, via text queries. Manual annotation is time consuming and provides a limited way to describe the surgical video content. Moreover, the content representation is restricted to predefined keywords/tags, and the same applies to the user's query. Hence, if, for example, one seeks for a certain surgical task that was not initially annotated, then one has to re-annotate the entire video database for this task. Therefore, indexing and retrieval of higher level concepts becomes inefficient. Moreover, recent studies have shown evidence that video analysis can improve surgeons' performance, highlighting the educational importance of video content analysis [3, 4].

Visual content representation and retrieval constitutes a major field of research that aims to support efficient image and video understanding. The main goal is to develop computational techniques able to extract higher level semantics by analyzing the video content, rather than searching for predefined keywords. Based on this idea, a plethora of applications has emerged across several domains, such as in sports video analysis, TV news analysis, intelligent management of movies, and video surveillance [5]. A major challenge is to bridge the semantic gap, the linking between low-level visual features and high-level concepts representing the visual content. In the multimedia domain, the generic framework for video content analysis includes: structure analysis (segmentation into structural units), feature extraction (for object/activity representation), data mining (using the features extracted), and classification/annotation (for building a semantic video index). The video database is then searched using the constructed index in conjunction with a distance similarity measure. To date, several techniques have been developed under this framework [6]. Recently, visual classification has moved from local features and standard machine learning tools, to the employment of deep convolutional neural networks (CNNs) for semantic concept detection [7].

The advances in multimedia content analysis have also inspired researchers from the field of surgical content representation. Initial works employed instrument signal data for reasons such as skills assessment [8], task recognition [9], and workflow analysis [10]. However, these approaches require the employment of specialized sensors, usually attached to the surgical instruments or surgeon's hands, which is cumbersome and may interfere with the operational process. Over the last few years, the advancement of efficient cost-effective imaging technologies, video storage hardware, and most importantly video analysis algorithms, has paved the way for developing more intriguing applications in MIS, using only the video signal from the endoscopic camera. This is a significant progress in the field of surgical technology considering that kinematic sensors provide only rough information about 'how a task is performed', rather than 'what is actually performed', which is an additional piece of information included in the visual data. Hence, by analyzing the video of a surgical procedure/task, one may develop applications which were otherwise impossible using only sensor signals, such as video annotation, task retrieval, concept detection, video summarization, and workflow analysis. For example, consider a surgeon that seeks to detect certain operational tasks or phases from a database of surgical videos, or to retrieve video segments similar to a video task query. As another case, a trainee desires to retrieve images with semantic concepts (e.g., tools, anatomy, etc.), similar to those included in a query image submitted to the database. In a more advanced level, an intelligent image analysis software able to recognize key-objects included in a surgical image/video, provides the trainee with contextual information about the depicted tissue organs or surgical tools. These are only a few hypothetical scenarios from an expanding list of applications that are currently under active investigation. It is believed that after systematic validation, research outcomes from the aforementioned application domains will constitute a significant part of the technological framework of the OR of the future [11].

Hence, prompted by the growing research activity in surgical video content analysis, in this paper we review recent developments and analyze future directions in this field. In particular, we include published research works that exploit video data for extracting higher level semantics about the surgical content of operations. The word 'content' here is used to describe semantic concepts such as activity, gesture, task, skill, event, and phase. To our knowledge, this is the first time that a structured literature survey in surgical video content analysis and representation is performed.

Methodology and outline

The articles included in this paper were collected from PubMed and Google Scholar using combinations of the following keywords: *(surgery) AND (video) AND (phase OR task OR skills OR event OR shot) AND (analysis OR retrieval OR detection OR classification OR recognition)*, resulting in 25 queries. The literature search was finalized in January 2017. The keywords were selected to describe the nature of the data employed and the technical goal sought. Each query returned a variable number of records; for example, 'surgery video phase analysis' returned 247 articles, whereas 'surgery video task detection' returned 28 articles, in PubMed.

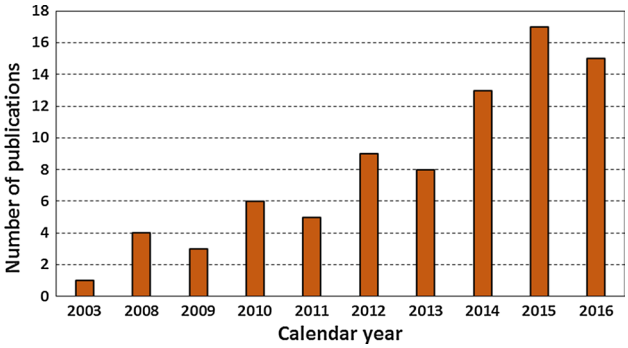

Fig. 1 Yearly number of publications in surgical video content analysis based on 81 articles retrieved from PubMed and Google Scholar

endoscopy and eye surgery. Table 1 presents a structured analysis of the articles collected. In the first category ('operation decomposition'), the operation is considered to follow the hierarchical decomposition: phases, steps, tasks/events, and gestures [8]. Eye surgery articles mostly fall under this category and hence are reviewed in a separate subsection. The second category ('content decomposition'), assumes that the video content is structured according to the descending hierarchy: shots, keyframes, and frames, similarly to that followed in multimedia content analysis [6]. For both categories the goal is to detect/classify video segments corresponding to a particular level of granularity. The third category includes works on the detection of operational tasks based on visual instrument recognition. The fourth category focuses on video-based assessment of surgical skills. The fifth category summarizes approaches towards the development of software tools for surgical video database management. Each section/subsection of this survey first provides a brief introduction about the main purpose of the papers included in the corresponding category/subcategory, followed by paper review and summary of the reported accuracy results. The final section discusses the main findings and analyzes possible directions for future research.

Table 1 Research articles categorization

Category	Sub-category	Articles
Operation decomposition (55%)	Phase recognition (11%)	[10, 18–25]
	Task, activity, gesture recognition (10%)	[28–32, 34–36]
	Surgery-event detection (9%)	[37–43]
	External camera views (10%)	[44–51]
	Eye surgery (phase, task, gesture detection) (16%)	[52–64]
Content decomposition (16%)	Shot detection (6%)	[65–69]
	Keyframe extraction (3%)	[70, 71, 73]
	Image, task, surgery retrieval (6%)	[74–78]
Instrument recognition (9%)	Instrument recognition (9%)	[79–85]
Skills assessment (15%)	Skills assessment (15%)	[86–97]
Software tools (5%)	Software tools (5%)	[72, 98–100]

Numbers in parenthesis denote percentage of articles in each category

References not indexed in these databases were followed to obtain a complete record of published data. The inclusion criterion was articles that described computational techniques for video-based content analysis and representation of surgical procedures. Articles on video-based 3D tissue reconstruction [12], instrument segmentation/tracking [13], or manual evaluation of operational videos [14, 15], were excluded. However, methods on instrument type recognition were included as the latter is closely related to surgical phase recognition. Moreover, we have excluded articles in which the surgical content representation/analysis was based on non-visual information. The primary information source for inclusion/rejection was the title and abstract, followed by the main body of the article in case the decision was ambiguous.

A total of 81 articles were found to satisfy the inclusion and exclusion criteria, with the majority of them being presented in international conferences. Figure 1 shows the yearly distribution of the articles, where it is clear that the publication activity is increasing, especially over the last 3 years. From the analysis it was found that most studies employ videos from two surgical specialties: surgical

Operation decomposition

As described earlier, surgical operations are considered to follow a descending order of granularity. This decomposition though is defined rather loosely since some levels may coincide or be further decomposed into additional sublevels. Most works on surgical video structure analysis aim at either segmenting a video into its top hierarchy levels (i.e., phases), or recognizing lower levels, such as tasks and

gestures. Apart from the aforementioned decomposition scheme, various research efforts have also been placed on the detection and classification of surgical episodes, activities and events. These semantic elements may be considered sublevels of the aforementioned hierarchy, although a formal definition is still missing from the literature.

The related literature in this category is mostly divided into techniques for video analysis of endoscopic and eye surgeries. The visual content between these procedures is significantly different and each one has its own characteristics. For example, in eye surgery, compared to endoscopic surgery, there is small tissue deformation, constant illumination, fixed camera view, and almost co-planar anatomy. Hence, although the research aim of some studies is similar, we review them separately due to the different purpose and technical challenges addressed.

Surgical phase detection

Detecting the main phases of an operation is usually defined as 'surgical workflow analysis' (SWA), an important topic of research with various applications such as skills assessment, automatic selection of teaching scenarios, and real-time workflow recognition. Especially the later would potentially offer the OR team members the opportunity to better prepare for the next case. Moreover, it could provide information about the progress of the operation to the clinical staff outside the OR, something that currently is performed manually and is prone to errors and delays.

