Surgical Tools Recognition and Pupil Segmentation for Cataract Surgical Process Modeling

David Bouget, Florent Lalys, Pierre Jannin

Surgical Tools Recognition and Pupil Segmentation for Cataract Surgical Process Modeling

David BOUGET[a, b, c, 1] , Florent LALYS[a, b, c], Pierre JANNIN[a, b, c]
[a] *INSERM, U746, Faculté de médecine CS 34317, F-35043 Rennes Cedex, France*
[b] *INRIA, VisAGeS Unité/Projet, F-35042 Rennes, France*
[c] *University of Rennes 1, CNRS, UMR 6074, IRISA, F-35042 Rennes, France*

Abstract. In image-guided surgery, a new generation of Computer-Assisted-Surgical (CAS) systems based on information from the Operating Room (OR) has recently been developed to improve situation awareness in the OR. Our main project is to develop an application-dependant framework able to extract high-level tasks (surgical phases) using microscope videos data only. In this paper, we present two methods: one method to segment the pupil and one to extract and recognize surgical tools. We show how both methods improve the accuracy of the framework for analysis of cataract surgery videos, to detect eight surgical phases.

Keywords. Computer vision, video analysis, Bag-of-words, mathematical morphology.

Introduction

In the context of surgical process modeling, being able to automatically retrieve low-level information from the Operating Room (OR) and then extract high-level tasks from these data is a growing need. In previous works [1,2], authors developed an application-dependant framework automatically able to extract surgical phases from microscope videos. They first extracted visual cues for each video frame using image-based analysis. Each frame is therefore composed of binary information forming a semantic signature. These time series are finally used as input for analysis and classification using either Hidden Markov Model or Dynamic Time Warping approaches. As outputs of these two time series analyses, a sequence of surgical phases is proposed.

In this paper, we present two methods: one method to segment the pupil and one to extract and recognize surgical tools. We show how both methods improve the accuracy of the framework for analysis of cataract surgery videos to detect eight surgical phases. Validation studies were performed with a dataset of twenty cataract surgeries.

[1] Corresponding Author; E-mail: david.bouget@irisa.fr.

1. Materials and Methods

1.1. Surgical Tools Recognition

In order to accurately model a surgical process, a vital information needing to be retrieved is the presence of instruments in the surgical field of view. Detecting and recognizing surgical tools can be relatively complex with image-based analysis because they have quite similar shapes. Moreover, they never appear with the same orientations, scales or under the same illumination. The method developed, based on machine learning (see Figure 1), was automatically able to detect any instrument appearing in the video and is composed of three steps:

- Segmentation
- Description
- Classification

For the segmentation step, the goal was to extract from the image as many regions of interest as surgical tools existing in the image. The better the regions of interest around the tools are, the better the identification will be. This step was based upon the fact there is a distinct color difference surgical tools and the image background. To do so, pre-processing operations were first performed on the input image to create a black and white mask containing all loud outlines. Those operations were respectively: a Gaussian blur transform, a Laplacian transform, along with threshold and dilatation operations. We then refined the mask by applying a connected component method (8 connexity) in order to remove every too small component. Indeed, we can assume that small components can't be outlines of tools and so are considered as noise. From this clean mask, containing only loud outlines, we retrieved the two largest remaining connected components. Indeed, no more than two instruments can be present at a same time within the surgical scene. Lastly, we applied separately each connected component mask extracted on the input image in order to obtain two regions of interest (see Figure 2). Those regions were the most likely to contain a surgical tool.

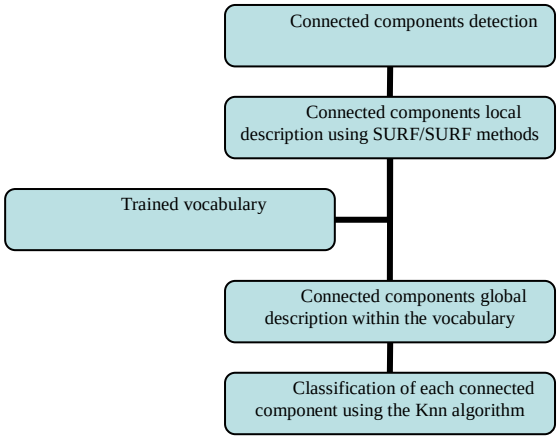

Figure 1. Process pipeline for the surgical tools detection module.

