An Application-Dependent Framework for the Recognition of High-Level Surgical Tasks in the OR

Florent Lalys[1,2,3], Laurent Riffaud[1,2,3,4], David Bouget[1,2,3], and Pierre Jannin[1,2,3]

[1] INSERM, U746, Faculté de Médecine CS 34317,
F-35043 Rennes Cedex, France
[2] INRIA, VisAGeS Unité/Projet, F-35042 Rennes, France
[3] University of Rennes I, CNRS, UMR 6074, IRISA, F-35042 Rennes, France
[4] Department of Neurosurgery, Pontchaillou University Hospital,
F-35043 Rennes, France

Abstract. Surgical process analysis and modeling is a recent and important topic aiming at introducing a new generation of computer-assisted surgical systems. Among all of the techniques already in use for extracting data from the Operating Room, the use of image videos allows automating the surgeons' assistance without altering the surgical routine. We proposed in this paper an application-dependent framework able to automatically extract the phases of the surgery only by using microscope videos as input data and that can be adaptable to different surgical specialties. First, four distinct types of classifiers based on image processing were implemented to extract visual cues from video frames. Each of these classifiers was related to one kind of visual cue: visual cues recognizable through color were detected with a color histogram approach, for shape-oriented visual cues we trained a Haar classifier, for texture-oriented visual cues we used a bag-of-word approach with SIFT descriptors, and for all other visual cues we used a classical image classification approach including a feature extraction, selection, and a supervised classification. The extraction of this semantic vector for each video frame then permitted to classify time series using either Hidden Markov Model or Dynamic Time Warping algorithms. The framework was validated on cataract surgeries, obtaining accuracies of 95%.

Keywords: Surgical phase, digital microscope, cataract surgeries, DTW.

1 Introduction

The field of surgical process analysis and modelling has recently gained much interest. Due to the technologically rich environment of the Operating Room (OR), a new generation of computer-assisted surgical (CAS) systems has appeared. As a result of these new systems, a better management, safety, and comprehension of the surgical process is needed. For such purposes, systems should rely on a context-aware tool, which knows the score to be played for adapting assistance accordingly. The challenge is therefore to assist surgery through the understanding of OR activities, which could be introduced in CAS systems. Clinical applications also include evaluation and training of surgeons, the creation of context-sensitive user interfaces, or the generation of automatic post operative reports.

G. Fichtinger, A. Martel, and T. Peters (Eds.): MICCAI 2011, Part I, LNCS 6891, pp. 331–338, 2011.

The goal is to collect signals from the OR and automatically derive a model. While it is possible to design such model manually, there are advantages of automating this process, mainly because manual work is time-consuming and can be affected by human bias and subjectivity. Due to the increasing number of sensors in the OR, the automatic extraction of data is now easier. Based on these signals, it's possible to recognize high-level tasks and hence avoid any additional installation of materials. Among all sensors, teams recently focused on videos coming from cameras already used in the clinical routine, which are a rich source of information. Compared to other data extraction techniques, it uses a source that does not have to be controlled by humans, automating the surgeons' assistance without altering the surgical routine.

Current work has made progress in classifying and automating the recognition of high-level tasks in the OR based on videos. Using laparoscopic videos, Speidel et al. [1] focused on surgical assistance by identifying 2 scenarios: one for recognizing risk situations and one for selecting adequate images for visualization. Their analysis was based on augmented reality and computer vision techniques. Lo et al. [2] used vision to segment the surgical episode. They used color segmentation, shape-from-shading techniques, and optical flows for tracking instruments. These features, combined with other low-level cues, were integrated into a Bayesian framework. Klank et al. [3] extracted image features for scene analysis and frame classification. A crossover combination was used for selecting features, while Support Vector Machines (SVMs) were used for the classification. Blum et al. [4] automatically segmented the surgery into phases. A Canonical Correlation Analysis was applied based on tool usage to reduce the feature space, and the modeling of resulting feature vectors was performed using Dynamic Time Warping (DTW) and Hidden Markov Model (HMM). Bhatia et al. [5] analyses overall OR view videos. After identifying 4 states of a common surgery, relevant image features were extracted and HMMs were trained to detect OR occupancy. Padoy et al. [6] also used external OR videos to extract low-level image features through 3D motion flows combined with hierarchical HMMs to recognize on-line surgical phases. In robotic using the Da Vinci, Voros and Hager [7] used kinematic and visual features to classify tool/tissue interactions. Similarly, Reiley and Hager [8] focused on the detection of subtasks for surgical skill assessment.

