Learning Shared, Discriminative Dictionaries for Surgical Gesture Segmentation and Classification

Shahin Sefati[1], Noah J. Cowan[2], and René Vidal[1]

[1] {shahin,rvidal}@cis.jhu.edu, Center for Imaging Science
[2] ncowan@jhu.edu, Mechanical Engineering,
Johns Hopkins University, Baltimore, MD 21218, USA

Abstract. We propose a surgical gesture segmentation and classification method based on shared, discriminative, sparse dictionary learning, which can be used to effectively analyze complex surgical gestures recorded by the da Vinci robotic surgical system. Rather than learning a separate dictionary for each gesture in an unsupervised manner, we propose an algorithm for jointly learning a common overcomplete dictionary for all gestures together with a multi-class linear support vector machine for classifying each gesture. Experiments on the JHU-ISI Gesture and Skill Assessment Dataset (JIGSAWS), which contains the motions from three surgical tasks, reveal that the proposed method performs on par or better than state-of-the-art methods based only on kinematic data.

Keywords: surgical gesture classification; surgical gesture segmentation, sparse dictionary learning; time series analysis

1 Introduction

The advent of robotic minimally invasive surgery (RMIS) has enabled the capture of rich, high-fidelity time-series data during the execution of surgical tasks by experts and trainees alike. The automatic segmentation and classification of surgical gestures from such data holds transformative promise, from surgical skill assessment to prognostic outcome prediction. Prior work on surgical gesture classification has been mainly based on kinematic data recorded by a surgical robot. These data include position of the robot's tools, the robot joint angles, and translational and rotational velocities of both joints and tooltips. Many prior studies have quantified and analyzed high-level parameters such as time to completion of a task [5, 8], distance travelled [8], and force and torque signatures [16, 23, 8] for classification of surgical tasks. These approaches are generally easy to implement, but they do not take advantage of the fact that a surgical task such as suturing can be decomposed into a number of lower-level surgical gestures.

In recent years, several studies have attempted to provide a more detailed description of surgical tasks by decomposing them into a set of pre-defined atomic gestures called *surgemes* [17, 15, 22, 21, 11]. Examples of different surgemes in a surgical task such as suturing include "reach needle", "insert needle," and "pull suture." (see Fig. 1). Automatic segmentation and classification of surgemes can facilitate automatic skill classification based on how well each of the surgical

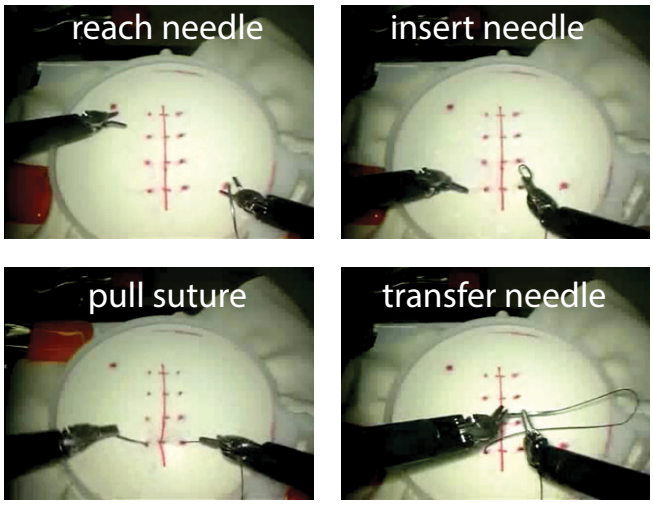

Fig. 1. Examples of four surgical surgemes in the suturing task.

gestures is performed in a particular surgical task. A number of methods have been proposed for classification [24, 2]. While these approaches perform very well in classifying gestures, they assume that the data is already segmented. On the other hand, a number of statistical models—including Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs)—have been proposed for the joint segmentation and classification of surgical gestures. The most widely used statistical models are HMMs and their variations [17, 11, 15, 21, 19, 20], which differ from each other on how the observations associated with each gesture are modeled. In particular, a Sparse-HMM (S-HMM) uses sparse dictionaries to model the HMM observations (as the name suggests). While this model achieves solid performance in segmentation and classification of surgical gestures, a separate dictionary of atomic surgical motions is learned for each surgical gesture.

In this paper, we propose a new sparse representation approach, called *Shared Discriminative Sparse Dictionary Learning* (SDSDL), where instead of learning a separate dictionary for each gesture, we jointly learn a common dictionary for all possible surgical gestures together with the parameters of a multi-class linear support vector machine (SVM). In this way, the learned dictionary of atomic motions is shared across all gestures, which results in more compact dictionaries. In addition, the dictionary is more discriminative, as it is learned together with the gesture classifiers. [3] The classification scores of SDSDL are then integrated into an HMM for gesture segmentation and classification. We test our proposed SDSDL method on the kinematic data from the JHU-ISI Gesture and Skill Assessment Dataset (JIGSAWS) [7]. Our experiments show that the proposed SDSDL method outperforms state-of-the-art methods for joint segmentation and classification of surgical gestures that use only kinematic data.

[3] Similar advantages of learning shared discriminative dictionaries for object classification in static images have been observed, e.g., in [12].

2 Methods

Our approach to modeling surgical gestures consists of two stages: 1) dictionary learning and sparse coding, and 2) training a multi-class linear support vector machine. These stages are integrated into a common learning procedure, as illustrated in Fig. 2. The details of each stage of the model, as well as the algorithms for inference and parameter learning, are discussed in the following sections.