Many techniques on SWA employ data indicating the tools used at each time step. These data may be extracted via manual annotation of videos acquired from the endoscopic camera and/or external cameras [11], RFID tags and electromagnetic (EM) sensors [16], or combinations among them [17]. Content-based video analysis for phase recognition is currently attracting growing interest. In [18], one of the first works on SWA, a feature extraction mechanism based on evolutionary reinforcement learning was presented. Endoscopic images were classified into the coarse surgical phases (6 out of 14) of laparoscopic cholecystectomy (LC) via support vector machines (SVMs). The recognition accuracy was about 50%, but the proposed features performed better than classic features such as color histograms and edge patterns. Around the same time, surgical phase recognition was approached by combining various data sources, such as instrument trajectories (obtained via infrared markers), audio signals (for detection of coagulation), as well as visual cues [10]. The purpose of visual processing was to detect the type of the instruments present at each step of the operation. The visual cues were extracted via the bag of words (BoW) model, which was applied on the SIFT features extracted from the endoscopic image. The reported accuracy for recognition of four instruments varied between 58 and 93%, whereas for phase recognition (based on all data sources) the accuracy was close to 93%. No results were provided for the visual cues alone.

A few years later, Blum et al. [19] investigated the potential of detecting all phases of a laparoscopic operation using endoscopic video data. The experiments targeted the recognition of 14 phases of LC, although the technique could be applicable to other surgeries. Simple features such as gradient magnitudes, histograms and color values were extracted, resulting in high-dimensional feature vectors. Dimensionality reduction was performed with canonical correlation analysis (CCA), or principal component analysis (PCA), using manually annotated signals of the type of instrument used at each time step (17 signals were extracted from an external camera). For phase segmentation, four different methods were compared, based on combinations of hidden Markov models (HMMs), dynamic time warping (DTW), PCA and CCA. DTW with CCA yielded the best performance (76.8%).

Dergachyova et al. proposed a method for surgical phase segmentation and recognition evaluated on a dataset from the MICCAI 2015 EndoVis challenge [20]. The dataset included videos and instrument usage annotations from 7 laparoscopic cholecystectomies. The method first employed a surgical process model (SPM), second a feature extraction process (using visual cues and instrument usage information), third an AdaBoost phase classifier, and finally a hidden semi-Markov model to generate the final decision. The visual descriptor vector (252 values) included color, shape, and texture information. Using only visual cues, the achieved accuracy was close to 68%, whereas a combination of visual and instrument information yielded an accuracy close to 90%.

In another study, phase border detection in LC videos was performed via instrument recognition. First, the video frames were split into instrument and non-instrument regions via color image analysis and binary image processing. After feature extraction, instrument type classification was performed via BoW and SVMs. Finally, the LC phases were recognized based on a set of rules with regard to the presence/absence of certain tools. Six different phase transitions were detected in a dataset of six LC videos [21].

Apart from the common feature extraction techniques, some researchers have recently started utilizing more advanced computational architectures for SWA. For example, Twinanda et al. proposed the EndoNet architecture, a CNN based on the AlexNet architecture, which was fine-tuned for online and offline tool and phase recognition tasks [22]. The method was validated on two main datasets of LC videos (Cholec80 and EndoVis), and also was compared with state-of-the-art methods, showing superior performance. The highest average accuracy was achieved for offline analysis: 92.2% (Cholec80) and 86% (EndoVis),

for phase and tool recognition, respectively. Lea et al. also employed a CNN, which captures object motion over short intervals, for offline surgical phase recognition [23]. Three different classifiers that take the learnt spatiotemporal features as input were compared with one another and also with two other methods. The validation was performed on two video datasets. In one of them (EndoVis), DTW provided the best accuracy: 91% when training was done using video and tools or only tools information, and 85% using only video information.

Research efforts for surgical phase detection have also been applied in neurosurgical interventions, where a surgical microscope is used to visualize the anatomical area [24, 25]. A dataset of 500 frames was selected from 16 pituitary surgeries. Feature vectors included information about: color (RGB and HSV histograms), texture (Haralick descriptors), and shape (spatial moments and discrete cosine transform (DCT), coefficients). After feature selection, multiclass SVMs and left–right HMMs were combined for recognizing the six phases of the procedure. Accuracy varied between 75 and 95%.

Task, activity and gesture recognition

For task, activity and gesture recognition, a lot of research has been based on modeling signals acquired from EM and optical sensors, usually attached to the surgeon's hands or instruments [9, 26, 27]. Video recognition of endoscopic surgical tasks has also received some attention. In the pioneering work of Zapella et al. [28, 29], several methods for gesture classification were proposed. Three robotic surgery (RS) tasks were pre-segmented into video clips of individual gestures (defined as surgemes). Among others, the authors proposed the use of: space–time interest points (STIPs), encoding techniques based on BoW, temporal modeling based on linear dynamical systems (LDSs), and classification schemes based on SVMs and multiple kernel learning (MKL). Using exclusively video data, the accuracy varied between 68 and 90%, depending on the combination of the techniques employed and the task examined.

With regard to video-based task segmentation and recognition in robotic surgery (RS), recent techniques showed that employing the video signal from the robot can be effective. In [30], a combined Markov/semi-Markov conditional random field (MsM-CRF) model for joint segmentation and recognition of gestures from kinematic and video data was proposed. The method was able to capture both local and global cues. The authors applied the technique on kinematic and video data on the same dataset used in [29]. The results were almost comparable in some tasks, although the technique in [30] did not assume known segmentation of the surgical tasks. In another related work [31], the authors proposed a Deformable Part Model to capture high-level features relating the robot and object parts in an image. Using a set of both kinematic and video features, an improved accuracy was achieved in extracting semantically meaningful features, as opposed to other techniques based on abstract features [30]. Recently, the same group proposed a model for RS action segmentation that combines a spatiotemporal CNN, which encodes low- and mid-level visual information, and a semi-Markov model that models high-level temporal information [32]. After evaluation on a dataset developed for RS training and assessment (JIGSAWS) [33], the proposed method yielded substantially improved performance ($\approx$74% accuracy), as compared to other recent baseline methods.

A method for unsupervised task segmentation in RS training tasks based on milestone learning was proposed in [34]. The goal was to identify milestones, regions in the state-space that denote transitions in task demonstrations, without knowledge of the labeled surgemes or a motion vocabulary. The data included kinematic features (robot pose) and two visual features (object grasp and surface penetration) extracted manually. Results showed that the identified milestones contained exactly one subtask transition in 74 and 66% of total milestones for the needle passing and suturing tasks, respectively. This method has recently been extended with automatically constructed visual features using deep CNNs [35]. In particular, the authors propose an unsupervised algorithm that leverages video and kinematic data for RS task segmentation. The algorithm finds regions of the visual feature space that mark transition events using features constructed from layers of pre-trained image classification CNNs. Using both kinematics and visual data, the algorithm matches manual annotations with up to $\approx$0.73 normalized mutual information for suturing and needle passing.

In another recent study, a CNN was trained to learn a model for the 3D position of the end effectors of the surgical robot, from video footage [36]. The video-predicted sensor values were further processed for surgical action recognition using the JIGSAWS dataset [33]. The recognition accuracy for three different training tasks was 60–77%, which was comparable to that obtained using the kinematic signals from the robot, and superior to the state-of-the-art video-based results.

Event and surgery classification

Compared to SWA, methods on video event detection and classification are limited. An early work by Lo et al. [37] investigated the detection of major events encountered in MIS. Cauterization was the event with the lowest recognition performance (<60%), whereas the experiments were performed on limited number of short segments (1–2 min). A method that analyzes laparoscopic images and indicates the possibility of an injury to the cystic artery was proposed

in [38], although no information was given about how depth data were extracted from the endoscopic camera.

In a more advanced work, Giannarou and Yang proposed a novel framework for content-based surgical scene representation by detecting key surgical episodes via probabilistic motion modeling [39]. An extended Kalman filter (EKF) was employed for tracking salient features detected by an affine-invariant anisotropic region detector. Episode borders were defined when feature tracking failed, signifying the appearance of a contrastingly different visual content. In addition, probabilistic motion modeling, via a Gaussian-like distribution, was employed for representing the motion of tracked features. Due to the lack of accepted benchmarking for video segmentation, the evaluation was focused on the motion pattern of the detected episodes. The proposed technique yielded statistically significant similarity with data of manually identified features, for surgical episodes detected in four RS videos: respiratory and camera motion, as well instrument–tissue interaction. The need for development of techniques for event-based annotation of endoscopic surgery videos was also highlighted in a recent study [40]. As a first attempt at retrieval of surgical events, a method for detecting the smoke produced by electrosurgery tasks was presented. Using ad hoc spatio-temporal features and one-class SVMs, the method was able to generate about 85% recognition accuracy.

In addition to surgical events, detecting irrelevant video frames plays an important role in terms of limiting the dataset only to those video sections that are relevant to the interval view of the patient. An unsupervised frame rejection technique was presented in [41], using a set of hard-thresholding color evaluation conditions. In another work, Munzer et al. proposed a supervised method based on color features processing, edge detection and fuzzy classification [42]. The preprocessing step included a series of steps to determine the relevant pixels that contribute to the differentiation between relevant and irrelevant frames. Based on a definition of three types of irrelevant frames (dark, blurry and out-of-patient), the method yielded an accuracy > 95% in a frame-based evaluation scheme.

A few research efforts have also targeted the automated classification of laparoscopic videos into surgery types. The works of Twinanda et al. were the first ones that addressed this problem [41, 43]. The proposed framework included a series of steps. After rejecting irrelevant frames using a color-based technique, feature vectors were extracted using various information sources such as color, saliency and gradients. After feature encoding based on vector quantization, a series of classifiers were compared (e.g., SVM, MKL), on a dataset of eight classes of abdominal surgery operations. An accuracy close to 90% was achieved, using a combination of the best: features, vector quantization techniques, and classification algorithms.