In a previous work [9], using neurosurgical videos, we proposed to extract surgical phases by combining a feature extraction process with HMM. In this paper, we extend this approach by proposing an application-dependent framework that can be adaptable to any type of surgeries. The idea is first to extract visual cues that can be helpful for discriminating high-level tasks. The visual cues are detected by specific image-based classifiers, obtaining a semantic signature for each frame. Then, these time series are aligned with a reference surgery using DTW algorithm to recognize surgical phases. Compare to traditional video understanding algorithms, this framework extracts application-dependant visual cues that are generic. The combination of image-based analysis and time series classification allows getting high recognition rates. We evaluated our framework with a dataset of cataract surgeries through cross-validation studies, and compared results of the DTW approach with the HMM classification.

2 Materials and Methods

2.1 Application-Dependant Visual Cues

Four classifiers based on different image processing tools were implemented (Fig. 1.). Each of these classifiers was related to one kind of possible visual cue. Visual cues recognizable through color were detected using a color histogram approach. For each shape-oriented visual cue such as the recognition of a specific object, a Haar classifier was trained. For texture-oriented visual cues, we used a bag-of-word approach using Scale Invariant Feature Transform (SIFT) descriptors, and finally for other visual cues that don't match these descriptions, we used an image classification approach including a feature extraction, selection and a classification with SVM.

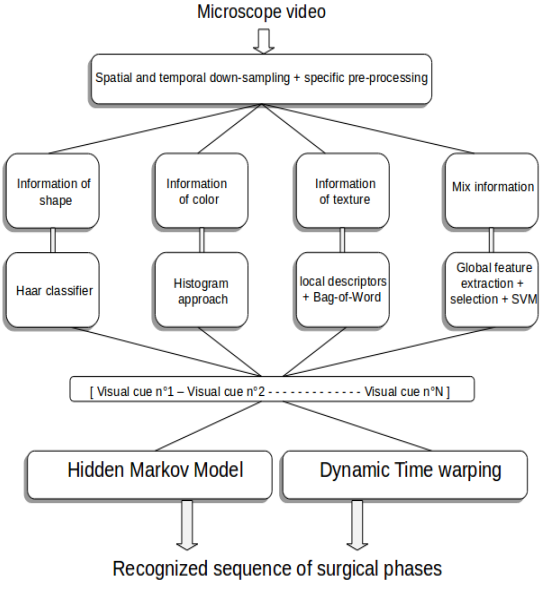

Fig. 1. Framework of the recognition system

Color-oriented visual cues: The color is one of the primary features used to represent and compare visual content. Especially, color histograms have a long history as a method for image description, and can also be used for identifying color shade through images. Here we used the principle of histogram classification to extract color-oriented visual cues, by creating a training image database composed of positive and negative images. Two complementary color spaces were extracted: RGB and HSV space. For quantifying similarities between histograms, we used the correlation.

Shape-oriented visual cues: We used here a Viola-Jones object detection framework [10], mainly used to detect specific object within images. The basic idea is to create a classifier based on features selected by AdaBoost. Weak learners of the algorithm are based on the Haar-like rectangular features, comparing the sum of intensities in adjacent regions inside a detection window. Then, strong learners are arranged into a classifier cascade tree in complexity order. The cascade classifier is therefore composed of stages, each one containing a strong learner. During the detection phase, a window looks through the image with different scales and positions. The idea is to determine at each stage if a given sub-window may be the searched object or not. The false positive rate and the detection rate are thus the product of each rate at each stage.

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 demonstrated impressive performances. 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 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 then expressed as a histogram recounting the number of occurrences of each pattern in a given image. For the texture analysis, we used the SIFT [11] descriptors.

Other visual cues: We already presented this approach in a previous paper [12]. Each frame was represented by a signature composed of low-level spatial features (RGB space, co-occurrence matrix with Haralick descriptors [13], spatial moments [14], and Discrete Cosine Transform (DCT) [15] coefficients). This signature was then reduced by feature selection. For that purpose, we fused a filter and a wrapper approach by using the union of both selection results. The RFE-SVM [16] and the mutual information (MI) [17] were chosen for the wrapper and the filter method respectively, keeping the 40 first features. Finally, a SVM was applied to extract the binary cue.

