A SEMI-AUTOMATIC EDITING METHOD FOR SURGERY VIDEOS

Zhiwei Fan, Congliang Chen, Tingting Jiang

National Engineering Laboratory for Video Technology, Cooperative Medianet Innovation Center,
School of Electronics Engineering and Computer Science, Peking University
Email:{fanzw,chcoliang,ttjiang}@pku.edu.cn

ABSTRACT

The editing of a raw surgery video is expensive and time-consuming, for it can take an editor with professional medical knowledge hours. We investigate the possibility of reducing the editing cost and propose a feasible semi-automatic editing method for surgery videos. With our method, the editor just needs to annotate a very small part of the video segments in the raw video. And then a model is trained with the partially labeled segments, which can be used to generate an edited version of the whole video according to the editor's criterion. An active learning strategy is adopted here to reduce the number of video segments that need to be annotated. To verify the function of our method, we build a dataset of two raw surgery videos with their edited versions. It shows that two edited versions of the same raw video can be very different because of different editing criteria. And simulation experiments show that our method is able to generate an edited video meeting the expected editing criteria with limited human annotations.

Index Terms— video editing, video summarization, active learning, surgery video

1. INTRODUCTION

With the technology of multimedia widely used in the medical fields, large amounts of videos are captured during surgeries nowadays. For example, in a laparoscopic surgery (a form of minimally invasive surgery), a camera is placed into the abdomen together with the surgical instruments, and records the whole process of the surgery. These raw videos are first-hand data of surgeries, and contain lots of information. However, as the video of a surgery can be hours long, and include many redundancies, it needs to be edited for real applications. It means that highlights of the raw surgery video should be selected to make up an edited version. Such work of editing is time-consuming and requires the editor's professional medical knowledge. And potential editors such as doctors are expensive human resources. It is impractical or wasteful to have a doctor spend hours to skim over the raw video and provide an edited version. In that case, some method to lighten the workload of manual editing is in need.

In surgery video editing, the editing criterion is different when the edited video is prepared for different purposes or is edited by different editors. If the edited video is used to record the whole process of the surgery, only the video segments captured when the camera is shaking and not focused on the surgery operation as shown in Fig. 2(a) should be removed. In a compact version of surgery video for teaching, only the significant steps of the surgery should be selected, while the common surgical operation actions like hemostasis (stopping bleeding) as shown in Fig. 1(a) may not be included. Besides, different editors have different opinions on editing criterion. For instance, some editors prefer to keep the video segments of searching for abnormal tissues as shown in Fig. 1(b), while others prefer to remove them. Thus, two edited versions of the same raw video can be very different. In that case, to more flexibly and accurately edit the surgery video, the editing criterion should be learnt. It is very difficult to learn all the editing criteria, because of the variety of the surgery videos and editing criteria. Therefore, it is challenging to edit surgery videos with a full-automatic method.

Some previous work on video summarization [1, 2, 3, 4, 5, 6, 7] tries to solve similar problems. Such work usual-

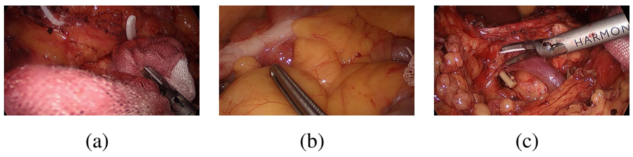


Fig. 1. Screenshots of a colonic surgery. Fig. 1(a) shows the image of hemostasis. Fig. 1(b) shows the image of the surgeon searching for the abnormal tissues. Fig. 1(c) shows the image of removing tissues.

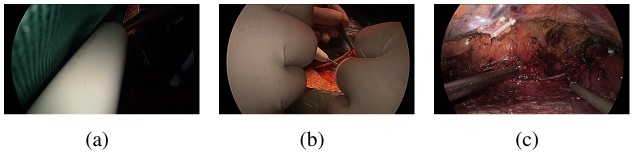


Fig. 2. Screenshots of a gastrectomy surgery. Fig. 2(a) shows the image captured when the camera is shaking and not focused on the suregery. Fig. 2(b) shows the image of opening a incision. Fig. 2(c) shows the image of removing tissues.

