Surgical Gesture Segmentation and Recognition

Lingling Tao, Luca Zappella, Gregory D. Hager, and René Vidal

Johns Hopkins University, 3400 North Charles Street, Baltimore, MD, 21218, USA

Abstract. Automatic surgical gesture segmentation and recognition can provide useful feedback for surgical training in robotic surgery. Most prior work in this field relies on the robot's kinematic data. Although recent work [1,2] shows that the robot's video data can be equally effective for surgical gesture recognition, the segmentation of the video into gestures is assumed to be known. In this paper, we propose a framework for joint segmentation and recognition of surgical gestures from kinematic and video data. Unlike prior work that relies on either frame-level kinematic cues, or segment-level kinematic or video cues, our approach exploits both cues by using a combined Markov/semi-Markov conditional random field (MsM-CRF) model. Our experiments show that the proposed model improves over a Markov or semi-Markov CRF when using video data alone, gives results that are comparable to state-of-the-art methods on kinematic data alone, and improves over state-of-the-art methods when combining kinematic and video data.

Keywords: surgical gesture segmentation, surgical gesture recognition, time series analysis, conditional random fields, structured output learning.

1 Introduction

Robotic minimally invasive surgery (RMIS) offers several advantages over traditional surgery, including better precision, smaller incisions, and quicker recovery time. However, reductions in the amount of one-on-one teaching [3], together with the steep learning curve of RMIS [4], advocate for the development of a new teaching paradigm where surgical skill is automatically assessed and timely feedback is automatically provided.

Advances in machine learning, computer vision, and speech processing can be exploited for this purpose. For instance, segmentation and recognition methods can be used to decompose a surgical task (e.g., suturing) into a sequence of gestures (e.g., grab needle, position needle, insert needle), and perform skill assessment based on how well a sequence of gestures is executed. As shown in [5], gesture recognition can also be used to accomplish shared or cooperative control tasks that are triggered based on what the user is doing. Hence, we could envisage robotic gesture recognition being used, for example, to trigger appropriate information displays. All these applications require segmentation and recognition of surgical gestures, which is the main focus of this paper.

Most of the prior work on automatic segmentation and recognition of surgical gestures relies on the kinematic observations of the robot's motion. The temporal evolution of such observations is typically modeled using a Hidden Markov Model (HMM), where each gesture corresponds to one or more states of the HMM and the transitions among consecutive gestures are modeled by the HMM transition probabilities. Different papers [6,7,8,9,10,11,12] use different models for the observations associated with

K. Mori et al. (Eds.): MICCAI 2013, Part III, LNCS 8151, pp. 339–346, 2013.

each gesture, including discrete HMMs, Gaussian HMMs, factor analyzed HMMs and Sparse HMMs. While generally successful, these methods rely mostly on *local cues* from a few frames, thus failing to capture *global cues* about the whole execution of a gesture. To address this issue, [11] uses Switched Linear Dynamical Systems (SLDSs), which model the dynamics of the whole execution of a gesture with an LDS. However, this comes at a steep computational cost because inference for SLDSs is intractable.

More recently, video-based solutions have drawn the attention of the research community. Most of the prior work focuses on the detection of the instruments used during surgery or in the operating room [13,14,15,16,17,18,19,20] using techniques such as dynamic time warping, support vector machines and HMMs. However, these techniques use only frame-level features, such as color, texture and shape-based cues. Moreover, the desired solution is a high-level recognition based on the presence of some tools, rather than fine-grained gesture recognition based on motion data.

To the best of our knowledge, the only existing works that use state-of-the-art computer vision algorithms for surgical gesture recognition from video data are [1,2]. These works show that video data can be as discriminative as kinematic data when appropriate features and algorithms are used. Specifically, [1,2] obtain very high gesture recognition rates using *global* bag-of-spatio-temporal features (BoSTF) and LDSs to model the whole execution of a gesture. However, this is possible only because the temporal segmentation of the video into gestures is assumed to be known. Recent work in computer vision addresses the joint segmentation and recognition of generic actions in videos using conditional Random fields (CRFs), where the sequence of gestures is obtained by minimizing the energy of the CRF. The work of [21] adopts a Markov CRF model whose energy depends on which objects are present in each frame and their interactions, while the work of [22] adopts a semi-Markov CRF model based on global features extracted from many frames. However, these methods have not been combined for joint gesture segmentation and recognition, nor have they been applied to surgical gestures.