2.2 Time Series Classification

A binary semantic signature was extracted from each frame, composed of the recognized visual cues. We used the DTW algorithm [18] to classify these time series in a supervised way. The objective of DTW is to compare two sequences by computing an optimal match. These sequences may be time-series composed of feature sequences sampled at equidistant points in time. A local cost measure is needed to compare features. We used here the Hamming distance. To compare each surgery, we created an average surgery with the method described in [19]. Every query surgery was first processed to extract visual cues, and then the time series were compared to the average one. Once warped, the phases of the average surgery are transposed to the query one. We also used the Itakura parallelogram global constraint that limits the warping path to be within a parallelogram.

2.3 Data-Set

Our framework was evaluated on cataract surgeries. 20 cataract surgeries from the Hospital of Munich were included to the study (mean surgical time: 15 min). Videos were recorded using the OPMI Lumera surgical microscope (Carl Zeiss) with a resolution of 720 x 576 at 25 fps. We downsampled the videos to 1 fps, and spatially downsampled by a factor 8 with a 5-by-5 Gaussian kernel. Eight surgical phases were defined (Fig. 2). Additionally, five binary visual cues were chosen: the pupil color range (orange or black), the presence of antiseptic, of the knife, of the IOL instrument, and the global aspect of the cataract. Combinations of these five binary cues are informative enough to discriminate all 8 phases. The pupil color range and the presence of the antiseptic were extracted using color histograms. The knife was recognized using a Haar classifier. The IOL instrument was not identifiable through only color or shape analysis, that's why we chose the fourth approach using global

spatial feature extraction and SVM classification. Finally, the global aspect of the cataract was recognized using the BVW approach. For this detection as well as for the pupil color range classification, a step of pupil segmentation was first applied, using preprocessing steps composed of dilation/erosion operations and a Hough transform.

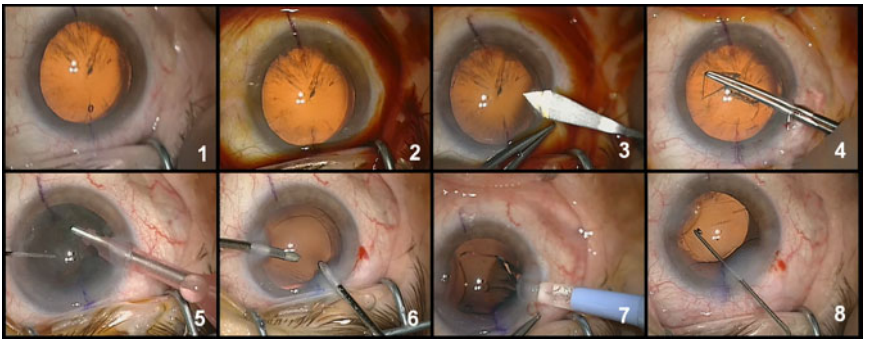

Fig. 2. Example of typical digital microscope images for the eight phases: 1-preparation, 2-betadine injection, 3-corneal incision, 4-capsulorhexis, 5-phacoemulsification, 6-cortical aspiration of the remanescent lens, 7-implantation of artificial IOL, 8-adjustment of the IOL

2.4 Cross-Validation

The initial work of phase and visual cue labeling was performed for each video. From each video, we randomly extracted 100 frames, getting an image database composed of 2000 labeled images. We then evaluated both aspects of our framework. First, every visual cue detection was assessed trough 10-fold cross-validation studies, by dividing the image database into 10 random subsets. Then, we evaluated the global framework with the same procedure. 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 validation, we computed the Frequency Recognition Rate (FRR). We also validated the added-value of the DTW algorithm, by comparing with an HMM approach, described in previous studies [9].

