Surgical video retrieval using Deep Neural Networks

Christos Varytimidis[1], Konstantinos Rapantzikos[1], Constantinos Loukas[2], Stefanos Kollias[1]

[1] National Technical University of Athens, Greece
[2] National and Kapodistrian University of Athens, Greece

Abstract. Although the amount of raw surgical videos, namely videos captured during surgical interventions, is growing fast, automatic retrieval and search remains a challenge. This is mainly due to the nature of the content, i.e. visually non-consistent tissue, diversity of internal organs, abrupt viewpoint changes and illumination variation. We propose a framework for retrieving surgical videos and a protocol for evaluating the results. The method is composed of temporal shot segmentation and representation based on deep features, and the protocol introduces novel criteria to the field. The experimental results prove the superiority of the proposed method and highlight the path towards a more effective protocol for evaluating surgical videos.

Keywords: surgical video retrieval, deep features, shot segmentation, shot representation

1 Introduction

Over the last two decades *minimally invasive surgery (MIS)* has gained tremendous popularity for an increasing number of surgical operations. Video acquisition in MIS is straightforward since the surgeon operates with an endoscopic camera inserted into the body. Currently, an increasing amount of surgical multimedia content is uploaded on video-sharing websites and dedicated web-based educational resources, so that it is accessible to educators and trainees. Cognitive training is performed prior to an operation, after retrieval of relevant videos based on a limited number of keywords. In most cases though, the trainee desires to concentrate on specific events rather than skimming the whole video stream. This is usually achieved by manual pre-annotation based on cues of potential interest, which is tedious and time consuming. Content-based video search is not applied to large scale search engines, mainly due to the diversity of the visual content and therefore the lack of a universal way to represent them. To automate and enhance this process, one must develop technologies to effectively index and retrieve surgical videos. Recently, image classification and detection have moved from exploiting local features (e.g. SIFT, SURF) combined with machine learning tools (e.g. SVMs, random forests), to employing deep convolutional neural

networks (CNN) for end-to-end object recognition and detection [15, 27]. The introduction of large scale datasets like Imagenet [5] and COCO [17], along with the advances in GPU accelerated parallel computing, have led to designing and training deep CNNs that are computationally fast and do not overfit, despite the millions of parameters learned. Pixel-based image analysis methods like segmentation have also adopted deep CNNs for providing fast and accurate results [19].

1.1 Related work

Despite the significant progress in the field of content-based image retrieval for medical applications [23], methods on detection and retrieval of surgical events (or shots) from MIS videos are limited. The potential of video-based analysis for surgery segmentation is first addressed in [3], where the goal is to extract the coarse phases of the operation. Lo et al. [18] propose an approach to MIS video event detection based on multiple visual cues. A shot detection technique in MIS videos based on motion vector analysis is described in [25]. Giannarou and Yang propose a novel framework for content-based surgical scene representation by detecting key surgical episodes via probabilistic motion modelling [8]. The use of instrument classification to enable semantic segmentation of laparoscopic videos is proposed in [26]. The importance of event-based annotation in MIS videos is highlighted in [20], where a method for the detection of tissue cauterization events is proposed. Recently, the same group proposed a spatiotemporal tracking technique using a variational Bayesian framework for shot border detection in laparoscopic cholecystectomy (LC) videos [21]. Shot borders are defined when tracking of Gaussian components detected along the video sequence fails, denoting a transition in the surgical scene. In another recent paper, a convolutional neural network (CNN) for joint phase recognition and tool detection in LC videos is proposed [30]. For tool detection, the confidence given by the network is used directly, whereas for phase detection the visual features extracted from the network are used in conjunction with SVMs and a hierarchical HMM.

Frequently, video retrieval methods use a video shot, i.e. the part marked by hard transitions (cuts), fades or dissolves [16], as the way to decompose the visual input into meaningful parts. However, MIS videos cannot be decomposed into shots based on cuts or dissolves, because the camera is continuously focusing on the area-of-interest and is –practically– free to move and rotate. Unavoidably, common shot detection methods lead to over- or under-segmentation. In order to cope with this issue, we propose a novel approach for shot detection that exploits local spatio-temporal changes to compute a global and robust measure of change that provides meaningful temporal segments.

