A framework for the recognition of high-level surgical tasks from video images for cataract surgeries.

Florent Lalys, Laurent Riffaud, David Bouget, Pierre Jannin

A framework for the recognition of high-level surgical tasks from video images for cataract surgeries

F. Lalys, L. Riffaud, D. Bouget, and P. Jannin

***Abstract*— The need for a better integration of the new generation of Computer-Assisted-Surgical (CAS) systems has been recently emphasized. One necessity to achieve this objective is to retrieve data from the Operating Room (OR) with different sensors, then to derive models from these data. Recently, the use of videos from cameras in the OR has demonstrated its efficiency. In this paper, we propose a framework to assist in the development of systems for the automatic recognition of high level surgical tasks using microscope videos analysis. We validated its use on cataract procedures. The idea is to combine state-of-the-art computer vision techniques with time series analysis. The first step of the framework consisted in the definition of several visual cues for extracting semantic information, therefore characterizing each frame of the video. Five different pieces of image-based classifiers were therefore implemented. A step of pupil segmentation was also applied for dedicated visual cue detection. Time series classification algorithms were then applied to model time-varying data. Dynamic Time Warping (DTW) and Hidden Markov Models (HMM) were tested. This association combined the advantages of all methods for better understanding of the problem. The framework was finally validated through various studies. Six binary visual cues were chosen along with 12 phases to detect, obtaining accuracies of 94%.**

***Index Terms*— Surgical workflow, surgical microscope, feature extraction, video analysis, surgical process model, DTW, HMM**

I. Introduction

Over the past few years, the increased availability of sensor devices has transformed the Operating Room (OR) into a rich and complex environment. With this technological emergence, the need of new tools for better resource support and surgical assessment increases [1]. In parallel, better management of all devices, along with improved safety is increasingly necessary. In order to better apprehend theses new challenges in the context of CAS systems and image-guided surgery, recent efforts have been made on the creation of context-aware systems. The idea is to understand the current situation, and automatically adapt the assistance functions accordingly. Being able to automatically extract information from the OR such as events, steps or even adverse events would allow the management of context-aware systems, but also the post-operative generation of reports or the evaluation of surgeons/surgical tools use. The goal of CAS systems based on surgery modelling is to retrieve low-level information from the OR and then to automatically extract high-level tasks from these data [2]. Consequently, it would be beneficial for surgeons performing the surgery, allowing better systems management, automatically reporting procedures, evaluating surgeons, increasing surgical efficiency and quality of care in the OR. This is why, in the context of CAS systems, the automatic extraction of high-level tasks in the OR has recently emerged.

To design such models, there are real advantages to automating the data extraction process, mainly because manual work is time-consuming and can be affected by human bias. Automatic data extraction is now easier, thanks to the high number of sensor devices. Among all sensors, teams have recently focused on videos from cameras already employed during the procedure, such as endoscopes or microscopes, which are a rich source of information. Video images provide information that can be processed by image-based analysis techniques, and then fused with other data from different sensors for creating models capturing all levels of surgery granularity. Moreover, computer vision techniques provide complex processing algorithms to transform images and videos into a new representation that can be further used for machine learning techniques through supervised or non-supervised classification. Compared to other data extraction techniques, the use of video not only eliminates the need to install additional material in the OR, but it is also a source of information that does not need to be controlled by humans, thus automating the assistance provided to surgeons, without altering the surgical routine.

Manuscript received July 18, 2011.

F. Lalys, L. Riffaud, D. Bouget and P. Jannin are with the U746 INSERM Institute and University of Rennes I, Faculté de Médecine, CS 34317, F-35043, Rennes Cedex France (phone: +33223233829; e-mail: florent.lalys@irisa.fr, david.bouget@irisa.fr, pierre.jannin@irisa.fr).

L. Riffaud is with the neurosurgical department of the University Hospital of Rennes, 35000 Rennes, France, (e-mail: laurent.riffaud@chu-rennes.fr).

Current work using surgical videos has made progress in classifying and automating the recognition of high-level tasks in the OR. Four distinct types of input data associated with their related applications have recently been investigated. Firstly, the use of external OR videos has been tested. Bhatia et al. [3] analysed overall OR view videos. After identifying 4 states of a common surgical procedure, relevant image features were extracted and HMMs were trained to detect OR occupancy. Padoy et al. [4] also used low-level image features through 3D motion flows combined with hierarchical HMMs to recognize on-line surgical phases. Secondly, the use of endoscope videos in Minimally Invasive Surgery (MIS) has been investigated. The main constraints in MIS range from the lack of 3D vision to the limited feedback. However, studies on the subject have recently shown that the use of videos in this context was relevant. Speidel et al. [5] focused on surgical assistance for the construction of context-aware systems. Their analysis was based on augmented reality and computer vision techniques. They identified two scenarios within the recognition process: one for recognizing risk situations and one for selecting adequate images for the visualisation system. Lo et al. [6] used vision and particularly visual cues to segment the surgical episode. They used colour segmentation, shape-from-shading techniques and optical flows for instrument-tracking. These features, combined with other low-level visual cues, were integrated into a Bayesian framework for classification. Klank et al. [7] extracted image features for further scene analysis and frame classification. A crossover combination was used for selecting features, while Support Vector Machines (SVMs) were used for the supervised classification process. Also in the context of endoscopic interventions, Blum et al. [8] automatically segmented the surgery into phases. A Canonical Correlation Analysis was applied based on tool usage to reduce the feature space, and resulting feature vectors were modelled using Dynamic Time Warping (DTW) and Hidden Markov Model (HMM). Thirdly, also based on videos but in the context of robotic assistance, with the Da Vinci robot, Voros and Hager [9] used kinematic and visual features to classify tool/tissue interactions in real-time. Similarly, Reiley and Hager [10] focused on the detection of subtasks for surgical skill assessment. Finally, our first work [11] proposed the extraction of surgical phases with microscope videos and validated it in the context of neurosurgical procedures. In this paper, we extend this approach by proposing aframework that can be adjusted (subject to further researches) to any type of surgery. The objective is to automatically detect surgical phases from microscope videos. Surgical procedures can be decomposed into four main levels of granularity: phases, steps, tasks and motions. Surgical phases are thus defined as sequences of steps performed by the surgeon at a high level of granularity. The idea of the framework is first to manually defined visual cues that can be helpful for discriminating high-level tasks. The visual cues are automatically detected by state-of-the-art image-based classifiers, obtaining a semantic signature for each frame. These time series are then aligned with a reference surgery using the DTW algorithm to recognize surgical phases. Compared to traditional video understanding algorithms, this framework extracts generic application-dependant visual cues. The combination of image-based analysis and time series classification enables high recognition rates to be achieved. We evaluated our framework with a dataset of cataract surgery videos through various cross-validation studies, and compared results of the DTW approach to the HMM classification.

