MVA2015 IAPR International Conference on Machine Vision Applications, May 18-22, 2015, Tokyo, JAPAN

# Temporal-Spatial Validation of Knot-Tying Procedures Using RGB-D Sensor for Training of Surgical Operation

Yoko Ogawa
Ritsumeikan University
Shiga, JAPAN
ogawa@i.ci.ritsumei.ac.jp

Nobutaka Shimada
Ritsumeikan University
Shiga, JAPAN

Yoshiaki Shirai
Ritsumeikan University
Shiga, JAPAN

Yoshimasa Kurumi
Shiga University of Medical Science
Shiga, JAPAN

Masaru Komori
Shiga University of Medical Science
Shiga, JAPAN

## Abstract

*We propose a method for validating surgical knot-tying motions for beginner's self-training. Our system observes trainee's hands by a RGB-D sensor and describes the point cloud as a SHOT feature. The features are used for matching an input image sequence to that of an expert by dynamic programming. The type of the knot-tying motion is recognized by the matching score. The matching enables the system to recognize type of input knot-tying motion and to validate each frame. The system specifies inappropriate frames by temporal validation based on the difference of the matched frames. Then the system detects and shows inappropriate parts in each inappropriate frame by utilizing the spatial structure of the feature. We tested our method on a motion dataset of novice trainees and achieved favorable performance.*

## 1 Introduction

Medical trainees usually learn surgical procedures under face-to-face teaching by skillful surgeons. The skillful surgeons are much busy for their own medical duties and have very limited time for teaching. For keeping the trainee's learning chances, an automatic system validating trainee's surgical procedure can be one of the solutions. Educational studies say that it is important to fast and effective remediation to indicate inappropriate or imperfect parts to the unskillful trainees. We first target knot-tying motions in suture operations and propose a method that automatically indicates inappropriate parts of trainee's motions. Each knot-tying motion is characterized by hand gestures making the knot. Therefore the knot-tying type is necessary to be recognized and then the inappropriate parts for that tying type should be specified. The proposed method recognizes the knot-tying type by modeling the skillful hand motion and matching an input to the models by frame-to-frame, then specifies the inappropriate frames and visualizes the detected inappropriate regions of the trainee's hands.

In general, recognition of hand gesture from a movie consists of feature extraction and matching the input to motion models. Hand feature extraction can be roughly divided into 3D parametric model-based approach and appearance-based approach.

In the 3D-parametric-model-based approach, hypotheses are generated from a model, e.g. a 3-D hand mesh with joint angle parameters, and then the feature is extracted as joint angle parameters tracked by fitting the hypotheses to the observation. Oikonomidis et al. [2] estimate two strongly interacting hand poses from RGB-D images using particle swarm optimization. The approach has few parameters, but it has worries in fitting and tracking accuracy. A drastic mistake of the fitting wreaks catastrophe for the tracking and the recognition.

The appearance-based approach extracts the hand regions from backgrounds and then extracts a motion feature of whole sequence of foreground regions. Li et al. [6] proposed sectioned-SIFT, that is temporally spliced Bag-of-Feature of 3D-SIFT, for recognition of hand actions during suture surgeries. Biswas et al. [3] use depth histograms and motion information of foreground regions.

The appearance-based approach is more robust to complicated hand shape with strong self occlusion than the model-based approach for the recognition. When the input motion partially includes an irregular motion like hesitating motions, however, whole motion cannot be matched to the appropriate model. Thus in order to recognize such partially irregular motions, a frame-wise appearance feature and the frame-wise matching should be employed.

Several frame-wise 3D appearance descriptors have been proposed. Spin Images [8] provides the spacial point distributions for hand regions by projecting 3D points to a cylindrical surface surrounding the hand region and counting the point histogram on the surface. FPFH [9] extracts a histogram of the local surface property around each 3D point of hands as the local distribution of the surface normals. Since these methods are based on histograms which coarsely encodes the spatial property of hand shape, it is difficult to adopt for the spatial validation. In this study, we employ a customized SHOT (Signature of Histograms of OrienTations) [10] feature for representing the 3D point distribution of both hands. Since the feature finer preserves 3D positions and surface normals of points, it is helpful for the spatial validation of the hand motions. We customize it and use for a frame-wise feature.