External camera views

The aforementioned works target video content analysis based on visual data obtained from the camera of the surgical endoscope to detect, segment or classify certain events, activities and tasks. Due to the complexity and variability of the surgical visual content, some researchers have employed external camera views. For example, Padoy et al. explored the potential of online recognition of the surgical phases based on information fusion: external video (for manual extraction of the instruments used at each time step), and endoscopic camera signal [44]. The latter was used to extract visual features based on color histograms. After PCA, Gaussian mixture models (GMMs) were used to model the color spaces of endoscopic images and outside images (when camera was temporarily removed), using an appropriate training set. The final outcome of this process was the extraction of a binary signal indicating the presence of clips, based on the analysis of the recognized internal camera frames. This signal was then fused with the instrument signals, which formed the input of an online HMM designed for phase segmentation. The error classification varied from 1 to 50%, depending on the surgical phase (14 in total).

Other research works were based on the use of video data from external cameras for recognizing events with different semantic content. For example, a system for determining the state (patient present or not, patient covered with a drape or not, etc.), of an ongoing operation was presented in [45]. Color features were combined with SVMs and HMMs to compute a sequence of OR states. In [46] a method was described for the detection of anomalous motion occurred in the OR. It adopted cubic higher order local auto-correlation (CHLAC) features, extracted from inter-frame differentials of the video data. The results were limited to examples of anomalous motion detection in one surgery. A multichannel audio-visual recording system for the (manual) identification of potential risks in surgical workflow was presented in [47]. The visual information was obtained from cameras views of the patient, surgical navigation console, anesthetist and scrub nurse, as well as bird-eye views of the OR from two different angles. The aim was rather to develop an integrated system synchronizing multiple signal sources.

In [48] a multi-view RGB-depth (RGBD) camera system was mounted on the ceiling of an OR. After visual feature extraction, from the recorded videos, and feature encoding-clustering, up to 15 actions were recognized with variable accuracy (24–93%). The RGBD camera setup has recently been used by the same group for surgical phase recognition (8 phases), in verteroblasty procedures [49]. Visual features were extracted using a CNN architecture similar to that in [22], and then passed to a recognition pipeline that consisted of SVMs and hierarchical HMMs (HHMMs). Based

on 37 surgeries, the recognition accuracy (offline and online) achieved was about 95%.

In another study, a multiple-camera setup composed of five cameras was used to capture seven phases of cholecystectomy in a surgical simulation environment [50]. Four cameras captured the surgical room, the rack, the trays, and the operation field, and the remaining one was the laparoscope. After optical flow extraction from labeled clips, HMM and latent Dirichlet allocation (LDA) algorithms were employed for training and classification.

In addition to color cameras, the potential for activity recognition from other camera systems has also been investigated. For example, Unger et al. employed an infrared thermal camera for recognizing surgical activities in endoscopic sinus surgery [51]. The acquired data included spatial information for hand temperatures. This datum was analyzed to perform recognition of 12 gestures, based on heat differences between the surgeon's warm hands and the colder background of the environment. The system achieved precision and recall rates of about 60%.

Eye surgery

In ophthalmologic surgery an increasing number of research methods on content-based video analysis has been presented, mostly by French research groups. These works may be categorized into two main groups: step/task retrieval and phase recognition, although the distinction between tasks and phases is not always clear. Usually the first group refers to low-level tasks (motion activities, etc.), whereas the second one refers to higher level tasks (surgical phases). However, as described earlier, the visual content and workflow of eye surgery is largely different and consequently the proposed methods are based on different conditions and constraints.

In many works, the challenge of video-based task retrieval is based on the assumption that the dataset has been in some way segmented (usually by a surgeon) into surgical steps/tasks and the goal is to retrieve those ones that match the query task. Initial works proposed motion tracking for generating visual features [52, 53]. Motion vectors from the MPEG-4 codec and Kalman filtering were employed for this purpose. Using a database of 20 videos annotated for three steps of epiretinal membrane surgeries (injection, coat and vitrectomy), a comparison between the query video and the video database was performed with DTW. The precision achieved was about 62%.

Using a similar dataset, Quellec et al. proposed a video retrieval technique using color–texture, optical flow, and corner-motion features [54]. The similarity of the target sequence with those included in the database was evaluated via the computation of a fast low-level squared Euclidian distance. Areas under the receiver operating characteristic (ROC) curves (A_z), ranged from 0.81 to 0.99. This work was later extended to combine instant feature vectors into a new vector that is unchanged by variations in duration and temporal structure among key subsequences of surgical tasks [55]. A novel algorithm to learn this mapping offline was described. This mapping allowed for real-time searches. The overall system was evaluated in the detection of three tasks in retinal surgery ($A_z > 0.91$), and nine tasks in cataract surgery ($A_z = 0.73$–0.98, depending on the task).

Quellec et al. have also proposed an approach to real-time task recognition in cataract surgery, based on the modeling of the motion content [56, 57]. The videos were first normalized for eye motion and zoom variations using pupil and scale factor tracking techniques. After feature extraction based on optical flow, spatiotemporal polynomials were employed for multiscale characterization of the features motion. Given a surgical task, the system was trained to recognize the spatiotemporal polynomials corresponding to this task. Using a supervised multiple-instance learning approach, the key polynomials were then used for recognizing the desired task. The method was tested for both task recognition and joint segmentation–recognition of tasks, resulting in an improved accuracy compared to that in [55].

Recently, the same group proposed a method for joint recognition of surgical tasks and phases [58, 59]. Two observation sources extracted from subsequences were investigated: manual annotations of tools, and image motion features (motion-based histograms analyzed with BoW). Various multilevel statistical models (based on combinations of Bayesian Networks (BN), HMMs, CRFs, and HHMMs) were tested on a dataset of 30 cataract surgeries. The best results were obtained for the combination BN + CRF, using as input the manual tool annotations ($A_z > 0.98$), whereas for the motion features the results were inferior ($A_z \approx 0.76$).

In another work, the same group studied the impact of eye motion and zoom scale variations in the extraction of motion features for task recognition in cataract surgery videos [60]. Video frames were normalized (i.e., refined) in various ways via some preprocessing steps presented in [61]. Using a similarity measure adapted from the field of video surveillance, the proposed refinement method was compared with other retrieval methods for two types of features (STIPs and motion histograms). After preprocessing, the retrieval performance was improved for most surgical tasks examined (nine in total).

A number of techniques have also been presented for the recognition of high-level surgical tasks. Lalys et al. presented a framework based on the extraction of semantic visual features with regard to the shape of the pupil and the identification of the instruments [62]. Problem-specific segmentation and classification algorithms were applied for this purpose. In addition, some global features related to the texture and color content of the video frame were used. After feature encoding with a BoW model, HMMs and DTW

were employed for phase recognition in cataract surgery. The framework was evaluated on test images (for pupil segmentation and instrument classification) and videos (for phase recognition). An overall 94% accuracy was achieved.

The same group also combined the aforementioned work with SPMs, for recognizing low-level tasks in cataract surgery [63, 64]. SPMs were defined as a set of activities towards the achievement of a surgical objective. The activities were defined by the triplet: (action, tool, and structure). Because the actions were very hard to identify, the focus was placed on the other two. The authors applied image analysis algorithms for tool detection and anatomical structure segmentation based on various visual features (color, SURF, etc.). Twelve actions, 13 surgical tools, and 6 structures were identified, the combination of which resulted in 18 possible activities. Based on the hypothesis that most activities occur in 1–2 phases (out of 8 in total), 25 possible pairs of activities were identified with a recognition rate of 65%.

Content decomposition

Instead of the decomposition scheme based on phases, tasks, etc., other researchers consider the hierarchy: scenes, shots, keyframes and frames. The central task is to segment the video into individual units with semantic content, leading to techniques such as shot boundary detection, scene segmentation and keyframe extraction. Additional applications may then be explored such as summarization, browsing, annotation and retrieval. For surgical videos, the aforementioned hierarchy has only recently been adopted by some works, which are reviewed below.

Shot detection

Video segmentation into structural units constitutes a fundamental task in video content analysis. The segmented shots may then be used for representation, indexing and retrieval of the visual information. An early effort on surgical video segmentation was presented in [65], which describes the architecture of such a system. Key components include: segmentation engine, retrieval engine, and assessment module. Shots are segmented by detecting the boundaries via color histogram differences and self-similarity matrix analysis, described in [66]. The authors also proposed key frame extraction (from the shots), based on standard techniques such as k-means clustering and salient region detection. Key frames of neurosurgical video sequences were presented. However, the system was evaluated for its retrieval performance in a non-surgical image database.

In endoscopic surgery, techniques on video shot detection are still limited. As a first attempt, Primus et al. proposed a method based on differences of motion [67]. Using the well-known Kanade–Lucas–Tomasi tracker, an aggregate movement vector was extracted separately for nine areas in each frame. Using a sliding temporal window, the border was detected by analyzing the spatiotemporal deviation of the length of these vectors. The performance of the algorithm was tested on 20 video segments (≈ 3 min long) of laparoscopic thyroid surgeries. Mean precision and recall was 86%. The authors also measured coverage (79%) and overflow (57%), where the former reflected the percentage of a truly detected shot, and the later denoted the percentage of a detected segment that falsely exceeds the length of a true segment.

