Video Based Assessment of OSATS Using Sequential Motion Textures

Yachna Sharma[1], Vinay Bettadapura[1], Thomas Plötz[2], Nils Hammerla[2], Sebastian Mellor[2], Roisin McNaney[2], Patrick Olivier[2], Sandeep Deshmukh[3], Andrew McCaskie[4], and Irfan Essa[1]

[1] Georgia Institute of Technology, [2] Culture lab, School of Computing Science, Newcastle University, United Kingdom, [3] Newcastle University, United Kingdom, [4] Cambridge University, United Kingdom.

Abstract. We present a fully automated framework for video based surgical skill assessment that incorporates the sequential and qualitative aspects of surgical motion in a data-driven manner. We replicate Objective Structured Assessment of Technical Skills (OSATS) assessments, which provides both an overall and in-detail evaluation of basic suturing skills required for surgeons. Video analysis techniques are introduced that incorporate sequential motion aspects into motion textures. We also demonstrate significant performance improvements over standard bag-of-words and motion analysis approaches. We evaluate our framework in a case study that involved medical students with varying levels of expertise performing basic surgical tasks in a surgical training lab setting.

Keywords: Surgical skill assessment, motion texture, bag-of-words.

1 Introduction

Surgical skill development, i.e., gaining proficiency in surgical procedures and techniques, is an essential part of medical training for surgeons. Learning surgical skills is a time-consuming process, and requires expert supervision and evaluation, merged with extensive practice, throughout all stages of the training procedure. Manual assessment of surgical skills by experts, is the prevalent practice, and poses substantial time and resource problems to medical schools and teaching hospitals. The assessment criteria used are typically domain specific and often subjective where even domain experts do not always agree on the assessment scores [1].

To alleviate the problem of subjectivity in manual assessments, structured manual grading systems, such as the Objective Structured Assessment of Technical Skills (OSATS) [2] are used in medical schools. OSATS covers a variety of evaluation criteria: respect for tissue (RT), time and motion (TM), instrument handling (IH), suture handling (SH), flow of operation (FO), knowledge of procedure (KP) and overall performance (OP). Since manual assessments are time consuming and prone to variations, automated analysis of surgical motion has received attention in recent years.

The field of surgical skill assessment is dominated by automated recognition of surgical gestures for robotic minimally invasive surgery (RMIS) [3, 4]. Some recent works [5, 6] have also proposed skill assessment based on the OSATS criteria for general surgical training. However, these efforts focus on either short sequential surgical actions or capture qualitative motion dynamics via fine texture analysis of surgical motion without any temporal or sequential information. Automated assessment of different OSATS criteria using a common framework remains challenging as some criteria such as "respect for tissue" depend upon qualitative motion aspects while others such as "knowledge of procedure" depend upon the sequential motion aspects. Thus, to provide assessments on diverse OSATS criteria using a common framework, it is essential to capture both the qualitative and the sequential motion aspects.

In this work, we propose sequential motion texture (SMT) analysis technique. Our technique captures both the sequential and qualitative motion aspects. In SMT, we evaluate motion texture features in sequential time windows, which are automatically obtained from motion dynamics. We demonstrate that by incorporating both qualitative and sequential information in the OSATS skill assessment framework, classification accuracy improves substantially as compared to the techniques that model either qualitative or sequential motion information.

2 Background

The state-of-the-art in computerized surgical skill evaluation is dominated by RMIS using robots such as *da-Vinci* [7]. Several kinematic features (with over a hundred variables) such as torques, forces etc. from robotic arms and actuators are used for the analysis. Local approaches [7], proposed for RMIS, decompose a surgical task into simpler gestures followed by modeling each individual gesture using kinematic data. The second domain is assessment of skills in medical schools and teaching hospitals [5, 6, 8].

