H-038

Fundamental Study of Recognizing the Surgeon's Action during Suture Surgery from the Video Sequence

李 イエ† 大谷 淳† 千葉 敏雄‡ 徐 栄† 山下紘正‡

Ye Li Jun Ohya Toshio Chiba Rong Xu Hiromasa Yamashita

1 Introduction

Shortage of nurses is a serious world problem. A robotic scrub nurse (RSN) can support surgeons during surgery to alleviate the shortage problem, and also can reduce human error in the operating room (OR). Jacob et al. [1] proposed a GestoNurse, which can pass tools to surgeons according to their gesture commands, but this method imposes extra works on surgeons. Bardram et al. [2] proposed a system to recognize the phase during surgeries using embedded and body-worn sensors. This method achieves good results, but is not practical. To solve these issues, we aim at a robot system that can automatically judge which tools should be passed to the surgeon, by recognizing surgical situations via video analysis. This paper reports a fundamental study of how to recognize the surgeon's action during suture surgery from the video sequence acquired by the camera that observes the surgery scene.

2 Suture Surgery

Suture is the simplest and most basic operation of surgery. As a first step towards the realization of autonomous surgery support robots, this paper deals with recognizing the surgeon's hand action during suture surgery.

A suture surgery consists of the three stages: preparation, suture and post-treatment. In the preparation stage, surgeons disinfect the wound, inject local anesthetic and wash the wound. In the suture stage, surgeons suture the wound and tie the knot. In the post-treatment, surgeons disinfect the wound again and pack it, where the surgical tools are returned to the original places.

These stages contain surgeons' five actions: disinfection, anesthesia, washing, suture and tying. Surgeons use cotton-tips to disinfect, injector to anesthetize, forceps and cotton-balls to wash the wound, as well as forceps and needle-holder to suture the wound and tie the knot.

The first three actions use different surgical tools; therefore, we might be able to detect the tools first and to recognize these three actions by tools. However, in case of suture and tying (Fig.1), we cannot use the tools to distinguish them, because, as described earlier, the two actions use the same tools. That is why we need to recognize the actions using information obtained from the hands. Different from the general action recognition, hand action during suture surgery includes slight and complicated movements.

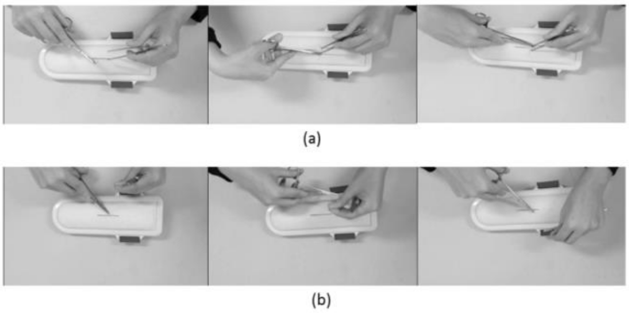


Fig.1 (a) suture (b) tying

3 Method

This paper proposes a method for classifying suture and tying. The proposed method consists of the following three steps (Fig.2): hand detection, feature computation and classification.

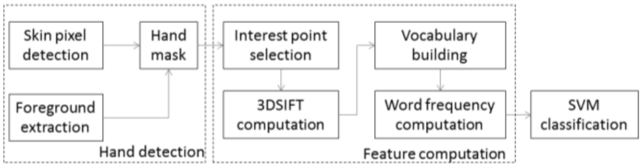


Fig.2 Flow chart

We detect the hand positions during the whole video and compute 3D SIFT descriptors in the hand area. After computation, we build a word vocabulary and compute the frequency of each word. Finally, SVM (Support Vector Machine) classifies actions.

3.1 Hand Detection

We use skin pixel detection [3] and foreground extraction [4] to build a 3D hand mask in the surgical video.

For skin pixel detection, we choose a certain subspace in HSV space to define the skin pixels. If a pixel's color components in HSV space are included in the chosen subspace, then this pixel is judged to belong to the skin pixel.

The skin pixel detection can extract the skin pixels from the video, but during the surgery, the wound's color is similar to the

† 早稲田大学大学院国際情報通信研究科 Waseda University, GITS

‡ 国立成育医療研究センター臨床研究センター NCCHD, CRC

skin (Fig.3(a)); therefore, we need to remove the wound from the hand mask. To solving this problem, we use foreground extraction. A pixel is judged to belong to foreground if its color is different from any background colors. Since only surgeon's hands move during the surgery and the patient does not move, we can assume the surgeon's hands correspond to the foreground to be extracted (Fig.3 (b)).