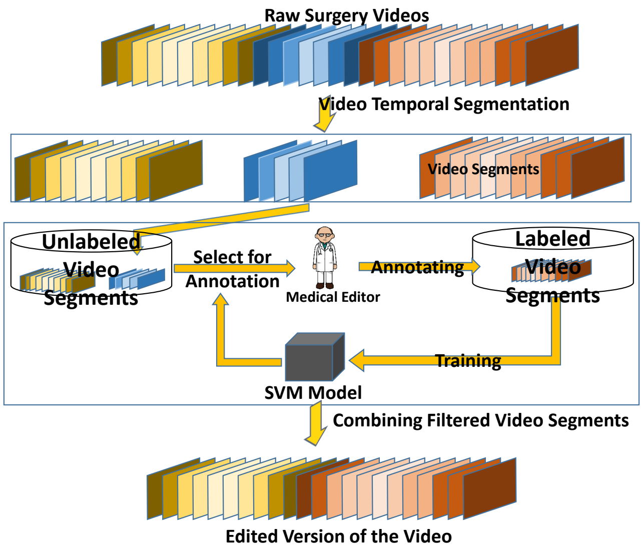


Fig. 3. Overview of the semi-automatic video editing method.

ly aims to provide a short summary of a video by selecting both interesting and representative shots or key frames in the video. The technology of video summarization has been applied to various kinds of videos, including casually captured user videos [2], edited videos such as TV news [7], *etc*, but it has never been used for medical or surgery videos. As the application of video summarization is quite limited to entertainment or daily life, the previous work usually focuses on interestingness of the video or recording representative activities. This is not appropriate for the editing of surgery videos, for all the details of a surgery can be very important depending on the purposes. Thus, the technology of video summarization can not be directly applied to surgery video editing.

Although it is difficult to learn all editing criteria together simultaneously, we can easily learn a specific editing criterion from a specific editor for a specific purpose. We implement a semi-automatic method as shown in Fig. 3 to adapt to different videos and different criteria. First, we cut a raw video into segments, and randomly choose a few segments for a medical editor to annotate. Then according to the annotation, we train an SVM classifier. With the SVM we select several other segments based on an active learning strategy. This process is repeated until the limit of annotation (e.g. medical editor's time) is reached. The SVM trained in the final turn will predict which segments should be kept. Our goal is to accurately predict each segment's label with as few human annotations as possible.

We make two contributions in this paper as follows.
i) A dataset of surgery videos. It contains the raw videos with their edited versions of 2 surgeries. The resolution of those videos is 1920 × 1080.
ii) A semi-automatic editing method for surgery videos. With only 5% of the video segments annotated, the precision of our results is at least 84% in our experiments.

The remainder of the paper is organized as follows. Sec. 2 discusses the related work. Sec. 3 concretely illustrates the proposed semi-automatic video editing method. Sec. 4 introduces the dataset of the surgery videos. Sec. 5 verifies our proposed method with experiments. And Sec. 6 is a conclusion of the whole paper.

2. RELATED WORK

2.1. Video Summarization

Works on summarization aim to output a shortened version to summarize the initial video. A general idea is to segment the initial video into shots, and then highlights among them are selected to make up the summary. For instance, Sun *et al.* [1] propose to rank the highlights in personal videos by analyzing edited videos. Gygli *et al.* [2] estimate the interestingness of video shots and select a summary from them using a 0/1-knapsack optimization. Later they [3] propose a method to optimize multiple objectives for finding interesting, representative and uniform video shots.

There are also some domain specific methods, which summarize a video of a particular category. Potapov *et al.* [4] produce high quality summaries by dealing with videos of a typical category such as "birthday party". Rather than simply optimizing a summary's interestingness or representativeness, Lu *et al.* [5] utilize the connectivity between the events in the video, and propose a method driven by the important people and objects. Gaze *et al.* [6] summarize ego-centric videos based on gaze tracking information.

Especially, there are some existing methods on endoscopy video summarization. M. Ismail *et al.* [8] partition the video frames into subsets and generate summary by few representative samples. Ahmed Z. Emam *et al.* [9] use different features to evaluate similarity between frames and remove similar frames. Both methods are unsupervised and can't adapt to different editions. Besides, the resolution of those videos they use is about 300×300, which is much lower than what we are going to handle.