Recently, the attention has shifted towards video based analysis in both RMIS and teaching domains. Most of the video analysis methods classify different surgemes or surgical phases. For example, Haro *et al.* [3] and Zapella *et al.* [4], employed both kinematic and video data for RMIS surgery. They used linear dynamical systems (LDS) and bag-of-features (BoF) for surgical gesture (surgeme) classification in RMIS surgery.

BoF, also known as bag-of-words (BoW) model, is a state-of-the-art technique for video-based activity recognition and is typically constructed using visual codebooks derived from local spatio-temporal features [9]. However, BoW do not capture the underlying structural information, neither of causal nor of sequential type that is inherent by the ordering of the words. A-BoW [5] attempts to overcome this limitation by modeling the motion as short sequences of events and encoding the temporal and structural information into BoW models. With the A-BoW technique, higher classification accuracy is reported as compared to standard BoW technique [5]. Some recent efforts have also focused on qualitative motion aspects instead of sequential information.

Most of the works on surgical skill assessment [3, 4, 10, 11] relate to either robot assisted or thoracoscopic/laparoscopic surgeries. Our work is different from RMIS works with respect to application domain, type of skill assessments and techniques. Our goal is to provide automated assessment in a general surgical training lab and we do not use robotic kinematic data. Instead of overall expertise classification into novice, intermediate and expert levels, we provide these assessments based on the standard OSATS criteria used in medical training. Although, it would be interesting to incorporate sequential information via HMMs as in [10, 11], in this paper, we do not focus on gesture based skill assessment. In [5], Bettadapura et. al have compared A-BoW against HMMs and they show that A-BoW is superior to HMMs. In this paper, we compare to A-BoW and demonstrate that our technique is superior to A-BoW for OSATS skill classification. So, we indirectly show that SMT works better than both A-BoW and HMMs for OSATS assessment.

Our work is related to [6] since we also use motion textures to represent qualitative motion aspects, however we extend the basic motion texture (MT) technique to include sequential information automatically in a data-driven manner. Only few works [5, 6] have reported automated OSATS assessments. We compare our SMT approach with these works and demonstrate that by including both sequential and qualitative motion aspects, OSATS skill classification accuracy improves substantially.

3 Framework for surgical skill assessment

In our skill assessment framework, the input to the system is a video recording of a trainee performing suturing task and the output is an automated skill assessment (expertise level) according to the most common and thus most relevant seven OSATS criteria. We achieve this goal by first computing the sequential motion texture (SMT) features to encode both the qualitative and sequential motion aspects, followed by feature selection and classification.

3.1 Sequential motion texture (SMT)

Our SMT technique involves following steps: 1) computing low-level motion features; 2) learning motion classes (corresponding to moving entities); 3) computing data-driven time windows, and 4) sequential encoding of motion dynamics. Figure 1 gives an overview of the proposed procedure.

We extend the basic motion texture (MT) analysis approach [6] to incorporate sequential information and demonstrate that doing so results in improved classification accuracy for all seven OSATS criteria. The main advantages of motion texture analysis is its view-independent representation via frame kernel matrices (also known as self-similarity matrix (SSM) [12]), its ability to handle high dimensional data (the frame kernel matrix will be $N \times N$, regardless of time series dimensionality, where N is the length of time series or number of frames in the video), and its ability to encode qualitative motion aspects via texture

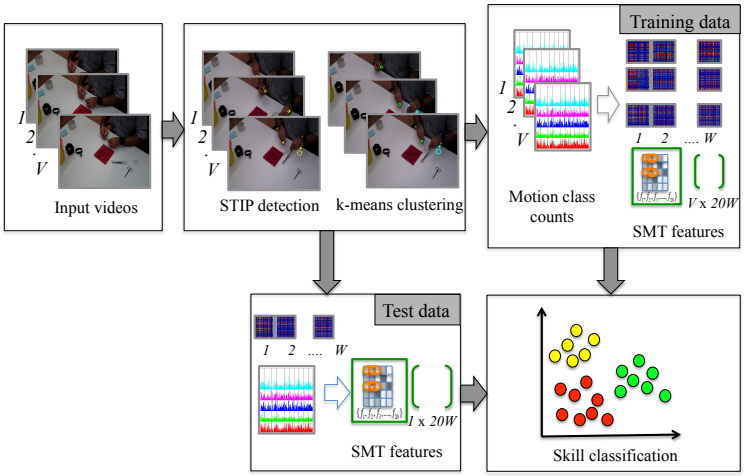

Fig. 1. Framework for sequential motion texture analysis.

