Medical Image Analysis 47 (2018) 203–218

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Monitoring tool usage in surgery videos using boosted convolutional and recurrent neural networks

Hassan Al Hajj[a], Mathieu Lamard[a,b], Pierre-Henri Conze[a,c], Béatrice Cochener[a,b,d], Gwenolé Quellec[a,*]

[a] Inserm, UMR 1101, Brest F-29200, France
[b] Univ Bretagne Occidentale, Brest F-29200, France
[c] Institut Mines-Télécom Atlantique, Brest F-29200, France
[d] Service d'Ophtalmologie, CHRU Brest, Brest F-29200, France

ARTICLE INFO

Article history:
Received 6 October 2017
Revised 5 March 2018
Accepted 3 May 2018
Available online 9 May 2018

Keywords:
Cataract and cholecystectomy surgeries
Tool usage monitoring
Video analysis
Convolutional and recurrent neural networks
Boosting

ABSTRACT

This paper investigates the automatic monitoring of tool usage during a surgery, with potential applications in report generation, surgical training and real-time decision support. Two surgeries are considered: cataract surgery, the most common surgical procedure, and cholecystectomy, one of the most common digestive surgeries. Tool usage is monitored in videos recorded either through a microscope (cataract surgery) or an endoscope (cholecystectomy). Following state-of-the-art video analysis solutions, each frame of the video is analyzed by convolutional neural networks (CNNs) whose outputs are fed to recurrent neural networks (RNNs) in order to take temporal relationships between events into account. Novelty lies in the way those CNNs and RNNs are trained. Computational complexity prevents the end-to-end training of "CNN+RNN" systems. Therefore, CNNs are usually trained first, independently from the RNNs. This approach is clearly suboptimal for surgical tool analysis: many tools are very similar to one another, but they can generally be differentiated based on past events. CNNs should be trained to extract the most useful visual features in combination with the temporal context. A novel boosting strategy is proposed to achieve this goal: the CNN and RNN parts of the system are simultaneously enriched by progressively adding weak classifiers (either CNNs or RNNs) trained to improve the overall classification accuracy. Experiments were performed in a dataset of 50 cataract surgery videos, where the usage of 21 surgical tools was manually annotated, and a dataset of 80 cholecystectomy videos, where the usage of 7 tools was manually annotated. Very good classification performance are achieved in both datasets: tool usage could be labeled with an average area under the ROC curve of $A_z = 0.9961$ and $A_z = 0.9939$, respectively, in offline mode (using past, present and future information), and $A_z = 0.9957$ and $A_z = 0.9936$, respectively, in online mode (using past and present information only).

1. Introduction

With the emergence of imaging devices in the operating room, the automated analysis of videos recorded during the surgery is becoming a hot research topic. In particular, videos can be used to monitor the surgery, for instance by recognizing which surgical tools are being used at every moment. An immediate application of surgery monitoring is report generation. If automatic reports are available for many surgeries, then the automatic analysis of these reports can help optimize the surgical workflow or evaluate surgical skills. Additionally, if we are able to generate such a report in real-time, during a surgery, then we could compare it with previous reports to generate warnings, if we recognize patterns often leading to complications, or recommendations, to help younger surgeons emulate more experienced colleagues based on their surgical reports (Quellec et al., 2015). With adequate image analysis techniques, tool usage could be monitored reliably in tool-interaction videos, such as endoscopic videos (in laparoscopic or retinal surgeries) or microscopic videos (in anterior eye segment surgeries). In the simplest scenario, we can consider that a tool is being used if it is visible in these videos. In a more advanced scenario, we can consider that it is in use if it is in contact with the tissue (as opposed to approaching the tissue, waiting to be used, etc.). Therefore, several tool detection techniques for tool-interaction videos have been proposed in recent years (Bouget et al., 2017). To compare these techniques, two tool

* Corresponding author at: LaTIM - IBRBS - CHRU Morvan - 12, Av. Foch, Brest CEDEX 29609, France.
E-mail address: gwenole.quellec@inserm.fr (G. Quellec).

https://doi.org/10.1016/j.media.2018.05.001
1361-8415/

detection challenges were organized recently. A first challenge, organized at the M2CAI 2016 workshop,[1] relied on endoscopic videos of cholecystectomy operations performed laparoscopically. We organized a second challenge for cataract surgery, the most common surgical procedure worldwide (Trikha et al., 2013).[2] It relied on videos recorded through a surgical microscope. Following the trend in medical image and video analysis (Shen et al., 2017), the best solutions of both challenges all relied on convolutional neural networks (CNNs) (Raju et al., 2016; Sahu et al., 2016; Twinanda et al., 2017; Zia et al., 2016; Roychowdhury et al., 2017; Hu and Heng, 2017; Maršalkaitė et al., 2017).

