Automated Surgical Step Recognition in Normalized Cataract Surgery Videos

Katia Charrière, Gwénolé Quellec, Mathieu Lamard, Gouenou Coatrieux, Béatrice Cochener and Guy Cazuguel

Abstract— **Huge amounts of surgical data are recorded during video-monitored surgery. Content-based video retrieval systems intent to reuse those data for computer-aided surgery. In this paper, we focus on real-time recognition of cataract surgery steps: the goal is to retrieve from a database surgery videos that were recorded during the same surgery step. The proposed system relies on motion features for video characterization. Motion features are usually impacted by eye motion or zoom level variations, which are not necessarily relevant for surgery step recognition. Those problems certainly limit the performance of the retrieval system. We therefore propose to refine motion feature extraction by applying pre-processing steps based on a novel pupil center and scale tracking method. Those pre-processing steps are evaluated for two different motion features. In this paper, a similarity measure adapted from Piciarelli's video surveillance system is evaluated for the first time in a surgery dataset. This similarity measure provides good results and for both motion features, the proposed pre-processing steps improved the retrieval performance of the system significantly.**

I. INTRODUCTION

Automated analysis of data recorded during video-monitored surgeries or examinations is an increasing research field. The developed methods could help the surgeons in different ways: automatic documentation and report generation [1], [2], fast retrieval of similar cases from a database [3] or educative video construction [4]. A substantial part of existing methods are developed for surgical skill evaluation [5]. Currently, several methods try to automatically segment videos into surgical tasks [2] or gestures [6] by using hidden Markov models (HMM), linear dynamical systems (LDS) or dynamic time warping (DTW) [7]. Some methods are also developed to automatically recognize different steps of surgical exams [8]. But, few methods are able to work in real time and most methods can only provide information after the end of the surgery or the examination. Although few methods are able to work in real time in the surgery field, many real-time content-based video retrieval (CBVR) systems are developed for video surveillance to mine behaviors and detect salient events in public scenes [9]. But video surveillance data are quite different from medical data, especially because they contain a stationary background.

A study has been initiated in the LaTIM laboratory to set up an alarm/recommendation generation system for video-monitored surgery [10]. This implies to have fast and robust methods able to recognize surgical tasks or gestures in real time. To obtain the best possible results, a first important step is to select the best way to represent video sequences, by pre-processing steps and feature extraction.

In this paper, we test several pre-processing steps based on a pupil center and scale tracking method presented in a companion paper submitted to this conference [11]. Effects of those pre-processing steps are tested for two kinds or extracted features: 1) bag of visual word (BoW) features based on STIP extraction [12] and 2) motion histograms (MH) extracted from the optical flow between two consecutive frames. A simple nearest neighbor search, using a method proposed by Piciarelli et al. [13] as similarity measure, is used to estimate which step is executed during a video sequence. This similarity measure, initially proposed for video surveillance systems, is evaluated for the first time in a surgery dataset hereafter.

II. CHARACTERIZATION OF VIDEO SEQUENCES

A. Pre-processing Steps

Feature extraction for video sequence characterization is an essential part of CBVR systems. Robust video characterizations must be found to obtain the best possible results in the retrieval step. As visual features are often based on gradient magnitude, motion or color information, some pre-processing steps could be applied to refine feature extraction. This improves video characterization without increasing feature vector sizes. In this paper we normalize video frames in three different ways by pre-processing steps based on a pupil center and scale tracking method presented in a companion paper submitted to this conference [11]. In that method, the pupil center and the image scale are tracked without explicitly segmenting the pupil or the iris. First, the pupil center is detected using the Hough transform: circle centers are detected in spatially and temporally smoothed 2-D accumulators. Since the pupil boundaries, the limbus and the lids have approximately the same center, their edge information will accumulate in the same region of the accumulator. So pupil center detection is quite reliable with this approach. To push performance further, only circles whose inside is darker than the outside are retained. Then, the zoom level

K. Charriere, Gouenou Coatrieux and G. Cazuguel are with INSTITUT TELECOM; TELECOM Bretagne; UEB; Dpt ITI, Brest, F-29200 France katia.charriere@telecom-bretagne.eu

G. Quellec is with INSERM UMR 1101, Brest, F-29200 France

M. Lamard and B. Cochener are with Univ Bretagne Occidentale, and and European University of Brittany, Brest, F-29200 France

B. Cochener is with CHU Brest, Service d'Ophtalmologie, Brest, F-29200 France

All authors are with LaTIM - INSERM UMR 1101, SFR ScInBioS (IFR 148), Brest, F-29200 France

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

is estimated by the size of the illumination pattern reflected on the cornea, which is mostly controlled (linearly) by the zoom factor.