2.2. Active Learning

Active learning [10] is widely used to reduce the amount of labels required for a learning based model in various fields, including annotation [11, 12], recognition [13], retrieval [14], *etc*. It achieves this aim by selecting informative samples for labeling.The strategy of selecting the most informative samples in active learning is usually driven by two measures [15, 16, 17], uncertainty measure and information density measure. The uncertainty measure is an "exploitation" strategy [15, 17], leading labelers to annotate the samples near the boundary, which in return helps to refine the boundary. The boundary here means the region in the feature space that the model is most uncertain about, such as the hyperplane in the problem of classification. The information density measure is an "exploration" strategy [15, 17], leading labelers to annotate samples in different regions of the feature space, which avoids the problem that outliers are always selected in a pure "exploitation" strategy [17].

3. PROPOSED METHOD

The proposed semi-automatic method works as shown in Fig. 3. It is made up of three components, video temporal segmentation, training a SVM classifier to predict the label of each video segment, and combining video segments to form an edited version.

Assume V represents a video of a surgery. Then $S_V = \{S_V^1, S_V^2, ..., S_V^N\}$ is a temperal segmentation of video V, where $S_V^i (1 \leq i \leq N)$ is a video segment of video V, and N is the number of video segments in video V. $L(S_V^i) \in \{0, 1\} (1 \leq i \leq N)$ represents the true label of S_V^i with regard to a edited version. $L(S_V^i) = 1$ means that S_V^i is kept in the edited version, while $L(S_V^i) = 0$ means that S_V^i is removed from the edited version. $\hat{L}(S_V^i) \in \{0, 1\} (1 \leq i \leq N)$ denotes the predicted value of $L(S_V^i)$. And $f(S_V^i)$ denotes the feature of segment S_V^i.

3.1. Video Temperal Segmentation

The aim of video segmentation is to segment videos into logical units of videos. An ideal segment of a video should contain one complete action of the main object in the video, and abrupt motion changes should be avoided in a video segment. As the raw surgery video is continuously captured by a single camera, it often contains only one single shot. Thus, the traditional approaches of video segmentation based on shot detection are not appropriate for our problem. Here we adopt the method of subshot segmentation proposed by Gygli *et al.* [2]. The main idea of the method is to cut the video when there is little motion [18].

For a video segment S_V^i, its energy function $E(S_V^i)$ is defined as Eqn. 1, measuring the quality of video segment S_V^i.

$$E(S_V^i) = \frac{1}{1 + \alpha M(S_V^i)} \times P_l(len(S_V^i)), \tag{1}$$

where $M(S_V^i)$ is the sum of the motion magnitude in the first and last frame in S_V^i, and $P_l(\cdot)$ is a length prior of video segments. $len(\cdot)$ denotes the length of a video segment in terms of frames. α is a controlled parameter adjusting the influence between the motion magnitude and the length prior.

With the energy function of a video segment defined, here comes the energy function of the whole video segmentation S_V as

$$E(S_V) = \sum_{i=1}^{N} E(S_V^i). \tag{2}$$

Thus, the optimal video segmentation S_V^* can be calculated as

$$S_V^* = argmax\ E(S_V). \tag{3}$$

Details: The motion magnitude in $M(S_V^i)$ is estimated by KLT [19]. The parameter α is simply set as 1 in all of our later experiments. The length prior $P_l(\cdot)$ can be learnt by fitting a log-normal distribution to a histogram of segment lengths of the human created video segmentation. The optimization of Eqn. 3 is solved by dynamic programming.

3.2. Model Training

3.2.1. Feature Representation

The feature of each video segment is represented as follows. We first extract the state-of-the-art dense trajectory motion features [20] of each video segment. And then the dimension of the dense trajectory motion features is reduced from 426 to 126. A Gaussian mixture model of 200 mixture components is learnt with all the reduced dense trajectory motion features of all the video segments. Thus, a fisher vector with a fixed dimension (50400) can be generated to represent a video segment.