Compared to other computer vision tasks, surgical tool usage annotation has several specificities. First, as opposed to many computer vision tasks, including the popular ImageNet visual recognition challenges,[3] the problem at hand is not multiclass classification (one correct label per image among multiple classes), but rather multilabel classification (multiple correct labels per image): the number of tools being used in each image varies (from zero to three in cataract surgery for instance). Therefore, multilabel CNNs should be used. Second, taking the temporal sequencing into account is important: knowing which tools have already been used since the beginning of the surgery greatly helps recognize which tools are currently being used. Therefore, multilabel recurrent neural networks (RNNs) (Hochreiter and Schmidhuber, 1997) may also be used advantageously. In fact, recent machine learning competitions clearly show that ensembles of CNNs outperform single CNNs (Russakovsky et al., 2015): multiple CNNs with different architectures are generally trained independently and their outputs are combined afterward using standard machine learning algorithms (decision trees, random forests, multilayer perceptrons, etc.). However, this simple strategy is suboptimal since difficult samples may be misclassified by all CNNs. And there are many difficult samples to classify in surgery videos: in particular, many tools resemble one other (e.g. two types of cannulae in cataract surgery). Building the ensemble of CNNs using a boosting meta-algorithm (Freund and Schapire, 1997) can theoretically design CNNs focusing specifically on challenging samples. Boosting an ensemble of RNNs would also make sense as there are difficult samples along the time dimension as well: in particular, some tools or tool usage sequences are very rare and temporal sequencing algorithms tend to misclassify those rare cases. Therefore, we propose to jointly boost an ensemble of CNNs and an ensemble of RNNs for automatic tool usage annotation in surgery videos. In the same way as CNN boosting (or RNN boosting) allows various CNNs (or RNNs) to be complementary, this general boosting solution allows CNNs to be complementary with RNNs. In that sense, it approximates the end-to-end training of a "CNN+RNN" network, which is theoretically ideal but not computationally tractable.

The remainder of this paper is organized as follows. Section 2 reviews the state of the art of video analysis, and surgery video analysis in particular. Sections 3 and 4 describe the proposed solution. Section 5 presents the video datasets and Section 6 reports the experiments performed on that dataset. We end with a discussion and conclusions in Section 7.

2. State of the art

2.1. Deep learning for video analysis

The automatic analysis of dynamic scenes through deep learning has become a very hot research topic (Simonyan and Zisserman, 2014; Wang et al., 2017; Donahue et al., 2017). Different strategies have been proposed for this task. A first strategy is to regard videos or video portions as 3-D images and therefore analyze them with 3-D CNNs (Ji et al., 2013), which is computationally expansive. A second strategy is to analyze 2-D images as well as the optical flow between consecutive images (Simonyan and Zisserman, 2014), with the disadvantage of only modeling short-term relationships between images. A third strategy is to combine a CNN, analyzing 2-D images, with a RNN analyzing the temporal sequencing (Donahue et al., 2017). The main advantage of this "CNN+RNN" approach, which is now the leading video analysis solution, is that long-term relationships between events can be taken into account efficiently. One application of "CNN+RNN" models, which is particularly relevant for our study, is video labeling: the goal is to assign one class label to each frame inside a video (Singh et al., 2016; Khorrami et al., 2016). Medical applications of this research, ranging from gait analysis (Feng et al., 2016) to surgery monitoring (Bodenstedt et al., 2017; Twinanda et al., 2016), are starting to emerge.

2.2. Temporal analysis of surgery videos

In the context of surgical workflow analysis, solutions have been proposed to recognize surgical phases in surgery videos (Lalys and Jannin, 2014; Charrière et al., 2017). In Primus et al. (2018), phases are recognized using one CNN processing the visual content of one frame plus the relative timestamp of that frame. However, most solutions rely on statistical models, such as Hidden Markov Models (HMMs) (Cadène et al., 2016), Hidden semi-Markov Models (Dergachyova et al., 2016; Tran et al., 2017), Hierarchical HMMs (Twinanda et al., 2017), Linear Dynamical Systems (Zappella et al., 2013; Tran et al., 2017) or Conditional Random Fields (Tao et al., 2013; Quellec et al., 2014; Lea et al., 2016a). Recently, solutions based on RNNs have also been proposed (Jin et al., 2016; Bodenstedt et al., 2017; Twinanda et al., 2016). Following the state-of-the-art video analysis strategy, these RNNs process instant visual features extracted by a CNN from images. In particular, Jin et al. (2016) applied a "CNN+RNN" network to a small sliding window of three images. Bodenstedt et al. (2017) applied a "CNN+RNN" network to larger sliding windows and copy the internal state of the network between consecutive window locations. As for Twinanda et al. (2016), they applied a "CNN+RNN" network to full videos. Interestingly, the CNN proposed by Twinanda et al. (2016), namely EndoNet, detects tools as an intermediate step. A challenge on surgical workflow analysis was also organized at M2CAI 2016:[4] two of the top three solutions relied on RNNs (Jin et al., 2016; Twinanda et al., 2016). It should be noted that successful works on the analysis of kinematics surgery data have also been reported, using a RNN (Dipietro et al., 2016) or a CNN along the temporal dimension (Lea et al., 2016b). In all these works, statistical models or RNNs were used to label surgical activities and phases. Given the strong correlation between surgical activities and tool usage, they can be expected to improve tool recognition as well.

2.3. Deep learning for surgical tool detection

As evidenced by the M2CAI 2016 and CATARACTS 2017 challenges, the state-of-the-art algorithms for tool detection in surgery videos are CNNs. The best solutions of these challenges rely on a transfer learning strategy: well-known CNNs trained to classify still images in the ImageNet dataset were fine-tuned on images extracted from surgery videos. For M2CAI 2016, Sahu et al. (2016) and

[1] http://www.camma.u-strasbg.fr/m2cai2016/index.php/tool-presence-detection-challenge-results/.

[2] https://www.cataracts.grand-challenge.org/.

[3] http://www.image-net.org/challenges/LSVRC/2017/index.php.

[4] http://www.camma.u-strasbg.fr/m2cai2016/index.php/workflow-challenge-results/.