Effect of Spatial Alignment in Cataract Surgical Phase Recognition

Xiang Xiang
Dept. Computer Science
Johns Hopkins University
xxiang@cs.jhu.edu

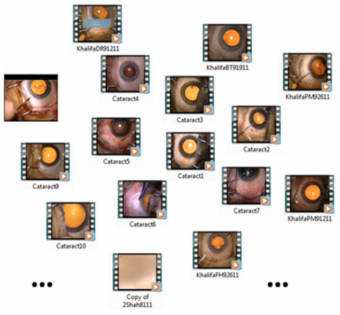


Figure 1. We aim to design systems for automatic recognition of surgical tasks using microscope videos analysis. We validate its use on cataract procedures and analyze spatial alignment effect.

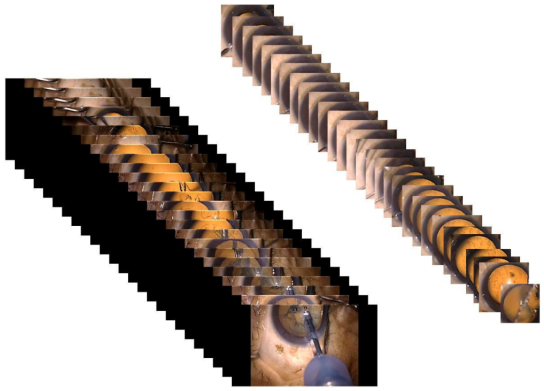


Figure 2. Results of the spatial alignment. Since the size of the limbus region in the video changes, the aligned image patches are not cropped according to the same size, yet the same aspect ratio.

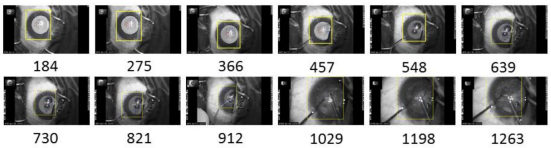


Figure 3. Pupil tracking results for video 1.

Abstract

The need for a better integration of the new generation of computer assisted surgical systems and automated surgery grading has been recently emphasized. In this abstract paper, we analyze the effect of spatial alignment in cataract surgical step recognition and present preliminary results.

1. Introduction

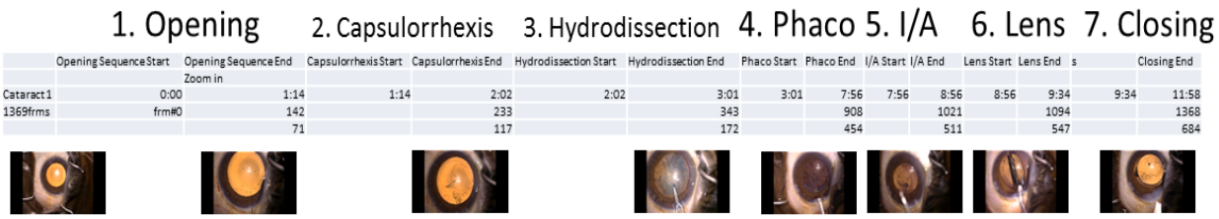
The need for a better integration of the new generation of computer assisted surgical systems and automated surgery grading has been recently emphasized. One necessity to achieve this objective is to retrieve data from the Operating Room (OR) with different sensors, then to derive models from these data. Recently, the use of videos from cameras in the OR has demonstrated its efficiency [1]. As shown in Fig. 1, we validate its use on cataract procedures, supported by the Flaum Eye Institute at URs Medical Center. Moreover, this course project aims to practice the techniques taught in lecture 4 (pattern recognition concepts), 5 (modern visual features), 6 (mid-level vision), and 11 (tracking).

2. Problem Statement and Data Acquisition

A standard cataract surgery is commonly divided into 7 steps, as shown in Fig. 2. A system is in need to automatically recognize each step, which is generally discriminative from others. We first try to well define the problem: step recognition is actually a multi-class classification task. Our dataset includes two parts: the videos and the manual labels indicating the starting and ending time of each step of the surgery. These labeled data will be used either for training classification algorithms or as ground truth data to validate the accuracy of the trained algorithms. Currently, we have received 10 labeled videos.

3. Approach to Surgical Phase Recognition

We decide to train a multi-class SVM in a 1-vs-the-rest manner [3]. A 1-vs-1 manner is also feasible, but it takes more training time than 1-vs-the rest manner [3]. For each sequence, we extract the GIST feature [4] for each frame and train them with labels using SVM, as shown in Fig. 3. GIST is a global feature descriptor, which gives a holistic representation of image. See Fig. 4 and Fig. 5 for illustrations of the GIST feature.



	Opening Sequence Start	Opening Sequence End	Capsulorrhexis Start	Capsulorrhexis End	Hydrodissection Start	Hydrodissection End	Phaco Start	Phaco End	I/A Start	I/A End	Lens Start	Lens End	s	Closing End
		Zoom in												
Cataract 1	0:00	1:14	1:14	2:02	2:02	3:01	3:01	7:56	7:56	8:56	8:56	9:34	9:34	11:58
1369frms	frm#0	142		233		343		908		1021		1094		1368
		71		117		172		454		511		547		684

Figure 4. Seven steps defined in a standard cataract surgery. The starting and ending time of each step is listed.