analysis of motion dynamics. We exploit these basic characteristics and by data-driven time windowing, we also achieve inclusion of sequential motion aspects. In the following, we will discuss the technical details of our SMT framework.

Step 1) Motion feature extraction: To obtain the frame kernel matrix, first we detect the spatio-temporal interest points (STIP) using Laptev detector. We use Laptev's STIP implementation[1], with default parameters and with the sparse feature detection mode. We compute the *HOG* (Histogram of Oriented Gradients) and *HOF* (Histogram of Optical Flow) on a 3D video patch around each detected STIP to get a 162 element HOG-HOF descriptor as described in [9].

Step 2) Learning motion classes: We collect all the detected STIPs and their corresponding HOG-HOF descriptors from two videos of an expert surgeon. We classify the STIPs into k distinct clusters by applying k-means clustering to the HoG-HoF descriptors.

Step 3) Classification of STIPs into motion classes: Each cluster of points (learnt in step 2) represents a distribution for a particular motion class in the data. We assign the STIPs from remaining videos to each of the learnt motion distribution based on minimum Mahalanobis distance of a given STIP point from the cluster distribution. The two expert videos in step 2 are only used to learn the motion class clusters. We do not use them for evaluating classification accuracy.

Step 4) Computing motion class counts: We process each video to compute motion class counts for each of the k classes in each frame. We represent these counts in a $k \times N$ matrix $\mathbf{X}$, where N is the number of frames in the video. Each element in $\mathbf{X}$, $x(p, q)$, represents the number of STIP points in qth frame and belonging to the pth motion class.

[1] http://www.di.ens.fr/%7Elaptev/download.html#stip

Step 5) Computing data-driven time windows: We divide the time frequency matrix $\mathbf{X}$ into temporal windows such that each window contains equal proportions of the STIPs corresponding to the largest motion class in a given video. For example, if the largest motion class has 1,000 STIPs in the whole video, then the time series can be divided into $W = 10$ equally sized windows with approximately 100 STIPs in each bin. Using equally sized bins; we group the motion energy into equivalent segments (or motion bursts) that replicate the repetitive and procedural behavior of surgical motion.

Step 6) Computing sequential motion texture features: For each time window, we calculate the frame kernel matrix $\mathbf{K_w}$ given by

$$\mathbf{K_w} = \phi(\mathbf{X_w})^T \phi(\mathbf{X_w}), \tag{1}$$

where each entry in $\mathbf{K_w}$ defines similarity between the motion class counts in the ith and jth frames (x_{wi} and x_{wj}) of the wth time window using a kernel function $\phi(x_{wi})^T \phi(x_{wj})$. We use the Gaussian kernel function given by $\kappa_{ij} = \exp(-\frac{\|x_{wi}-x_{wj}\|^2}{2\sigma^2})$ where σ is the standard deviation.

To encode the qualitative motion dynamics in each time window, we apply Gray Level Co-occurrence Matrices (GLCM) texture analysis. For each $\mathbf{K_w}$, we compute $L \times L$ dimensional GLCM, where L is the number of gray levels. We compute the GLCMs for eight directions ($0^\circ - 360^\circ$ in steps of 45° at a spatial offset of 1 pixel. After averaging (and normalizing) over the GLCM, we compute twenty GLCM based motion texture features as in [6] and proposed in [13–16].