Extracting and grouping descriptors in order to index video shots to support similarity-based matching leads to increased query cost [14]. Douze et al. [6] propose a method for encoding the descriptors extracted from all video frames in a temporal representation in order to efficiently perform both video retrieval and temporal alignment of video shots. Action recognition algorithms also cope with extracting compact representations from videos. Sun et al. [29] create trajectories by matching SIFT local features between two consecutive frames and

combine the SIFT descriptors to create a trajectory descriptor. Using the bag-of-words (BoW) model they efficiently match video segments. Our baseline method and its improvements are related to this work. Recently, image region proposal algorithms have been exploited for action recognition in videos. Jain et al. [10] extend the proposals provided by selective search [31] from images to videos, in order to select candidate cubes in the spatio-temporal space, that contain actions. Gkioxari et al. [9] filter out image region proposals of selective search that are not motion salient, and extract region descriptors with two CNNs, one based on appearance, and one on motion.

Deep CNNs have been exploited as general purpose feature extractors for tasks other than classification [27]. Karpathy et al. [13] extract deep features from small video clips for video classification, while Yue-Hei Ng et al. [33] combine raw pixel intensities with optical flow to extract video descriptors with deep CNNs.

In this paper we propose a novel method for automated video shot detection and retrieval and apply it to laparoscopic videos. Driven by the content, we split videos in shots when the region changes significantly, and extract localized and global temporal descriptors for representation. Our contribution is threefold. First, we exploit object proposals in a global per-frame basis, monitoring changes across frames for shot detection. Second, we investigate using local features and temporal feature trajectories in order to create shot descriptors for performing video retrieval. Finally, we further improve the retrieval performance by extracting frame descriptors using deep CNNs, and grouping them in a single global descriptor per shot. We investigate using different network architectures and layers for descriptor extraction, and evaluate the proposed framework in retrieving *relevant* videos shots in MIS videos. To the best of our knowledge this is the first time that shot detection and retrieval in MIS videos receives such a detailed investigation, combining the application of objectness models (shot detection), and local feature tracking as well as deep CNNs (shot retrieval).

As will be further described below, "relevance" here refers to the agreement in the tool types present in the query shot and the shots retrieved, although alternative criteria may also be explored (e.g. surgical tasks, phases). In this respect, our work is conceptually different to [30, 26] that address tool/phase detection but not shot retrieval, and to the works [21, 25] that address solely shot border detection (neither shot retrieval, nor tool type detection).

2 Surgical video content

For performing surgical video retrieval, we created a new dataset consisting of LC videos. LC is a common operation in abdominal MIS with great educational value; usually it is the first laparoscopic surgery performed by residents. In brief, the operation consists of three major phases that have to be performed sequentially: division of adhesions involving the gallbladder and adjacent organs (gallbladder dissection), division of the cystic artery and duct (clipping and cutting), and separation of the gallbladder from the liver bed followed by extraction (liver bed coagulation).

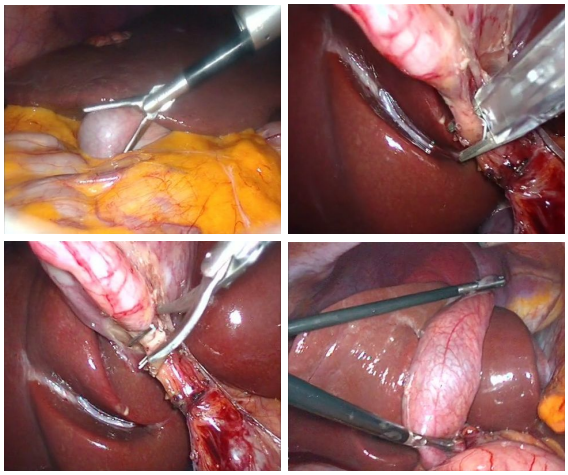

Fig. 1. Example frames from the LC operation showing various laparoscopic tools in interaction with the gallbladder.