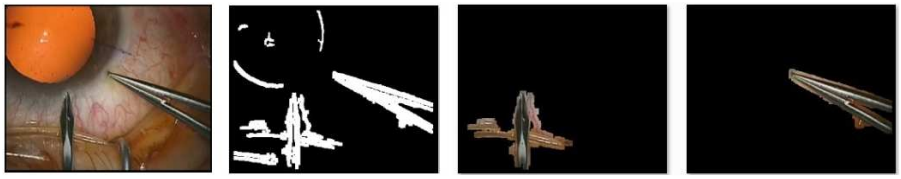

Figure 2. Different detection steps in the tools recognition module. From left to right: input image, mask, final connected components images.

For the description step, the goal was to provide a robust and reproducible way to describe each region of interest previously segmented. Those descriptions were mandatory in order to subsequently perform the classification. Given the fact that the goal of this method was to identify surgical tools, we were willing to extract features invariant to scale, rotation, illumination variations.

In order to create a global description, we first need to obtain locals descriptors (or: key points). We have compared four local features detection methods: Harris [3], SIFT [4], SURF [5], STAR [6]. All of these methods provide a sample of key points, but those key points don't have the same behavior regarding the previously described variations. After detection, a key point (describing a patch of the image) was expressed in a formal way. It is represented by a descriptor vector whose length is variable and depends on the chosen method. Here again, among others, two descriptors methods are generally used: SIFT and SURF descriptors. A SIFT descriptor vector is of length 128 and a SURF descriptor vector is of length 64. In the study, we have decided to use SURF key points as long as SURF descriptors because of the good trade-off between accuracy and computing time.

Using those local descriptors, a global descriptor for a region of interest will be created. To do so, the bag-of-visual-words approach [7], representing an image as an orderless collection of local features, is used. The bag-of-words model can be defined as an histogram representation based on independent features. This approach usually includes following three steps: feature detection, feature description and codebook generation. From a collection of images, a representative set of patches (or key points) is selected and each is transformed into a description vector. This set of vectors characterizes objects appearing in the images collection. Each vector is called a word and the whole set is called a vocabulary or codebook.

To obtain the global descriptor, the image (represented as a bag of key points) is expressed in the vocabulary space. This bag of key points is expressed as an histogram recounting the number of occurrences of each word in the image.

For the classification step, a database containing the occurrence histograms of each surgical tool needing to be found was used. A smooth classification was then performed by applying a k-nearest neighbor algorithm (k=20). At the end of the classification, for a region of interest, we had its probabilities to contain respectively every surgical tool of the database. A class "not a tool" is also considered and called "background class".

1.2. Pupil Segmentation

In the context of cataract surgery, a step of pupil segmentation can turn out to be very useful to identify some specific visual cues. Those more specific visual cues inside the pupil will improve surgical phases recognition.

This procedure can be divided in three parts and is based upon the assumption of a color difference between the pupil and everything else. The first part led to the creation of a black and white outline mask from the input image, which has been expressed, into the YUV color space. This mask was computed by smoothing, thresholding and performing morphological operations to the input image. Then, we tried to determine circles through the mask using the Hough transform. Sometimes, incomplete circles outlines in the mask may occur, leading to Hough circle detection failure. To tackle the problem, an iterative search was performed on the mask to identify the most probable circular zone. This search was based both on pixel counting and circle radius assumption. Finally, a normalization was performed. We kept the circle center found previously and we set the region of interest radius to a constant value. Following this procedure (Figure 3), a region of interest around the patient pupil has been retrieved.