II. MATHERIALS AND METHODS

A. Application-dependant visual cues

The proposed framework (Fig. 1) was created to be adapted, if needed, to other types of surgery. Therefore, five sub-systems based on different image processing tools were implemented. Each of these sub-systems is related to one type of visual cue: visual cues recognizable through colour were detected with simple histogram intersection. For shape-oriented visual cues such as object recognition, a Haar classifier was trained. For texture-oriented visual cues, we used a bag-of-words approach using local descriptors, and finally for all other visual cues we used a conventional image classification approach including a feature extraction process, a feature selection process and a supervised classification with SVM. In all cases, the features were considered to be representative of the appearance of the cues to be recognized.

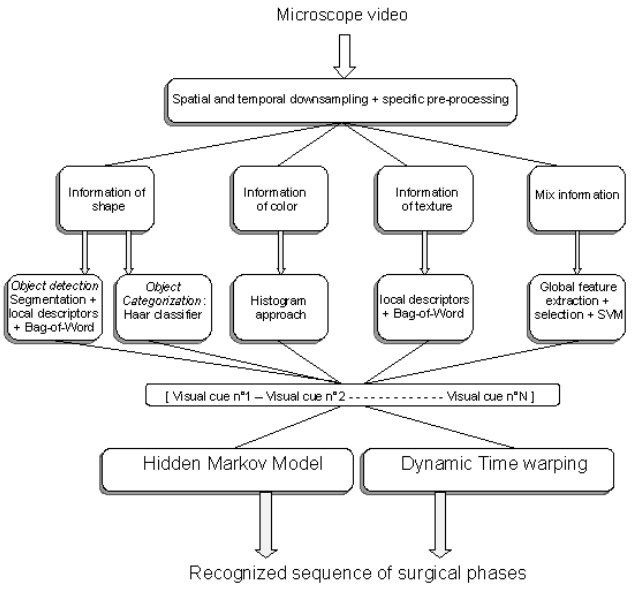


Fig. 1. Framework of the recognition system.

Preliminary pupil segmentation step

In the context of cataract surgery, some visual cues are identifiable only inside the pupil. The regions around the

pupil may therefore bias the detection and a preliminary segmentation step is needed to ensure good detection. Detecting this Region Of Interest (ROI) will allow the retrieval of more specific visual cues, consequently improving surgical process analysis and modelling. The pupil segmentation procedure can be divided into three steps and is based on the colour difference between the pupil and the remaining eye regions. The first step allows the creation of an outline mask from the input image (Fig. 2-a) transformed into the YUV colour space. Within this first step, smoothing, thresholding and morphological operations were performed to the input image, obtaining a mask (Fig. 2-b). Using the mask, the second step consists in determining circles through the image using the Hough [12] transform (Fig. 2-c). The third step can be considered as a verification step. Incomplete circle outlines in the image mask may occur, leading to Hough circle detection failure. In order to tackle this problem, an iterative search is performed on the mask to identify the most probable circular zone, based on both the counting of corresponding pixels and circle diameters. Following this procedure, the ROI around the patient pupil can be retrieved (Fig. 2-d).

Fig. 2. Different steps of the pupil segmentation. a) input image, b) 1st step: creation of the mask, c) 2nd step: Hough transform computation, d) 3rd step: final segmentation of the pupil.

Colour-oriented visual cues

Colour is one of the primary visual features used to represent and compare visual content [13]. In particular, colour histograms have a long history as a method for image description, and can also be used for identifying colour shade through images. Here we used the principle of histogram intersection to extract colour-oriented visual cues, by creating a training image database composed of positive and negative images. Two complementary colour spaces [14] were extracted: RGB space (3 x 16 bins) along with Hue (32 bins) and Saturation (32 bins) from HSV space. For classifying visual cues, we used a KNN classifier with the correlation distance to compare histograms composed of feature vectors.

Texture-oriented visual cues

For whole-image categorization tasks, bag-of-visual-words (BVW) representations, which represent an image as an orderless collection of local features, have recently demonstrated impressive levels of performance along with relative simplicity of use. The idea of BVW is to treat images as loose collections of independent patches, sampling a representative set of patches from the image, evaluating a visual descriptor vector for each patch independently, and using the resulting distribution of samples in descriptor space as a characterization of the image. A bag of keypoints is expressed as a histogram recounting the number of occurrences of each particular pattern in an image. Given the occurrence histograms of positive and negative regions of a training database, a classifier can be trained. Considering the objective of getting binary visual cues, we used a SVM classifier with a Gaussian kernel.

In order to find these keypoints, we tested 4 keypoints detection methods: SIFT [15], SURF [16], Harris [17] and STAR [18], providing access to local image information. All of them provided a similar result, which is a sample of keypoints, though they differed radically in the methods used to obtain them and by the nature of the keypoints found. After detection, a keypoint is then described as a local, rectangular or circular, patch of the image and is represented in a formal way. They are thus represented by a descriptor vector whose length is variable and highly correlated to the chosen descriptor method. Two main descriptors are generally used: SIFT and SURF descriptors. In this study, we focused on SURF descriptors, for computational reasons. Indeed, the vector space dimension is reduced by a half (from 128 to 64) when switching from SIFT to SURF descriptors.

Shape-oriented visual cues

The presence of instruments in the surgical layout is a vital piece of information to access a lower granularity-level in surgical process analysis. The main limitation is that instruments frequently have similar shapes and are therefore very difficult to recognize through image-based analysis only. Two methods were thus implemented: one for recognizing and categorizing a specific instrument (i.e. “specific instrument categorization”), and one for detecting the presence of a surgical instrument without being able to categorize it (i.e. “Detection of other instruments”).

Specific instrument categorization: For this we used a Viola-Jones object detection framework [19]. It is mainly used to detect specific objects in an image, such as human faces. We chose this approach for computational reasons, getting a robust method that minimizes computation time while achieving high detection accuracy. The basic idea is to create a feature-based classifier based on features selected by AdaBoost [20]. AdaBoost is a method of combining weighted weak learners to generate a strong learner. Here, weak learners of the algorithm are based on the Haar-like rectangular features [21]. It is based on comparing the sum of intensities in adjacent regions inside a detection window. Strong learners are then arranged in a classifier cascade tree in order of complexity. The cascade classifier is therefore composed of stages each containing a strong learner. The strong learners, computed during the learning stage, are optimized based on negative images composed of background, and positive images of the object. During the detection phase, a window looks through the

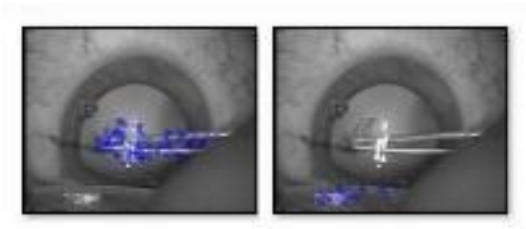
image with different scales and positions for detecting the object. The idea is to determine, at each stage of the cascade, whether the given sub-window is not or may be the searched object. This enables a high number of sub-windows to be very rapidly rejected. As a result, the false positive rate and the detection rate are the product of each rate of the different stages. This algorithm is known to work well for rigid objects. We applied it in the case of surgical instrument categorization.