Detection of the inappropriate frames and parts needs to match an input to models frame-to-frame. For the matching, Hidden Markov Model [4], CDP (Continuous Dynamic Programming) [5], and Neural Network [6] are often used. Li et al. [6] proposed a NN-based method for recognizing hand actions such as "suturing"and "tying" in suture operations. Their method concentrates to recognition of action types but treats neither indication of inappropriate parts nor temporal-spatial matching. For spotting and frame-level validation of tying motions, we take two-step approach.

First, candidate sections of knot-tying motion are detected from an input sequence which consists of many tying motions. Next, the candidates and models are matched frame-to-frame by modified CDP. We utilize the spatial structure of SHOT feature as the frame-wise feature of the input and model motions. After the frame-wise matching to every model, the best-matched model to the input is determined.

Our method temporally evaluates appropriateness of each input frame based on the corresponding model frame, and then spatially evaluates appropriateness of each part in the inappropriate frames. Finally, it visualizes the inappropriate parts of the trainee's procedure. This paper shows favorable performance of our method by testing on an our knot-tying motion dataset.

## 2 Structure of Knot Tying Motion

A typical knot-tying motions in suture operations consists of three knot-tying motions. Each knot-tying motion makes a half hitch. The three half hitches are alternately tied with obverse and reverse. Surgical operators should tie as fast and correct as possible in surgical operations. The correctness of a knot-tying motion is evaluated based on both making knot and tightening knot techniques. We divide a knot-tying motion into the following three steps.

**closing** closing and closing the strand by both hands

**tying** tying the strand

**tightening** tightening the strand tension

Figure 1 shows a knot-tying motion and the corresponding steps. The hand motions in the second step decides the knot type (obverse or reverse), and these hand postures in the second step are the most important for evaluation of making knot technique. For evaluation of tightening knot technique, tightening direction in the third step is the most important.

In this paper, we use the tying step for recognizing type of tying motions and validation of the motions.

## 3 Detection of Tying Section Candidates

The system observes point clouds by a RGB-D sensor. The observed points have RGB color information and XYZ position information. In order to describe the hand motion, hand regions should be extracted from the input point clouds. In this paper, we use simple thresholding of the input position and color which converted from RGB to HSV for the extraction. These thresholds are determined experimentally.

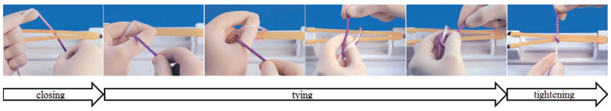

Figure 1. Knot-tying motion flow: The upper images are cited from [1]

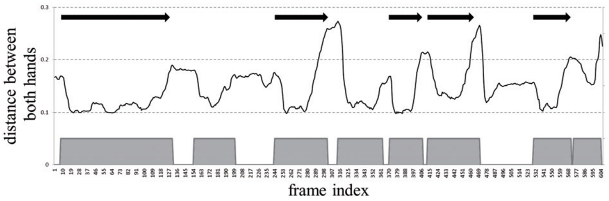

Figure 2. Transition of distances between both hands : Arrows on the top line represent sections of actual knot-tying motions containing whole steps. Boxes on the bottom line represent detected sections.

Table 1. Type of knot-tying motion

| Type | primary hand | directions of strand heads | direction of hand motion | knot type (obverse /reverse) |
|---|---|---|---|---|
| 1 | both hands | left : far side<br>right : near side | backward | reverse |
| 2 | both hands | left : near side<br>right : near side | backward | obverse |
| 3 | both hands | left : near side<br>right : near side | forward | reverse |
| 4 | left hand | left : near side<br>right : near side | backward | obverse |
| 5 | left hand | left : near side<br>right : near side | forward | reverse |
| 6 | right hand | left : near side<br>right : near side | backward | reverse |

We use a transition of the distance between both hands for detection of tying section candidates. In knot-tying motions, the hand distance decrease during closing step. The distance keep constant during tying step. In tightening step, the distance increase. We divide the hand point cloud to left and right hands by k-means++ [11], and treat the distance between two centroids as the distance between both hands. Figure 2 shows the transition of distances between both hands in a sequence that actually consists of 5 times of knot-tying motions.

In each actual knot-tying motion, the distance between both hands roughly makes a valley. In order to detect these sections, we first detect lower bottoms and then search for higher peaks before and after each bottom. We treat these sections between the peaks before and after the bottoms as tying section candidates.

