Event Understanding in Endoscopic Surgery Videos

Mario Guggenberger
mg@itec.aau.at

Mathias Lux
mlux@itec.aau.at

Institute of Information Technology
Alpen-Adria-Universität Klagenfurt
9020 Klagenfurt am Wörthersee, Austria

Michael Riegler
michael@simula.no

Pål Halvorsen
paalh@simula.no

Media Performance Group
Simula Research Laboratory AS
Norway

ABSTRACT

Event detection and understanding is an important area in computer science and especially multimedia. The term event is very broad, and we want to propose a novel event based view on endoscopic surgeries. Thus, with the novel view on surgery in this paper, we want to provide a better understanding and possible way of segmentation of the whole event *surgery* but also the included sub-events. To achieve this sophisticated goal, we present an annotation tool in combination with a thinking aloud test with an experienced surgeon.

Categories and Subject Descriptors

H.5 [**Information Interfaces and Presentation (e.g., HCI**]: [Miscellaneous]

Keywords

annotations; events; video; endoscopy; event understanding

1. INTRODUCTION

Understanding events can leverage the development of automatic algorithms for learning, detection, or classification to a high degree. When hearing the word *event*, people usually think of high-level events like concerts and parties, but even a surgery procedure on the heart can be seen as an event. This obviously leads to the conclusion that events are hidden everywhere. In this paper, we take an event-based look at endoscopic surgeries, or more specifically, the annotation of videos of laparoscopic surgeries. Our findings should, however, be applicable to different types of endoscopic interventions.

Endoscopic surgeries can be seen as a special type of human centred event since they involve the participation of multiple people. They are very complex, and a lot of expertise is necessary to annotate recordings of the operations, which is the reason why surgeons usually do it themselves. The rationale to annotate events varies, but it is usually for documentation or training purposes. The problem is that the currently provided tools are very often too complex for surgeons and make it hard to capture the important information in a fast way, or there are no tools provided at all. Since surgeons are often under a lot of time pressure, they want to annotate their surgeries in an intuitive way and as fast as possible. Therefore, it is important to assess their requirements as soon as possible and include their expertise in the design phase to decide which functions and results are important and would support them later in production use. Fancy features, like instrument detection, do not help much if they do not provide useful information to the surgeons and doctors that will finally work with the results.

To tackle these problems, we designed an annotation tool that is able to support the doctors to annotate surgeries naturally. We then performed a thinking aloud test with a

Figure 1: Experimental setup of the thinking aloud test.

HuEvent'14, November 7, 2014, Orlando, FL, USA.

http://dx.doi.org/10.1145/2660505.2660509.

world renowned surgeon working in the field of laparoscopy to obtain first-hand information about what such an annotation tool needs to provide in order to be suitable for and usable by surgeons and lead to a better understanding. The thinking aloud test is a proven method to test user interaction with a system. It requires a special setup including the recording of the interaction with the program and the reaction of the person itself. Figure 1 shows this experimental setup. One camera was used to record the doctors reactions (small image in the upper right corner) and one to record his interaction with the annotation tool (main image). Our work shows that different kinds of events in surgeries ask for different kinds of annotations, distinguished by their level of detail. We also present an event-based segmentation model of endoscopic surgeries, based on the analysis of our expert's information. We believe that our work will help researchers to collaborate and get information from surgeries more efficiently. The main contributions of our work therefore are:

- Providing a general event-based model that is valid for different types of endoscopic surgeries.
- Providing detailed information about which functions an annotation tool for surgeries should include.
- Presenting an advanced prototype of the annotation tool.
- Providing a better understanding of endoscopic surgeries with the help of an expert.

Note at this point that the annotation and understanding of the surgeries is just a first step. The collected information will be used for machine learning and computer vision techniques, and in the best case lead to automatic detection or classification of events or sub-events.