Detection of other instruments: This method was meant to be able to automatically detect any instrument appearing in the microscope's field of view. For this detection we used a pixel-based approach composed of 3 steps: segmentation, description and classification. For the segmentation step, the goal was to create as many ROIs as instruments in the image. The better the ROIs around the instruments, the better the identification will be. Pre-processing operations were first performed to create a mask composed of all strong outlines. This approach is based upon the fact that there is a distinct colour difference between instruments and the background. A Gaussian blur transform, a Laplacian transform, along with threshold and dilatation operations were applied to the input image (Fig. 3-a) to create the mask. By applying a connected component method to the mask, we were able to detect and remove all small connected components which were assumed to be noise. The final mask is then clean, and only contains strong outlines (Fig. 3-b). Lastly, we retrieved the two largest remaining connected components respectively (no more than 2 instruments can be present at a same time within the surgical scene), and created a mask for each one. At this stage, these selected connected components are very likely to be the instruments. By applying these masks to the input image, we obtained two different images, each one with only a ROI of the input image (Fig. 3-c, Fig. 3-d).

a) b) c) d)

Fig. 3. Different stages of the segmentation step for the detection of instruments a) the input image, b) the clean mask, c) the region of interest corresponding to the first connected component, d) the ROI corresponding to the second connected component.

For the description and the segmentation step, the aim was to provide a robust and reproducible method for describing the ROIs that have been isolated by the segmentation step, and then to identify the ROI as an instrument or not using the descriptors. Given the fact that the goal was to detect instruments, we were willing to extract scale and rotation invariant features. We used the same approach as that described in subsection "Texture-oriented visual cues" in II.A, with local description of features using a BVW approach, and finally a supervised classification using SVM. Fig. 4-a and Fig. 4-b show an example of the extraction of local features for the input image of Fig. 3-a. If the same instrument appears with various scales and orientations, we will be able to extract the same feature points with the same descriptors. Using this approach, ROIs can be classified as belonging to an instrument or not.



a) b)

Fig. 4. SURF features detected on image from Fig. 3-a., and shown as blue circles. a) SURF points on the first connected component, b) SURF points on the second connected component.

Alternative method

This approach can be seen as conventional image classification. Each frame is represented by a signature composed of global spatial features. The RGB and the HSV spaces, the co-occurrence matrix along with Haralick descriptors [22], the spatial moments [23], and the Discrete Cosine Transform (DCT) [24] coefficients were all computed. Each signature was finally composed of 185 complementary features, that had to be reduced by feature selection. We combined a filter and a wrapper approach [25] using the method described by Mak and Kung [26] using the union of both results for improving the final selection. The Recursive Feature Elimination (RFE) SVM [27] was chosen for the wrapper method and the mutual information (MI) [28] was chosen for the filter method. Finally, 40 features were kept, and a SVM was applied to extract the desired visual cue. This particular procedure has been presented and validated in Lalys et al. [29].

Once every visual cue of a particular surgery has been defined and detected, creating a semantic signature composed of binary values for each frame, the sequences of frame signatures (i.e. the time series) must be classified using appropriate methods. Two different approaches were tested here: the HMM modelling and the classification by DTW alignment.

Hidden Markov Model

HMMs [30] are statistical models used for modelling non-stationary vector times-series. An HMM is formally defined by a five-tuple (S, O, Π, A, B), where $S = \{s_1 \ldots s_N\}$ is a finite set of N states, $O = \{o_1 \ldots o_M\}$ is a set of M symbols in a vocabulary, $\Pi = \{\pi_i\}$ are the initial state probabilities, $A = \{a_{ij}\}$

the state transition probabilities and $B = b_i(o_t)$ the output probabilities. Here, outputs of the supervised classification were treated as observations for the HMM. States were represented by the surgical phases that generated a left-right HMM (Fig. 5.). Transition probabilities from one state to its consecutive state were computed for each training video, and then averaged. If we set one probability to α, the probability of remaining in the same state is then $1-\alpha$. Output probabilities were computed as the probability of making an observation in a specific state. Training videos were applied to the supervised classification for extracting binary cues and output probabilities were obtained by manually counting the number of occurrences for each state. Then, given the observations and the HMM structure, the Viterbi algorithm [31] identifies the most likely sequence of states. HMM training, like feature selection process, must be performed once only for each learning database.

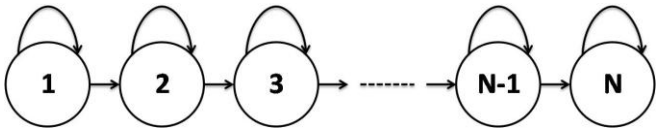


Fig. 5. Left-right HMM, where each state corresponds to one surgical phase

Dynamic Time Warping

We used the DTW algorithm [32] as a method to classify the image sequences in a supervised manner. DTW is a well-known algorithm used in many areas (e.g. handwriting and online signature matching, gesture recognition, data mining, time series clustering and signal processing). The aim of DTW is to compare two sequences X := (x1 , x2 , . . . , xN) of length N and Y := (y1 , y2 , . . . , yM) of length M. These sequences may be discrete signals (time-series) or, more generally, feature sequences sampled at equidistant points in time. To compare two different features, one needs a local cost measure, sometimes referred to as local distance measure, which is defined as a function. Frequently, it is simply defined by the Euclidean distance. In our case however, considering that we are using binary vectors, we will use the Hamming distance, well-adapted for this type of signature. In other words, the DTW algorithm finds an optimal match between two sequences of feature vectors which allows for stretched and compressed sections of the sequence. To compare each surgery, we created an average surgery based on the learning dataset using the method described by Wang and Gasser [33]. Each "query" surgery is first processed in order to extract semantic information, and then the sequence of image signature is introduced in the DTW algorithm to be compared to the average surgery. Once warped, the phases of the average surgery are transposed to the unknown surgery in a supervised way. Additionally, global constraints (also known as windowing functions) can be added to the conventional algorithm in order to constrain the indices of the warping path. With this method, the path is not allowed to fall within the constraints window. For surgery modelling, we used the Itakura parallelogram [34]. This prevents the warping path from straying too far away from the diagonal path.