1.3. Validation Studies

In this study, twenty cataract surgeries performed by three different surgeons from the University Hospital of Munich were available. They were recorded using the OPMI Lumera surgical microscope (Carl Zeiss) with an initial resolution of 720x576 at 25fps. In order to speed up processing, each video has been down-sampled to 1 fps and each frame has been spatially down-sampled by a factor 4 with a 5x5 Gaussian kernel, leading to a final resolution of 360x288. Eight surgical phases were defined (Figure 4).

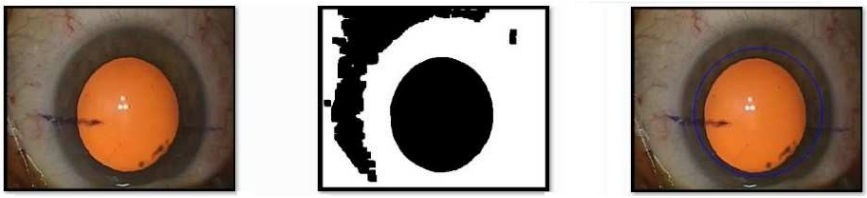

Figure 3. Pupil segmentation steps. From left to right: input image, mask, final segmentation (blue circle).

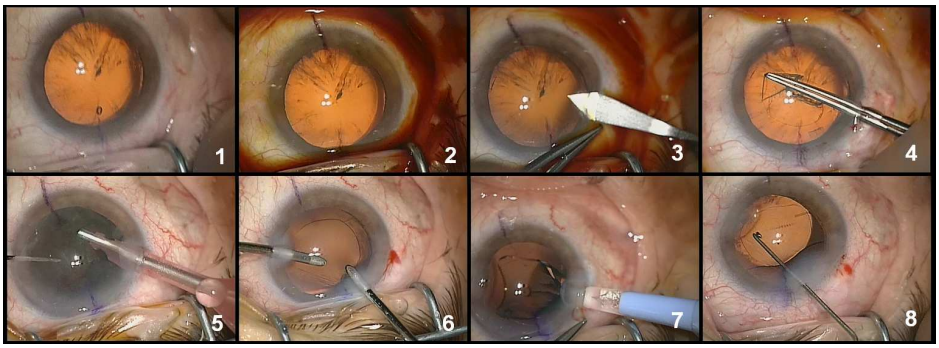

Figure 4. Typical images for the eight surgical phases: 1-Preparation, 2-Betaisodona injection, 3-Corneal incision, 4-Capsulorhexis, 5-Phaco-emulsification, 6-Cortical aspiration, 7-IOL implantation, 8-IOL adjustment and wound sealing.

As mentioned in the introduction, this work is a part of a higher-level study [1,2]. The aim of this framework is to automatically detect phases of a cataract surgery. Each video of the database has been labeled by surgeons, where they have defined the phases' transitions.

The segmentation procedure has been individually and manually validated over the entire video database by testing each frame of each video. The pupil was considered correctly segmented if it was mainly in the segmented region of interest.

For the surgical tools recognition, we first compared the four local key points detection methods along with the optimal number of words. This test was necessary to optimize the bag-of-words approach. Then, we studied results from this module independently from the global framework. To do so, we manually labeled each frame of each video and compared recognized tools and labeled tools. We defined seven classes for the classification, six for the surgical tools and one for everything that is not a tool (also called background class). Each class has been built with 100 representative images.

Finally, we compared results from the whole surgery modeling framework (i.e. detection of the surgical phases) with and without these two modules. The pupil segmentation has been integrated in order to improve the detection of the pupil color shade through histogram-based approach. With this new approach, the recognition of particular color shade within the pupil is now easier. The recognition of surgical tools has been added to the global framework by directly detecting surgical instruments of the surgery. New visual cues were created, corresponding to each instrument to be detected, increasing the number of information composing each semantic frame signature.

2. Results

For the pupil segmentation module alone, the pupil was correctly detected within 95% of all images (in Table 1). Best result was a segmentation accuracy of 99% and worse result was of 78%.