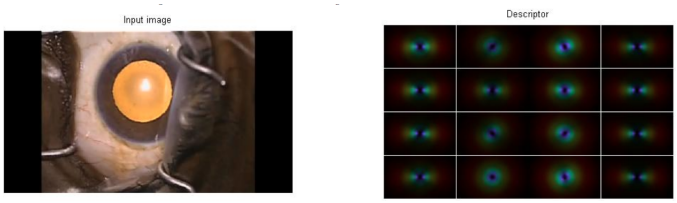


Figure 5. The classification is done using SVM inputted with GIST features. Shown for one frame in the cataract surgery video.

Steps	1	2	3	4	5	6	7
Prec	100	83.33	71.05	91.12	71.79	80	93.01
Recall	100	97.83	98.18	98.23	98.25	100	97.08
F-sco	100	90	82.44	94.65	82.96	88.89	95

Table 1. Performance (%) evaluation for validation experiments.

4. Approach to Spatial Alignment

The motivation to perform the spatial alignment is to rule out distraction. We aim to focus on the limbus or pupil region. Except the instruments, the region out of the limbus generally appears similar. Therefore, we think that the outside region is not so discriminative. The desired alignment results are displayed in Fig. 6.

We attempt to achieve the spatial alignment through limbus region tracking. We use a state-of-the-art tracker named TLD (Tracking-Learning-Detection) [2]. Fig. 7 displays its tracking results for the 1st video [1]. Then, the enlarged image patches will be used for training and testing.

5. Experiments and Evaluation

We compute the precision, recall and F-measure score for each testing. First, we perform validation experiments for video 1 to verify the correctness of the implementation. Namely, we sample both training data and testing data from video 1. While training data and testing data are from the same video, there is no overlapping. The performance is presented in Table 1. We conduct a comparison between the method without tracking and the method with tracking, as illustrated in Figure 6. Moreover, we analyze each testing in detail. In this report, we present the analysis for step 2, the example frame of which is displayed in Fig. 8.

6. Conclusion

From the experimental results and performance evaluations, we can draw two tentative conclusions: Tracking does not necessarily mean spatial alignment. Accurate tracking means that. Long-term accurate tracking is hard. TLD is generally robust, while the results (position and size) are still not accurate enough. Tracking may need top-down guidance as well. However, we still believe that spatial alignment helps, while it is hard. How about getting back to look at the frames holistically yet cleverly? When we focus too much, we may lose the spatial context, such as the instrument, which is discriminative across steps.

[1] See video results at https://youtu.be/3_CzI4l3nL4

	Step1	Step2	Step 3	Step 4	Step 5	Step 6	Step 7
Precision (no track)	60.87%	17.49%	**6.95%**	67.21%	**11.72%**	7.01%	**14.76%**
Precision (tracking)	**94.58%**	**22.18%**	5.12%	**91.65%**	0	**11.94%**	0.19%
Recall (no track)	**57.38%**	**41.94%**	**20%**	**44.16%**	**42.11%**	21.11%	**35.43%**
Recall (tracking)	43.74%	**41.91%**	18.38%	16.57%	0	**77.94%**	0.12%
F-score (no track)	**59.07%**	24.69%	**10.32%**	**53.30%**	18.34%	10.52%	**20.84%**
F-score (tracking)	**59.82%**	**29.01%**	8.01%	28.07%	0	**20.71%**	0.15%

Figure 6. Performance evaluation on the full sequence of video 2. The method without tracking and the method with tracking are compared, with better performance highlighted. Step 2 will be analyzed in detailed below.

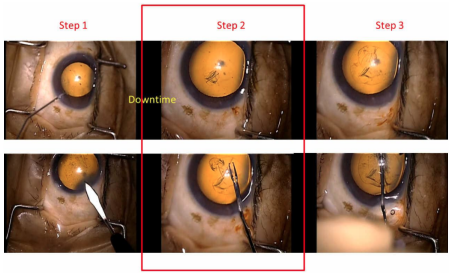


Figure 7. Step 2 is visually different with the adjacent step 1 and step 3, mostly in the surgical instrument and the pupils inside area.

References

[1] F. Lalys, L. Riffaud, D. Bouget, and P. Jannin. A framework for the Recognition of High-Level Surgical Tasks from Video Images for Cataract Surgeries. IEEE TBME, 00885, 2011.

[2] Z. Kalal, J. Matas, and K. Mikolajczyk. P-N Learning: Bootstrapping Binary Classiers by Structural Constraints. In IEEE CVPR, 2010.

[3] C. Bishop. Multiclass SVMs. Pattern Recognition and Machine Learning, pp. 338 339, Springer.

[4] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. IJCV, 42(3): 145-175, 2001.