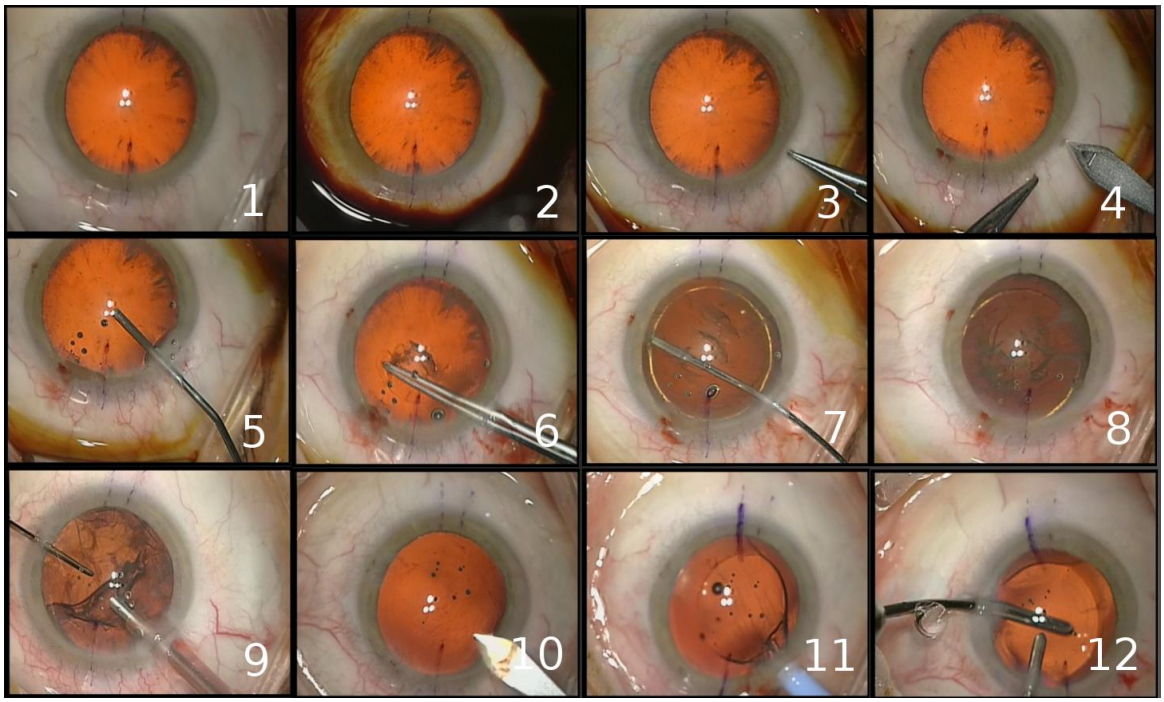


Fig. 6. Typical digital microscope frames for the 12 surgical phases: 1-preparation, 2-betadine injection, 3-lateral corneal incision, 4-principal corneal incision, 5-viscoelastic injection, 6-capsulorhexis, 7-phacoemulsification, 8-cortical aspiration of the big pieces of the lens,

9- cortical aspiration of the remanescent lens, 10-expansion of the principal incision, 11-implantation of the artificial IOL, 12- adjustment of the IOL+ wound sealing

B. Data

Video data-set

Our framework was evaluated on cataract surgeries (eye surgeries). The principle of cataract surgery is to remove the natural lens of the eye and insert an artificial one (referred as an IntraOcular Lens, or IOL) in order to restore the lens' transparency. In this project, twenty cataract surgeries from the University Hospital of Munich were included (with mean surgical time of 15 min). Three different surgeons performed these procedures. All videos were recorded using the OPMI Lumera surgical microscope (Carl Zeiss) with an initial resolution of 720 x 576 at 25 fps. Considering the goal of recognizing only high-level tasks surgical tasks, each video was down-sampled to 1 fps. Original frames were also spatially down-sampled by a factor 4 with a 5-by-5 Gaussian kernel. Twelve surgical phases were defined (Fig. 6.).

Definition of the visual cues

Six pieces of binary visual cues were chosen for discriminating the surgical phases. The pupil colour range, defined as being orange or black, was extracted using a preliminary segmentation of the pupil along with a colour histogram analysis. Also analysing only the pupil after the segmentation step, the global aspect of the cataract (defined as parcelled out or not) was recognized using the BVW approach with local spatial descriptors. The presence of antiseptic, recognizable by virtue of its specific colour, was detected using colour histogram analysis, but on the entire image. Concerning the detection of surgical instruments, only one had a characteristic shape, the knife. We trained a Haar classifier using 2000 negative images and 500 positive images for its detection. All other instruments have very similar shapes and are very difficult to categorize. For this reason, we chose to detect the presence of an instrument as a particular visual cue. Lastly, the IOL instrument was not readily identifiable through only colour or shape analysis and we chose a classical approach using many spatial features along with a SVM classification to detect this visual cue.

C. Validation studies

Initial indexing was performed by surgeons for each video, by defining the phase's transitions, along with all visual cues for each video. With this labelled video database, we were able to evaluate both aspects of our framework, i.e. detection of the different visual cues and the global recognition rate of the entire framework. From each video, we randomly extracted 100 frames in order to create the image database, finally composed of 2000 labelled images, each being associated to one video. This image database was used for the assessment of visual cue detection, whereas the videos, along with their corresponding frames from the image database, were used to assess the entire framework.

One step of our procedure didn't require any training stage: the preliminary step of pupil segmentation. This step was simply validated over the entire video database by testing each frame of each video. During this validation, a pupil was considered correctly segmented if and only if the circle found by the Hough transform precisely matched the pupil. A percentage was then obtained corresponding to the accuracy of the segmentation.

The second aspect of our framework that was assessed was the recognition of all visual cues. Independently, for knife recognition, the training stage was performed using manually selected positive and negative images for better object training. For the validation, 1000 test images were used and global recognition accuracy was computed. The four other visual cues classifiers were the assessed through 10-fold cross-validation studies. The image database was therefore divided into 10 subsets, randomly selected from the 20 videos. Nine were used for training while the prediction was made on the 10^{th} subset. One test subset was consequently composed of frames from two videos, while the validation sets was composed of the rest of the frames from the 18 other videos. This procedure was repeated 10 times and results were averaged. After evaluating each classifier, we validated their use compared to a traditional image-based classifier, i.e. compared to feature extraction, selection and classification as performed by the conventional classifier. Additionally, we also compared the four local keypoint detectors presented in subsection “Texture-oriented visual cues” in II.A, along with the optimal number of visual words, for both the texture-oriented classifier and the detection of instrument presence. This validation, performed under the same conditions as the detection of the visual cues, was necessary to optimize the detection of texture and shape-oriented information by the BVW approach.

Lastly, we evaluated the global framework, including visual cue recognition and the time series analysis with the same type of cross-validation. Similarly to the visual cues recognition, at each stage, 18 videos (and their corresponding frames from the image database) were used for training and recognitions were made on the 2 others. For this assessment, the criterion chosen was the Frequency Recognition Rate (FRR), defined as the percentage of frames correctly recognized over a video by the recognition framework.

III. RESULTS

Taking all frames from each video (at 1 fps), the pupil was correctly extracted with an accuracy of 95% (Tab. 1.). The worse video was very difficult to segment, with 78% of all frames correctly segmented. The best video, on the other hand, had almost its entire frame correctly segmented (99%).

Table 1. Mean accuracy, minimum and maximum of the segmentation of the pupil over the entire video database.

	Accuracy (Std)	Minimum	Maximum
Detection	95,00 (6)%	78,00%	99,00%

Table 2. Parameters of the classification algorithms used for extracting visual cues.