By knowing the pupil center and the zoom level in each frame of the video, we can pre-process them in order to balance the effects of eye motion, zoom or level variations before the feature extraction step. Pre-processing steps are presented below:

- **Registration:** Motion is a relevant feature in videos, but disruptive motions could appear, induced by the camera or eye motion in the case of cataract surgery for instance. We balance eye motion by registering all frames on the same pupil center. A simple coordinate system change is applied, which places the iris center at the image center. This should eliminate motion induced by the eye or the camera and make instrument motion most prominent.
- **Region of interest:** All relevant actions should appear in a region close to the iris location. In that case, it is not useful to extract visual features in all the camera's field of view. this is why a circular mask centered on the iris center is applied to select a region of interest.
- **Scaling:** The effects of zoom level variations can be balanced by scaling all frames at the same scale level. After this last pre-processing step, all irises should have the same radius size.

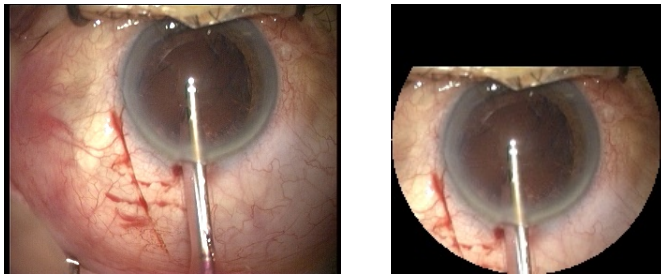
Before each pre-processing step, video frames were spatially downsampled by a factor of 2. First, the effects of those three different normalizations are evaluated independently and then, we test the combination of those three normalizations together. Example of frames before and after the three pre-processing steps are presented in Fig. 1.

B. Feature Extraction

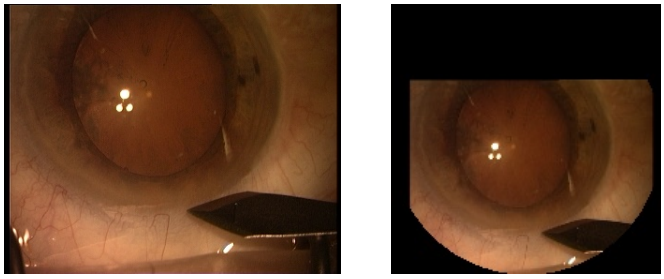
We evaluate video normalizations proposed in section (§II-A) for two kinds of features. The first one is based on Space-Time Interest Points (STIP), as proposed by Laptev et al. [12]. Histograms of oriented gradient (HOG) and histograms of optical flows (HOF) are extracted from a cube centered around each STIP point and concatenated. Those feature vectors are used to build a dictionary of visual words. Then, a bag of visual words is extracted for each video frame in order to characterize them.

A second kind of visual features are extracted to compare video sequence is based on the optical flow. Strong corners are first detected and selected. Then, the optical flow between two consecutive frames is computed at each strong corner by the Lucas-Kanade iterative method [14]. The OpenCV 2 library[1] was used to select strong corners and compute the optical flow. Finally, the motion is characterized by one 8-bin amplitude histogram, two 8-bin amplitude-weighted spatial histograms (one for the x-coordinates and one for the y-coordinates) and one 8-bin amplitude-weighted directional histogram.