## 4 Recognition of Knot-tying Motion Type

Validation of knot-tying motions in suture operations requires to discriminate the knot types (obverse or reverse) of three half hitches respectively. While the direct observation of a small knot is quite difficult the knot type can be estimated by observing hands and identifying their knot-tying motion type. We manually labeled types of knot-tying motion based on three attributes; "primary hand" (making knot), "direction of strand heads at the beginning of the knot-tying motion", and "direction of the hand motion."

For recognition of knot-tying motion type, we construct models of the knot-tying motions using only tying step sequences. In this study, we focus on typical 6 types of knot-tying motions in Table 1.

### 4.1 Feature Description

In this paper, we employ a SHOT feature to describe shapes of both hands in each frame. Figure 3

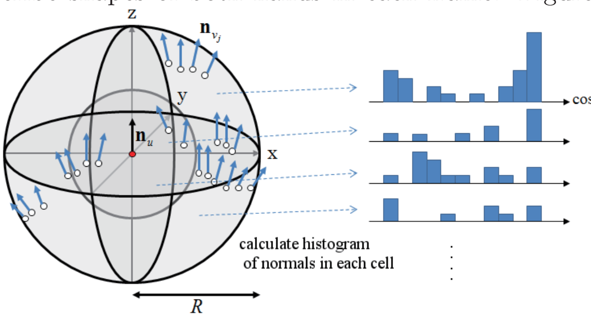

Figure 3. Structure of SHOT feature

shows the feature structure. The feature consists of relative histograms of normals of a point cloud. The feature consists of Local Reference Frame (LRF) and SHOT descriptor. LRF is a local coordinate system for the feature description. SHOT descriptor shows local shapes based on normals of the points $\mathbf{p}_j$ in sphere of radius $R$ around the origin of LRF $\mathbf{p}$. The descriptor has 32 spatial bins corresponding spatial partition, resulting from 8 azimuth divisions, 2 elevation divisions and 2 radial divisions (only 4 azimuth divisions are shown in Figure 3). Each spatial bin consists of a local normal histogram which has 11 bins. SHOT descriptor consists of jointed these histograms. L1 norm of the descriptor is normalized to 1.

We assume that the person performing the knot-tying is facing the camera and that the strand position relative to the person's body is almost fixed. In this study, we assign a gravity center of the hand points in every frame to $\mathbf{p}$, and we fix the coordinate system of LRF to the camera's one. Furthermore, we set $R$ to $150mm$ in order to enclose both hands.

### 4.2 Model Construction

We prepare model sequences of each type in the Table 1. A model performer randomly try the 6 types of knot-tying motion. The tying section candidates are detected from the captured sequence by the method which has been described in section 3. Then we manually choose sections for the model sequences from the candidates based on smoothness of the motion.

### 4.3 Frame-to-Frame Matching and Recognition of Knot-Tying Motion Type

The system detects several tying section candidates from an input sequence and describes them as SHOT feature sequences. We define the frame difference $d$ between an input and a model as follows.

$$d = \sum_k |f_k^I - f_k^M| \tag{1}$$

$f_k^I$ is the value of the $k$-th dimension of the feature vector of the input frame. $f_k^M$ is the one of the model frame. The system matches each tying section candidate with the models using CDP matching. The conventional CDP matching minimizes accumulated costs which is normalized by the model frame length by allowing only stationary or forward transitions for model frames. Nishimura et al. had proposed non-monotonic CDP [5] in order to recognize hesitated gestures. Although their method temporally decays the costs in the accumulation, we minimizes the average cost along the path regardless of time passage in order to evaluate equally all frames in the tying motion.

The given path represents the corresponding frame pairs and the average cost along the path is treated as the matching cost of the input and model sequences. If the minimal cost is more than a threshold $T_c$, we regard the tying section candidate as not matched to the model. We assume that the tying step lies at the center of the actual tying section, and adopt constraints for the start frame and the finish frame. If the length of the tying section candidate is $n$ frames, we allow the frames before $(\frac{n}{2} - 1)$-th frame to be a start frame, and the frames after $\frac{n}{2}$-th frame to be a finish frame. The best matched model is derived by minimization of the matching cost under the above conditions. If there is no model satisfying the condition, we treat the section as a non-knot-tying motion section.