Recently, Loukas et al. proposed a spatiotemporal tracking technique for shot border detection [68]. The video sequence was first decomposed into consecutive clips. Color-motion feature vectors extracted from each clip were modeled with GMMs using a variational Bayesian (VB) methodology. The estimated components were then matched along the clip sequence via the Kullback–Leibler distance. Shot borders were defined when component tracking failed, signifying a different visual appearance of the surgical scene. Using a dataset of 53 laparoscopic cholecystectomy videos, the method was compared with that in [67] and GMMs, resulting in higher performance in most assessment metrics (precision/recall $> 80\%$ and coverage $= 84\%$).

In another recent paper, Varytimidis et al. described a novel approach to surgical video retrieval using deep CNNs [69]. Laparoscopic cholecystectomy videos were split into shots when the region changed significantly, which was determined by the variation of an objectness model. Deep CNNs were used as global frame descriptors, which were aggregated into a single shot descriptor to allow for fast retrieval of similar videos. The authors also proposed novel criteria for method evaluation and provided statistical results on a retrieval framework. Evaluating the performance of different network topologies and layers, the method exceeded the state-of-the-art using local features and temporal trajectories. The reported shot retrieval accuracy, based on a tool type recognition criterion, was $> 80\%$.

Keyframe extraction

Keyframe extraction usually follows shot segmentation. Given the redundancies existing among the frames of the same shot, the main goal is to extract a small number of frames that best represent the shot content. Indicative features used for keyframe extraction include: color, texture, edges, and motion information. In MIS, where videos contain quite irregular movements, noise, and blurred frames, keyframe extraction has been explored by a limited number of works. An effort for video summarization of arthroscopic procedures was presented in [70]. The proposed tool generated a keyframe-based summary by clustering similar

frames. Five different combinations of features and dissimilarity metrics were employed.

Schoeffmann et al. proposed a keyframe extraction method for endoscopic surgery videos [71]. The video was first divided into a fixed number of clips, and then keyframes were extracted from each clip. Using ORB (Oriented FAST and Rotated BRIEF) descriptors, nearby frames were matched to determine frames of significant content change (candidate keyframes). These keyframes were inspected further, and blurry frames as well as frames that were similar to already selected keyframes, were removed. Using a dataset of ten segments, the method was evaluated by external observers, based on the 'appropriateness' of the extracted keyframes, using a three-point Likert-scale. The average rating of the method was higher than that of other standard methods as well as a prior technique based on color histogram similarity [72].

The problem of selecting too few or too many keyframes was addressed in another recent study from the same group [73]. Using a set of keyframes extracted when motion tracking failed, hierarchical clustering was performed resulting in a dendrogram, which was used to assign a priority to each keyframe. The result was a binary dendrogram for a set of representatives. A browser was proposed that employed timestamps for timeline-based visualization of the representatives. The keyframes are pinned to the timeline based on their temporal position. Dynamic browsing may be performed with the aid of the mouse.

Video retrieval

Previously, it was shown that most works in video analysis of eye surgeries are concerned with the retrieval of surgical tasks and activities. For endoscopic surgery, the papers in this area are still limited, and most of them have been published recently. A few studies have addressed the problem of linking an image (acquired during the operation), with the video segment that it was captured from. This problem was first addressed by Roldan-Carlos et al. in [74]. Three different methods for video retrieval were implemented based on global features (color, edges, etc.) and the local SIMPLE descriptor. The methods were tested on >1200 video shots and 600 query images, and the resulting accuracy in image linking was close to 80%. However, significant processing was required for creating ranked lists for each separate feature, which were then fused into a unified similarity metric. Beecks et al. proposed another method for endoscopic surgery video retrieval using a signature-based approach for linking query images with video segments [75, 76]. The method was based on an adaptive-binning feature signature model, for feature selection, and a variant of the signature matching distance, for image-video linking. The method required the creation of a database of features extracted from the target pool of video segments. The features of the query image were then compared to each one from the database. On average the recall performance was about 80%.

Twinanda et al. tackled the problem of retrieving the time boundaries of a task in a laparoscopic video [77]. After rejection of irrelevant frames (based on color histogram thresholding), feature descriptors based on histogram of gradients (HOG) were encoded via Fisher Kernel (FK). The query feature representations were compared with those in the target video using a novel coarse-to-fine temporal search to find the time boundaries. The underlying assumption was that the task was present in every target video. Using a dataset of 79 surgeries, the lowest–highest precision and recall achieved among four query tasks was: 23–79% and 26–78%, respectively. Several other configurations based on BoW and DTW were also performed, but with lower accuracy.

A novel approach to instructive video retrieval from a database of laparoscopic training tasks (peg transfer) was proposed in [78]. The idea was based on the selection of those videos that are 'educationally similar' to the query video. Three primitive query actions were defined for this task (lift, transfer, and place). The first step was to evaluate the 'instructiveness' of the action video based on relative attribute learning (three skill attributes were proposed). Then, a hybrid ranking SVM method was presented for video retrieval. The retrieval accuracy, in terms of selecting videos with higher instructiveness, across the three actions varied from 75 to 90%.

Instrument type recognition

As shown in the previous sections, a significant number of papers focused on applications related to video structure analysis of surgeries. Given that in many operations the recognition of a surgical phase/task is largely related to the type of the instrument used, some research efforts have focused on image-based tool type recognition.

An early study presented a method based on template matching of the tip using pre-constructed 3D virtual models [79]. The 3D pose of the tool was obtained using color segmentation of images obtained from a stereo endoscope. A 3D optical tracking device was also used. The experimental system was evaluated on single images in a simulation training environment. In RS, a multiple tool detection framework was presented in [80]. Different detectors for each tool were learnt using a novel object detector. The detector captured the object shape via deformable part model (DPM), which consists of a pictorial model that links root of the model to its parts using deformable springs. The model captured articulation and allowed for learning a detector for different configurations. HOG features and latent SVMs were used for tool classification in a simulation training environment.

Results on tool recognition were not reported though as the primary aim was to identify tool attributes such as open/close.

Clinical data for instrument classification in laparoscopic surgery were used in [81, 82]. The authors proposed the use of instrument classification to enable semantic segmentation of laparoscopic videos. The technique was based on standard algorithms for feature extraction (SURF, SIFT, ORB), encoding (BoW) and classification (SVMs), using a training set of images with six different instruments. For the best combination, the average accuracy achieved was 80–90%.

In eye surgery, a three-step method for surgical tool detection was presented in [83]. In the first step, standard image analysis algorithms (smoothing, thresholding, dilation) were used for color segmentation of the tool. Next, various feature descriptors (SURF, SIFT, etc.), were extracted from the segmented areas. After creating a global descriptor for each image using BoW, the final step was classification, which was based on a k-nearest neighbor algorithm. The SURF/SIFT combination generated an accuracy of 84%. In a recent study [84], a deep conventional neural network was learned offline using training patches from four eye surgery tools. The overall accuracy for tool recognition was 94% in eye phantom images, and 87% in real microsurgery datasets.

In the field of neurosurgery, a method for joint tool detection and pose estimation from microscope images was proposed in [85]. The first step aimed to label each pixel into 'tool' and 'background' using gradient, color, texture and position features. The training set included about 3000 images of 8 different tools. The second step aimed to derive the global shape of the tool present in the image by evaluating a tool-specific template on top of the labeling results. An overall tool labeling accuracy close to 86% was achieved.

Skill assessment

Objective evaluation of surgical skills has been a major topic of research for many years. Among standard parameters such time and errors, surgical performance may also be assessed using instrument/hand motion. Many studies have showed an inverse correlation between expertise level and motion parameters such as instrument pathlength and number of movements [8]. However, to generate these metrics one needs to employ specialized sensors, which is not always possible because of modification of the training setup, sterilization issues, and potential interference in task performance.

Video-based assessment of surgical skills has been introduced as an alternative with significant advantages, since video recording of an MIS task is straightforward and does not require employment of additional sensors. Significant research efforts, such as the observational clinical human reliability analysis (OCHRA) tool, have been performed recently, highlighting the importance of video-based skills assessment [14, 15]. However, these tools require an expert to review and analyze the entire video of the operation, which is labor intensive and time consuming. Skills assessment based on computational video analysis provides significant advantages over manual evaluation, although it requires the resolve of several technical challenges and the extraction of high-level semantic information. One basic hypothesis is that the tool shaft/tip movement may reveal salient features that can be extracted by the endoscopic camera. Hence, by detecting these image features one may indirectly assess the tool movement inside the box trainer [86].

In laparoscopic surgical training, initial research focused on detecting salient features (STIPs) from the endoscopic video, feature encoding with BoW, and temporal modeling with a hidden Markov process [87]. The classification of the motion expertise level for a testing sequence was based on choosing a model that maximized the likelihood of the given sequence. Subjects performed the 'peg transfer' task in an FLS-specific training environment. The total accuracy achieved was 86–89%. The model also provided meaningful insights about the motion patterns of experts and novices. The same group also proposed a novel formulation (termed Relative HMMs), for video-based evaluation of motion skills in laparoscopic training [88]. The method makes the reasonable assumption that the trainees improve their skills over time, so the video sequences are relatively ranked based on the time performance. Using the feature extraction method described in [87], the proposed algorithm learns a model from the training data so that the attribute under consideration is linked to the likelihood of the input, hence supporting comparison of new sequences.