Type of visual cues	Algorithm	Parameters
Color-oriented	Color histogram intersection	Type of color space: RGB, HSV Classifier: KNN Distance: correlation
Texture-oriented	BVW approach	Classifier: SVM with Gaussian kernel Interest points detectors: SIFT Feature representation: SURF Codebook generation: KNN
Instrument categorization	Viola-Jones approach	Features: Haar-like rectangular Negative images: 2000 Positive images: 500
Detection of other instruments	BVW approach	Classifier: SVM with Gaussian kernel Interest points detectors: SURF Feature representation: SURF Codebook generation: KNN
Alternative method	Global features classification	Spatial features: RGB, HSV spaces, Haralick descriptors, DCT, spatial moments Wrapper method: RFE-SVM Filter method: MI Classifier: SVM with Gaussian kernel

Fig. 7-a,b. shows the BVW study for choosing the best parameters for both the detection of instrument presence and the texture-oriented classifier respectively. Surprisingly, for both figures, the number of visual words did not appear to be a major parameter to be enhanced. Indeed, the accuracy didn't vary significantly from 1 to 60 visual words, and this result was true for the 4 keypoint detectors and the two BVW studies. For both studies, the best accuracy was still obtained for a number of visual words equal to 12. On the contrary, the influence of keypoint detectors was significant. For the detection of instruments presence (Fig. 7-a), the SURF keypoint detector showed best recognition accuracies (with 12 visual words: 86%), whereas for the detection of the cataract aspect, the SIFT keypoint detectors shows best results (with 12 visual words: 83%). Tab. 2. gives all parameter values used in the 6 detection methods of visual cues.

The results of the cross-validation study for the recognition of all visual cues (Tab. 3.) showed that very good detection accuracies were obtained, which outperform the standard classifier in all cases. The best recognition was obtained for the presence of the Knife (Haar classifier), achieving a recognition rate of 96.7%, whereas the lower rate was obtained for the recognition of the instrument presence. Similarly, the other visual cue detected using a BVW approach (the aspect of the cataract) has not a very high accuracy (87.2). Histogram approaches shows good results (96.2% for the pupil colour range detection and 96.1% for the antiseptic detection), whereas the IOL instrument has also a good recognition rate of 94.6%, even detected with the conventional classifier.

Table 3. Mean accuracy (standard deviation) for the recognition of the 6 binary visual cues.

	Pupil colour range	Presence antiseptic	Presence Knife	Presence IOL instrument	Cataract aspect	Presence instrument
Specific image-based classifier (%)	96.2 (3.6)	96.1 (0.7)	96.7 (3.4)	94.6 (1.1)	87.2 (5.4)	84.1 (8.6)
Classical approach (%)	94.1 (4.6)	95.6 (0.4)	88.5 (4.3)	X	54.1 (3.6)	58.7 (6.1)

Table 4. Mean, minimum and maximum FER of the DTW

	Accuracy (Std)	Minimum	Maximum
HMM (%)	91.4 (6.4)	80.5	99.7
DTW (%)	94.4 (3.1)	90.6	96.4

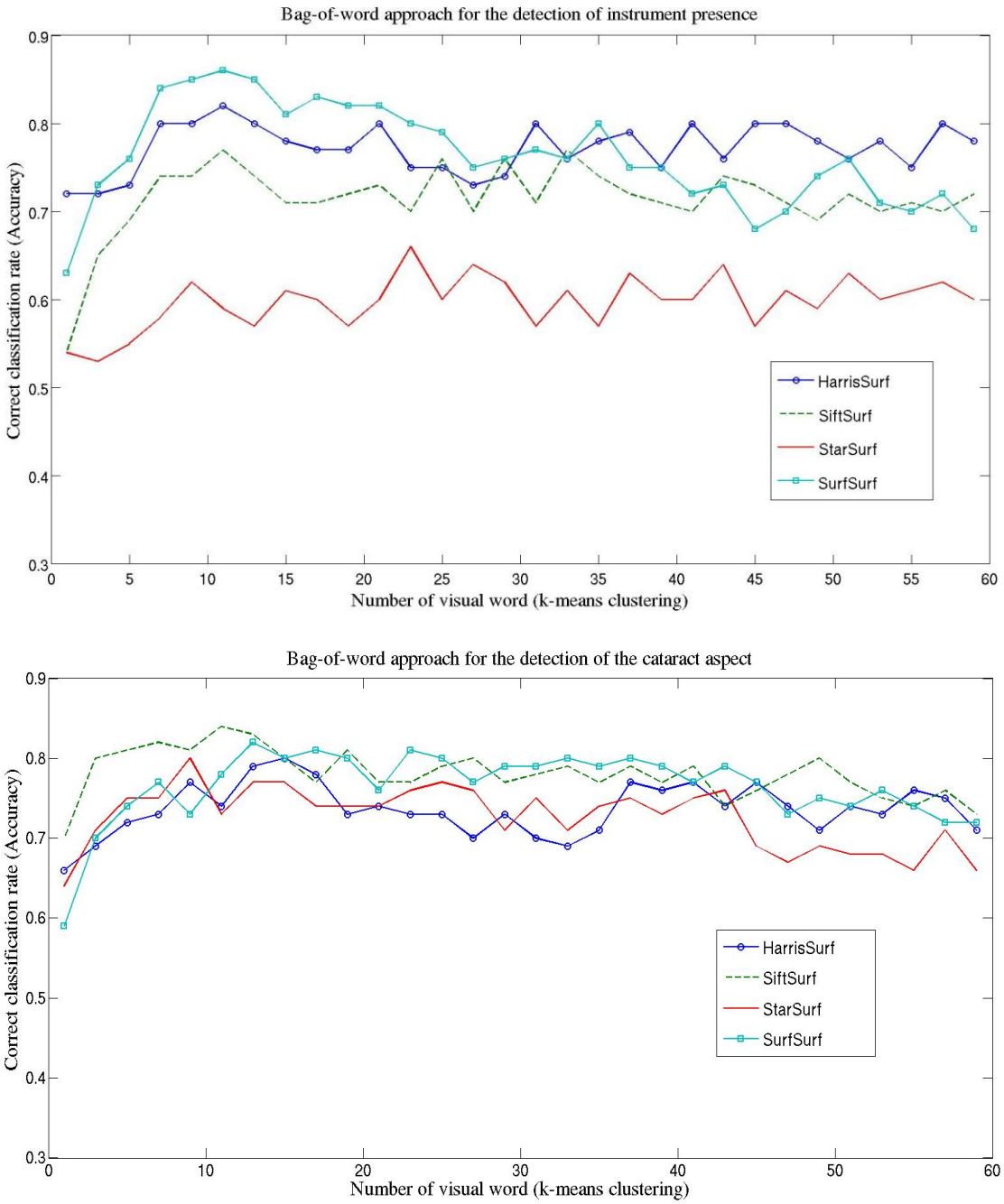


Fig. 7. BVW validation studies comparison of accuracies with different number of visual words and different keypoints detectors: a) Detection of the instruments presence, b) Recognition of the cataract aspect.