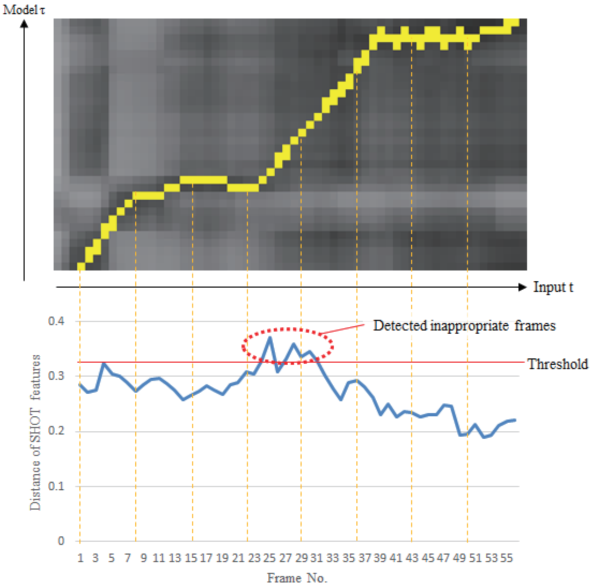

Figure 4. (Top) Distance map of SHOT features and minimum cost path. (Bottom) Distance of the features on the path.

## 5 Validation of Knot-Tying Motion

The validation of knot-tying motion is divided into temporal and spacial validation. In this paper, we define "parts" as the spacial partitions of SHOT descriptor. The specifying inappropriate frames is based on the frame-to-frame matching. We use the feature difference of the matched frames for the validation. If a distance of a matched frame pair is more than threshold $T_f$, the input frame is detected as an inappropriate frame. We define the difference of each spatial bin $d_i$ as follows.

$$d_i = \sum_{k \in q_i} |f_k^I - f_k^M| \tag{2}$$

$q_i$ is a set of dimension's indices corresponding to the $i$-th spatial bin. In the inappropriate frames, if $d_i < T_s$, the $i$-th partition is detected as an inappropriate part.

## 6 Experimental Results

We evaluated our method based on the following two cases using ASUS$^{\mathrm{TM}}$Xtion$^{\mathrm{TM}}$Pro Live sensor as the RGB-D sensor.

1. Identical person performs every model and an input sequence.
2. A person performs every model sequence and another person performs an input sequence.

### 6.1 Result of Recognition Type of Knot Tying Motion

In case 1, a performer tries randomly the 6 types of knot-tying motion. We picked up manually 6 tying section candidates as model sequences. The input sequence of case 1 actually consists of 14 times of knot-tying motion. 25 tying section candidates were detected in the input sequence automatically. We set the threshold $T_d$ to $40mm$ and the threshold of DP output $T_c$ to 0.35. The section which can be determined visually what motion was tried should be identified the motion type even if the motion was interrupted. In the above conditions, the correct recognition rate was 92% (23 sections), the false-negative rate was 4% (1 section) and the false-positive rate was 4% (1 section). The

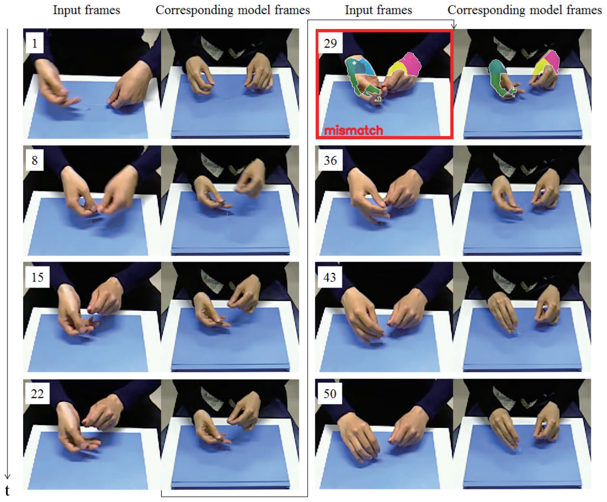

Figure 5. Examples of matched frames: The red rectangle shows inappropriate frame. The colored regions in the hands of the inappropriate frame show inappropriate parts.

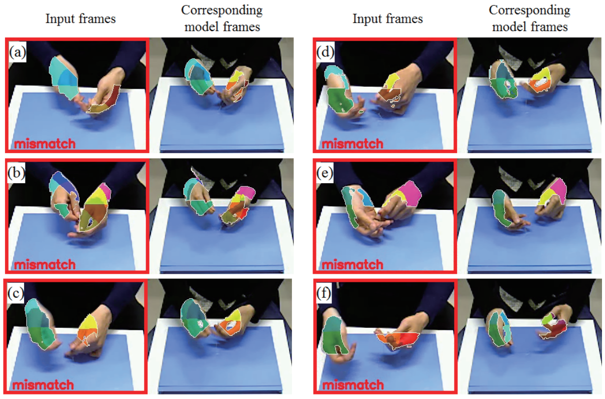

Figure 6. Inappropriate frames and parts: The colored regions show inappropriate parts. (a),(b),(c): Matched frames which has more difference between input and model left-hand's shapes and relative position of both hands. (d),(e),(f): Matched frames which has more difference between input and model hand's shapes of both hands.

false-negative was detected for interrupted motion section, and the false-positive was detected for re-holding strand motion section.