Other researchers have considered external cameras to monitor the trainee's movements during task performance, such as in [89–92]. In [89], the feature extraction/encoding process was similar to that in [87]. The learning algorithm utilized CCA to discover the latent relationship between different video streams that captures different scenes and movements arising from the same physical process. Classification was based on SVMs. Using video data from performances of the peg transfer task, the accuracy in recognizing experts from novices was $>90\%$. The same task was also used in [90], where video data from two external cameras were analyzed using optical flow algorithms. A 'hand movement' parameter, representing the smoothness of optical flow, was extracted with a custom software, but no further details were provided about its computation. A commercial video analysis software was used in [91], allowing the computation of parameters, such as angles and distances, directly from the video of the training task. However, neither results from the extracted measurements nor analysis about how these measurements were computed are provided. A compact surgical

skill training and evaluation tool based on video analysis from three camera sources was recently presented in [92]. Two webcams recorded hand movement, and a third one the tool movement. Color segmentation and motion history image (MHI) were used to compute various parameters (smoothness, acceleration, etc.) of the movement pattern of the tool and the hands. The system was validated for its educational value on three FLS tasks.

Instead of the common classification task into experts and novices, some studies have investigated automated assessment of the OSATS (objective structured assessment technical skills) criteria used in surgical training. An augmented BoW technique was introduced in [93], where time and motion were modeled as short sequences of events. The underlying local and global structural information was then encoded into BoW models, resulting in an accuracy 65–75% across the seven OSATS criteria. Sharma et al. proposed a novel framework based on sequential motion texture analysis [94]. The technique involved: (a) STIP detection and HOG extraction, (b) motion class learning and classification, (c) data-driven time window computation, and (d) sequential motion texture feature extraction. To encode the qualitative motion dynamics in each time window, gray-level co-occurrence matrices (GLCM) texture analysis was applied. The percentage of correctly classified videos with respect to the OSATS criteria ranged from 83 to 96%. The same group later presented another technique for representing periodic motion elements inherent in basic surgical tasks. Frequency coefficients (DFT and DCT) extracted from the time series analysis of the video-extracted visual features were employed [95]. The technique increased the classification accuracy to 91–100%. The same group recently compared different video-based OSATS assessment techniques for surgical skill evaluation [96]. These techniques capture the motion information at a coarser granularity using symbols or words, extract motion dynamics using textural patterns in a frame kernel matrix, and analyze fine-grained motion information using frequency analysis. Results showed that frequency features outperform other feature types previously reported in the literature.

In another recent study, Loukas et al. studied the application of four feature detector descriptors and two temporal models for laparoscopic skills assessment [86]. Two different setups were designed: static and dynamic video-histogram analysis. STIP-HOG yielded the best performance in classifying expert's and novice's performance, independent to the employed temporal model. Important differences were found between the two groups with respect to the underlying dynamics of the video-histogram sequences.

In the field of eye surgery, Zhu et al. have proposed a vision-based approach to evaluate cataract surgery videos, in specific the capsulorhexis step, performed on a simulator [97]. First, computer vision algorithms were used to obtain keyframes, spatial measures, and optical flow magnitude curves for each surgery. Based on these measurements, three metrics were defined (spatiality, duration, and motion), and applied in linear regression and linear SVM models to assign grades. The results showed that the method was in reasonable agreement with the experts' opinion.

Software applications

A few research efforts have been placed towards the development of software tools and small-scale applications to aid visualization and analysis of the surgical video content. One of the first studies towards this direction focused on the binding of MeSH (Medical Subject Headings) terms into the structure description of surgical videos [98]. The authors described the development of an annotation tool that creates descriptions in the MPEG-7 metadata standard using the MeSH classification. MPEG-7 metadata classes were mapped to MeSH sub-categories (e.g., *Why* to F3: Mental Disorders, *WhatAction* to E1–E6: General techniques, *How* to: E7-Equipment and supplies). Using the software developed, an example describing the annotation process for a video from a thoracoscopy procedure was presented.

A content-based surgical video analysis and management system that provides convenience in accessing the relevant content was presented in [72]. The system has two main components: temporal video segmentation into shots and content-based retrieval. Video analysis techniques included color histogram analysis, clustering, and Euclidian distance comparison. The system was implemented as a web-based application and may be used in mobile devices.

An Austrian group has recently presented its efforts towards the development of novel tools for event understanding, annotation and learning of semantically relevant segments from surgical endoscopy videos [99, 100]. In [99], the authors described a visual–vocal annotation tool with various editing functionalities such as hand drawing of annotations in anatomic areas of interest, recording spoken audio notes, bookmark setting, and annotation visualization in the video timeline. The tool was evaluated by an expert surgeon and a general event model of surgery was derived by identifying relationships between the granularity of an event and the type of its annotation. Further efforts on the recognition of relevant segments from endoscopic surgery videos were presented in [100]. The author provided prospects about the development of a social network-based platform that integrates relevant experts' knowledge. The platform will have video editing capabilities to allow experts to collaboratively edit endoscopic video contents and share them among each other. The interaction with the video editing tool will be monitored in the background, and then interpreted so that relevant information could be derived. The author

also suggested ways for data interpretation, learning of user interaction patterns, and evaluation of results.

Discussion and future directions

In this paper, we have presented recent developments in surgical video content analysis and representation. The articles that matched the inclusion criteria were categorized into theme groups and subgroups based on their technical focus. In most articles the surgical operation was assumed to follow the 'operation decomposition' scheme; eye surgery papers were the majority in this category. The main goal was the detection and retrieval of surgical phases and tasks. For endoscopic surgery, the data originated from various sources, such as simulation training tasks, actual operations, and external camera views. The employment of different data sources was mainly due to the greater technical challenges encountered, such as the frequent camera movement, the variable illumination, and the significant deformation and transition of tissue structures.

Another notable remark is that the different granularity levels (phases, tasks, etc.), were not defined consistently. For example, there is no general consensus on what defines/differentiates a surgical task, or which ones constitute the phase of an operation. These semantic concepts are closely related to the type of a surgical operation, and the same task may be met in more than one operations. Hence, ontologies and tools to describe the structure elements of a surgery are important. Important steps have already been made towards this direction [101]. The development of a formal framework of ontological definitions would signify a key step towards this direction. In addition, there is a great need for a generally acceptable benchmark dataset to allow researchers compare and evaluate their techniques. To date, JIGSAWS[1] is the only publicly available working set for gesture and skills assessment [33]. Among others, it includes video data and gesture annotations for three RS tasks performed by surgeons on a bench-top model: suturing (SU), knot-tying (KT), and needle-passing (NP). All three tasks are typically part of surgical skills training curricula. In overall, the JIGSAWS dataset consists of 39 trials of SU, 36 trials of KT, and 28 trials of NP. The video part of the dataset includes stereo video data captured from both endoscopic cameras of the da-Vinci surgical system.

The availability of a similar dataset from real surgical operations would be of significant value for the evaluation of techniques targeting other surgical video content analysis applications. Hence, a bold step towards this direction was performed last year, with the release of a phase- and instrument-annotated video dataset of seven laparoscopic cholecystectomies (EndoVis). This dataset was employed in the MICCAI 2015 challenge for surgical phase detection.[2] Recently, in the 2016 M2CAI workshop, further data were included for two separate challenges: surgical phase (m2cai16-workflow dataset) and instrument detection (m2cai16-tool dataset).[3] In overall, the dataset includes 41 cholecystectomy videos with ground truth annotations of the phases, and 15 cholecystectomy videos with ground truth binary annotations of the present tools. All videos are annotated on a 'per-frame' basis. The phase annotation includes: trocar placement, preparation, calot triangle dissection, clipping and cutting, gallbladder dissection, gallbladder packaging, cleaning and coagulation and gallbladder retraction (eight phases). The tool annotation includes: grasper, hook, clipper, bipolar, irrigator, scissors and specimen bag (seven tools). The same group has also released a similarly annotated, but bigger, dataset containing 80 videos of cholecystectomy surgeries performed by 13 surgeons (Cholec80).[4] To date, these are the most comprehensive annotated video datasets of surgical operations, which are publicly available for academic research. Due to their superior quality and resolution, the videos may also be used for other video content analysis applications (e.g., retrieval, recognition, summarization etc.), from researchers that have no access to such a specialized dataset. Development of similarly annotated MIS datasets for other levels of video granularity (e.g., tasks, shots, keyframes, etc.) would be of significant value.

In the 'content decomposition' category, most research papers have been published within the last 2–3 years. This is a new area of research that has great potentialities, considering the already extensive research in multimedia content analysis (e.g., news, sports, movies, etc.). Hence, techniques and ideas already proposed for video segmentation and database management may well be adapted in the surgical domain. An apparent issue is that a surgical video is composed of a single shot, since the captured area does not change. Nevertheless, important changes may be detected, such as the insertion or removal of a surgical tool, manipulation of an organ, or the viewpoint change. Moreover, one may combine techniques from both categories (task and shot detection), to explore other intriguing applications, such as automated video summarization of surgical operations, modeling of user's preferences during video browsing, suggestions for video tasks with similar content and presentation of information related to objects recognized in the video.