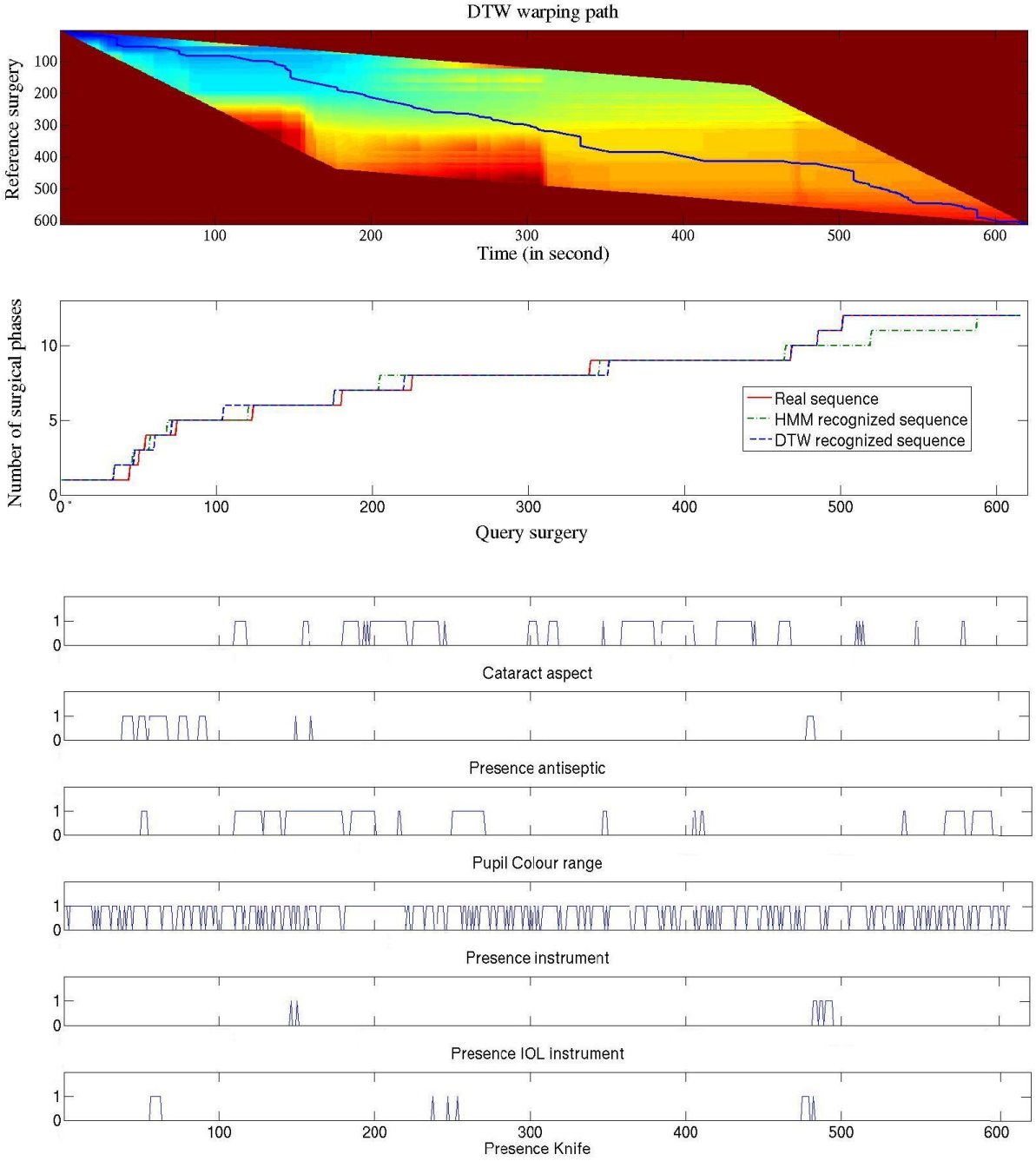


Fig. 8. Distance map of two surgeries and dedicated warping path using the Itakura constraint (up), along with transposition of the surgical phases (middle), and the visual cues detected by the system (down).

From Tab. 4., the time series study showed better results using the DTW approach than with HMM classification. With HMM, a mean FRR of 91.4% was obtained with a high maximum of 99.7%, and a low minimum of 80.5%. The DTW approach shows a mean FRR of 94.4% with quite high minimum (90.6%) and particularly low maximum (96.4) compared to its mean FRR.

Fig. 8. shows an example of video recognized by the system (by HMM classification and with the DTW approach respectively) along with the warping path from the DTW approach. On this particular recognized video, DTW shows the best recognition compared to the HMM classification.

IV. DISCUSSION

In this paper, we proposed a framework that automatically recognizes surgical phases of cataract surgery. The recognition relies on data provided by microscope videos, which is a novel way of addressing situation recognition issues. Twelve phases were defined by surgeons, and their recognition was based on the detection of 6 visual cues, along with time series analysis. This combined approach allowed a high degree of automatic

recognition system accuracy, achieving accuracies of approximately 94% using DTW modelling.

A. Adaptation to other surgical procedures

The detection framework is based on the recognition of visual cues within the surgical scene. The recognition of such information allows the modelling of the surgical procedure and final surgical phase recognition. Due to the varying facilities between departments, the numbers of phases, as well as the colours, tools and shapes will differ. Consequently, considering that surgical environments are different in each hospital, one recognition system should be created for each department. The adopted solution was to create a framework using specific image-based sub-systems in order to be as generic as possible, and to provide as many tools as possible for being exhaustive. This way, our method addresses the issue of system adaptability.

Even though the framework was created to be adaptable, each kind of surgical environment has its own particularities and characteristics. This is why preliminary pre-processing steps may be mandatory in order to tune the recognition system according to the type of surgery. For instance, in the case of low-resolution video images, the purpose would be to improve image quality for further processing. In the context of cataract surgery, the microscope field of view is precisely delineated, thus enabling the use of a preliminary step of segmentation to restrict the search for a specific visual cue within a ROI defined by the pupil outlines. This is the only step that is specific to cataract surgery. For the adaptation to other surgical procedures, this segmentation step could be either adapted or even removed. Taking as example neurosurgical procedures and specifically hypophyse surgeries that we used in our previous work [11], segmentation would not be necessary as the field-of-view is already zoomed and adapted for image-based analysis and visual cues extraction. Pre-processing steps for image quality enhancement would also not be required because of the high-resolution of neurosurgical microscopes, neither intensity corrections nor specular reflection removal. This example would be true for this specific type of easy and reproducible surgical procedures. However, other adaptations could be conducted. Dealing with more complex surgeries would involves further researches on the pre-processing step, on the segmentation of surgical tools before their categorizations and possibly on the definition of other sub-systems for the detection of visual cues. We proposed in this paper a complete framework that we tested on one surgical procedure, but the evolution to other surgical procedures should be experimented.