We encode the sequential motion information by adding the GLCM features sequentially (corresponding to the order of the windows in the time series) to obtain a $20W$ feature vector for each video, where W is the number of windows. Due to equal proportions of STIPs in each time window, the windows with more movement will be shorter in duration as compared to the windows with less movement. Despite varying window duration, the qualitative motion aspect for each window is encoded into twenty texture features which when ordered corresponding to window locations, also captures the sequential variation of motion textures. This is important because, an expert surgeon might finish a task faster as compared to a novice surgeon. However, their overall motion dynamics can be represented equivalently using same number of windows (and same number of SMT features) even though the overall task duration and the actual STIP counts are different.

3.2 Feature selection and skill classification

Some of the GLCM texture statistics are highly correlated with one another and may be redundant. Also, some features might be noisy and irrelevant for the skill classification task. In addition, the MT texture analysis yields a 20-element feature vector while SMT has $(20 \times W)$-element feature vector. To derive skill relevant features and to compensate for the effect of more features (overfitting) in SMT as compared to MT, we perform feature selection for both MT and SMT. We use Sequential Forward Feature Selection (SFFS) [17] to select a

subset of relevant features for each OSATS criteria. We use a Nearest-Neighbor (NN) classifier with cosine distance metric as a wrapper function for SFFS and select the feature subset with minimum classification error in leave-one-out cross-validation (LOOCV). Other classification algorithms can also be used. However, we want to compare with BoW and A-BoW and thus use same classifier (i.e. 1-NN with cosine distance metric) as reported in [5].

We compared our method with state-of the-art BoW models (built directly from the HoG-HoF descriptors), that are typically used for video-based action recognition [9] and have also been used for surgical gesture recognition [3, 4]. We also compare with the motion texture (MT) and A-BoW techniques [5, 6].

4 Experimental Evaluation

To test our SMT technique and compare with published OSATS works, we used the same data as described in [5, 6]. We briefly describe the data acquisition, expertise level of the participants in our case study and our experiments to demonstrate the significance of including both sequential and qualitative motion aspects for automated OSATS assessments.

4.1 Data Acquisition

Video data was collected from sixteen participants. Every participant performed suturing activities involving tasks such as stitching, knot tying, etc. thereby using a needle-holder, forceps and the tissue suture pads. These training sessions were recorded using a standard video camera (50fps, 1280×720 pixels), which was mounted on a tripod. Fifteen participants performed two sessions of a suturing task. Each session was recorded in a separate video. An expert surgeon also performed three sessions giving a total of thirty-three videos. The average duration of the videos is 18 minutes. Each subject performed the sub-tasks in order involving a knot tying and running suturing.

The expert surgeon provided ground truth annotation based on the OSATS scoring scheme. We group the participants into three categories according to their expertise: novice (OSATS score ≤ 2), intermediate ($2 <$ OSATS score ≤ 3.5) and expert ($3.5 <$ OSATS score ≤ 5). With availability of more samples in future, the number of classes could be increased to five. However, with our small sample size and few samples with very low and high OSATS scores, grouping the participants into three categories ensures that we have sufficient samples for each category. Table 1 shows the number of videos used in our study corresponding to three expertise levels for each OSATS criteria. For example, we have 9, 15 and 7 subjects (novice, intermediate, and expert respectively) for the "time and motion" OSATS criterion. For "suture handling", on the other hand, we have 10, 15 and 6 participants at novice, intermediate, and expert levels respectively. Since, a given participant may not be an expert (or novice) on all OSATS criteria, individual assessment of each OSATS criteria is essential for training and feedback.

Table 1. Number of samples for different expertise levels

	RT	TM	IH	SH	FO	KP	OP
Novice	2	9	8	10	3	8	6
Intermediate	14	15	16	15	16	9	17
Expert	15	7	7	6	12	14	8

Abbreviations: RT: Respect for Tissue, TM: Time and Motion, IH: Instrument Handling, SH: Suture Handling, FO: Flow of Operation, KP: Knowledge of Procedure, OP: Overall Performance.