[1]http://opencv.org/



(a) Before (b) After



(c) Before (d) After

Fig. 1: Example of frames before and after the combination of the three pre-processing steps

III. SIMILAR VIDEO SEQUENCE RETRIEVAL

A. Video Sequence Similarity Measurement

Similarity measurements between two video sequences need to take care of time distortions between two surgeries. A senior surgeon can perform a surgery very fast while a novice surgeon needs more time. The well-known dynamic time warping distance handles those time distortions but doesn't allow real time comparisons, in the sense that the surgery videos need to be available in full before they can be matched [2]. To deal with this problem, we adapted the progressive similarity measure proposed by Piciarelli et al. [13]. Distance measure between two videos is defined as follows:

$$D(V_1, V_2) = \sum_{i=1}^{n} d(v_{1i}, V_2)$$

where

$$d(v_{1i}, V_2) = \min_j(dist(v_{1i}, v_{2j}),\ j \in \{[(i-\delta)i]...[(i+\delta)i]\}$$

and $dist(v_{1i}, v_{2j}$ is the Bhattacharyya distance between the visual feature vectors of two frames. The distance from a new sequence to another sequence in the dataset is the average distance between one frame in the new sequence and its nearest neighbor in the other video, found inside a sliding temporal window centered in j. The sliding temporal window has a variable increasing size. This distance can be computed *in-line*, as the frames are collected.

B. Sequence Categorization

We aim to determine which step is executed in a new video sequence. Given a video database in which each video sequence represents one surgical step, a simple nearest

neighbor search is performed using the similarity measure introduced in the previous section. The 5 nearest neighbors of the new video sequence are used to estimate the probability that this video sequence belongs to each possible step.

IV. EXPERIMENTS

A. Brest Cataract Dataset

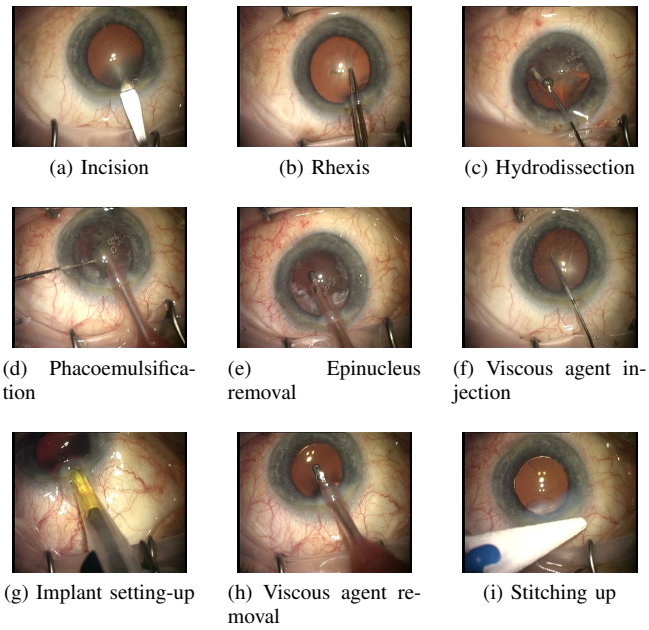
A Dataset of 30 cataract surgery videos, recorded at Brest University Hospital in 2011, were used in this experiment. Videos were acquired in DV format with a resolution of 720x576 pixels. Surgeries were performed by 7 different surgeons. The nine following steps were manually delimited by a surgeon in all videos: incision, rhexis, hydrodissection, phacoemulsification, epinucleus removal, viscous agent injection, implant setting-up, viscous agent removal and stitching up. The full surgery videos were then segmented into shorter video sequences with respect to those delimitations: 303 sequences were obtained.

B. Results

Algorithm parameters were set empirically: the number of visual words for BoW features was set to 1000 and the δ parameter of the sliding temporal window for similarity measurement was set to 0.1. Those values were chosen by looking only at videos that had not been pre-processed. For image scaling normalization, the mean iris radius was set to 93 pixels, which is the mean iris radius measured in the database. The circular mask for the ROI selection was designed with a radius equal to the iris radius plus 50 pixels. Pre-processing effects are evaluated for two kinds of features : BoW computed from STIP points and Motion Histograms (MH) computed from optical flow extraction. We first evaluated the effects of the three pre-processing steps independently and then, we evaluated the combination of the three pre-processing steps. The performance of the system is measured for each surgery step and each pre-processing step in terms of A_z, the area under the Receiver Operating Characteristic (ROC) curve. For BoW characterization, a training step is necessary for visual word dictionary building. The 303 surgeries video subsequence were divided randomly into two sets of approximately equal size. One of these sets was used as test set, and the other one as training set. The results are presented in TABLE I and processing times are presented on TABLE II.