In the area of visual instrument recognition, the number of published papers is limited, especially in terms of

[1] https://cirl.lcsr.jhu.edu/research/hmm/datasets/jigsaws_release/

[2] http://grand-challenge.org/site/endovissub-workflow/data/

[3] http://camma.u-strasbg.fr/m2cai2016/index.php/program-challenge/

[4] http://camma.u-strasbg.fr/datasets

employing data from real surgical operations. Most works employ visual features and a classifier trained on labeled images. The number of papers that combine in vivo endoscopic views with virtual models of instruments is limited. Moreover, although the instruments used are closely related to the phases of a surgical operation [11], few papers have exploited video data for this purpose. These research directions certainly deserve further attention in the future.

For video-based assessment of surgical skills, endoscopic as well as external cameras have been employed. The general framework that is followed includes visual feature extraction, encoding, and classification based on a training set of different skill levels. A disadvantage of this approach is that it does not provide information about certain errors committed by the trainee, but it rather generates abstract performance parameters correlated to those extracted from a predefined skills group. Very few studies so far have attempted to correlate visual features with absolute measures of performance [94]. A potential extension to this approach would be the employment of additional cameras and pre-constructed 3D models of the training setup. Hence, using instrument tracking, image registration, and object recognition algorithms, one would be able to derive absolute performance parameters, such as for example unsuccessful attempts for cutting/peg placement, and number of times that a needle came in contact with the wall. In addition to skills recognition, assessment of the trainee's learning curve is also important. Thus, given that trainees perform a training task several times before reaching a plateau, a video-based framework that provides information about the performance level with respect to the learning curve would be of great educational value. Other ideas for potential research include visual recognition of erroneous actions and video-based image-guided surgical training.

With regard to software application development, a few sporadic approaches have been proposed for surgical video database management. Most of them described small-scale systems designed as research tools for visual evaluation of video analysis algorithms. An interesting system was proposed in [98], which describes an interface for surgical video organization based on terms derived from a medical lexicon. Future research efforts could be placed in the development of semi-automated video management systems that combine similar approaches with automated video content analysis algorithms. Due to the tedious process of manual annotation, the development of a private social network, such as the one proposed in [100], with global, timeline-based, and image-based editing and annotation capabilities would be an important milestone. For general purpose images, online annotation tools for building image databases for computer vision research are already in use (e.g., *LabelMe*), so the surgical community may be inspired by these efforts.

A potential limitation of the present survey is that it does not include a comparison among the various methods used in each research subtopic (e.g., phase recognition, skills assessment, etc.). As described in the Introduction, the primary goals were to present, for the first time, a structured literature survey of the published research activity, and to identify the various research trends, in the field of surgical video content analysis. Given the great diversity of the video analysis algorithms, research goals and experimental datasets, a direct correlation among the various methods employed could not be performed under the structure of the present survey. In the future we aim to perform a critical review of the most active research subtopics, such as surgical phase recognition, task recognition, and skills assessment. Given the growing research interest, it is expected that in the near future the publication capacity in these subtopics will be sufficient enough to perform an authoritative critical review. Currently, the broad field of surgical video content analysis is still in its infancy, but with great potential for exploration of its various research directions.

In conclusion, content-based video analysis of surgical tasks and procedures constitutes a rapidly expanding scientific field. Several future prospects for research exist, such as shot boundary detection, keyframe extraction, video summarization, pattern discovery, and video annotation. Software applications for efficient management and organization of surgical video databases would be a useful tool for surgeons and clinical educators. Moreover, the public availability of benchmark datasets for evaluation and comparison of the implemented algorithms is essential. We hope that the findings of this survey will be inspiring not only for the advancement of the current techniques but also for the discovery of additional novel applications.

Disclosures Dr. Constantinos Loukas has no conflicts of interest or financial ties to disclose.

References

1. Henken KR, Jansen FW, Klein J, Stassen LPS, Dankelman J, van den Dobbelsteen JJ (2012) Implications of the law on video recording in clinical practice. Surg Endosc 26:2909–2916. doi: 10.1007/s00464-012-2284-6
2. Turnbull AMJ, Emsley ES (2014) Video recording of ophthalmic surgery-ethical and legal considerations. Surv Ophthalmol 59:553–558. doi: 10.1016/j.survophthal.2014.01.006
3. Dimick JB, Varban OA (2015) Surgical video analysis: an emerging tool for improving surgeon performance. BMJ Qual Saf 24:490–491. doi: 10.1136/bmjqs-2015-004439
4. Bonrath EM, Gordon LE, Grantcharov TP (2015) Characterising "near miss" events in complex laparoscopic surgery through video analysis. BMJ Qual Saf 24:516–521. doi: 10.1136/bmjqs-2014-003816
5. Snoek CGM, Worring M (2007) Concept-based video retrieval. Found Trends Inf Retr 2:215–322. doi: 10.1561/1500000014

6. Hu W, Xie N, Li L, Zeng X, Maybank S (2011) A survey on visual content-based video indexing and retrieval. IEEE Trans Syst Man Cybern C 41:797–819. doi: 10.1109/TSMCC.2011.2109710
7. Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y (2013) OverFeat: integrated recognition, localization and detection using convolutional networks. In: International conference on learning representations (ICLR 2014), Banff, pp 1–16
8. Reiley CE, Lin HC, Yuh DD, Hager GD (2011) Review of methods for objective surgical skill evaluation. Surg Endosc 25:356–366. doi: 10.1007/s00464-010-1190-z
9. Dosis A, Bello F, Gillies D, Undre S, Aggarwal R, Darzi A (2005) Laparoscopic task recognition using hidden Markov Models. Stud Health Technol Inform 111:115–122
10. Weede O, Dittrich F, Worn H, Jensen B, Knoll A, Wilhelm D, Kranzfelder M, Schneider A, Feussner H (2012) Workflow analysis and surgical phase recognition in minimally invasive surgery. In: 2012 IEEE international conference on robotics and biomimetics (ROBIO)—conference digest, pp 1068–1074. doi: 10.1109/ROBIO.2012.6491111
11. Padoy N, Blum T, Ahmadi S-A, Feussner H, Berger M-O, Navab N (2012) Statistical modeling and recognition of surgical workflow. Med Image Anal 16:632–641. doi: 10.1016/j.media.2010.10.001
12. Lin B, Sun Y, Qian X, Goldgof D, Gitlin R, You Y (2016) Video-based 3D reconstruction, laparoscope localization and deformation recovery for abdominal minimally invasive surgery: a survey. Int J Med Robot Comput Assist Surg 12:158–178. doi: 10.1002/rcs.1661
13. Bouget D, Allan M, Stoyanov D, Jannin P (2017) Vision-based and marker-less surgical tool detection and tracking: a review of the literature. Med Image Anal 35:633–654. doi: 10.1016/j.media.2016.09.003
14. Foster JD, Miskovic D, Allison AS, Conti JA, Ockrim J, Cooper EJ, Hanna GB, Francis NK (2016) Application of objective clinical human reliability analysis (OCHRA) in assessment of technical performance in laparoscopic rectal cancer surgery. Tech Coloproctol 20:361–367. doi: 10.1007/s10151-016-1444-4
15. Miskovic D, Ni M, Wyles SM, Parvaiz A, Hanna GB (2012) Observational clinical human reliability analysis (OCHRA) for competency assessment in laparoscopic colorectal surgery at the specialist level. Surg Endosc 26:796–803. doi: 10.1007/s00464-011-1955-z
16. Bardram JE, Doryab A, Jensen RM, Lange PM, Nielsen KLG, Petersen ST (2011) Phase recognition during surgical procedures using embedded and body-worn sensors. In: IEEE international conference on pervasive computing and communications, pp 45–53. doi: 10.1109/PERCOM.2011.5767594
17. Bouarfa L, Jonker PP, Dankelman J (2011) Discovery of high-level tasks in the operating room. J Biomed Inform 44:455–462. doi: 10.1016/j.jbi.2010.01.004
18. Klank U, Padoy N, Feussner H, Navab N (2008) Automatic feature generation in endoscopic images. Int J Comput Assist Radiol Surg 3:331–339. doi: 10.1007/s11548-008-0223-8
19. Blum T, Feussner H, Navab N (2010) Modeling and segmentation of surgical workflow from laparoscopic video. Lect Notes Comput Sci 6363:400–407
20. Dergachyova O, Bouget D, Huaulmé A, Morandi X, Jannin P (2016) Automatic data-driven real-time segmentation and recognition of surgical workflow. Int J Comput Assist Radiol Surg 11:1081–1089. doi: 10.1007/s11548-016-1371-x
21. Primus MJ, Schoeffmann K, Böszörmenyi L (2016) Temporal segmentation of laparoscopic videos into surgical phases. In: 14th international workshop on content-based multimedia indexing, pp 1–6
22. Twinanda AP, Shehata S, Mutter D, Marescaux J, de Mathelin M, Padoy N (2017) EndoNet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans Med Imaging 36:86–97. doi: 10.1109/TMI.2016.2593957
23. Lea C, Choi JH, Reiter A, Hager GD (2016) Surgical phase recognition: from instrumented ORs to hospitals around the world. In: Medical image computing and computer-assisted intervention M2CAI—MICCAI workshop, pp 45–54
24. Lalys F, Riffaud L, Morandi X, Jannin P (2010) Automatic phases recognition in pituitary surgeries by microscope images classification. Lect Notes Comput Sci 6135:34–44. doi: 10.1007/978-3-642-13711-2
25. Lalys F, Riffaud L, Morandi X, Jannin P (2011) Surgical phases detection from microscope videos by combining SVM and HMM. Lect Notes Comput Sci 6533:54–62
26. Megali G, Sinigaglia S, Tonet O, Dario P (2006) Modelling and evaluation of surgical performance using hidden Markov models. IEEE Trans Biomed Eng 53:1911–1919. doi: 10.1109/TBME.2006.881784
27. Loukas C, Georgiou E (2011) Multivariate autoregressive modeling of hand kinematics for laparoscopic skills assessment of surgical trainees. IEEE Trans Biomed Eng 58:3289–3297. doi: 10.1109/TBME.2011.2167324
28. Zappella L, Béjar B, Hager G, Vidal R (2013) Surgical gesture classification from video and kinematic data. Med Image Anal 17:732–745. doi: 10.1016/j.media.2013.04.007
29. Haro BB, Zappella L, Vidal R (2012) Surgical gesture classification from video data. Lect Notes Comput Sci 7510:34–41
30. Tao L, Zappella L, Hager GD, Vidal R (2013) Surgical gesture segmentation and recognition. Lect Notes Comput Sci 8151:339–346
31. Lea C, Hager GD, Vidal R (2015) An improved model for segmentation and recognition of fine-grained activities with application to surgical training tasks. In: IEEE winter conference on applications of computer vision, Waikoloa, pp 1123–1129
32. Lea C, Reiter A, Vidal R, Hager GD (2016) Segmental spatiotemporal CNNs for fine-grained action segmentation. Lect Notes Comput Sci 9907:36–52. doi: 10.1007/978-3-319-46487-9_3
33. Gao Y, Vedula SS, Reiley CE, Ahmidi N, Varadarajan B, Lin HC, Tao L, Zappella L, Bejar B, Yuh DD, Chen CCG, Vidal R, Khudanpur S, Hager GD (2014) The JHU-ISI gesture and skill assessment dataset (JIGSAWS): A surgical activity working set for human motion modeling. In: Medical image computing and computer-assisted intervention M2CAI—MICCAI workshop
34. Krishnan S, Garg A, Patil S, Lea C, Hager G, Abbeel P, Goldberg K (2016) Transition state clustering: unsupervised trajectory segmentation of multi-modal demonstrations with deep learning. In: IEEE international conference on robotics and automation, Genova, Italy, pp 1–8
35. Murali A, Garg A, Krishnan S, Pokorny FT, Abbeel P, Darrell T, Goldberg K (2016) TSC-DL: Unsupervised trajectory segmentation of multi-modal surgical demonstrations with deep learning. In: IEEE international conference on robotics and automation, Stockholm, Sweden, pp 1–8
36. Rupprecht C, Lea C, Tombari F, Navab N, Hager GD (2016) Sensor substitution for video-based action recognition. In: 2016 IEEE/RSJ international conference on intelligent robots and systems, pp 5230–5237. IEEE
37. Lo BPL, Darzi A, Yang G (2003) Episode classification for the analysis of tissue/instrument interaction with multiple visual cues. In: 6th international conference on medical imaging and computer-assisted intervention, Montréal, pp 230–237
38. Lahane A, Yesha Y, Grasso M, Joshi A, Park A, Lo J (2012) Detection of unsafe action from laparoscopic cholecystectomy video. In: Proceedings of the 2nd ACM SIGHIT international health informatics symposium—IHI 2012. ACM Press, New York, pp 315–322