In the reminder of the paper we will at first give an overview and discuss related work in the area of events and endoscopic surgeries. Then we present the methodology split into an annotation tool and a thinking aloud test. In the evaluation section we present the analysis and findings of the experiment in a conceptual and technical point of view. Finally we draw a conclusion and discuss about ongoing and possible future work based on our insights of this paper.

2. RELATED WORK

Philosophy defines an event as a special incident over a specific time span involving one or more objects and happening at a specific place, which can be described by an encasing term - the name of the event. For example, there are general events like endoscopic surgeries, birthdays or funerals, and sports events like football, basketball or soccer games. An event can also consist of many sub-events. For instance, a soccer game has goals and fouls as sub-events and an endoscopic surgery consists of sub-events like injections or cuts. Moreover, an image may depict an event, but it is usually just a snapshot and therefore only covers one time instant of the event's time span [8].

In the context of event processing, detection is an important field of research. It is widely employed in computer vision and classification, because to classify an event, it needs to be detected before. Reuter et al. [13] are using the concept of events to classify multimedia streams automatically into corresponding events. To achieve this, they use a two-step approach. At first, they retrieve event candidates, and secondly, they use machine learning to assign new event candidates to existing ones or to new ones. Petkos et al. [12] try to tackle social event detection by presenting an algorithm that uses multimodal clustering and multimodal fusion to combine different features that can be helpful for event detection in a clever way. A similar approach is presented in [15] by Zeppelzauer et al. They use an unsupervised clustering method to cluster events based on time, user and geo-location information. All these approaches are well performing state-of-the-art methods, and they show that event-based segmentation or classification are promising directions for multimedia content exploring.

Another important direction in the research field of events is event synchronization. Event synchronization combines data of different sources to form an overall picture of a specific event, which usually includes pictures and videos. This can help to get a better understanding of events. Actual work form this area is presented in [4] and [7]. In the first paper the authors try to analyze the content of the images in different photo collections to synchronize them into homogeneous events. The second paper describes an approach to synchronize streams of photos based on to which events they are belong. This is done by a scalable message-passing based optimization framework. Additionally, there are initiatives like the MediaEval Benchmark[1] with tasks like social event detection and multi-user event media synchronization, which shows that the consideration of events is a promising and interesting field of research.

To the best of our knowledge, there is no work that approaches endoscopic videos as a flow of events like we propose in this paper. In [11], Münzer et al. take a low-level bottom-up approach by detecting three classes of irrelevant segments in endoscopic videos (dark, out-of-patient, blurry). Transitions between those classes and the *relevant* class can be seen as low-level sub-events, e.g., the start of the actual surgery when the first out-of-patient segment transitions to an in-patient segment. In this work, we facilitate the detection of high-level sub-events in surgeries through annotations by the actual surgeon. As stated before, regarding a surgery as an event with hierarchical sub-events can help to understand the surgery better, and it can provide an easy view on a complicated topic that is understandable by both surgeons and computer scientists. We anticipate that this increased understanding can lead to the development of better annotation tools, which in turn can provide better information for computer vision, machine learning and classification approaches.

Overall, there have not been many attempts to use multimedia content like images or video for a better understanding of events in the medical sector. Battles et al. [1] presented an event reporting system for blood transfusions. Their system was designed to detect, select, describe, classify, compute, interpret and locally evaluate the event of blood transfusion. It was shown that such a system can improve the health-care results positively, but a good system strongly needs input from both end-users and external experts. The reason is that doctors often have their own techniques of handling multimedia material in their hospitals, which is hardly ever the most sophisticated or effective way. Due to a lack of knowledge in computer science, they often underestimate or do not know the capabilities of

[1] `http://www.multimediaeval.org/mediaeval2014/`

techniques like machine learning and computer vision. An interdisciplinary working expert with extensive knowledge in the area is therefore desired as a source of information, ideas and ready-to-use techniques.