3 Results

Results of the cross-validation studies (Tab. 1.) showed that very good accuracies were obtained for visual cues with quite low standard deviations. The best recognition was obtained for the detection of the antiseptic, with a recognition rate of 98.5%, whereas the lower rate was obtained for the IOL instrument recognition (94.8%). An example of a DTW computation is shown on Fig. 3. Small errors occur in the phase transitions, but the global FRR stay high (~93%). Tab. 2. shows the global accuracy of the framework, using DTW or HMM approach. The global validation study with DTW showed a **mean FRR of 94.8%**, with a **min of 90.5%** and a **max of 98.6%.**

Table 1. Mean accuracy and standard deviation (Std) for the recognition of the 5 binary visual cues, computed on the entire video dataset

	Pupil color range	Presence antiseptic	Presence Knife	Presence IOL instrument	Cataract aspect
Accuracy (%)	96,5	98,5	96,7	94,8	95,2
Std (%)	3,7	0,9	3,4	1,1	1,8

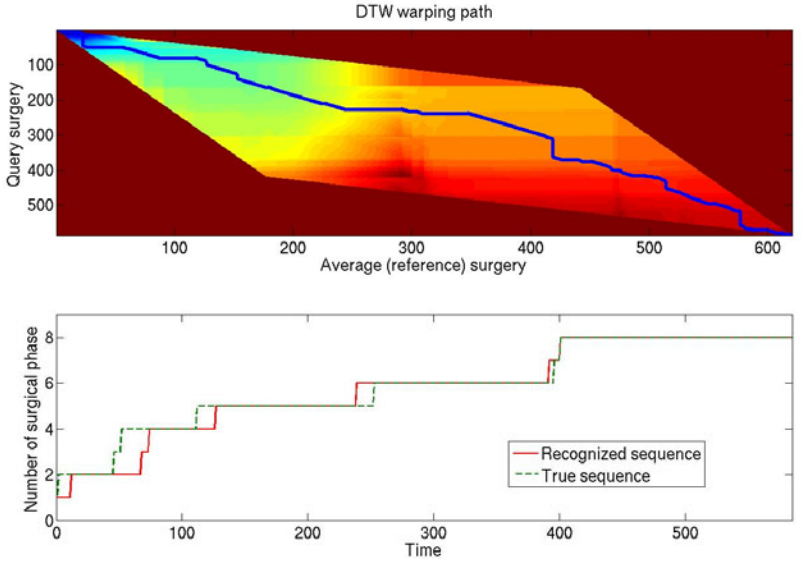

Fig. 3. Distance map of two surgeries and dedicated warping path using the Itakura constraint (up), along with the transposition of the surgical phases (down)

Table 2. Mean, minimum and maximum FRR of the HMM and DTW studies

	FRR (Std)	Minimum (%)	Maximum (%)
HMM (%)	92,2 (6,1)	84,5	99,8
DTW (%)	94,8 (3,7)	90,5	98,6

4 Discussion

4.1 Visual Cues and DTW

Combining with state-of-the-art techniques of visual cues recognition, DTW showed very good performance and allows further promising works on high-level tasks recognition in surgery. The comportment in color, texture and shape of the visual cues are intuitively known, allowing the classifiers to be effective. This approach turns out to be as generic as possible, and adaptable to any type of surgery. However, one limitation of the DTW algorithm is that it can't be used on-line, because the entire surgery is needed in order to find the optimal path. However, first results showed that the DTW algorithm was quite better for classifying times series data than the HMM.

4.2 Microscope Video Data

The real added value of the project lies in the use of microscope videos. This device is not only already installed in the OR, but it has also not to be monitored by the staff. Compared to other additional sensors, this allows the recognition to be fully automatic and independent. Moreover, microscope video data are reproducible within a same surgical environment and image features are invariant to task distortion [20]. Due to facilities differences between surgical departments, the system could not be flexible. The solution would be to train dedicated databases for each department, which would be adapted to the corresponding surgical environment and microscope scene layout.

4.3 Clinical Applications

The automatic recognition of surgical phases might be helpful for various applications. Purposes are generally to bring an added value to the surgery or to the OR management. This work could be integrated in an architecture that would take in real-time the microscope videos as input and transform it into information helping the decision making process, and driving context-sensitive user interfaces. Off-line, surgical videos would be very useful for learning and teaching purposes given their automatic indexation. Moreover, we could imagine the creation of pre-filled post operative reports that will have to be completed by surgeons. The recognition of lower level information, such as gestures, is difficult with microscope videos only. In future works, lower-level information such as surgeon's gestures will have to be detected to create multi-layer architectures.