In this paper, inspired by the use of graphical models in speech processing [23], we propose a combined Markov/semi-Markov conditional random field (MsM-CRF) model for joint segmentation and recognition of surgical gestures from kinematic and video data. Our MsM-CRF model captures both local cues (thanks to the Markov component) and global cues (thanks to the semi-Markov component). Moreover, we exploit the high accuracy of the BoSTF approach from [1,2] to model the unary potentials of the CRFs using two kinds of spatio-temporal features. Our experiments on a typical surgical training setup show that our model improves over a Markov or semi-Markov CRF model on video data alone, gives state-of-the-art results on kinematic data alone, and improves over state-of-the-art methods by combining kinematic and video data.

2 Joint Segmentation and Recognition of Surgical Gestures

2.1 MsM-CRF Model

Let $\mathcal{V} = \{I_t\}_{t=1}^{T}$ be a sequence of observations, where I_t represents the observations at frame t. Let $\mathcal{C} = \{1, \ldots, C\}$ be the set of possible gesture labels. Our goal is to find the sequence of frame-level gestures $\mathcal{Y} = \{Y_t^F\}_{t=1}^{T}$, where $Y_t^F \in \mathcal{C}$ denotes the gesture label at frame t. Since each gesture may span a few consecutive frames, we can divide

the time interval $[1, T]$ into M segments, where the i-th segment is $[t_{i-1}, t_i)$, such that the frame gesture label within a segment does not change. Here t_i is the last frame of i, $t_0 = 1$, and $t_M = T$. We can then define the sequence of segment-level gestures as $\{\mathbf{Y}_i^S\}_{i=1}^M$, where $\mathbf{Y}_i^S = Y_t$ for all $t \in i$. In what follows, we will use the superscripts F and S to denote variables at the frame and segment levels, respectively.

We represent $\mathcal{V}$ with a graphical model $\mathcal{G} = (\mathcal{N}^F, \mathcal{E}^F, \mathcal{N}^S, \mathcal{E}^S)$. In the case of CRFs, each node $N_t^F \in \mathcal{N}^F$ denotes a frame t, hence $|\mathcal{N}^F| = T$. In the case of semi-CRFs, each node $N_i^S \in \mathcal{N}^S$ denotes the collection of frames in segment i, hence $|\mathcal{N}^S| = M$. In both cases, an edge $e_j \in \mathcal{E}$ denotes the connection between consecutive nodes N_j and N_{j+1}. We model the conditional distribution of the sequence of labels $\mathcal{Y}$ given $\mathcal{V}$, with a Gibbs distribution: $p(\mathcal{Y}|\mathcal{V}) \propto \exp(E(\mathcal{Y}, \mathcal{V}))$, where $E(\mathcal{Y}, \mathcal{V})$ is an energy function defined on the cliques of $\mathcal{G}$. In our MsM-CRF model, we have:

$$\begin{aligned} E(\mathcal{Y}, \mathcal{V}) = \lambda^{FU} \sum_{t=1}^{T} \psi_t^{FU}(Y_t^F; \mathcal{V}) + \lambda^{FP} \sum_{t=1}^{T-1} \psi_{t,t+1}^{FP}(Y_t^F, Y_{t+1}^F; \mathcal{V}) + \\ \lambda^{SU} \sum_{i=1}^{M} \psi_i^{SU}(\mathbf{Y}_i^S; \mathcal{V}) + \lambda^{SP} \sum_{i=1}^{M-1} \psi_{i,i+1}^{SP}(\mathbf{Y}_i^S, \mathbf{Y}_{i+1}^S; \mathcal{V}), \end{aligned} \tag{1}$$

where ψ^{FU} and ψ^{FP} are the CRF unary and pairwise potentials, while ψ^{SU} and ψ^{SP} are the semi-CRF unary and pairwise potentials, each one weighted by its own λ factor.