3.2.2. Model Training with active learning

With the raw video segmented, we need to collect the labels of all the segments to form an edited version. A basic idea is to have medical editors annotate each video segment by judging if it should be included in the edited version. But as a surgery can last for hours, it is time-consuming to annotate every video segment. To reduce the amount of human annotation, we actively select some video segments for annotation, and train a SVM classifier with these labeled segments. Then we can predict the labels of segments of the raw video with the trained model.

Here we adopt the active learning strategy of *density-based re-ranking* proposed by Zhu *et al.* [17] in our method. In our problem, the density measure of a video segment, $D(S_V^i)$, is defined as

$$D(S_V^i) = \frac{\sum_{x \in N_K(S_V^i)} \cos(f(S_V^i), f(x))}{K}, \tag{4}$$

$$\cos(f_i, f_j) = \frac{f_i \cdot f_j}{||f_i|| \cdot ||f_j||}, \tag{5}$$

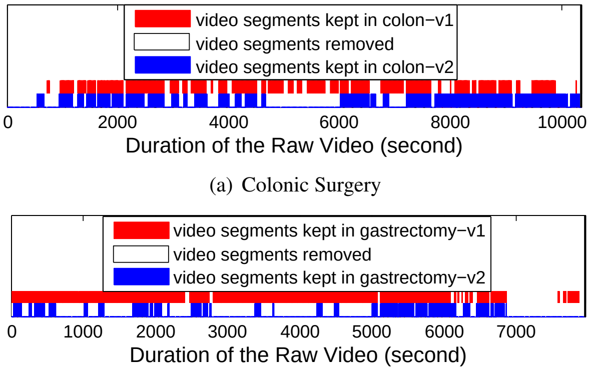
where $N_K(S_V^i)$ is the set of K unlabeled video segments that are most similar to S_V^i. $\cos(f_i, f_j)$ is used to measure the similarity between two samples. The larger its value is, the larger the similarity is. Thus, the larger $D(S_V^i)$ is, the more samples similar to S_V^i are unlabeled.

As our system is based on SVM, there is no probabilistic output. Thus, the uncertainty of a video segment, $U(S_V^i)$, is measured as the the margin between the sample and the classification hyperplane of the current SVM model.

With the uncertainty measure $U(S_V^i)$ and the density measure $D(S_V^i)$ defined, the active learning strategy of *density-based re-ranking* is made up of two steps. First, it selects the top Q samples of the maximum uncertainty. Second, it selects the sample of the maximum density among the selected

Table 1. Information of the Videos of Two Surgeries

Surgery	Length	Resolution	Frame Ratio (fps)	Edited Version			Video Segmentation		
				Version	Length	Difference Ratio	Segment Num (frame)	Mean Len (frame)	Std Len (frame)
Colon	172min 36sec	1080x1920	25	Colon-v1	105min 30sec	30.09%	4746	47.4	11.1
				Colon-v2	101min 42sec				
Gastrectomy	132min 33sec	1080x1920	25	Gastrectomy-v1	107min 54sec	51.78%	3824	47.7	11.5
				Gastrectomy-v2	43min 58sec				



(a) Colonic Surgery

(b) Gastrectomy Surgery

Fig. 4. The difference between two edited versions of two raw videos. Fig. 4(a) and Fig. 4(b) respectively show the difference between the two versions of the colonic surgery and the gastrectomy surgery.

Q samples. Parameter Q is used to balance the influence of uncertainty and density.

4. DATASET

We collect the raw videos of two surgeries. One is a laparoscopic surgery for colon carcinoma, and the other is a surgery of gastrectomy. The colonic carcinoma surgery is a laparoscopic surgery, which means that most of its video is captured inside the abdomen of the patient. The gastrectomy surgery is a open surgery, which means that its video is captured outside the patient's body. And the screenshots of the two videos are respectively shown in Fig. 1 and Fig. 2.

For each of the two raw videos, we invite two doctors to edit it with their own criterion in seconds. Thus, we get two edited versions for each of the raw videos. The left 7 columns of Table. 1 illustrates these videos in detail. And Fig. 4 shows how each edited version is generated from the raw, including which segments are kept and which segments are removed. We can see that for the same raw surgery video, doctors can provide very different edited versions. 30.09% of the raw colon surgery and 51.78% of the raw gastrectomy surgery are edited differently in their two edited version. That is because the doctors' judgements and the applications of the edited versions are different.