Since endoscopy covers a large number of different surgeries like gastroscopy, mediastinoscopy, rhinoscopy, colonoscopy, laparoscopy, and arthroscopy, it is still a rather unexplored field in multimedia. The most explored sub-type are colonoscopic surgeries. The latest state of the art for colonoscopy images and videos is 3D reconstruction of the colon like discussed in [5, 6]. Current related work regarding videos of endoscopic surgeries in general can be found in [2] where the authors automatically segment a surgery into distinct phases. Furthermore, Münzer et al. [10] are concerned with the detection of the typical circle that is framing the view of an endoscope. We expand this knowledge with a description of the general procedure of an endoscopic surgery event and its encompassing sub-events.

3. METHODOLOGY

The methodology that we were employing was a thinking aloud test setup as described in [3]. It consisted of two stages, stage one being a hands-on experience by a surgeon who used our annotation tool as he would wish to use it in his daily work. Stage two was an open interview reflecting his experience with the tool, and an interview following a prepared exit questionnaire where we asked specific questions that came up during the creation of the concept and the implementation of the tool.

We did this test with only one expert because it is a highly specialized domain where experts are scarce resources and hospital doctors in general have very limited time. Our expert from a regional hospital is a lead technology user in this area who has been recording, documenting, and even live-broadcasting his surgeries overseas since many years. Due to storage demands, he does not always record the full coverage of a surgery (the 1080p format with constant bit rate as recorded by the employed equipment has a high storage demand), but sometimes limits the recordings to cover only the most important phases of a surgery. To mark important moments, he additionally saves single frames as pictures. Both pictures and videos are not only used for the hospital's internal documentation, but also to explain surgeries to patients, to present and discuss interesting cases with colleagues and at conferences, and to school student trainees. Currently, he does not have any means to annotate his recordings and to store them. Persisted annotations would not only save him time to repeat an explanation, but make it more tangible to presentees, enable iterative improvements of annotations, and automatically build up a library of annotated videos. Such a library will then help computer scientists to analyze and classify events, detect similar events in unannotated videos, segment videos any hopefully even semantically synchronize them.

3.1 Annotation Tool

The annotation tool as seen in Figure 2 is an improved version of the annotation tool described in [9]. The most important requirement of the tool was to make its usage as simple as possible, and at the same time, extract as much information as possible from the video. The tool is a tablet computer with a video player that loads a recording of a surgery, and offers four main functions: (i) drawing visual annotations by hand or by using a touchscreen pen, (ii) recording spoken audio notes from the microphone, (iii) setting bookmarks for easier navigation in the video, and (iv) providing a video timeline that visualizes existing annotations and also enables navigation inside the video.

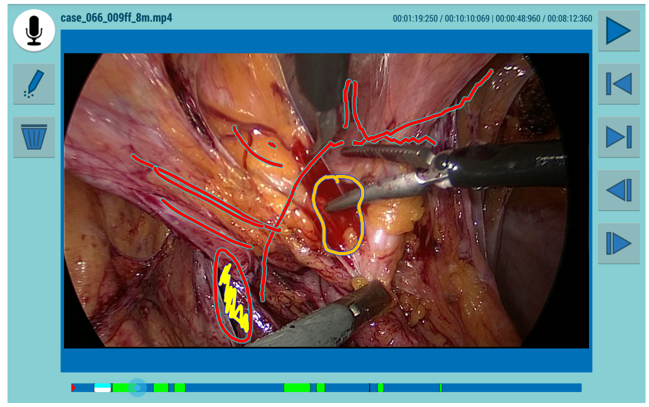

Figure 2: The interface of the annotation tool including annotations of the doctor. The buttons on the right side are for controlling the video, like seeking forward or backward, or switching between existing annotations. The buttons on the left side are used to make annotations or delete them. By pressing and holding the annotation button (pencil), the doctor can start the annotation. While the button is pressed, the annotator can draw annotations in the video and the voice is recorded. The annotation can be stopped by releasing the button. This can be done while the video is playing (moving annotations) and while the video is paused (still annotations). The timeline at the bottom can be used for seeking and also shows already existing annotations and their duration.