CRF Unary. This potential gives the score of assigning a gesture label to a single frame. For kinematic data, the score is computed from the output of an SVM classifier, with an RBF kernel, trained for each possible gesture on the raw data of each frame. For video data, this score is obtained from the output of a classifier applied to a histogram of features extracted from a neighborhood of the frame. Specifically, during training the spatio-temporal video features are clustered by K-means to form a dictionary of visual words. Each frame is then represented with a histogram of words and these histograms are used to train an SVM classifier with a χ^2-RBF kernel. In both cases, the logarithm of the probability returned by regression of the SVM output is used as a unary score.

Semi-CRF Unary. This potential gives the score of assigning a gesture label to a segment, thereby capturing global features related to the overall gesture. For kinematic data, we train an SVM classifier with RBF kernel for each gesture on the average of the raw data within each segment. For video data, we represent each segment by the histogram of words accumulated over all the frames that correspond to the segment using the same dictionary of visual words described before. These histograms are then used to train an SVM classifier with χ^2-RBF kernel for each gesture. Hence, we use the logarithm of the probability returned by regression of the SVM output as our unary term. This way of computing the most likely label for each segment corresponds exactly to the approach followed in [1,2].

Spatio-temporal Features. We use two kinds of spatio-temporal features. The first one is a concatenation of histograms of oriented gradients (HOG) and histograms of optical flows (HOF) extracted from a cuboid centered around each STIP point [24]. Since STIP points tend to be sparse in space, we also use the dense features presented in [25], which

consist of HOG, HOF, and histograms of motion boundaries and velocities (in term of x and y coordinates) computed around dense trajectories.

CRF Pairwise. This potential captures the probability of switching from gesture label g_k to g_j when moving from one frame to the next. Since a gesture is composed of many frames, this potential encourages the frame labels to be temporally coherent. We capture the relationship between adjacent frames using the transition probability

$$P^F_{g_k,g_j} = \frac{\text{\# frames switching from } g_k \text{ to } g_j}{\text{\# frames with label } g_k} \tag{2}$$

computed from the training set. We then set $\psi^{FP}_{t,t+1}(Y^F_t, Y^F_{t+1}; \mathcal{V}) = \log(P^F_{Y^F_t, Y^F_{t+1}})$.

Semi-CRF Pairwise. This potential captures the probability of switching from gesture label g_k to g_j when moving from one segment to the next. Since each segment represents a single instance of a gesture, two consecutive segments should not have the same label. Thus, this potential encourages a switch from one gesture label to a different one. We capture the relationship between adjacent segments using the transition probability

$$P^S_{g_k,g_j} = \frac{\text{\# segments switching from } g_k \text{ to } g_j}{\text{\# segments with label } g_k} \tag{3}$$

computed from the training set. We then set $\psi^{SP}_{i,i+1}(\mathbf{Y}^S_i, \mathbf{Y}^S_{i+1}; \mathcal{V}) = \log(P^S_{\mathbf{Y}^S_i, \mathbf{Y}^S_{i+1}})$.

2.2 Efficient Inference and Learning

Inference. The energy in (1) can be re-written as $E(\mathcal{Y}, \mathcal{V}) = \mathbf{w}^\top \Psi(\mathcal{Y}; \mathcal{V})$, where

$$\mathbf{w} = \begin{pmatrix} \lambda_{FU} \\ \lambda_{FP} \\ \lambda_{SU} \\ \lambda_{SP} \end{pmatrix} \quad \text{and} \quad \Psi(\mathcal{Y}; \mathcal{V}) = \begin{pmatrix} \sum_{t=1}^{T} \psi^{FU}_t(Y^F_t; \mathcal{V}) \\ \sum_{t=1}^{T-1} \psi^{FP}_{t,t+1}(Y^F_t, Y^F_{t+1}; \mathcal{V}) \\ \sum_{i=1}^{M} \psi^{SU}_i(\mathbf{Y}^S_i; \mathcal{V}) \\ \sum_{i=1}^{M-1} \psi^{SP}_{i,i+1}(\mathbf{Y}^S_i, \mathbf{Y}^S_{i+1}; \mathcal{V}) \end{pmatrix}. \tag{4}$$

Given the MsM-CRF model parameters in $\mathbf{w}$ and a test video $\mathcal{V}$, we can perform joint gesture segmentation and recognition by solving the inference problem $\mathcal{Y}^* = \operatorname{argmax}_{\mathcal{Y}} E(\mathcal{Y}, \mathcal{V})$. One can show that the energy in (1) is equivalent to an energy that depends only on the segment labels $\{\mathbf{Y}^S_i\}$. The maximization of the resulting energy can be done by a Viterbi-like dynamic programming algorithm, as described in [22].