The results of the bag-of-word approach optimization regarding surgical tools recognition are the following. On the one hand, the number of words did not have a lot

of influence on the classification accuracy. However, a number of words ranged between 5 and 15 seems to be the best. The choice of this number will be highly correlated with the key points detection. On the other hand, there are major accuracy variations depending on the key points/description method combination. Of all the combinations tested, the SURF/SURF one seems to give the best classification accuracy (86% for 12 words).

For our study, we decided to use the SURF/SURF combination along with a 15-word vocabulary. Within the videos dataset, a surgical tool has been correctly identified as a tool with an accuracy of 84,1% (in Table 2). Given the fact that the classification returns a probability for each tool, no validation study has been conduct to verify if a found tool as been classify as the correct one.

Finally, we have added the new modules to the framework and we now obtain an overall recognition accuracy of 94.4%, results have been slightly improved (in Table 3).

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

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

Table 2. Tool detection percentage with the tool recognition module alone.

	Accuracy (Std)	**Std**
Detection	84.1%	8.6%

Table 3. Percentage of surgical phases correctly recognized with and without the new modules.

	Average (%)	**Std (%)**	**Min (%)**	**Max (%)**
Without modules	90.2	8.4	78.1	99.9
With modules	94.4	3.1	90.6	99.9

3. Discussion and Conclusion

In this paper, we proposed to add two new modules of visual cues detection to a framework able to automatically recognize surgical phases of cataract surgery. Even though the framework was created to be application-dependant, each kind of surgical environment has its own particularities and characteristics. As a consequence, the framework has to be tuned for a specific type of surgery in order to be as competitive as possible.

The pupil segmentation method, composed of image-based analysis, detects a region of interest containing the pupil with an accuracy of 95%. This can be considered as preliminary step in order to detect specific visual cues within the pupil. In order to avoid any further detection that could be done within the pupil, we decided to define a constant circumference value for every region of interest. As a drawback, the region of interest is not always perfectly centered on the middle of the pupil. Moreover, automatic segmentation turns out to be difficult when there are interferences in the microscope field of view. For instance, sometimes the pupil is not completely in the image or pupil outlines are too distorted. Retractors can also be too wide and fill too much space, or surgeon's fingers can appear in the field of view.

The surgical tool recognition method, as the pupil segmentation one, has been tuned for this type of surgery. Indeed, a training step is required before the utilization of the framework for each surgical tool likely to appear. Results obtained are promising, we obtain 84,1% of good recognition over all the videos. Some tools are easier to recognize because of their bigger size or because of color gradients more important. Surgical tools information slightly enhanced the framework but they can be far more useful for surgery modeling at a lower granularity level. Surgical tool presence can be directed linked to surgeon's activities. As a drawback, connected components obtained during the first stage of the method do not always contain whole surgical tool. This incomplete detection induces lower recognition rates. Moreover, it is quite difficult to build a complete background class so we improved it as much as possible.

To conclude, the addition of these two modules within the framework leads to better cataract surgery phases recognition. Other visual cues could be extracted in order to further improve phases recognition results. However, in future work, it would be more interesting to focus on lower level information. Surgical tools information could be used to detect surgeon's gestures and thus extract activities within major surgical phases of surgeries.

4. Acknowledgments

The authors would like to acknowledge the financial support of Carl Zeiss Surgical GmbH, as well as Dr. L. Riffaud for his help on the clinical aspect.

References

[1] 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).

[2] Lalys, F., Bouget, D., Jannin, P. An application-dependant framework for the recognition of high-level tasks in the OR. MICCAI 2011 (To be published).

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

[4] Lowe, DG. Object recognition from scale-invariant features. ICCV'99, 2, 1150-1157 (1999).

[5] Bay, H., Tuytelaars, T., Van Gool, Luc. SURF: Speeded Up Robust Features. Computer Vision - ECCV (2006).

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

[7] André, B., Vercauteren, T., Buchner, A.M. Endomicroscopic video retrieval using mosaicing and visual words. Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on.