V. DISCUSSION

Regarding motion histogram features, the use of each processing step has a significant impact on the mean A_z for each pre-processing step. Incision, rexhis an hydrodissection step recognition are essentially improved by registration of the iris center at the image center. It could be assumed that each motion are induced mainly by the patient and not by surgical gestures. On the other hand, implant setting-up step recognition is not improved by registration: for this step, motion induced by surgical gestures provides a useful information. The selection of a ROI and of the scaling normalization improves the mean A_z with contrasted results



(a) Incision (b) Rhexis (c) Hydrodissection

(d) Phacoemulsification (e) Epinucleus removal (f) Viscous agent injection

(g) Implant setting-up (h) Viscous agent removal (i) Stitching up

Fig. 2: Cataract surgery steps.

for each surgical step. An interesting point could be to analyze the impact of ROI radius variations on A_z. For example, in the incision step, where instrument positions and gestures predominantly take place close to the iris border, a too small ROI can lead to a loss of information. For all surgical steps, the retrieval performance is significantly improved by the combination of all three proposed pre-processing steps, with the exception of Rhexis where A_z is similar with or without pre-processing.

Regarding BoW features, for all surgical steps except Rhexis and Viscous agent removal, the retrieval performance is significantly improved by using a circular mask to select a region of interest centered on iris center. Resizing the frame to obtain the same iris size also provides relevant results. But, no significant effects are obtained by registering video frames. This can be explained by the fact that this representation does not care about the position of the detected motions, unlike the histogram-based representation.

The average processing time was between 20 ms and 54 ms per image for pre-processing steps. But video normalizations help speed up the computation time for feature extraction. Even if features based on STIP outperform MH features in terms of retrieval performance, STIP extraction requires more computation time. This can be explain by the search of interest points not only in the spatial domain but also in the temporal domain, by extension of the Harris method with a temporal component.

VI. CONCLUSION

A similarity measure based on the method proposed by Picarelly et al. [13] for video surveillance systems was tested on a surgery dataset and provide good results even without normalization, especially for BoW features. Also,

TABLE I: Categorization performance (A_z) for each surgery step and pre-processing

Surgical steps	MH	MH / Registration	MH / ROI	MH / Scale	MH / Reg ROI Scale
Incision	0.623213	**0.698801**	0.602917	0.654751	0.685886
Rhexis	0.72851	**0.75293**	0.644505	0.668437	0.720818
Hydrodissection	0.5	**0.637931**	0.576139	0.63252	0.611566
Phacoemulsification	0.744078	0.740049	0.678266	0.840171	**0.893468**
Epinucleus removal	0.710989	0.763004	0.840781	0.751221	**0.906349**
Viscous agent injection	0.956565	0.98096	0.974669	0.965734	**0.99053**
Implant setting-up	0.715873	0.632234	0.741087	0.682234	**0.763248**
Viscous agent removal	0.761233	0.851343	0.787485	0.848901	**0.909402**
Stitching up	0.848779	0.767033	0.846032	0.818559	**0.855189**
Mean A_z	0.732137778	0.758253889	0.743542333	0.762503111	**0.815161778**
Standard error of the mean	0.04272642	**0.035865538**	0.043650246	0.037614073	0.041740541

Surgical steps	BoW	BoW / Registration	BoW / ROI	BoW / Scale	BoW / Reg ROI Scale
Incision	0.763415	0.786179	0.78374	**0.823577**	0.801355
Rhexis	0.726947	0.623207	0.72541	**0.778945**	0.658555
Hydrodissection	0.807209	0.63922	**0.811177**	0.801257	0.79828
Phacoemulsification	0.781478	0.885027	0.911765	0.899368	**0.935586**
Epinucleus removal	0.831707	0.821951	**0.895664**	0.871003	0.855556
Viscous agent injection	0.967857	0.956006	0.973864	0.966234	**0.974351**
Implant setting-up	0.877592	0.881912	**0.906682**	0.885081	0.883065
Viscous agent removal	0.834787	0.799891	0.826063	**0.86096**	0.803708
Stitching up	0.563364	0.790323	**0.868664**	0.806452	0.766705
Mean A_z	0.794928444	0.798190667	**0.855892111**	0.854764111	0.830795667
Standard error of the mean	0.037127279	0.036537665	0.02540899	**0.019562752**	0.031483357