Learning. Given B training videos $\{\mathcal{V}_i\}_{i=1}^B$ and their corresponding labelings $\{\bar{\mathcal{Y}}_i\}_{i=1}^B$, we learn the parameters $\mathbf{w}$ using a method based on structural SVM [26]. Specifically, let us refer to any labeling of $\mathcal{V}_i$ that is different from $\bar{\mathcal{Y}}_i$ as a negative example, and denote the set of negative examples for a video $\mathcal{V}_i$ as $\mathcal{Y}^-_i$. Given $\mu > 0$, we learn the parameters $\mathbf{w}$ by solving the following optimization problem:

$$\{\mathbf{w}^*, \{\xi^*_i\}_{i=1}^B\} = \underset{\mathbf{w}, \{\xi_i\}_{i=1}^B}{\arg\min} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{\mu}{B}\sum_{i=1}^{B} \xi_i, \qquad \text{subject to} \tag{5}$$

$$\begin{aligned} &\text{(a) } \forall i = 1, \ldots, B,\ \forall \mathcal{Y} \in \mathcal{Y}^-_i,\ \mathbf{w}^\top \big(\Psi(\bar{\mathcal{Y}}_i; \mathcal{V}_i) - \Psi(\mathcal{Y}; \mathcal{V}_i)\big) \geq \ell(\bar{\mathcal{Y}}_i, \mathcal{Y}) - \xi_i \\ &\text{(b) } \forall i = 1, \ldots, B,\ \xi_i \geq 0 \quad \text{and} \quad \text{(c) } \mathbf{w} \geq \mathbf{0}. \end{aligned} \tag{6}$$

The intuition behind the first inequality is that we want the energy at the ground truth labeling $\mathbf{w}^\top \Psi(\bar{\mathcal{Y}}_i; \mathcal{V}_i)$ to be greater than the energy of any wrong labeling $\mathbf{w}^\top \Psi(\mathcal{Y}; \mathcal{V}_i)$ by the loss $\ell(\bar{\mathcal{Y}}_i, \mathcal{Y})$ while allowing some slack ξ_i. The loss function $\ell(\bar{\mathcal{Y}}_i, \mathcal{Y})$ measures the error in the labeling $\mathcal{Y}$ as the fraction of misclassified frames. Since the number of constraints is exponentially large, we use the cutting plane algorithm [27] to find $\mathbf{w}$.

3 Experiments

Data. We evaluate our approach using the dataset in [28], which contains three different surgical tasks, suturing (SU), needle passing (NP) and knot tying (KT), each performed by 8 surgeons with three different skill levels (expert, intermediate and novice). Each surgeon performs around 3 to 5 trials for each task, which gives 39 trials for SU, 26 trials for NP and 36 trials for KT. Each trial lasts, on average, 2 minutes, and both kinematic and video data are recorded at a rate of 30 frames per second. Kinematic data consists of 78 motion variables (positions, rotation angles, and velocities of the master/patient side manipulators), whereas video data is taken from the first person view point of the robot and consists of jpeg images of size 320×240. The data is manually labelled with 15 surgical gestures (e.g., reach needle, position needle, etc.) described in Figure 1.