4.2 OSATS skill assessment

We present the results using MT and SMT techniques as percentage of correctly classified videos using seven nearest neighbor (NN) classifiers trained for each OSATS criteria. All results are compared against the ground truth provided by the expert surgeon. We select the parameters (number of gray levels in GLCM L, number of motion classes k, and time windows W in SMT) using standard grid search. First, we briefly describe effect of the time windowing and feature selection on classification accuracy.

Time windowing and feature selection: Figure 2(a) shows sample motion class counts for $k = 5$ clusters grouped into $W = 10$ time windows for SMT. Note that some time windows (*e.g.* second and fourth from left) may be longer in duration if there is less motion (quantified in terms of STIP class counts for the largest motion class (cyan)) in these windows. Despite varying motion and duration of videos, SMT features encode both the qualitative and sequential motion in a given video into a fixed size feature vector of dimension $20W$, where W is the number of windows. It is important to note that time windowing does segment the time series, however, the segments obtained are not related or identified as gestures. We just incorporate the qualitative motion in each segment sequentially into a feature vector.

To test the level of sequential granularity, we vary the number of time windows. We also want to test if varying the number of selected features improves classification accuracy as in general classification techniques. Figure 2(b) shows the effect of varying the number of selected features and number of time windows on average (over seven OSATS criteria) classification accuracy. With only two windows, lower classification rate is observed as compared to higher classification rates with increasing number of windows and selected features. With $W > 6$, and $f_s > 10$, where f_s is the number of selected features, we observe that the average classification rate is greater than 85%. Thus, increased sequential granularity and including more features improves classification accuracy for all OSATS criteria.

Comparison with state-of-the-art techniques: Table 2 shows the results using different techniques. Note that BoW and A-BoW works [5] have used sixty-three videos since they used both the long range and close up videos for each participant. We used only long-range videos to ensure that the moving entities (hands, instruments etc.) exist in most of the frames. By using only thirty-one videos, we have less training data as compared to [5].

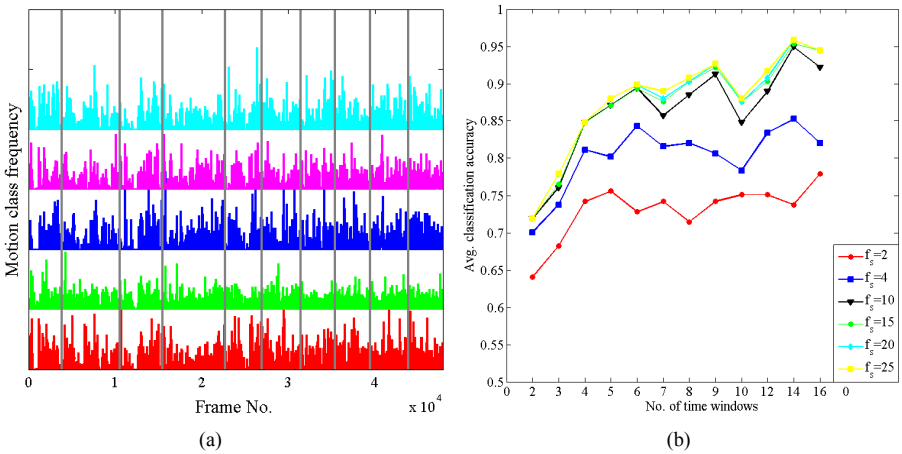

Fig. 2. (a) Motion class counts for with W=10 time windows for k=5 motion classes. Note that the time windows are of varying duration depending on the motion counts; (b) Effect of varying the number of selected features (f_s) and time windows.