39. Giannarou S, Yang G (2010) Content-based surgical workflow representation using probabilistic motion modeling. Lect Notes Comput Sci 6326:314–323
40. Loukas C, Georgiou E (2015) Smoke detection in endoscopic surgery videos: a first step towards retrieval of semantic events. Int J Med Robot Comput Assist Surg 11:80–94. doi: 10.1002/rcs.1578
41. Twinanda AP, Marescaux J, de Mathelin M, Padoy N (2015) Classification approach for automatic laparoscopic video database organization. Int J Comput Assist Radiol Surg 10:1449–1460. doi: 10.1007/s11548-015-1183-4
42. Munzer B, Schoeffmann K, Boszormenyi L (2013) Relevance segmentation of laparoscopic videos. In: IEEE international symposium on multimedia. IEEE, Anaheim, pp 84–91
43. Twinanda AP, Marescaux J, de Mathelin M, Padoy N (2014) Towards better laparoscopic video database organization by automatic surgery classification. Lect Notes Comput Sci 8498:186–195
44. Padoy N, Blum T, Feußner H, Berger M-O, Navab N (2008) Online recognition of surgical activity for monitoring in the operating room. In: 20th national conference on innovative applications of artificial intelligence (IAAI 2008), pp 1718–1724
45. Bhatia B, Oates T, Xiao Y, Hu P (2007) Real-time identification of operating room state from video. In: 19th international conference on innovative applications of artificial intelligence. Vancouver, pp 1761–1766
46. Sakabe F, Murakawa M, Kobayashi T, Higuchi T, Otsu N (2009) Anomalousness detection for surgery videos using CHLAC feature. In: Symposium on bio-inspired, learning, and intelligent systems for security (BLISS 2009). IEEE, Edinburgh, pp 66–68
47. Suzuki T, Sakurai Y, Yoshimitsu K, Nambu K, Muragaki Y, Iseki H (2010) Intraoperative multichannel audio-visual information recording and automatic surgical phase and incident detection. In: International conference of the IEEE engineering in medicine and biological society, pp 1190–1193
48. Twinanda AP, Alkan EO, Gangi A, de Mathelin M, Padoy N (2015) Data-driven spatio-temporal RGBD feature encoding for action recognition in operating rooms. Int J Comput Assist Radiol Surg 10:737–747. doi: 10.1007/s11548-015-1186-1
49. Twinanda AP, Winata P, Gangi A, De M (2016) Multi-stream deep architecture for surgical phase recognition on multi-view RGBD videos. Medical image computing and computer-assisted intervention M2CAI—MICCAI workshop, pp 25–34
50. Tran D, Sakurai R, Lee J (2015) An improvement of surgical phase detection using latent dirichlet allocation and hidden Markov model. In: Innovation in medicine healthcare. Springer, Cham, pp 249–261
51. Unger M, Chalopin C, Neumuth T (2014) Vision-based online recognition of surgical activities. Int J Comput Assist Radiol Surg 9:979–986. doi: 10.1007/s11548-014-0994-z
52. Droueche Z, Lamard M, Cazuguel G, Quellec G, Roux C, Cochener B (2011) Content-based medical video retrieval based on region motion trajectories. In: 5th european conference of the international federation for medical and biological engineering, Budapest, pp 622–625
53. 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: IEEE engineering in medicine and biology society, San Diego, pp 4962–4965
54. Quellec G, Lamard M, Cazuguel G, Droueche Z, Roux C, Cochener B (2011) Real-time retrieval of similar videos with application to computer-aided retinal surgery. In: International conference of the IEEE engineering in medicine and biological society, Boston, pp 4465–4468
55. Quellec G, Charrière K, Lamard M, Droueche Z, Roux C, Cochener B, Cazuguel G (2014) Real-time recognition of surgical tasks in eye surgery videos. Med Image Anal 18:579–590. doi: 10.1016/j.media.2014.02.007
56. Quellec G, Lamard M, Droueche Z, Cochener B, Roux C, Cazuguel G (2013) A polynomial model of surgical gestures for real-time retrieval of surgery videos. Lect Notes Comput Sci 7723:10–20. doi: 10.1007/978-3-642-36678-9_2
57. Quellec G, Lamard M, Cochener B, Cazuguel G (2015) Real-time task recognition in cataract surgery videos using adaptive spatiotemporal polynomials. IEEE Trans Med Imaging 34:877–887
58. Charriere K, Quellec G, Lamard M, Martiano D, Cazuguel G, Coatrieux G, Cochener B (2016) Real-time multilevel sequencing of cataract surgery videos. In: 14th international workshop on content-based multimedia indexing, pp 1–6
59. Charrière K, Quellec G, Lamard M, Martiano D, Cazuguel G, Coatrieux G, Cochener B (2016) Real-time analysis of cataract surgery videos using statistical models. arXiv:1610.05465
60. Charriere K, Quellec G, Lamard M, Coatrieux G, Cochener B, Cazuguel G (2014) Automated surgical step recognition in normalized cataract surgery videos. In: International conference of the IEEE engineering in medicine and biology society, Chicago, pp 4647–4650
61. Quellec G, Charriére K, Lamard M, Cochener B, Cazuguel G (2014) Normalizing videos of anterior eye segment surgeries. In: International conference of the IEEE engineering in medicine and biology society, pp 122–125
62. Lalys F, Riffaud L, Bouget D, Jannin P (2012) A framework for the recognition of high-level surgical tasks from video images for cataract surgeries. IEEE Trans Biomed Eng 59:966–976. doi: 10.1109/TBME.2011.2181168
63. Lalys F, Bouget D, Riffaud L, Jannin P (2013) Automatic knowledge-based recognition of low-level tasks in ophthalmological procedures. Int J Comput Assist Radiol Surg 8:39–49. doi: 10.1007/s11548-012-0685-6
64. Lalys F, Riffaud L, Bouget D, Jannin P (2011) An application-dependent framework for the recognition of high-level surgical tasks in the OR. Int Conf Med Image Comput Comput Interv 14:331–338
65. Mendi E, Cecen S, Ermisoglu E, Bayrak C (2010) Automated neurosurgical video segmentation and retrieval system. J Biomed Sci Eng 3:618–624. doi: 10.4236/jbise.2010.36084
66. Mendi E, Bayrak C (2011) Shot boundary detection and keyframe extraction from neurosurgical video sequences. Imaging Sci J 60:90–96. doi: 10.1179/1743131X11Y.0000000005
67. Primus MJ, Schoeffmann K, Böszörmenyi L (2013) Segmentation of recorded endoscopic videos by detecting significant motion changes. In: 11th international workshop on content-based multimedia indexing, Veszprem, pp 223–228
68. Loukas C, Nikiteas N, Schizas D, Georgiou E (2016) Shot boundary detection in endoscopic surgery videos using a variational Bayesian framework. Int J Comput Assist Radiol Surg 11:1937–1949. doi: 10.1007/s11548-016-1431-2
69. Varytimidis C, Rapantzikos K, Loukas C, Kollias S (2016) Surgical video retrieval using deep neural networks. In: Medical image computing and computer-assisted intervention M2CAI—MICCAI workshop, pp 4–14
70. Lux M, Marques O, Schöffmann K, Böszörmenyi L, Lajtai G (2009) A novel tool for summarization of arthroscopic videos. Multimed Tools Appl 46:521–544. doi: 10.1007/s11042-009-0353-1
71. Schoeffmann K, Del Fabro M, Szkaliczki T, Böszörmenyi L, Keckstein J (2014) Keyframe extraction in endoscopic video. Multimed Tools Appl 74:11187–11206. doi: 10.1007/s11042-014-2224-7
72. Mendi E, Bayrak C, Cecen S, Ermisoglu E (2013) Content-based management service for medical videos. Telemed e-Health 19:36–41. doi: 10.1089/tmj.2011.0239