Hand drawn and spoken annotations are supported for video playback (moving annotations) and for still frames when the video is paused (still annotations). Annotations on still frames extend the total runtime of the video when played back because the annotations are animated in the same way as the surgeon did draw them and the spoken annotation is replayed correspondingly. Technically, both hand drawn and spoken annotations are recorded in parallel, i.e., it is not possible to explicitly select one or the other, but the annotator can still freely decide his means to annotate. For each stroke of the hand drawn annotations, the color can be selected from a predefined color palette. Bookmarks can be set by shaking the tablet quickly to mark important frames in the video. When playing back the video with its annotations, the hand drawn strokes are animated like they were drawn, and the audio is overlaid. We also have an HTML5 player capable of playing back the annotated video on the web, which has been extended for a different use case [14].

3.2 Hands-On

For the hands-on experience, we handed a Nexus 7 tablet running our annotation tool to the surgeon. We asked him to use it and annotate the videos as he envisions, and to speak out his thoughts while doing this. Since we had access to the recordings of his surgeries, we randomly selected three videos recorded during the past 12 months, out of a collection of hundreds of recordings. Two of them are clips of 8 and 12 minutes runtime and the third is a full surgery recording with a runtime of about 2 hours. We are sure that the thinking aloud protocol, where the user assesses the tool while using it, is much more effective than a usual expert interview because it exposed problems that nobody of us had thought of yet, and it enabled a qualitative investigation of the tool and its annotations. The session was recorded with two video cameras, one over his shoulder capturing his mechanical interaction with the tool, the other from the front, capturing the whole scene including his face and voice.

3.3 Interview

The interview was also recorded by the same cameras and started with a discussion of the hands-on experience. It gave us a lot of insight on the expert's expectations of such a tool and the chance to discuss possible solutions to arisen problems. We then concluded it with a prepared exit questionnaire, where we tried to assess his satisfaction with the tool, possible usability features, use-cases, video processing methods, navigation patterns, and its market or *everyday use* potential.

4. EVALUATION

The video recordings of our thinking aloud session have been investigated by a group of three people to learn the most of the session. While an evaluation with a single expert cannot be fully valid for all people in this medical domain, it was still very productive and definitely showed us a precise direction we need to take. We define a model that is a valuable base for a study of larger scale. We divide our insights into conceptual findings that apply to the whole area of surgical event annotation, and technical findings that apply to the implementation of our tool, but which can still be valuable to developers of similar tools.

4.1 Conceptual Findings

A totally unexpected, but perhaps the most interesting insight that we got from our evaluation is the idea of a general event model of endoscopic surgeries, where the granularity of the events is directly connected to the type of annotation. We observed that our test candidate followed a pattern on all videos, where he always annotated the same kind of event with the same kind of annotation. The model is shown in Figure 3. A surgery can be split into different hierarchical sub-events. The first two sub-events can help to segment an operation in-patient and out-of-patient. If the camera is outside the patient, the segment is not interesting and does not carry any medical information. The surgeon did never perform an annotation when the camera was outside the patient. Therefore, it makes sense to segment a video based on that first.

When the camera is inside the body of the patient, there are three possible sub-events. A surgery is usually started with an overview of the concerning area, to document the actual status of the body and objects of interest, followed by the actual surgery. It is concluded by another overview after the surgery is finished, which leads to a documented before-and-after comparison.

During the surgery, we can first distinguish between general actions. These are moving around, which leads to blurry

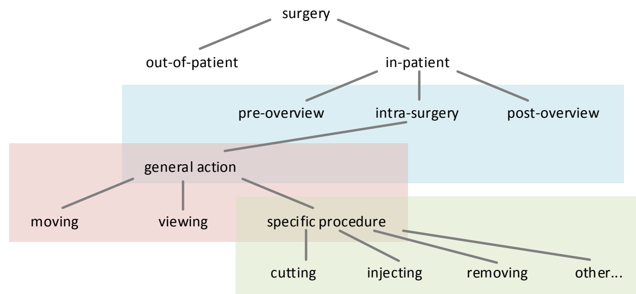

Figure 3: Model of the endoscopic surgery event. The higher the event in the hierarchy, the lower is the level of detail and granularity required for the annotations. The blue box marks low detail events (interval-marking annotations), the red box marks medium detail events (speech annotations to running video), and the green box marks high detail events (speech and hand drawn annotations to still frames).