5 Conclusion

We proposed in this paper a framework that automatically recognizes surgical phases from microscope videos. The first step of the framework is the definition of several visual cues for extracting semantic information and therefore characterizing every frame. Then, time series models allow an efficient representation of the problem by modeling time varying data. This association permits to combine the advantages of all methods for better modeling. We tested the framework on cataract surgeries, where 8 phases and 5 visual cues were defined by an expert, getting global accuracies of 95%. This recognition process is a first step in the construction of context-aware surgical systems, opening perspectives for a new generation of CAS systems.

Acknowledgements. The authors would like to acknowledge the financial support of Carl Zeiss Meditec.

References

1. Speidel, S., Sudra, G., Senemaud, J., Drentschew, M., Müller-stich, B., Gun, C., Dillmann, R.: Situation modeling and situation recognition for a context-aware augmented reality system. In: Progression in Biomed. Optics and Imaging, vol. 9(1), p. 35 (2008)
2. Lo, B., Darzi, A., Yang, G.: Episode Classification for the Analysis of Tissue/Instrument Interaction with Multiple Visual Cues. In: Ellis, R.E., Peters, T.M. (eds.) MICCAI 2003. LNCS, vol. 2878, pp. 230–237. Springer, Heidelberg (2003)

3. Klank, U., Padoy, N., Feussner, H., Navab, N.: Automatic feature generation in endoscopic images. Int. J. Comput. Assist. Radiol. Surg. 3(3-4), 331–339 (2008)
4. Blum, T., Feußner, H., Navab, N.: Modeling and Segmentation of Surgical Workflow from Laparoscopic Video. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A., et al. (eds.) MICCAI 2010. LNCS, vol. 6363, pp. 400–407. Springer, Heidelberg (2010)
5. Bhatia, B., Oates, T., Xiao, Y., Hu, P.: Real-time identification of operating room state from video. In: AAAI, pp. 1761–1766 (2007)
6. 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)
7. Voros, S., Hager, G.: Towards "real-time" tool-tissue interaction detection in robotically assisted laparoscopy. Biomed. Robotics and Biomechatronics, 562–567 (2008)
8. Reiley, C., Hager, G.: Decomposition of robotic surgical tasks: an analysis of subtasks and their correlation to skill. In: M2CAI Workshop, MICCAI 2009 (2009)
9. Lalys, F., Riffaud, L., Morandi, X., Jannin, P.: Surgical phases detection from microscope videos by combining SVM and HMM. In: Menze, B., Langs, G., Tu, Z., Criminisi, A. (eds.) MICCAI 2010. LNCS, vol. 6533, pp. 54–62. Springer, Heidelberg (2011)
10. Viola, P. and Jones, M.: Rapid real-time face detection. IJCV, 137-154 (2004)
11. Lowe, D.G.: Object recognition from scale-invariant features. In: ICCV 1999, vol. 2, pp. 1150–1157 (1999)
12. Lalys, F., Riffaud, L., Morandi, X., Jannin, P.: Automatic phases recognition in pituitary surgeries by microscope images classification. In: Navab, N., Jannin, P. (eds.) IPCAI 2010. LNCS, vol. 6135, pp. 34–44. Springer, Heidelberg (2010)
13. Haralick, R., Shanmugam, K., Dinstein, I.: Textural features for image classification. IEEE Trans. on Systems, Man, and Cybernetics 3(6), 61–621 (1973)
14. Hu, M.: Visual pattern recognition by moment invariants. Trans. Inf. Theory 8(2), 79–87 (1962)
15. Ahmed, N., Natarajan, T., Rao, R.: Discrete Cosine Transform. IEEE Trans. Comp., 90–93 (1974)
16. Guyon, I., Weston, J., Barhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machine. Machine Learning 46, 389–422 (2002)
17. Hamming, R.W.: Coding and Information Theory. Prentice-Hall Inc., Englewood Cliffs (1980)
18. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken work recognition. In: IEEE Trans. Acoustic, Speech and Signal Proc., vol. 26(1), pp. 43–49 (1978)
19. Ahmadi, S.-A., Sielhorst, T., Stauder, R., Horn, M., Feussner, H., Navab, N.: Recovery of surgical workflow without explicit models. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4190, pp. 420–428. Springer, Heidelberg (2006)
20. Bouarfa, L., Jonker, P., Dankelman, J.: Discovery of high-level tasks in the operating room. J. Biomedical Informatics 44(3), 455–462 (2010)