During the procedure, six different surgical tools were employed (e.g. scissors, dissector, grasper). To measure the performance of the retrieval algorithm, the detected shots were pre-annotated with regard to the tool(s) used in each shot (see Fig. 1). Pre-annotation also considered three more tags: no-tools, trocars (a tubular component used for tool insertion), and clips (metallic clips applied by the surgeon). A shot may have been annotated with more than one tag, in case it contains multiple tools (e.g. scissors and dissector). From the nine available tags, the minimum and maximum number of tags used in each shot was one and five respectively (median = 2).

Very recently, two new datasets of laparoscopic surgery videos were released for the tool detection and surgical flow challenges of M2CAI[1]. We exploit the tool detection dataset, together with the provided annotation, in order to perform shot retrieval among different videos.

3 Video shot detection

Intuitively, a full MIS video is composed of a single shot, since the captured area, namely the area beneath the liver, does not change. Nevertheless, important changes like the insertion or removal of a surgical tool, manipulation of the gallbladder or the mild change of viewpoint (e.g. from surrounding tissue to the gallbladder) can be captured.

We focus on the appearing/disappearing tools and train an objectness model to highlight their presence. Recently, the concept of objectness has been used in

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

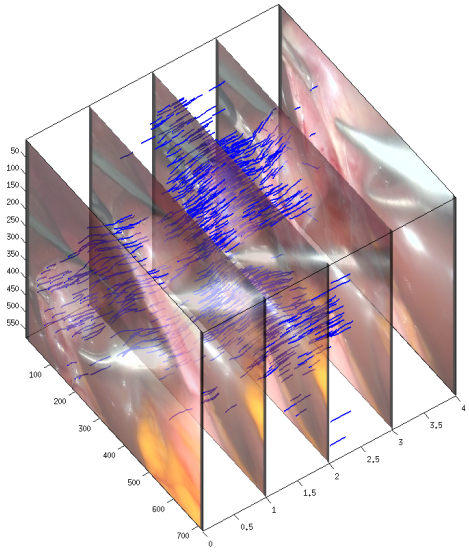

Fig. 2. Local feature trajectories: selection of local features that are matched between at least 5 consecutive video frames.

many object detection algorithms in order to locate generic objects disregarding their identity [1, 31]. Given the objectness of different image subwindows, these algorithms focus on proposing candidate image regions fully enclosing distinct objects. Objectness is related to visual saliency [4], and has also been extended to video sequences [2]. For our method, we adapt the objectness approach of [1]. We run the model on every frame, fuse the results and compute a single-valued measure for every frame. We suggest that changes in the global objectness of video frames correlate well with meaningful shot changes. The single-valued output is produced by averaging the objectness response across the frame. The variance of the measure is monitored across the video, and we mark shots as the temporal periods between two consecutive outliers (sudden jumps or falls).

4 Video shot description

We extract a single high-dimensional descriptor for each shot in order to perform nearest-neighbor based search. The representation is compact, can scale up well and minimizes the query time for fast video retrieval in large datasets. For each shot, we first extract video frames with a constant time interval between frames, set to $200ms$ in all our experiments.

4.1 Local features and descriptors

Our first approach considers extracting video shot descriptors exploiting spatio-temporally matching local features. For each video frame we detect SIFT [22]

and extract the corresponding descriptors. We build a visual vocabulary of $100K$ visual words, by clustering a random subset of the extracted descriptors using approximate k-means [24]. In order to maximize the effectiveness of the global descriptor, we fuse only the matching features between subsequent frames, i.e. the features that –ideally– correspond to the same patch across frames. Starting from the aligned features, we aggregate their labels and create a single histogram of labels for each shot. We use the bag-of-words (BoW) representation, which is a L^2-normallized histogram of the labels. The size of the final shot descriptor is equal to the size of the vocabulary, i.e., $100K$.

Baseline (R): The baseline system aggregates the labels of matching local features between subsequent frames in order to create a single histogram of visual labels. We select matching features by running RANSAC [7] to spatially align adjacent frames, and discard labels of non-matching features.