A-BoW captures the temporal and sequential motion aspects and performs better than standard BoW. MT (with SFFS feature selection) captures the qualitative motion aspects, which results in higher classification accuracy of 83.8% (an increase of 10% from A-BoW) for qualitative OSATS criteria such as "respect for tissue". However, for sequential OSATS such as "knowledge of procedure", A-BoW performs better than both MT and standard BoW. For "time and motion", MT approach performs slightly better with 80.6% correctly classified videos (an increase of 6% over A-BoW) possibly due to finer analysis of motion dynamics. For other OSATS, both A-BoW and MT show comparable performance but better than standard BoW. SMT, without feature selection provides comparable performance as obtained with other methods, however SMT with feature selection significantly outperforms all previous methods.

Since feature selection is used for both MT and SMT (Table 2, column 2 and 6), the improved performance of SMT is due to encoding of sequential information into SMT features. Figure 3 shows the confusion matrices corresponding to SMT (SFFS) results in Table 2 (column 6). Despite varying expertise for different OSATS criteria, most of the participants are correctly classified. The classification accuracy for the expert level is lower than novice and intermediate skill levels, especially for the IH, FO, OP criteria. It is known that experts may not use all the steps in a task as reported in [18] and might develop their own style for performing a specific task.

5 Discussion and future work

In this work, we demonstrated the significance of including both qualitative and sequential motion aspects for automated OSATS assessments. Our results show

Table 2. Percentage of correctly classified videos using different techniques

OSATS	MT (SFFS)	BoW	A-BoW	SMT	SMT (SFFS)
Respect for Tissue	83.8% (26/31)	66.6% (42/63)	73.0 (46/63)	70.9% (22/31)	**96.7% (30/31)**
Time and Motion	80.6% (25/31)	50.7% (32/63)	74.6% (47/63)	80.6% (25/31)	**90.3% (28/31)**
Instrument Handling	70.9% (22/31)	50.7% (32/63)	68.2% (43/63)	70.9% (22/31)	**83.8% (26/31)**
Suture Handling	74.1% (23/31)	69.8% (44/63)	73.0% (46/63)	61.2% (19/31)	**83.8% (26/31)**
Flow of Operation	70.9% (22/31)	49.2% (31/63)	66.6% (42/63)	64.5% (20/31)	**83.8% (26/31)**
Knowledge of Procedure	61.2% (19/31)	60.3% (38/63)	80.9% (51/63)	70.9% (22/31)	**93.5% (29/31)**
Overall Performance	74.1% (23/31)	52.3% (33/63)	71.4% (45/63)	77.4% (24/31)	**93.5% (29/31)**

Abbreviations: OSATS: Objective Structured Assessment of Technical Skills, MT: Motion texture, BoW: Bag-of-Words, A-BoW: Augmented Bag-of-Words, SMT: Sequential Motion Texture SFFS: Sequential Forward Feature Selection (SFFS).

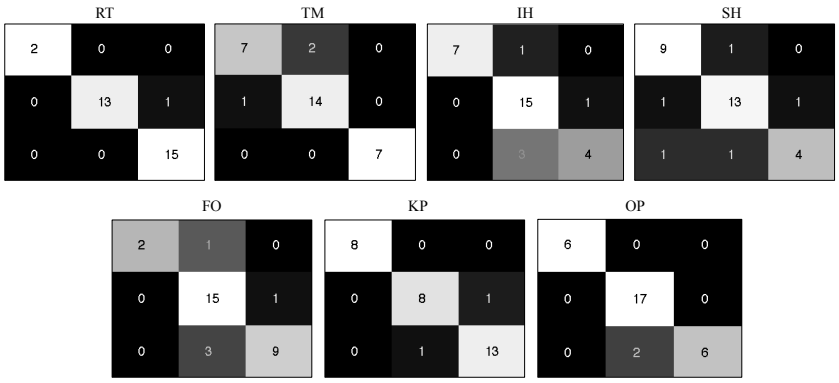

Fig. 3. Confusion matrices for seven OSATS criteria using SMT with SFFS feature selection (Table 2, column 6)

that SMT approach outperforms previously proposed techniques for video-based OSATS assessment of surgical skills (MT, BoW, A-BoW).