The main difference between Colon-v1 and Colon-v2 is the treatment of the video segments of hemostasis as shown in Fig. 1(a) and the video segments of searching for abnormal tissues as shown in Fig. 1(b). Colon-v1's editor keeps segments of hemostasis because they are of high video quality and effective surgery actions. However, Colon-v2's editor regards hemostasis as common surgery steps. He deems that it is not necessary to include these segments into the edited version if the raw video is to be edited as the teaching material of colonic surgery. Colon-v1's editor thinks that searching for abnormal tissues are not effective surgery actions, in which case, he removes the segments of searching for abnormal tissues. As for Colon-v2's editor, he keeps these segments for he thinks that they are instructive.

Gastrectomy-v2 is almostly a compact version of Gastrectomy-v1. Gastrectomy-v1 is used to record the whole process of the surgery, in which case, only video segments captured when the camera is not focused on the surgery operation as Fig. 2(a) are removed. Gastrectomy-v2 is edited to extract the important steps of the surgery as Fig. 2(c), which means only the essence of the raw video is kept, and that common surgery actions such as opening a incision as Fig. 2(b) are not included in Gastrectomy-v2.

5. EXPERIMENT

We process both of the two raw videos introduced in Sec. 4 with our proposed semi-automatic video editing method. We first segment the raw videos into temporal segments. The ground truth of each segment's label is set according to the edited version. And the simulation experiments show that with a small part of the segments labeled, our method is able to accurately predict the labels of other segments.

5.1. Evaluation Measure

The performance of our proposed semi-automatic video editing method is based on the quality of the edited video it produces, which can also be seen as its similarity with the targeted edited video. Here we take the edited versions provided by doctors as the targeted edited videos. Recall that $L(S_V^i) \in \{0, 1\}$ and $\hat{L}(S_V^i) \in \{0, 1\}$ are respectively the true label and the predicted label of S_V^i with regard to the targeted edited version. Thus, the quality of the generated edited video can be measured as the precision P, recall R and F-measure F of the trained model as Eqn. 6, 7, and 8. Larger values of them mean better quality.

$$P = \frac{\sum_{i=1}^{N} L(S_V^i) \times \hat{L}(S_V^i)}{\sum_{i=1}^{N} \hat{L}(S_V^i)} \quad (6)$$

$$R = \frac{\sum_{i=1}^{N} L(S_V^i) \times \hat{L}(S_V^i)}{\sum_{i=1}^{N} L(S_V^i)} \quad (7)$$

$$F = \frac{2P \times R}{P + R} \quad (8)$$

5.2. Video Temporal Segmentation

The results of video segmentation are shown in the right 3 columns of Table. 1. According to our observation, the results are quite logical, and meet our expectations.

Each video segment of a raw video has a label with regard to a particular edited version, which shows if the segment is included in the edited version. This label can not be directly transmitted from the edited version, for doctors edit the raw videos in seconds. In that case, we first obtain each frame's label according to the edited version, and then set a video segment's label as the label of the majority of the frames it contains. Thus, the label of each video segment with regard to a particular edited version is obtained. In fact, the percentages of video segments that contain frames of both labels are respectively 1.33%, 0.99%, 1.05%, 2.20% with regard to Colon-v1, Colon-v2, Gastrectomy-v1, and Gastrectomy-v2. The percentages of these video segments are so small, that whether keep them or remove them in the edited version has little influence on the quality of the edited version.

5.3. Model Training and Results

We simulate the process of model training on the dataset we collect as follows. For a raw video and its targeted edited version, we first randomly select 2% of its segments to initialize the SVM model. And then we iteratively select segments and update the SVM model through the active learning strategy illustrated in Sec. 3.2. To investigate the function of active learning, we also simulate the process of model training the same as the above, except that all the labeled segments are randomly selected.