Iterative Ransac (R_{iter}): In order to remove the single-plane restriction of RANSAC, we employ a simple, yet effective approach. Similar to [32, 12], we match adjacent frames by running RANSAC iteratively. In each iteration, we discard the already matched features, until no good matches are found. This allows us to employ low spatial matching tolerance in each iteration, and yet select features that were missed by the baseline system.

Local Feature Trajectories (LFT): We create local feature trajectories by matching local features using either RANSAC, or iterative RANSAC. We select only the trajectories that are at least 5 frames long and therefore further remove noisy features that are either appearing for a very short time (occlusions), or are mistakenly matched by RANSAC or iterative RANSAC (see Fig. 2).

Table 1. Evaluation results on our MIS video dataset, using local features.

method	precision (%)
SIFT, R	64.6
SIFT, R_{iter}	65.9
SIFT, R, LFT	76.8
SIFT, R_{iter}, LFT	**78.3**

4.2 Deep features

The proposed representation employs deep convolutional features. Initially, convolutional and pooling layers are applied to the frames, corresponding to visual filters and local aggregators, followed by fully connected layers. The latter fuse the responses of the convolutional layers, in order to create distinctive image representations.

The extracted video frames are resized and passed through the convolutional network. We extract the responses (feature maps) of the fully connected layers as frame descriptors. We concatenate the frames' descriptors along the temporal

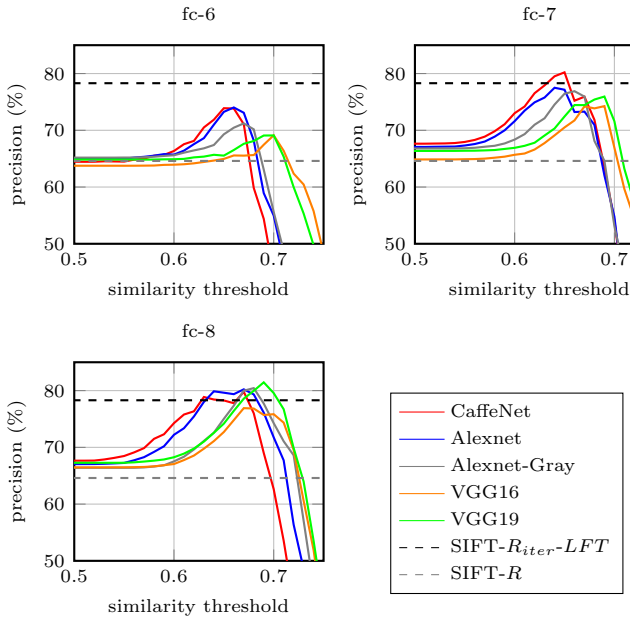

Fig. 3. Evaluation results on our MIS video dataset, using deep features.

dimension and finally apply a low-pass temporal filter to smooth out fluctuations. The aggregation to a fixed-size shot descriptor is performed by a temporal max-pooling layer over the descriptors. The size of the shot descriptor is equal to the size of the fully connected layer used to extract feature maps. In the network topologies we evaluate in Section 5, this is equal to 4096 for "fc-6" and "fc-7", and 1000 for "fc-8".

5 Experiments

From a 2 hour MIS video we extracted 186 shots using the method in Section 3. Each shot was manually annotated by a field-expert, focusing on the tools appearing in the shots (as described in Section 2). We measure the average precision metric for evaluating retrieval. A retrieved result is considered a true positive if it contains at least one of the tools depicted in the query shot.

The video retrieval pipeline consists of the following steps: a) video segmentation by monitoring the global objectness measure, b) descriptor extraction for each shot using either local features or deep CNNs, and c) retrieval of similar shots. Since the experimental evaluation is not of large scale, we search for the exact nearest neighbors in the descriptor space.

First, we evaluate the performance of local feature-based approaches. We evaluate the performance of the baseline system using local features with RANSAC

(R), and test the performance improvement when employing R_{iter}, local feature trajectories (LFT), and both. The results in Table 1 show that employing LFT selection of local features improves the results by approximately 19%. Also, employing iterative RANSAC further improves the results in both cases (with or without LFT selection).