and useless material, and viewing around. Moving around is for example when the doctor moves the camera in the body or through the colon to a specific area. Viewing around is when the surgeon looks around in a specific part of the patient like in the stomach or in an area of the colon where for example a disease is located. For these two events, the surgeon used only the annotation by speech. It could make sense to differentiate between these two types, in order that moving-around or looking-around parts in the videos can for example automatically be skipped, stored in a lower resolution or bit rate, or replayed with higher speed. The highest-granularity event that we identified during a surgery is a specific medical procedure. This can be an action like cutting a cyst, injecting liquid, removing a polyp, but also detecting exceptionally normal or abnormal looking organs (e.g., a liver without any disease and one with cancer). When one of these events occurred, the surgeon always paused the video and described it with both hand-drawn annotations and speech.

These findings clearly show that a division of the surgery into sub-events makes sense. For hierarchically high events, annotations do not have to be very detailed, but even imprecise descriptions can provide important information for classification or segmentation algorithms that have the potential of helping surgeons. In contrast, specific sub-events like cutting and injecting require very detailed annotations. This provides three very important advantages, namely the surgeon's usage for teaching purposes, the potential of automatically generated summaries, and the researchers' usage for training specific algorithms like cancer detection or instrument detection.

Our tested tool offered just one kind of annotation which could be universally applied in every situation: recorded speech combined with hand drawings, which are both recorded but optional, that ultimately covers a segment of the video. Since it seemed cognitively demanding of our test candidate to decide which annotation to take, we propose to shift to an event-based paradigm offering three kinds of annotations that cover three levels of detail: (i) marking intervals as the lowest detail level, (ii) recording speech to the running video as medium detail, and (iii) recording speech with hand-drawn schemes on still frames for the highest detail level.

As seen in [11], low-level annotations can already be generated automatically, and we suggest it as a preprocessing step, of which the resulting intervals should automatically be integrated as annotations into the tool's timeline. Additionally, it is planned that interesting events will already be marked during the surgery, which makes them much easier to be found and annotated afterwards. Also interesting is the fact that the surgeon is not interested in classical image-processing methods like measuring the sharpness of the picture, the intensity of movement, dominant colors or surgical instrument detection. The only feature he would be interested in is the out-of-patient detection. He explained that recording this kind of surgeries works analogue to a movie script: the best and most interesting shots are usually deliberately orchestrated, meaning a stable camera, good lighting and no instruments blocking sight.

Finally we want to point out that, for the doctor, annotations on a playing video were useless in the current version (i.e., the surgeon always paused the video before starting a new annotation). They would be a very important feature if the annotation system would support the surgeon during the annotation process with object tracking that automatically repositions the drawings according to the camera movements. Otherwise he would have to do a nearly frame-by-frame-wise correction of the annotations, which would be too tedious.

4.2 Technical Findings

On the technical side, a big issue was the tablet size of 7 inches, which is great for portability reasons but a trade-off between user interface widgets and the video, i.e., the annotation drawing area size. Our tester indicated that a tablet with at least a 10 inch screen would be preferable. An additional improvement can be achieved by elaborating a usability concept where the video is drawn over the entire screen and control elements overlaid when needed, like it is typical for video players in full screen mode.

The next major issue is the drawing of annotations by fingers. The first thing our test candidate asked for when beginning the test was a touchscreen pen. This might be a question of individual taste, but this candidate definitely had problems drawing with his fingers. On the small tablet, a pen has the advantage that it does not occlude as much of the screen as fingers do. We discussed the usage of proprietary technology like the Samsung S-Pen[2], which would enable advanced drawing techniques like thickness adjustments of drawn lines by pressure. However, at the same time, it takes away the tablet screen's multi-touch ability that our tool is currently designed for.