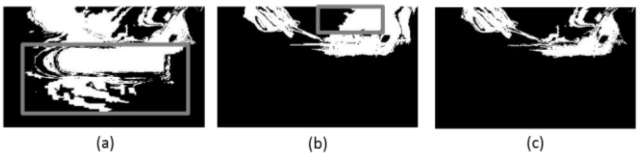


Fig.3 Result of (a) skin pixel detection (b) foreground extraction (c) a+b

3.2 Feature Computation and Classification

After the hand detection, we select points at random in the hand area to compute the 3D SIFT descriptors [5]. 3D SIFT descriptor uses both space and time information and has robustness to noise and orientation.

We use the SIFT descriptor to build a word vocabulary. Because of the instability of K-means, we utilize the improved clustering algorithm, Hierarchical K-means [6] as well as K-means to cluster descriptors. Specifically, first K-means is performed for clustering the data several times, and then hierarchical clustering is applied to cluster the centroid points of the data obtained from K-means. Finally we use the centroid points obtained from hierarchical clustering as the initial points to be clustered by K-means.

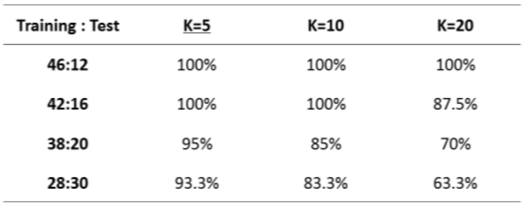
As a result of the clustering, we obtain some words that correspond to the centroids of the clusters. The frequency of each word is computed in each video and is accumulated into a histogram.

Finally, the frequency histograms are used as feature vectors for the classification by SVM.

4 Experiment

We collected 29 videos of suture and 29 videos of tying. The resolution of each from of the videos is 320×240 pixels. Since the speed of action is not uniform, the length of suture and tying videos range from 370 to 750 frames and from 270 to 1,270 frames, respectively. We extracted the hand areas in each video and selected 100 interest points in the area randomly. We use the 2×2×2 configurations of sub-histogram to compute the 3D SIFT descriptors. Consequently, we obtain 5800 descriptors, and the length of each descriptor is 640.

We divide these 5800 descriptors into the training set and test set according to different proportions, where the suture and tying in each set contain the same number of descriptors. We cluster the descriptors of the training set into K (pre-specified) clusters and compute the frequency of K centroid words in each suture and tying video (e.g. when the training set contains 23 suture and tying videos, respectively, if the number of clusters K=5, then 4600 descriptors in the training set are clustered into the five clusters, and finally we have gotten 5-dimensional feature vectors of suture and tying). The recognition results for different proportions and cluster numbers are shown in the Table 1.



Training : Test	K=5	K=10	K=20
46:12	100%	100%	100%
42:16	100%	100%	87.5%
38:20	95%	85%	70%
28:30	93.3%	83.3%	63.3%

Table.1 Recognition results (K: number of clusters)

As a result, we can find that training sets with larger number of descriptors tend to achieve good recognition accuracies. If the number of clusters is 5, we achieve the highest recognition rate.

5 Conclusion

This paper has proposed a method for recognizing surgeon's actions during suture surgery using hand detection and 3D SIFT descriptor. Recognition rates higher than 90% are achieved.

However, in the real surgery, besides suture and tying, there are not only many other actions, but also negative actions, which are meaningless, noisy movements of the hand. How to remove such a negative action and to recognize only significant actions during a long video sequence are our next work.

References

[1] M. Jacob et al. Gestonurse: A multimodal robotic scrub nurse. HRI'12, Boston.

[2] J. Bardram et al. Phase recognition during surgical procedures using embedded and body-worn sensors. PerCom'11, Seattle.

[3] C. Conaire et al. Detector adaptation by maximising agreement between independent data sources. IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1-6.

[4] K. Philip et al. A low-cost performance analysis and coaching system for tennis. MM'10, Florence.

[5] P. Scovanner et al. A 3-Dimensional SIFT descriptor and its application to action recognition. MM'07, Augsburg.

[6] K. Arai et al. Hierarchical K-means: an algorithm for centroids initialization for K-means. Reports of the Faculty of Science and Engineering, Saga University, Vol. 36, No.1, 2007.