For the proposed representation we evaluate the performance of pre-trained models on the Imagenet dataset [5]. Specifically, we use the pre-trained models from the Caffe library [11], namely *AlexNet* [15], the *CaffeNet* network (which is almost identical to AlexNet), and *VGG-nets* [28]. Note that all these models use color images as input, while in the local features based approach color information was discarded. In order to evaluate the importance of color, we train an adapted *AlexNet* on grayscale images from Imagenet, and use all the fully connected layers as output to form the descriptors. The results of the evaluation are depicted in Figure 3. The x axis corresponds to the similarity threshold for selecting similar shots. In all tested network topologies, "fc-6" is the first fully connected layer, fusing information from the convolutional layers; "fc-7" is the second fully connected layer, while "fc-8" is the last layer of the network.

Using the fc-6 layer to extract frame descriptors we achieve 74.0% precision, which is superior to the baseline method, but does not reach the best result achieved using local features. Using fc-7 or fc-8 exceeds the local features methods, achieving 80.2% and 81.5% respectively. Multiple fully connected layers provide better frame descriptors, as more neurons are used to fuse filtering responses from the preceding convolutional layers. The best performing descriptor is extracted by the fc-8 layer of VGG19 network, the deepest of the evaluation, using 19 layers. The fc-8 layer corresponds to the final layer, which in our case is trained to classify images to the Imagenet's 1000 categories. The fact that this layer is top-performing is due to the size of the Imagenet dataset, and the diversity of the included categories.

Comparing the performance of AlexNet and Alexnet-gray, we see that exploiting color information is useful in the first fully connected layer (fc-6). However, the performance gap is filled by the fusion of fc-7 and fc-8.

We also perform shot detection and retrieval in the tool detection dataset of M2CAI challenge. We split videos in shots using the proposed method and create per shot annotations, indicating the presence of tools in each shot. The top-performing networks from the previous experiment (i.e. CaffeNet, AlexNet and VGG19) are used to extract video shot descriptors. The retrieval results for two videos of the dataset are depicted in Fig. 4. AlexNet using layers fc-7 and fc-8 provides better performance, 81.7% and 81.2% respectively, followed closely by VGG19.

6 Discussion

In this paper we propose a novel video representation method that fits well with retrieval of surgical videos. The shot decomposition we propose is based on a per-frame global objectness measure, while the final representation is based on

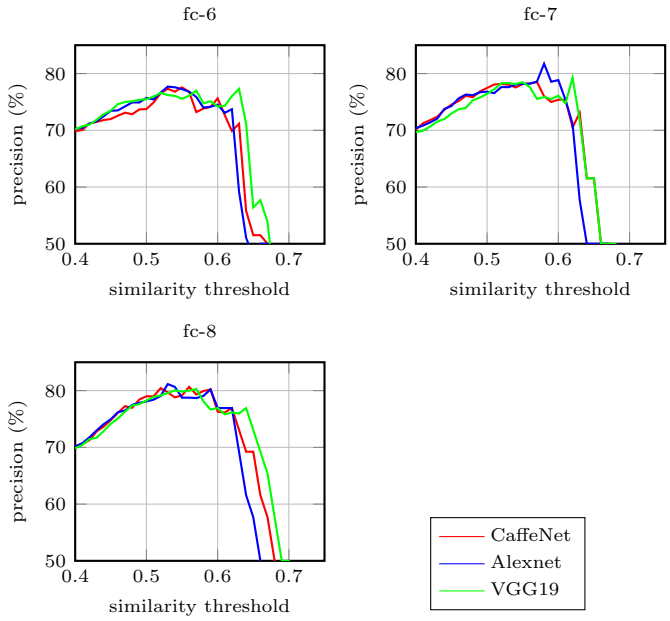

Fig. 4. Evaluation results on the M2CAI challenge dataset, using deep features.

CNN features extracted from frames and aggregated over time. The results are quite promising and we plan to apply this approach to a larger dataset of raw MIS videos.