We repeat both of the two processes of model training for 40 times. The mean values and variances of the generated edited video's quality are shown in Fig. 5 when the labeled segments accumulate. We can see that the values of the quality measures (P, R, F) of Colon-v1, Colon-v2 and Gastrectomy-v1 are at least 0.85 and 0.9 when 5% and 10% video segments are labeled. Some of the values are even quite close to 1. The values of the quality measures of Gastrectomy-v2 are lower, but not very bad. The values of its P, R and F are 0.92, 0.75 and 0.82, when 10% segments are labeled. Thus, our work does provide a feasible method to generate an edited version of a raw surgery video with limited human annotation. And an example of a raw surgery video with its edited version when 10% of its segments are labeled is involved in the supplemental material.

Comparing the quality measures' mean values and variances when active learning strategy used and not used, we can see that the active learning strategy does improve the performance of the leant model and make a contribution to the stability.

The quality of the generated video of Gastrectomy-v2 is markedly worse than that of the other 3 edited versions. This is probably because that Gastrectomy-v2 is a very compact version. Among many similar video segments, Gastrectomy-v2 just selects a small number of them as representatives. In that case, the relationship between video segments should be considered, and the editing cannot be easily treated as a problem of classification.

6. CONCLUSION AND FUTURE WORK

In this work, we investigate the possibility of reducing the high cost of manully editing surgery videos, and propose a feasible semi-automatic video editing method. We first segment the raw video into temporal video segments. Then the problem of video editing is transmitted into a problem of binary classification. With a small part of the video segments annotated whether they should be kept in the edited version, a SVM classifier can be learnt to predict whether the left unlabeled segments should be kept. Thus, by combining all the kept video segments orderly, we obtain an edited version of the video. To evaluate the function of our method, we build a dataset of two surgery videos with their edited versions. The evaluation of our method shows that with the assistance of our method, we can obtain an edited video meeting a particular editing criterion with a small part of the raw video labeled.

As the evaluation is now based on a simulation experiment, we will conduct a real test to further verify our method. Besides, we will improve our method to handle the problem of compact version in the future.

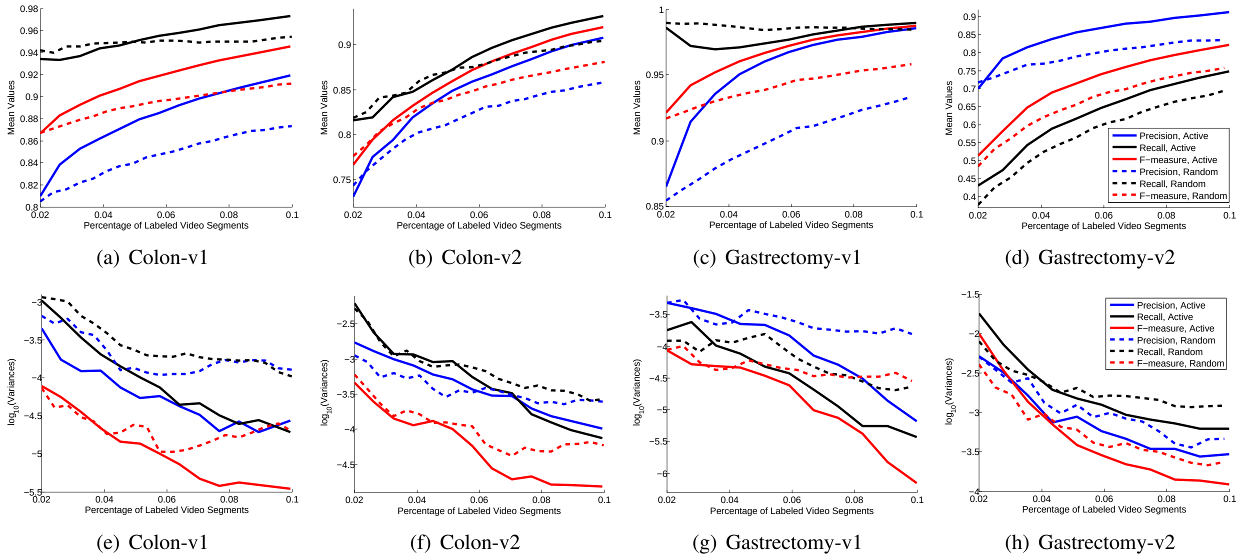
7. ACKNOWLEDGEMENT

This work was partially supported by National Basic Research Program of China (973 Program) under contract 2015CB351803 and the Natural Science Foundation of China under contracts 61572042, 61390514, 61421062, 61210005, 61527084.