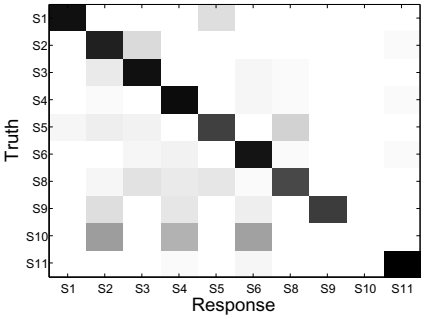
Setup. Following [1,2], we use two different test setups. The first one is the *leave-one-super-trial-out* (LOSO), where super trial i (i.e., trial i from each user) is held out for testing. The second, and more changeling setup, is the *leave-one-user-out* (LOUO), where all the trials from user i are held-out for testing. We compute the video CRF unary terms using a neighborhood of 25 frames, and the CRF and semi-CRF unaries using a dictionary of 300 words. To speed up the computation of the Viterbi-like algorithm, we perform inference every 10 frames, and assume that the maximum length of a segment is 400 frames. We compare the performance of the proposed MsM-CRF model with that of a CRF and a semi-CRF model. Note that the semi-CRF model with STIP features can be seen as an extension of the method presented in [1,2] to the case where the segmentation is unknown. The average percentage of correctly classified frames is shown in Table 1. For each technique, the type of features used is indicated within parenthesis. For MsM-CRF, the first feature refers to the CRF unaries and the second one to the semi-CRF unaries. For instance, MsM-CRF(kin-STIP) means that kinematic data are used for the CRF unary, while STIP features are used for the semi-CRF unary.

Results on Video Data. Notice that the combination of local and global features from video data used by the MsM-CRF model always improves over the CRF and semi-CRF models, for all tasks and test setups. Notice also that using dense features leads to better results than using of STIP features. This should be related to the fact that the features of [25] include motion boundaries and velocities in addition to HOG and HOF. Finally, notice that the results from [1] are 2-12% better, but they assume known segmentation.

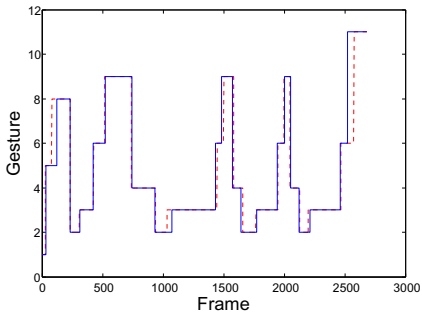
Results on Kinematic Data. In this case, the CRF model outperforms the semi-CRF model. Arguably, this is because the feature used in the unary term of the semi-CRF model (average of the data in a temporal window) is too simple. This is also reflected on the MsM-CRF results, which are similar to those of the CRF. The results of sparse-HMMs [12] are better than those of MsM-CRF(dense-dense) in the LOSO setup. However, in the more challenging LOUO setup MsM-CRF(dense-dense) seems to generalize

Table 1. Average percentage of correctly classified frames on the dataset in [28]. Best results for methods that do not assume known segmentation are highlighted in boldface.

	Method	LOSO			LOUO		
		SU	KT	NP	SU	KT	NP
Video	1) CRF(STIP)	70.86%	68.33%	55.12%	61.12%	62.63%	52.58%
	2) semi-CRF(STIP)	67.91%	67.07%	56.49%	51.71%	40.04%	45.03%
	3) MsM-CRF(STIP-STIP)	73.32%	70.95%	63.31%	66.28%	66.53%	58.85%
	4) CRF(dense)	76.51%	69.16%	62.23%	68.80%	60.17%	54.52%
	5) semi-CRF(dense)	65.83%	44.82%	56.22%	59.41%	41.46%	46.89%
	6) MsM-CRF(dense-dense)	79.04%	72.04%	68.81%	71.76%	66.94%	**60.39%**
	7) BoSTF(STIP,known segment.) [1]	84.87%	84.03%	72.16%	75.72%	79.05%	61.73%
Kinematic	8) CRF(kin)	81.62%	81.06%	74.56%	68.65%	67.38%	46.44%
	9) semi-CRF(kin)	63.20%	39.20%	54.15%	62.24%	44.28%	38.36%
	10) MsM-CRF(kin-kin)	80.99%	79.39%	74.85%	69.03%	64.28%	52.39%
	11) sparse-HMM [12]	81.1%	**82.6%**	76.1%	67.8%	65.7%	59.3 %
Mix	12) MsM-CRF(kin-STIP)	82.49%	80.50%	76.41%	70.09%	68.43%	54.41%
	13) MsM-CRF(kin-dense)	**82.81%**	81.10%	**76.82%**	**72.60%**	**68.83%**	57.08%



(a) Confusion matrix, darker colors mean higher percentage.