73. Lokoc J, Schoeffmann K, del Fabro M (2015) Dynamic hierarchical visualization of keyframes in endoscopic video. Lect Notes Comput Sci 8936:291–294
74. Roldan-Carlos J, Lux M, Giró-i-Nieto X, Muñoz P, Anagnostopoulos N (2015) Visual information retrieval in endoscopic video archives. In: International workshop on content-based multimedia indexing, Prague, pp 109–114
75. 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: Del Bimbo A, Chen S-C, Wang H, Yu H, Zimmermann R (eds) IEEE proceedings of the international symposium on multimedia, Miami, pp 1–6
76. Schoeffmann K, Beecks C, Lux M, Seran M, Seidl T (2016) Content-based retrieval in videos from laparoscopic surgery. In: SPIE medical imaging: image-guided procedures, robotic interventions, and modeling, San Diego, pp 1–10
77. Twinanda AP, de Mathelin M, Padoy N (2014) Fisher kernel based task boundary retrieval in laparoscopic database with single video query. Med Image Comput Comput Interv 17(3):409–416. doi: 10.1007/978-3-319-10443-0_52
78. Chen L, Zhang P, Li B (2014) Instructive video retrieval based on hybrid ranking and attribute learning a case study on surgical skill training. In: 22nd ACM international conference on multimedia (ACM MM), Orlando, pp 1045–1048
79. Speidel S, Benzko J, Krappe S, Sudra G, Azad P, Müller-Stich BP, Gutt C, Dillmann R (2009) Automatic classification of minimally invasive instruments based on endoscopic image sequences. In: SPIE medical imaging: image-guided procedures, robotic interventions, and modeling, pp 72610A–72610A1
80. Kumar S, Narayanan MS, Misra S, Garimella S, Singhal P, Corso JJ, Krovi V (2013) Vision based decision-support and safety systems for robotic surgery. In: Proceedings of workshop on medical cyber physical systems
81. Primus MJ, Schoeffmann K, Böszörmenyi L (2015) Instrument classification in laparoscopic videos. In: International workshop on content-based multimedia indexing, Prague, pp 1–6
82. Beecks C, Schoeffmann K, Lux M, Uysal MS, Seidl T (2014) Segmentation and indexing of endoscopic videos. In: ACM international conference on multimedia (ACM MM), Orlando, pp 659–662
83. Bouget D, Lalys F, Jannin P (2012) Surgical tools recognition and pupil segmentation for cataract surgical process modeling. Stud Health Technol Inform 173:78–84
84. Alsheakhali M, Eslami A, Navab N (2015) Microscopic surgical tool type detection. In: Proceedings of MICCAI workshop on interventional microscopy, Munich, pp 1–8
85. Bouget D, Benenson R, Omran M, Riffaud L, Schiele B, Jannin P (2015) Detecting surgical tools by modelling local appearance and global shape. IEEE Trans Med Imaging 34:2603–2617. doi: 10.1109/TMI.2015.2450831
86. Loukas C, Georgiou E (2016) Performance comparison of various feature detector-descriptors and temporal models for video-based assessment of laparoscopic skills. Int J Med Robot Comput Assist Surg 12:387–398. doi: 10.1002/rcs.1702
87. Zhang Q, Li B (2011) Video-based motion expertise analysis in simulation-based surgical training using hierarchical dirichlet process hidden markov model. In: International ACM workshop on medical multimedia analysis and retrieval. ACM Press, New York, pp 19–24
88. Zhang Q, Li B (2015) Relative hidden Markov models for video-based evaluation of motion skills in surgical training. IEEE Trans Pattern Anal Mach Intell 37:1206–1218. doi: 10.1109/TPAMI.2014.2361121
89. Zhang Q, Chen L, Tian Q, Li B (2013) Video-based analysis of motion skills in simulation-based surgical training. In: SPIE, multimedia content and mobile devices, Burlingame, pp 86670A–86770A
90. Gray RJ, Kahol K, Islam G, Smith M, Chapital A, Ferrara J (2012) High-fidelity, low-cost, automated method to assess laparoscopic skills objectively. J Surg Educ 69:335–339. doi: 10.1016/j.jsurg.2011.10.014
91. Suzuki T, Egi H, Hattori M, Tokunaga M, Sawada H, Ohdan H (2015) An evaluation of the endoscopic surgical skills assessment using a video analysis software program. Surg Endosc 29:1804–1808. doi: 10.1007/s00464-014-3863-5
92. Islam G, Kahol K, Li B, Smith M, Patel VL (2016) Affordable, web-based surgical skill training and evaluation tool. J Biomed Inform 59:102–114. doi: 10.1016/j.jbi.2015.11.002
93. Bettadapura V, Schindler G, Ploetz T, Essa I (2013) Augmenting bag-of-words: data-driven discovery of temporal and structural information for activity recognition. In: IEEE conference on computer vision and pattern recognition. IEEE, pp 2619–2626
94. Sharma Y, Bettadapura V, Ploetz T, Hammerla N, Mellor S, McNaney R, Olivier P, Deshmukh S, Mccaskie A, Essa I (2014) Video based assessment of OSATS using sequential motion textures. In: Proceedings of M2CAI, 2014
95. Zia A, Sharma Y, Bettadapura V, Sarin EL, Clements MA, Essa I (2015) Automated assessment of surgical skills using frequency analysis. In: Navab N, Hornegger J, Wells WM, Frangi AF (eds) Lecture notes in computer science (MICCAI 2015). Springer, Cham, pp 430–438
96. Zia A, Sharma Y, Bettadapura V, Sarin EL, Ploetz T, Clements MA, Essa I (2016) Automated video-based assessment of surgical skills for training and evaluation in medical schools. Int J Comput Assist Radiol Surg 11:1623–1636. doi: 10.1007/s11548-016-1468-2
97. Zhu J, Luo J, Soh JM, Khalifa YM (2015) A computer vision-based approach to grade simulated cataract surgeries. Mach Vis Appl 26:115–125. doi: 10.1007/s00138-014-0646-x
98. Kononowicz AA, Wiśniowski Z (2008) MPEG-7 as a metadata standard for indexing of surgery videos in medical e-learning. Lect Notes Comput Sci 5103:188–197
99. Guggenberger M, Lux M, Riegler M, Halvorsen P (2014) Event understanding in endoscopic surgery videos. In: 1st ACM international workshop on human centered event understanding from multimedia, Orlando, pp 17–22
100. Xhura D (2014) Learning recognition of semantically relevant video segments from endoscopy videos contributed and edited in a private social network categories and subject descriptors. In: ACM international workshop on multimedia, Orlando, pp 663–666
101. Lalys F, Jannin P (2014) Surgical process modelling: a review. Int J Comput Assist Radiol Surg 9:495–511. doi: 10.1007/s11548-013-0940-5