Medical Image Analysis 18 (2014) 579–590

Contents lists available at ScienceDirect

Medical Image Analysis

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

Real-time recognition of surgical tasks in eye surgery videos

Gwénolé Quellec [a,*], Katia Charrière [b,a], Mathieu Lamard [c,a], Zakarya Droueche [b,a], Christian Roux [b,a], Béatrice Cochener [c,a,d], Guy Cazuguel [b,a]

[a] *Inserm, UMR 1101, Brest F-29200, France*
[b] *INSTITUT Mines-Télécom, TELECOM Bretagne, UEB, Dpt ITI, Brest F-29200, France*
[c] *Univ Bretagne Occidentale, Brest F-29200, France*
[d] *CHRU Brest, Service d'Ophtalmologie, Brest F-29200, France*

ARTICLE INFO

Article history:
Received 28 November 2012
Received in revised form 7 February 2014
Accepted 17 February 2014
Available online 26 February 2014

Keywords:
CBVR
Real-time
Surgical task recognition
Eye surgery

ABSTRACT

Nowadays, many surgeries, including eye surgeries, are video-monitored. We present in this paper an automatic video analysis system able to recognize surgical tasks in real-time. The proposed system relies on the Content-Based Video Retrieval (CBVR) paradigm. It characterizes short subsequences in the video stream and searches for video subsequences with similar structures in a video archive. Fixed-length feature vectors are built for each subsequence: the feature vectors are unchanged by variations in duration and temporal structure among the target surgical tasks. Therefore, it is possible to perform fast nearest neighbor searches in the video archive. The retrieved video subsequences are used to recognize the current surgical task by analogy reasoning. The system can be trained to recognize any surgical task using weak annotations only. It was applied to a dataset of 23 epiretinal membrane surgeries and a dataset of 100 cataract surgeries. Three surgical tasks were annotated in the first dataset. Nine surgical tasks were annotated in the second dataset. To assess its generality, the system was also applied to a dataset of 1,707 movie clips in which 12 human actions were annotated. High task recognition scores were measured in all three datasets. Real-time task recognition will be used in future works to communicate with surgeons (trainees in particular) or with surgical devices.

1. Introduction

Nowadays, many surgeries are video-monitored. We believe real-time video monitoring may be useful to automatically communicate information to the surgeon in due time. Typically, whenever the surgeon begins a new surgical task, relevant information about the patient, the surgical tools, etc., in connection to this task, may be communicated to him or her (either visually or phonically). The advantage is obvious for the less experienced surgeons. Recommendations on how to best perform the current or the next task, given the patient's specificities, may be communicated to them: these recommendations would derive from the experience of their peers in similar surgeries. A first step towards that goal is presented in this paper: we describe an algorithm to detect surgical tasks in real-time during the surgery.

In recent years, a few systems have been presented for the automatic recognition of surgical tasks or gestures. A first group of systems assumes that the surgical tasks or gestures follow a predefined order: the goal is to find when each task or gesture ends and when the next one begins. Blum et al. (2010) proposed a system to segment such surgical tasks in laparoscopic videos. During the training phase, tool usage is analyzed to perform dimension reduction on visual features, using canonical correlation analysis. At the end of the surgery, the video is registered to a manually segmented average surgery, using Dynamic Time Warping (DTW). Note that a similar system was presented by Padoy et al. (2012): the main difference is that it processes tool usage directly as observations, rather than visual features. Lalys et al. (2011) also proposed a similar system for microscope videos. Many visual features are extracted from images, including color histograms, Haar-based features and SIFT descriptors. Then, the surgery is temporally segmented in surgical tasks using the DTW. In a second group of systems, the DTW is replaced by a Hidden Markov Model (HMM) in order to relax the 'predefined order' hypothesis, although transitions between surgical tasks or gestures that are not seen in training will have a null probability. Blum et al. (2010), Padoy et al. (2012) and Lalys et al. (2011) proposed a variation on their technique described above, where the DTW is replaced by a HMM. Tao et al. (2012) proposed a system for segmenting a surgical task into a sequence of gestures, in

[*] Corresponding author. Address: LaTIM, Bâtiment 1, CHU Morvan, 5, Av. Foch 29609 Brest CEDEX, France. Tel.: +33 2 98 01 81 29; fax: +33 2 98 01 81 24.
E-mail address: gwenole.quellec@inserm.fr (G. Quellec).

http://dx.doi.org/10.1016/j.media.2014.02.007
1361-8415/© 2014 Elsevier B.V. All rights reserved.

laparoscopic videos. The system relies on *sparse HMMs*, whose observations are sparse linear combinations of elements from a dictionary of basic surgical motions; a dictionary is learnt for each gesture. A third group of systems assumes that the tasks or gestures have already been segmented and the goal is to classify each segmented task or gesture without contextual information. In that case, the predefined order hypothesis is completely relaxed. Haro et al. (2012) evaluated two approaches to surgical gesture classification in video clips. The first one is based on a linear dynamical system; the other is based on the Bag-of-Words (BoW) model (Harris, 1954; Huang et al., 2012; Tamaki and Yoshimuta, 2013; Lalys et al., 2011). These two approaches combined perform equally well as gesture classification based on kinematic data (Haro et al., 2012). Finally, note that several systems have been designed for related tasks: surgical tool detection and tracking (Cano et al., 2008), surgical task detection without categorization (Cao et al., 2007; Giannarou and Yang, 2010), surgical skill evaluation (Reiley and Hager, 2009), etc. Like the third group of recognition methods, the proposed system does not assume that the surgical tasks follow a predefined order and it requires segmented surgical tasks for training. After training, the proposed system can detect, in real-time, key video subsequences that typically occur during a given task, but not during other tasks. This detection does not require any segmentation. The proposed system can also categorize a task as a whole in real-time. But in that case, like the third group of methods, it assumes that the task is segmented.