We have not correlated the motion information in time windows with surgical gestures. However, our approach can be extended to develop data-driven gesture vocabularies by correlating the time windows with expert segmented gestures. Given the very encouraging assessment results of our case study, we believe that automatic OSATS assessment has the potential to have a positive impact on real-world training settings in medical schools and teaching hospitals.

References

1. Awad, S., Liscum, K., Aoki, N., Awad, S., Berger, D.: Does the subjective evaluation of medical student surgical knowledge correlate with written and oral exam performance? Journal of Surgical Research **104**(1) (2002) 36–39
2. Martin, J., Regehr, G., Reznick, R., MacRae, H., Murnaghan, J., Hutchison, C., Brown, M.: Objective structured assessment of technical skill (osats) for surgical residents. British Journal of Surgery **84**(2) (1997) 273–278
3. Haro, B.B., Zappella, L., Vidal, R.: Surgical gesture classification from video data. In: MICCAI 2012. Springer (2012) 34–41
4. Zappella, L., Béjar, B., Hager, G., Vidal, R.: Surgical gesture classification from video and kinematic data. Medical Image Analysis (2013)
5. Bettadapura, V., Schindler, G., Plötz, T., Essa, I.: Augmenting bag-of-words: Data-driven discovery of temporal and structural information for activity recognition. In: CVPR. (2013)
6. Sharma, Y., Plötz, T., Hammerla, N., Mellor, S., Roisin, M., Olivier, P., Deshmukh, S., McCaskie, A., Essa, I.: Automated surgical OSATS prediction from videos. In: ISBI, IEEE (2014)
7. Lin, H., Hager, G.: User-independent models of manipulation using video contextual cues. In: International Workshop on Modeling and Monitoring of Computer Assisted Interventions (M2CAI)-Workshop. (2009)
8. Moorthy, K., Munz, Y., Sarker, S.K., Darzi, A.: Objective assessment of technical skills in surgery. BMJ: British Medical Journal **327**(7422) (2003) 1032
9. Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C., et al.: Evaluation of local spatio-temporal features for action recognition. In: BMVC. (2009)
10. Varadarajan, B., Reiley, C., Lin, H., Khudanpur, S., Hager, G.: Data-derived models for segmentation with application to surgical assessment and training. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2009. Springer (2009) 426–434
11. Rosen, J., Hannaford, B., Richards, C.G., Sinanan, M.N.: Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills. Biomedical Engineering, IEEE Transactions on **48**(5) (2001) 579–591
12. Junejo, I., Dexter, E., Laptev, I., Pérez, P.: View-independent action recognition from temporal self-similarities. PAMI (2011)
13. Haralick, R., Shanmugam, K., Dinstein, I.: Textural features for image classification. Systems, Man and Cybernetics **3**(6) (1973) 610–621
14. Soh, L., Tsatsoulis, C.: Texture analysis of sar sea ice imagery using gray level co-occurrence matrices. Geoscience and Remote Sensing **37**(2) (1999) 780–795
15. Clausi, D.: An analysis of co-occurrence texture statistics as a function of grey level quantization. Canadian Journal of Remote Sensing **28**(1) (2002) 45–62
16. Dean, C.: Quantitative Description and Automated Classification of Cellular Protein Localization Patterns in Fluorescence Microscope Images of Mammalian Cells. PhD thesis, Carnegie Mellon University (1999)
17. Pudil, P., Novovičová, J., Kittler, J.: Floating search methods in feature selection. Pattern Recognition Letters **15**(11) (1994) 1119–1125
18. Reiley, C., Hager, G.: Decomposition of robotic surgical tasks: an analysis of subtasks and their correlation to skill. In: MICCAI. (2009)