References

1. Alexe, B., Deselaers, T., Ferrari, V.: Measuring the objectness of image windows. PAMI 34(11), 2189–2202 (2012)
2. Van den Bergh, M., Roig, G., Boix, X., Manen, S., Van Gool, L.: Online video seeds for temporal window objectness. In: ICCV (2013)
3. Blum, T., Feußner, H., Navab, N.: Modeling and segmentation of surgical workflow from laparoscopic video. In: Medical Image Computing and Computer-Assisted Intervention, pp. 400–407 (2010)
4. Chang, K., Liu, T., Chen, H., Lai, S.: Fusing generic objectness and visual saliency for salient object detection. In: ICCV (2011)
5. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
6. Douze, M., Revaud, J., Verbeek, J., Jégou, H., Schmid, C.: Circulant temporal encoding for video retrieval and temporal alignment. IJCV pp. 1–16 (2015)
7. Fischler, M., Bolles, R.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM 24(6), 381–395 (1981)

8. Giannarou, S., Yang, G.: Content-based surgical workflow representation using probabilistic motion modeling. In: Medical Imaging and Augmented Reality, pp. 314–323 (2010)
9. Gkioxari, G., Malik, J.: Finding action tubes. In: CVPR (2015)
10. Jain, M., Van Gemert, J., Jégou, H., Bouthemy, P., Snoek, C.: Action localization with tubelets from motion. In: CVPR (2014)
11. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)
12. Kanazawa, Y., Kawakami, H.: Detection of planar regions with uncalibrated stereo using distributions of feature points. In: BMVC (2004)
13. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR (2014)
14. Karpenko, A., Aarabi, P.: Tiny videos: a large data set for nonparametric video retrieval and frame classification. PAMI 33(3), 618–630 (2011)
15. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates, Inc (2012)
16. Lienhart, R.: Comparison of automatic shot boundary detection algorithms. In: Electronic Imaging'99. pp. 290–301 (1998)
17. Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.: Microsoft coco: Common objects in context. In: ECCV (2014)
18. Lo, B., Darzi, A., Yang, G.: Episode classification for the analysis of tissue/instrument interaction with multiple visual cues. In: Medical Image Computing and Computer-Assisted Intervention, pp. 230–237 (2003)
19. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
20. Loukas, C., Georgiou, E.: Smoke detection in endoscopic surgery videos: a first step towards retrieval of semantic events. The International Journal of Medical Robotics and Computer Assisted Surgery 11(1), 80–94 (2015)
21. Loukas, C., Nikiteas, N., Schizas, D., Georgiou, E.: Shot boundary detection in endoscopic surgery videos using a variational bayesian framework. International Journal of Computer Assisted Radiology and Surgery (in press)
22. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV 60(2), 91–110 (2004)
23. Müller, H., Michoux, N., Bandon, D., Geissbuhler, A.: A review of content-based image retrieval systems in medical applicationsclinical benefits and future directions. International journal of medical informatics 73(1), 1–23 (2004)
24. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: CVPR (2007)
25. Primus, M., Schoeffmann, K., Boszormenyi, L.: Segmentation of recorded endoscopic videos by detecting significant motion changes. In: CBMI (2013)
26. Primus, M., Schoeffmann, K., Boszormenyi, L.: Instrument classification in laparoscopic videos. In: CBMI (2015)
27. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., Lecun, Y.: Overfeat: Integrated recognition, localization and detection using convolutional networks. In: ICLR (2014)
28. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
29. Sun, J., Wu, X., Yan, S., Cheong, L., Chua, T., Li, J.: Hierarchical spatio-temporal context modeling for action recognition. In: CVPR (2009)

30. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., de Mathelin, M., Padoy, N.: Endonet: A deep architecture for recognition tasks on laparoscopic videos. arXiv preprint arXiv:1602.03012 (2016)
31. Uijlings, J., van de Sande, K., Gevers, T., Smeulders, A.: Selective search for object recognition. IJCV 104(2), 154–171 (2013)
32. Vincent, E., Laganiére, R.: Detecting planar homographies in an image pair. In: Image and Signal Processing and Analysis. pp. 182–187 (2001)
33. Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: Deep networks for video classification. In: CVPR (2015)