8. REFERENCES

[1] M. Sun, A. Farhadi, T. H. Chen, and S. Seitz, "Ranking highlights in personal videos by analyzing edited videos," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5145–5157, 2016.

[2] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *ECCV*, 2014, pp. 505–520.



(a) Colon-v1 (b) Colon-v2 (c) Gastrectomy-v1 (d) Gastrectomy-v2

(e) Colon-v1 (f) Colon-v2 (g) Gastrectomy-v1 (h) Gastrectomy-v2

Fig. 5. The quality measures, including precision (blue), recall (black) and F-measure (red) of the generated edited videos (best view in color). Fig. 5(a), 5(b), 5(c), 5(d) respectively show the mean values (full line) of the quality measures of Colon-v1, Colon-v2, Gastrectomy-v1 and Gastrectomy-v2. Fig. 5(e), 5(f), 5(g), 5(h) respectively show the variances (dotted line) of the quality measures of Colon-v1, Colon-v2, Gastrectomy-v1 and Gastrectomy-v2 across the 40 turns. "Active" means that the labeled segments are actively selected, while "Random" means that the labeled segments are randomly selected.

[3] M. Gygli, H. Grabner, and L. Van Gool, "Video summarization by learning submodular mixtures of objectives," in *CVPR*, 2015, pp. 3090–3098.

[4] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *ECCV*, 2014, pp. 540–555.

[5] Z. Lu and K. Grauman, "Story-driven summarization for egocentric video," in *CVPR*, 2013, pp. 2714–2721.

[6] J. Xu, L. Mukherjee, Y. Li, J. Warner, J. M. Rehg, and V. Singh, "Gaze-enabled egocentric video summarization via constrained submodular maximization," in *CVPR*, 2015, pp. 2235–2244.

[7] M. A. Smith and T. Kanade, "Video skimming and characterization through the combination of image and language understanding techniques," in *CVPR*, 1997, pp. 775–781.

[8] Ouiem Bchir M. Maher Ben Ismail and Ahmed Z. Emam, "Endoscopy video summarization based on unsupervised learning and feature discrimination," 2013.

[9] Mohamed M. Ben Ismail Ahmed Z. Emam, Yasser A. Ali, "Adaptive features extraction for capsule endoscopy (ce) video summarization," 2015.

[10] B. Settles, "Active learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 6, no. 1, pp. 1–114, 2012.

[11] A. Kapoor, Gang Hua, A. Akbarzadeh, and S. Baker, "Which faces to tag: Adding prior constraints into active learning," in *ICCV*, 2009, pp. 1058–1065.

[12] R. Yan, J. Yang, and A. Hauptmann, "Automatically labeling video data using multi-class active learning," in *ICCV*, 2003, pp. 516–523.

[13] P. Jain and A. Kapoor, "Active learning for large multi-class problems," in *CVPR*, 2009, pp. 762–769.

[14] S. Vijayanarasimhan and K. Grauman, "Large-scale live active learning: Training object detectors with crawled data and crowds," in *CVPR*, 2011, pp. 1449–1456.

[15] T. Osugi, Deng Kim, and S. Scott, "Balancing exploration and exploitation: a new algorithm for active machine learning," in *ICDM*, 2005.

[16] S. Ebert, M. Fritz, and B. Schiele, "Ralf: A reinforced active learning formulation for object class recognition," in *CVPR*, 2012, pp. 3626–3633.

[17] J. Zhu, H. Wang, B. K. Tsou, and M. Ma, "Active learning with sampling by uncertainty and density for data annotations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1323–1331, 2010.

[18] J. V. Mascelli, "The five C's of cinematography," *Cine / Grafic Publications*, 1965.

[19] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *IJCAI*, 1981, pp. 674–679.

[20] H. Wang, A. Kläser, C. Schmid, and C. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.