Once the framework has been tuned a dedicated surgical procedure, its use is fully automatic and will work with any microscope video of this type of surgery in its environment. Similarly, other variability factors may affect recognition, such as the manner in which surgeons operate. With this system, a training stage is necessary for each surgical department, assuming that the surgeons within the department use identical materials and follow the same sequence of phases during the procedure, making image features invariant to task distortion.

B. Pupil segmentation

Using an adapted method composed of image-based analysis, the segmentation of the pupil returns highly accurate results. For 95% of the frames, the ROI correctly contains the entire pupil. Moreover, to avoid distorting any further detection that could be done within the pupil we decided to define a constant circumference value. Thus, each time a region of interest is detected, the centre is kept and the circumference value is reset to the default value. Due to its high accuracy over the entire video database, it allows all potential colour-associated noise to be removed from around the pupil for further recognition. The very low accuracy obtained for one video can be explained by the presence of the retractors, rendering the field of view very narrow. Automatic segmentation turns out to be difficult when the retractors, or even the surgical instruments, around the eye, occupy too much space within the field of view.

As a drawback, our approach, which always returns a ROI, is not always perfectly centred on the middle of the pupil. We can explain this issue by the fact that the pupil is not always completely inside the microscope's field of view. Sometimes the pupil outlines are too distorted due to surgical tools or the surgery itself. Sometimes the retractor is as wide as the pupil and sometimes the surgeon's fingers are in the field of view. In that case, it's difficult to extract the exact position of the pupil and its outlines and to adjust an intensity threshold accordingly. If the surgical microscope had a permanent position, or if we could precisely estimate the position of the pupil in each image, it would be possible to adjust a threshold for the segmentation automatically.

C. Application-dependant visual cues

Before the visual cue recognition training stage, the user will need to choose the visual cues and the associated image-based recognition classifier. In image classification problems, users usually do not think in terms of low-level features, resulting in poor recognition of the high-level semantic content of the images. Here, during the stage of visual cue definition, the colour, texture and shape behaviour of the visual cues are often intuitively known, allowing the most effective classifiers to be chosen. When visual cue is unknown or undocumented, the solution proposed is to choose the generic approach, integrating a large number of image features. This approach, combining global spatial features and SVM, may therefore be adapted to the recognition of any type of cue. The feature selection step allows the user to select discriminatory features and remove unsuitable ones, which is the intended objective. To improve recognition, however, the three other specific

classifiers seem to be well-adapted when the behaviour of the visual cue is well perceived.

Generally, the main drawback of a global colour histogram representation is that information concerning object shape, and texture is discarded. In our case, however, it is only used for colour-related visual cue detection. Similarly, the main drawback of shape-based approaches is the lack of well-defined outlines. The Haar classifier was used in our framework for specific objects only, e.g. the knife, which is significantly different from all others instruments used in cataract surgery. The use of this approach to categorize other instruments, such as the cannula, was tested, but gave very poor results due to the narrow field of view and the difficulty in discriminating that specific instrument from the others. For this reason, we chose to use a second approach for object recognition, allowing the system to gain an information concerning object presence, without categorizing it. This type of information was still relevant for phase detection and allowed complete image signatures to be achieved using other information with a different level of granularity. The use of a BVW approach combined with local descriptors was also validated. Local descriptor comparisons (Fig. 5.) enabled selection of the most appropriate features, and application with the recognition of the global aspect of the cataract gave very promising results.

With the exception of the Haar classifier, the three other classifiers are all based on a training image database. The power of discrimination of the image database is thus vital. We can easily imagine that accuracy may decrease sharply if the images do not efficiently represent all phases or all scene possibilities within the phases. Additionally, the training stage is time-consuming and requires human efforts. In our particular case, the best method, used here in our validation studies, was to annotate surgical videos before randomising the initial sample. The randomisation process is thus no longer performed on all frames, but on each video independently, extracting the same number of frames per video.

D. Time series analysis

Combined with state-of-the-art computer vision techniques, time series analysis displayed very good performance, opening the way for further promising work on high-level task recognition in surgery. Without this step of time series analysis, results are less accurate (~80%). This can be explained because some visual cues don’t appear only during one particular phase, and the information of sequentiality is needed. For instance, the knife always appears twice during cataract surgery: once during phase n°4 (principal corneal incision), and once during phase n°10 (expansion of the principal incision). All other visual cues are not present during these two phases. The discrimination of both phases appears to be possible with an information of time only that the HMM or the DTW can bring.

In particular, DTW captures the sequentiality of surgical phases and is well-adapted to this type of detection. The cost function between 2 surgical procedures with the same sequence of phases, but with phase time differences, will be very low. The advantage is that it can accurately synchronize two surgical procedures by maximally reducing time differences. The main limitation concerning the use of DTW, however, can be seen phase sequence differences appear between two surgeries. The warping path would not correctly synchronize the phases and errors would occur. In the context of cataract surgery, the procedure is standardized and reproducible, justifying the very good results of the recognition. But we can imagine that for surgeries that are not completely standardized, DTW would not be adapted. In this case, HMM could be used by adding bridges between the different states of the model (the transition matrix should be adapted in that case), allowing the sequence to be resumed and perform the same phase multiple times. The state machine would not be a left-right structure but would include more complex possibilities with many bridges between states. As a drawback, the complexity of such HMMs could completely affect the recognition accuracy. For each surgical procedure, the HMM structure should be created by minimizing the possibilities of transitions from states to states not to affect the classification phase. In the particular case of cataract surgery, the results showed that the DTW algorithm was, not surprisingly, quite better than HMM. Compared to HMM, the other limitation of the DTW algorithm is that it cannot be used on-line, as the entire procedure is required to determine the optimum path. For on-line applications, the HMM classification should be used.

E. Use in clinical routine

The automatic recognition of surgical phases is very useful for context-awareness applications. For instance, there is a need for an analysis methodology specifying which kind of information needs to be displayed for the surgeon's current task. The use of videos, such as endoscope or microscope videos allows automating the surgeons' assistance without altering the surgical routine. Moreover, it can also support intra-operative decision making by comparing situations with previously recorded or known situations. This would result in a better sequence of activities and improved anticipation of possible adverse events, which would, on the one hand optimize the surgery, and on the other hand improve patient safety. Because of the time-series methods used in the framework that are not fully on-line algorithms, the on-line use of the system remains a long-term objective. In its present form, the computation time of the recognition process for one frame (visual cues detection + DTW/HMM classification) was evaluated to around 3s on a standard 2-Ghz computer. It could be introduced into clinical routine for post-operative video indexation and creation of pre-filled reports. Surgical videos are increasingly used for learning and teaching

purposes, but surgeons often don't use them because of the huge amount of surgical videos. One can imagine a labelled database of videos with full and rapid access to all surgical phases for easy browsing. The video database created would contain the relevant surgical phases of each procedure for easy browsing. We could also imagine the creation of pre-filled reports that would need to be completed by surgeons.