The biggest software issue we encountered during our test was the video player provided by the Android 4.3 API, which does not support seeking to exact frames, but rather to the nearest sync frame. During our internal tests, we did not notice this issue as our video's group of pictures size was small, and errors of a few frames did not stand out. The surgeon, however, noticed even misplacements of single frames which, according to him, were a great distraction making his precise annotations worthless.

Regarding general usability, care must be taken that surgeons are usually not that computer savvy and do not have a lot of experience with different apps and platforms, which means that they are not used to common user interface paradigms. As an example, indicators of the recording status in the Android action bar, or Android toast messages as action feedback, often went unnoticed by our test candidate. This needs to be addressed with new concepts of greater visual and maybe even tactile impact. The video timeline as shown before is also not intuitively useable with long running videos, as annotation markers tend to shrink too small and get packed together. The solution might be a separate zoomed section of the timeline around the current position, to preserve a good level of overview detail.

To draw annotations, our tool offered different colors for one single type of stroke. It turned out that our tester used the colors very sparingly and only changed them randomly for no conscious reason, as he told us. We observed though, that the types of usage of the stroke can be classified into three different actions: (1) marking borders with a solid stroke, (2) marking areas with dots or hatched lines, optionally surrounded by a solid stroke, and (3) indicating directions of actions by drawing arrows from solid strokes. These actions always directly relate to the spoken annotation. We figured that it would be more helpful to provide different drawing tools with a single color each, instead of one tool with multiple colors. These could be a thin pen, a thicker felt pen, and a very thick semitransparent marker to highlight areas.

We also discussed the possibility to completely separate voice from drawn annotations, since they are now intertwined and, e.g., deleting an annotation deletes both the audio and the strokes, which is not always desired. This would however lead to two different annotation timelines that both want to be mapped to the video timeline and bring up many open questions. What if the annotator records his voice to the running video, then in parallel pauses the video for a drawn still-frame annotation, and later deletes this drawn annotation? Should the tool display a paused frame for no obvious reason, should this interval be cut out from the voice track to retain synchrony with the remainder of the voice-annotated video, or can we afford to lose synchronization between the voice annotation and the running video when purging the paused interval? There are several other cases leading to such situations.

There are also several potential additional features that our candidate indicated as helpful. Zooming into the video is anticipated since the interesting action often happens in a limited area in the video, and the assistant filming the surgery does not always correctly zoom in. This feature would go without a lot of image quality loss on the relatively small tablet screen as the videos are usually recorded in 1080p format. It would help focusing on the important area, drawing annotations more precise and also generate interesting metadata for analysis. He also mentioned the possibility to export and share annotated still frames from the videos, to fast-forward unimportant segments, and play back important ones in slow motion. He also wished for a function to render the annotations into a standalone video file. Lastly, he envisioned a multimedia integration of external image material from x-ray, ultrasonic and magnetic resonance therapy, to use the tool for multimedia presentations. He also thinks about usages for measurements in standardized recording settings, e.g., measuring size and area or analyzing structure and color of liquids and tissue.

[2] `http://developer.samsung.com/s-pen-sdk`

5. CONCLUSION

We presented an evaluation of a tool to visually and vocally annotate videos from endoscopic surgeries, evaluated with the help of a high-class surgeon in the field and on his own recordings. We deduced a general event model of such a surgery and identified a direct relationship between the granularity of an event and the type of its annotation. We also provided many insights, ideas, and a better understanding of endoscopic surgeries. They can help to develop appropriate annotation tools, which then in turn yield several interesting data and metadata for the analysis, classification, and post-production of endoscopic videos.