TABLE II: Computing time for pre-processing steps and feature extraction in second

	Without normalization	Registration	Scale	ROI	Reg Scale ROI
Pre-processing	0	0.0206	0.054	0.0200	0.0299
Optical flow extraction	0.0176	0.0031	0.0031	0.0032	0.0010
STIP extraction	1.0190	0.3173	0.3145	0.3347	0.1561
Bow extraction	0.0049	0.0029	0.0030	0.0033	0.0017

this work shows that video normalization significantly improves retrieval results when using visual features based on motion extraction. The combination of three normalisations (iris position, iris size et region of interest) provides good improvement for features based on optical flow extraction. Normalizations based on iris size or ROI selection improves performance for BoW features. This will be used for real time recognition of surgical steps in entire surgeries, with the aim to be able to provide recommendations, or video sequence examples during the surgery.

REFERENCES

[1] S. R. Stanek, W. Tavanapong, J. Wong, J. H. Oh, and P. C. De Groen, "Automatic real-time detection of endoscopic procedures using temporal features," *Computer methods and programs in biomedicine*, vol. 108, no. 2, pp. 524–535, 2012.

[2] F. Lalys, L. Riffaud, D. Bouget, and P. Jannin, "A framework for the recognition of high-level surgical tasks from video images for cataract surgeries," *Biomedical Engineering, IEEE Transactions on*, vol. 59, no. 4, pp. 966–976, 2012.

[3] B. André, T. Vercauteren, A. M. Buchner, M. B. Wallace, and N. Ayache, "Learning semantic and visual similarity for endomicroscopy video retrieval," *Medical Imaging, IEEE Transactions on*, vol. 31, no. 6, pp. 1276–1288, 2012.

[4] Y. Cao, S.-H. Liu, M. Li, S. Baang, and S. Hu, "Medical video event classification using shared features," in *Multimedia, 2008. ISM 2008. Tenth IEEE International Symposium on*. IEEE, 2008, pp. 266–273.

[5] C. E. Reiley and G. D. Hager, "Task versus subtask surgical skill evaluation of robotic minimally invasive surgery," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2009*. Springer, 2009, pp. 435–442.

[6] L. Zappella, B. Béjar, G. Hager, and R. Vidal, "Surgical gesture classification from video and kinematic data," *Medical image analysis*, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361841513000522

[7] N. Padoy, T. Blum, S.-A. Ahmadi, H. Feussner, M.-O. Berger, and N. Navab, "Statistical modeling and recognition of surgical workflow," *Medical Image Analysis*, vol. 16, no. 3, pp. 632–641, April 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361841510001131\#

[8] Y. Cao, D. Liu, W. Tavanapong, J. Wong, J.-H. Oh, and P. C. de Groen, "Computer-aided detection of diagnostic and therapeutic operations in colonoscopy videos." *IEEE Trans. Biomed. Engineering*, vol. 54, no. 7, pp. 1268–1279, 2007.

[9] T. Hospedales, S. Gong, and T. Xiang, "Video behaviour mining using a dynamic topic model," *International journal of computer vision*, vol. 98, no. 3, pp. 303–323, 2012. [Online]. Available: http://www.springerlink.com/content/4244528725h33111/

[10] G. Quellec, K. Charrière, M. Lamard, Z. Droueche, C. Roux, B. Cochener, and G. Cazuguel, "Real-time recognition of surgical tasks in eye surgery videos," *Medical Image Analysis*, 2014.

[11] G. Quellec, K. Charrière, M. Lamard, B. Cochener, and G. Cazuguel, "Normalizing videos of anterior eye segment surgeries," in *Proc Int Conf IEEE Eng Med Biol Soc*, 2014.

[12] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[13] C. Piciarelli and G. L. Foresti, "On-line trajectory clustering for anomalous events detection," *Pattern Recognition Letters*, vol. 27, no. 15, pp. 1835–1842, 2006.

[14] B. D. Lucas, T. Kanade, *et al.*, "An iterative image registration technique with an application to stereo vision." in *IJCAI*, vol. 81, 1981, pp. 674–679.