V. CONCLUSION

The goal of context-aware assistance is to collect information from the OR and automatically derive a model that can be used for advancing CAS systems. To provide this type of assistance, the current situation needs to be recognized at specific granularity levels. A new kind of input data is more and more used for detecting such high-level tasks: the videos provided by existing hardware in the OR. In this paper, our analysis is based exclusively on microscope video data. We proposed a recognition system based on application-dependant image-based classifiers and time series analysis, using either an HMM or DTW algorithm. Using this framework, we are now able to recognize the major surgical phases of every new procedure. We have validated this framework with cataract surgeries, where twelve phases were defined by an expert, as long as six visual cues, achieving a global recognition rate of around 94%. This recognition process, using a new type of input data, appears to be a non-negligible progression towards the construction of context-aware surgical systems. In future work, lower-level information, such as the surgeon's gestures, will need to be detected in order to create more robust multi-layer architectures.

REFERENCES

[1] Cleary, K. Chung, H.Y. Mun, S.K.: OR 2020: The operating room of the future. Laparoendoscopic and Advanced Surgical Techniques, vol. 15, no 5, pp. 495-500, 2005.

[2] Jannin, P. and Morandi, X. Surgical models for computer-assisted neurosurgery. Neuroimage, vol. 37, no. 3, pp. 783-91, 2007.

[3] Bhatia, B., Oates, T., Xiao, Y., Hu, P.. Real-time identification of operating room state from video. AAAI, pp. 1761-1766, 2007.

[4] Padoy, N., Blum, T., Feuner, H., Berger, M.O., Navab, N. On-line recognition of surgical activity for monitoring in the operating room. In: Proc's of the 20th Conference on Innovative Applications of Artificial Intelligence, 2008.

[5] Speidel, S., Sudra, G., Senemaud, J., Drentschew, M., Müller-stich, BP., Gun, C., Dillmann, R. Situation modeling and situation recognition for a context-aware augmented reality system. Progression in biomedical optics and imaging, vol. 9, no.1, pp. 35, 2008.

[6] Lo, B., Darzi, A., Yang, G. Episode Classification for the Analysis of Tissue-Instrument Interaction with Multiple Visual Cues. International Conference on Medical Image Computing and Computer-Assisted Intervention, 2003.

[7] Klank, U., Padoy, N., Feussner, H., Navab, N. Automatic feature generation in endoscopic images. Int J Comput Assist Radiol Surg, vol. 3, no. 3-4, pp. 331-339, 2008.

[8] Blum, T., Feussner, H., Navab, N. Modeling and Segmentation of surgical workflow from laparoscopic video. International Conference on Medical Image Computing and Computer-Assisted Intervention, 2010.

[9] Voros, S. and Hager, GD. Towards "real-time" tool-tissue interaction detection in robotically assisted laparoscopy. Biomed Robotics and Biomechatronics, pp. 562-567, 2008.

[10] Reiley, C.E. and Hager, G.D. Decomposition of robotic surgical tasks: an analysis of subtasks and their correlation to skill. M2CAI workshop. MICCAI, London, 2009.

[11] Lalys, F., Riffaud, L., Morandi, X., Jannin, P. Surgical phases detection from microscope videos by combining SVM and HMM. Medical Comp Vision Workshop, MICCAI, Beijing, 2010.

[12] Swain, M. and Ballard, D. Color indexing. International Journal of Computer Vision, vol. 7, no. 1, pp. 11--32, 1991.

[13] Hough, VC. Machine Analysis of Bubble Chamber Pictures. Proc. Int. Conf. High Energy Accelerators and Instrumentation, 1959.

[14] Smeulders, A., Worrin, M., Santini, S., Gupta, A., Jain, R. Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell, vol. 22, no. 12, pp. 1349-1380, 2000.

[15] Lowe, DG. Object recognition from scale-invariant features. ICCV'99,vol. 2, pp. 1150-1157, 1999.

[16] Bay, H., Tuytelaars, T., Van Gool, Luc. SURF: Speeded Up Robust Features. Computer Vision – ECCV, 2006.

[17] Harris, C., Stephens, M. A combined corner and edge detector. Alvey vision conference, 1988.

[18] Agrawal, M. and Konolige, K. CenSurE: Center surround extremas for realtime feature detection and matching. European Conf Comput Vision, ECCV'08, vol. 5305, pp. 102-115, 2008.

[19] Viola, P. and Jones, M. Rapid real-time face detection. IJCV, pp. 137-154, 2004.

[20] Freund, Y., Schapire, RE. A decision-theoretic generalization of on-line learning and an application to boosting. Proc. of the 2nd European conf on Computational Learning Theory, 1995.

[21] Papageorgiou, CP., Oren, M., Poggio, T. A general framework for object detection. Int Conf on Computer Vision, 1998.

[22] Haralick, RM., Shanmugam, K., Dinstein, I. Textural features for image classification. IEEE Trans on Systems, Man, and Cybernetics, vol. 3, no.6, pp. 61-621, 1973.

[23] Hu M. Visual pattern recognition by moment invariants. Trans Inf Theory, vol. 8, no. 2, pp. 79-87, 1962.

[24] Ahmed N, Natarajan T, Rao R. Discrete Cosine Transform. IEEE Trans Comp. pp. 90-93, 1974.

[25] Duda RO and Hart PE. Pattern classification and scene analysis. Wiley. New York, 1973.

[26] Mak, M.W. and Kung, S.Y. Fusion of feature selection methods for pairwise scoring SVM. Neurocomputing, vol. 71, pp. 3104-3113, 2008.

[27] Guyon, I., Weston, J., Barhill, S., Vapnik, V. Gene selection for cancer classification using support vector machine. Machine Learning, vol. 46, pp. 389–422, 2002.

[28] Hamming, R.W. Coding and Information Theory. Prentice-Hall Inc., 1980.

[29] Lalys, F., Riffaud, L., Morandi, X., Jannin, P. Automatic phases recognition in pituitary surgeries by microscope images classification. 1th Int Conf Inform Proc Comp Assist Interv, IPCAI'2010, Geneva, Switzerland, 2010.

[30] Rabiner, LR. A tutorial on Hidden Markov Models and selected applications in speech recognition, Proc of IEEE, vol. 77, no. 2, 1989.

[31] Viterbi, A. Errors bounds for convolutional codes. IEEE TIT, vol. 13, no. 2, pp. 260-269, 1967.

[32] Keogh, EJ., Pazzani, MJ. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. Prediction of the future: AI approaches to time-series problems, pp. 44-51, 1998.

[33] Wang, K. and Gasser, T. Alignment of curves by dynamic time warping. Annals of Statistics, vol. 25, num. 3, pp. 1251-1276, 1997.

[34] Niennattrakkul V. and Ratanamahatana, CA. Learning DTW global constraint for time series classification. Artificial Intelligence papers, 1999.