In case 2, we reuse the model sequences of case 1. The input sequence of case 2 actually consists of 18 times of knot-tying motion. We set $T_d$ to $50mm$ and $T_c$ to 0.45. 23 tying section candidates were detected. The correct recognition rate was 91% (21 sections), and the false recognition rate was 9% (2 sections).

### 6.2 Result of Matching Frame-to-Frame

Figure 4 shows the results of the frame-to-frame matching. Top of Figure 4 is the distance map of the features. The vertical axis is the frame number of the model sequence, and the horizontal one is that of the input sequence. The line on the distance map indicates the minimum cost path provided by CDP matching. The vertical broken lines correspond frame pairs in Figure 5. Bottom of Figure 4 shows the distances of the features on the path. Figure 5 shows the matched frames and the results of the temporal-spatial validation. We set $T_f$ to 0.33 and $T_s$ to 0.2.

Figure 6 shows examples of the inappropriate frames and parts. Our method detected both the differences of hand shapes and those of relative positions.

## 7 Conclusion

In this paper, we propose a system which temporally and spatially validate knot-tying procedure for self-training of surgical procedure. Our system recognizes the type of the knot-tying motion, and matches an input sequence and the 6 models frame-to-frame using modified CDP matching and SHOT feature. And we utilize SHOT feature for the spatial validation of the input procedures. Our future work includes detection of unknown motions.

## References

[1] Ethicon Inc.: "Knot Tying Manual, " Available at: http://academicdepartments.musc.edu/surgery/education/resident_info/supplement/suture_manuals/knot_tying_manual.pdf, 2005.

[2] I. Oikonomidis, N. Kyriazis, and A.A. Argyros: "Tracking the articulated motion of two strongly interacting hands, " Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012. pp. 1862–1869.

[3] K. K. Biswas, and S. K. Basu: "Gesture Recognition using Microsoft Kinect®, " Automation, Robotics and Applications (ICARA), 2011 5th International Conference on. IEEE, 2011. pp. 100–103.

[4] j. Yamato, J. Ohya, and K. Ishii: "Recognizing human action in time-sequential images using hidden markov model, " Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on. IEEE, 1992. pp. 379–385.

[5] T. Nishimura, T. Mukai, and R. Oka: "Non-monotonic continuous dynamic programming for spotting recognition of hesitated gestures from time-varying images, " Computer Vision-ACCV'98. Springer Berlin Heidelberg, 1997. pp. 734–741.

[6] Y. Li, J. Ohya, T. Chiba, R. Xu, and H. Yamashita: "Study of Recognizing Hand Actions from Video Sequences during Suture Surgeries Based on Temporally-Sectioned SIFT and Sliding Window Based Neural Networks, " IEICE Tech. Rep., vol. 113, no. 493, March 2014. pp. 151–156.

[7] J. Wang and Y. Wu: "Learning Maximum Margin Temporal Warping for Action Recognition, " Computer Vision (ICCV), 2013 IEEE International Conference on. IEEE, 2013. pp. 2688–2695.

[8] A. E. Johnson, and M. Hebert: "Using spin images for efficient object recognition in cluttered 3D scenes," Pattern Analysis and Machine Intelligence (PAMI), IEEE Transactions on 21.5 (1999). pp. 433–449.

[9] R. B. Rusu, N. Blodow, and M. Beetz: "Fast point feature histograms (FPFH) for 3D registration," Robotics and Automation, 2009. ICRA'09. IEEE International Conference on. IEEE, 2009. pp. 3212–3217

[10] F. Tombari, S. Salti, and L. D. Stefano: "Unique signatures of histograms for local surface description, " Computer Vision-ECCV 2010. Springer Berlin Heidelberg, 2010. pp. 356–369.

[11] D. Arthur and S. Vassilvitskii: "k-means++: The advantages of careful seeding, " Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, 2007. pp. 1027–1035.