Similar systems, without the real-time constraint, have been proposed outside the scope of surgery videos. They all rely on a collection of videos containing instances of the target actions for supervision. Piriou et al. (2006) proposed an action recognition framework for sport video indexing. A global probabilistic motion model is trained for each target action. To process a new video, the camera motion is first estimated and removed. Then, the residual motion is analyzed to classify the current action by maximum a posteriori estimation. Duchenne et al. (2009) presented a weakly supervised action recognition framework for movie clip indexing. First, spatiotemporal interest points are detected in videos. Then, video subsequences are characterized using a BoW model. Finally, subsequences containing the target action are detected using a weakly supervised SVM classifier. Xu and Chang (2008) proposed an event recognition framework for news video indexing. A BoW representation was also adopted to characterize varying-sized video subsequences. To compare two sequences, a variation on the EMD distance between subsequence characterizations was used. These frameworks were primarily designed for offline indexing of broadcast video, so they do not need to be run in real-time.

In order to detect key subsequences of surgical tasks, the proposed system relies on the Content-Based Video Retrieval (CBVR) paradigm. Given a video query, CBVR systems search for similar video contents in a video archive. Initially popularized in broadcasting (Naturel and Gros, 2008) and video surveillance (Smeaton et al., 2006; Hu et al., 2007), the use of CBVR is now emerging in medical applications (André et al., 2010; Syeda-Mahmood et al., 2005). We present a novel CBVR system able to perform real-time searches. In this paper, it is used to recognize the current surgical task by analogy reasoning. Section 2 presents the state of the art of CBVR and discusses the specific challenge of real-time CBVR.

The proposed system is applied to eye surgery. In those surgeries, the surgeon wears a binocular microscope and the output of the ophthalmoscope is recorded. Two of the most common eye surgeries are considered in this paper: epiretinal membrane surgery (Dev et al., 1999) and cataract surgery (Castells et al., 1998). Recently, Lalys et al. (2012) adapted their general system (Lalys et al., 2011) for segmenting cataract surgery videos. In the improved system, visual features are only extracted within the pupil only; an automatic pupil segmentation procedure is presented. Good temporal segmentation performances were measured (Lalys et al., 2012). However, that system does not allow real-time recognition of the surgical tasks. It needs to process the entire surgical video before segmenting it, which implies that the segmentation is only available after the end of the surgery. To our knowledge, this paper is the first attempt to recognize eye surgical tasks in real-time.

2. State of the art of content-based video retrieval

Many CBVR systems have been presented in the literature. These systems differ by the nature of the objects placed as queries. First, queries can be images (Patel et al., 2010). In that case, the goal is to select videos containing the query image in a reference dataset; these systems are very similar to image retrieval systems. Second, queries can be video shots (Naturel and Gros, 2008; Dyana et al., 2009). In that case, the goal is to find other occurrences of the query shot (Naturel and Gros, 2008), or similar shots (Dyana et al., 2009), in the reference dataset. Third, queries can be entire videos (André et al., 2010; Syeda-Mahmood et al., 2005). In that case, the goal is to select the most similar videos, overall, in the reference dataset.

CBVR systems also differ by the way videos or video subsequences are characterized. Several systems rely mainly on the detection and characterization of key frames (Juan and Cuiying, 2010; Patel et al., 2010). Others characterize videos or video subsequences directly (Dyana et al., 2009; Gao and Yang, 2010). In the system by Dyana et al. (2009), video shots are characterized by shape parameters and by the evolution of motion vectors over time. In the system by Gao and Yang (2010), spatiotemporal salient objects (i.e. moving objects) are detected in videos and characterized individually; videos are then compared using the Earth-Mover's Distance (EMD), which may be time consuming. The combination of multimodal (visual, audio and textual) information in a retrieval engine has also been proposed (Hoi and Lyu, 2007; Bruno et al., 2008).

Finally, CBVR systems differ by how flexible the distance metrics should be. First, several systems have been proposed to find objects that are almost identical to the query. For instance, Douze et al. (2010) proposed a copy detection system to protect copyrighted videos. In this system, images are compared individually and their temporal ordering is checked after hand. Another system has been proposed by Naturel and Gros (2008) to detect repeating shots in a video stream, in order to automatically structure television video content. However, in most CBVR systems, we are interested in finding videos or video subsequences that are semantically similar but whose visual content can significantly vary from one sequence to another (Juan and Cuiying, 2010; Xu and Chang, 2008; André et al., 2010). In other words, we need distance metrics able to bridge the so-called semantic gap (Smeulders et al., 2000).

In this paper, we present a CBVR system able to detect key subsequences in a video stream and also to categorize surgical tasks. Short video subsequences extracted from the video stream play the role of the query objects. A flexible distance metric is needed. As mentioned above, similar methods have been presented in the literature to solve this problem (Piriou et al., 2006; Duchenne et al., 2009; Xu and Chang, 2008). When searching for similar video subsequences, and not simply video files as a whole, the number of items that should be compared to the query item explodes. And, as opposed to above methods, the proposed system needs to run in real-time. In order to meet the real-time constraint, a very fast similarity metric must therefore be used to compare video subsequences. In particular, the use of temporally flexible distance metrics such as DTW (Sakoe and Chiba, 1978; Xu and Chang, 2008) is prohibited for time reasons. An alternative solution is