(b) Ground-truth "- -" and prediction "-" for one suturing trial.

Fig. 1. Results for the MsM-CRF(kin-dense) algorithm on the Suturing-LOSO setup. S1 reaching for needle with right hand. S2 positioning needle. S3 pushing needle through tissue. S4 transferring needle from left to right. S5 moving to center with needle in grip. S6 pulling suture with left hand. S7 pulling suture with right hand. S8 orienting needle. S9 using right hand to help tighten suture. S10 loosening more suture. S11 dropping suture at end and moving to end points. S12 reaching for needle with left hand. S13 making 'C' loop around right hand. S14 right hand reaches for suture. S15 both hands pull.

better. Overall, our method is able to achieve state-of-the-art results on kinematic data, and its performance on video data is comparable to that on kinematic data.

Results on Video and Kinematic Data. We also exploit the flexibility of the MsM-CRF framework to combine both kinematic data (in the CRF-unaries) and video data (in the semi-CRF unaries). The resulting MsM-CRF(kin-dense) model performs almost always better than any other technique that uses only kinematic or video data.

Fig. 1(a) shows the confusion matrix for the MsM-CRF(kin-dense) method on the Suturing-LOSO setup, while Fig. 1(b) shows the predicted and ground-truth sequence of gestures for one suturing trial. From these figures it is possible to appreciate the quality of the results produced. Note that gesture 10 is never classified correctly. This is due to the fact that this gesture appears very rarely, hence, it is not possible to learn a robust classifier. Fig. 1(b) shows that most of the errors appear around the switching times. Typically, the prediction switches either too early or with some delay. This might be due to the fact that for speed convenience we perform inference every 10 frames, hence, the results could be slightly improved by sacrificing some computational time.

Computing Time. A final note on the computational complexity. For completing one held-one-out experiment, in which usually around 30 trials are used for training and 8 for testing, the training stage took around 4 hours with a Matlab implementation on a x86@3.33GHz processor, and testing for one trial usually required around 1 minute.

4 Conclusions

We have proposed a combined Markov/semi-Markov CRF model for temporal gesture segmentation and recognition. Our model can capture local features and interactions between frames, as well as global characteristics of each gesture and interactions between gestures. We have shown on a typical surgical dataset that the MsM-CRF model always improves with respect to the CRF or semi-CRF frameworks used alone. We have also observed that the MsM-CRF model based on dense features is more robust in the LOUO setup. Moreover, thanks to the flexibility of the MsM-CRF, we were able to present a hybrid solution where kinematic and video data were both used. Such a hybrid solution achieved results that are similar or superior to those of state-of-the-art algorithms.

Acknowledgments. Work funded by Sloan Foundation, and NSF 0931805 and 0941362.

References

1. Béjar Haro, B., Zappella, L., Vidal, R.: Surgical gesture classification from video data. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part I. LNCS, vol. 7510, pp. 34–41. Springer, Heidelberg (2012)
2. Zappella, L., Béjar, B., Hager, G., Vidal, R.: Surgical gesture classification from video and kinematic data. Medical Image Analysis (2013)
3. Barden, C., Specht, M., McCarter, M., Daly, J., Fahey, T.: Effects of limited work hours on surgical training. Obstetrical & Gynecological Survey 58(4), 244–245 (2003)
4. Lenihan, J., Kovanda, C., Seshadri-Kreaden, U.: What is the learning curve for robotic assisted gynecologic surgery? J. of Minimally Invasive Gynecology 15(5), 589–594 (2008)
5. Padoy, N., Hager, G.D.: Human-machine collaborative surgery using learned models. In: IEEE Conference on Robotics and Automation, pp. 5285–5292 (2011)
6. Rosen, J., Hannaford, B., Richards, C., Sinanan, M.: Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills. IEEE Trans. Biomedical Eng. 48(5), 579–591 (2001)
7. Reiley, C.E., Hager, G.D.: Task versus subtask surgical skill evaluation of robotic minimally invasive surgery. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009, Part I. LNCS, vol. 5761, pp. 435–442. Springer, Heidelberg (2009)
8. Loukas, C., Georgiou, E.: Surgical workflow analysis with Gaussian mixture multivariate autoregressive (GMMAR) models: a simulation study. Computer Aided Surgery (2013)