Regarding future work, we want to iteratively develop our tool to a state where it will be productively used by at least our collaborating surgeon but hopefully by his colleagues as well. Through this, we hope to collect a huge pool of surgical event annotations that we want to analyze and hope to use for the automatic detection of similar events, and ultimately for the retrieval and the semantic synchronization of similar video recordings.

Furthermore, we want to test more sophisticated approaches in combination with the annotation tool. For example, we are testing annotations supported by object tracking in real time and a frame-by-frame based annotation where the video can be slowed down. We develop these two approaches based on web technologies. They are currently in an experimental stage, and we are discussing the applicability with several surgeons at a large hospital. Once they are ready for testing we also would like to perform a parallel thinking aloud test with a larger number of doctors.

Our findings will help to improve both the quality of our annotation tool and the data generated from it, and we are sure they will help other researchers working in this area.

6. ACKNOWLEDGEMENTS

We would like to thank Prim. Univ.-Prof. Dr. Jörg Keckstein from the Villach Regional Hospital for taking the time to work with us. This work was supported by Lakeside Labs GmbH, Klagenfurt, Austria, and funding from the European Regional Development Fund (ERDF) and the Carinthian Economic Promotion Fund (KWF) under grant 20214/22573/33955 and by the iAD center for Research-based Innovation (project number 174867) funded by the Norwegian Research Council.

7. REFERENCES

[1] J. B. Battles, H. Kaplan, T. Van der Schaaf, and C. Shea. The attributes of medical event-reporting systems. *Arch Pathol Lab Med*, 122(3):132–8, 1998.

[2] T. Blum, H. Feußner, and N. Navab. Modeling and segmentation of surgical workflow from laparoscopic video. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010*, pages 400–407. Springer, 2010.

[3] T. Boren and J. Ramey. Thinking aloud: reconciling theory and practice. *Professional Communication, IEEE Transactions on*, 43(3):261–278, Sep 2000.

[4] M. Broilo, G. Boato, and F. G. De Natale. Content-based synchronization for multiple photos galleries. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 1945–1948. IEEE, 2012.

[5] N. J. Durr, G. González, and V. Parot. 3d imaging techniques for improved colonoscopy. *Expert review of medical devices*, 11(2):105–107, 2014.

[6] D. Hong, W. Tavanapong, J. Wong, J. Oh, and P. C. de Groen. 3d reconstruction of virtual colon structures from colonoscopy images. *Computerized Medical Imaging and Graphics*, 38(1):22–33, 2014.

[7] G. Kim and E. P. Xing. Jointly aligning and segmenting multiple web photo streams for the inference of collective photo storylines. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 620–627. IEEE, 2013.

[8] J. Kim. Events as property exemplifications. In *Action theory*, pages 159–177. Springer, 1976.

[9] M. Lux and M. Riegler. Annotation of endoscopic videos on mobile devices: a bottom-up approach. In *Proceedings of the 4th ACM Multimedia Systems Conference*, pages 141–145. ACM, 2013.

[10] B. Munzer, K. Schoeffmann, and L. Boszormenyi. Detection of circular content area in endoscopic videos. In *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on*, pages 534–536. IEEE, 2013.

[11] B. Munzer, K. Schoeffmann, and L. Boszormenyi. Relevance segmentation of laparoscopic videos. In *Multimedia (ISM), 2013 IEEE International Symposium on*, pages 84–91, Dec 2013.

[12] G. Petkos, S. Papadopoulos, and Y. Kompatsiaris. Social event detection using multimodal clustering and integrating supervisory signals. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, page 23. ACM, 2012.

[13] T. Reuter and P. Cimiano. Event-based classification of social media streams. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, page 22. ACM, 2012.

[14] M. Riegler, M. Lux, V. Charvillat, A. Carlier, R. Vliegendhart, and M. Larson. Videojot: A multifunctional video annotation tool. In *Proceedings of International Conference on Multimedia Retrieval*, page 534. ACM, 2014.

[15] M. Zeppelzauer, M. Zaharieva, and M. Del Fabro. Unsupervised clustering of social events. In