9. Rosen, J., Solazzo, M., Hannaford, B., Sinanan, M.: Task decomposition of laparoscopic surgery for objective evaluation of surgical residents' learning curve using hidden Markov model. Computer Aided Surgery 7(1), 49–61 (2002)
10. Leong, J.J.H., Nicolaou, M., Atallah, L., Mylonas, G.P., Darzi, A., Yang, G.Z.: HMM assessment of quality of movement trajectory in laparoscopic surgery. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4190, pp. 752–759. Springer, Heidelberg (2006)
11. Varadarajan, B.: Learning and inference algorithms for dynamical system models of dextrous motion. PhD thesis, Johns Hopkins University (2011)
12. Tao, L., Elhamifar, E., Khudanpur, S., Hager, G.D., Vidal, R.: Sparse hidden Markov models for surgical gesture classification and skill evaluation. In: Abolmaesumi, P., Joskowicz, L., Navab, N., Jannin, P. (eds.) IPCAI 2012. LNCS, vol. 7330, pp. 167–177. Springer, Heidelberg (2012)
13. Wolf, R., Duchateau, J., Cinquin, P., Voros, S.: 3D tracking of laparoscopic instruments using statistical and geometric modeling. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part I. LNCS, vol. 6891, pp. 203–210. Springer, Heidelberg (2011)
14. Mung, J., Vignon, F., Jain, A.: A non-disruptive technology for robust 3D tool tracking for ultrasound-guided interventions. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part I. LNCS, vol. 6891, pp. 153–160. Springer, Heidelberg (2011)
15. Richa, R., Bó, A.P.L., Poignet, P.: Robust 3D visual tracking for robotic-assisted cardiac interventions. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010, Part I. LNCS, vol. 6361, pp. 267–274. Springer, Heidelberg (2010)
16. 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. (eds.) MICCAI 2010, Part III. LNCS, vol. 6363, pp. 400–407. Springer, Heidelberg (2010)
17. 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)
18. 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)
19. Lalys, F., Riffaud, L., Bouget, D., Jannin, P.: An application-dependent framework for the recognition of high-level surgical tasks in the OR. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part I. LNCS, vol. 6891, pp. 331–338. Springer, Heidelberg (2011)
20. Lalys, F., Riffaud, L., Bouget, D., Jannin, P.: A framework for the recognition of high-level surgical tasks from video images for cataract surgeries. IEEE Transactions on Biomedical Engineering 59(4), 966–976 (2012)
21. Fathi, A., Farhadi, A., Rehg, J.M.: Understanding egocentric activities. In: IEEE International Conference on Computer Vision, pp. 407–414 (2011)
22. Shi, Q., Cheng, L., Wang, L., Smola, A.J.: Human action segmentation and recognition using discriminative semi-markov models. Int. Journal of Computer Vision 93(1), 22–32 (2011)
23. Andrew, G.: A hybrid markov/semi-markov conditional random field for sequence segmentation. In: Conf. on Empirical Methods in Natural Language Processing, pp. 465–472 (2006)
24. Laptev, I.: On space-time interest points. International Journal of Computer Vision 64(2-3), 107–123 (2005)
25. Wang, Y., Tran, D., Liao, Z.: Learning hierarchical poselets for human parsing. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1705–1712 (2011)
26. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. J. of Machine Learning Research 6, 1453–1484 (2005)
27. Joachims, T., Finley, T., Yu, C.N.J.: Cutting-plane training of structural svms. Machine Learning 77(1), 27–59 (2009)
28. Reiley, C.E., Lin, H.C., Varadarajan, B., Vagolgyi, B., Khudanpur, S., Yuh, D.D., Hager, G.D.: Automatic recognition of surgical motions using statistical modeling for capturing variability. In: Medicine Meets Virtual Reality, pp. 396–401 (2008)