2.1 Dictionary learning and sparse coding

Let $\{\mathbf{x}_k\}_{k=1}^N$ be a sequence of observations (or features extracted from the observations), where $\mathbf{x}_k \in \mathbb{R}^p$ represents the observation at frame k (see Section 3.1 for dataset description). Let $\{y_k\}_{k=1}^N$ be their corresponding gesture labels and $\mathcal{C} = \{1, 2, \ldots, L\}$ be the set of all possible gestures.

Given n_{train} training trials $\{X^i\}_{i=1}^{n_{\text{train}}} \in \mathbb{R}^{p \times N_i}$ (where X^i denotes the concatenation of all observations in trial i, and N_i is the number of observations in trial i), and their corresponding labelings $\{\mathcal{Y}^i\}_{i=1}^{n_{\text{train}}}$, let

$$\mathcal{X} = [X^1, X^2, \ldots, X^{n_{\text{train}}}] = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N_t}] \in \mathbb{R}^{p \times N_t} \tag{1}$$

denote the concatenation of the observations from all training trials, where $N_t = \sum_{i=1}^{n_{\text{train}}} N_i$. We initialize our method by learning a common overcomplete dictionary of surgical atomic motions, $\Psi \in \mathbb{R}^{p \times m}$ $(m > p)$, for the entire training set and compute an initial sparse representation for each $\mathbf{x}_i$ with respect to Ψ by considering the following optimization problem:

$$\min_{\Psi, \mathcal{U} \in \mathbb{R}^{m \times N_t}} \frac{1}{N_t} \sum_{i=1}^{N_t} \Big(\frac{1}{2} \|\mathbf{x}_i - \Psi \mathbf{u}_i\|_2^2 + \lambda_u \|\mathbf{u}_i\|_1 \Big), \tag{2}$$

where λ_u is a regularization parameter, and

$$\mathcal{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{N_t}] \in \mathbb{R}^{m \times N_t}, \tag{3}$$

is the concatenation of all sparse coefficients. To prevent the entries of the dictionary from being arbitrarily large (which would cause the sparse coefficients to be very small), it is common to constrain the columns of the dictionary to have a unit l_2 norm. While the optimization problem in equation (2) is jointly non-convex, it is convex with respect to each of Ψ and $\mathcal{U}$ when the other one is fixed, and can be solved using existing sparse dictionary learning algorithms [13].

2.2 Multi-class linear support vector machine

While sparse codes have proven to be effective features for classification of static data (see [14] references therein), in this paper we are dealing with time-series data. Therefore, instead of using the sparse codes directly for classification, we propose the following two steps for computing features from the sparse codes.

First, given the sparse coefficients, $\mathbf{u}_i \in \mathbb{R}^m$ for $i = \{1, 2, \ldots, N_t\}$, which have been computed from (2), we split the positive and negative components of the sparse codes into two vectors:

$$\mathbf{u}_i^+ = \max(\mathbf{0}, \mathbf{u}_i), \quad \text{and} \quad \mathbf{u}_i^- = \min(\mathbf{0}, \mathbf{u}_i), \tag{4}$$

and form a feature vector of size $2m$ as follows:

$$\boldsymbol{\alpha}_i = \begin{bmatrix} \mathbf{u}_i^+ \\ \mathbf{u}_i^- \end{bmatrix}, \forall i = \{1, 2, \ldots, N_t\}. \tag{5}$$

Second, at each frame i, we construct a histogram by sum-pooling the sparse codes in a temporal window centered around the frame:

$$\mathbf{z}_i = \frac{1}{\mathcal{I}} \sum_{j \in I_i} \boldsymbol{\alpha}_i, \tag{6}$$

where I_i represents the set of indices centered around the ith frame, and $|I_i| = \mathcal{I}$ is the window size. The histogram carries local statistics of the surgical gestures.

Given a batch of training data $\{(\mathbf{z}_i, y_i)\}_{i=1}^{N_t}$, we initially train a multi-class linear support vector machine (SVM) by minimizing the following cost:

$$\min_{\mathbf{w}} \frac{\lambda_w}{2} \|\mathbf{w}\|_F^2 + \frac{1}{N_t} \sum_{i=1}^{N_t} \ell\big(\mathbf{w}; (\mathbf{z}_i, y_i)\big), \tag{7}$$

where $\mathbf{w} \in \mathbb{R}^{m \times L}$ is the matrix of classifiers that is formed by concatenating L linear classifiers corresponding to the L surgical gestures:

$$\mathbf{w} = \begin{bmatrix} w_{(1)} & w_{(2)} & \cdots & w_{(L)} \end{bmatrix}, \tag{8}$$

λ_w is the regularizer for the classifier, $\mathbf{z}_i$ is the feature vector for data point i in the training set, y_i is the corresponding class label, and N_t is the number of data points in the training set. The loss function is the hinge loss:

$$\ell(\mathbf{w}; (\mathbf{z}_i, y_i)) = \max(0, 1 - (w_{(y_i)}^T \mathbf{z}_i - w_{(y_i')}^T \mathbf{z}_i)), \tag{9}$$

where $y_i' = \arg\max_{l \in \mathcal{C}, l \neq y_i} w_{(l)}^T \mathbf{z}_i$.

2.3 Discriminative dictionary learning

In the learning approach described so far we first learn the dictionary of atomic motions Ψ by solving the optimization problem in (2). This is done in an unsupervised manner, i.e., without any knowledge about y. Then, given Ψ, we learn the gesture classifiers $\mathbf{w}$ by minimizing the empirical cost in (7).

In this section, we propose an alternative approach (called SDSDL) in which Ψ and $\mathbf{w}$ are learned jointly. Our rationale is that the optimization problem in (7) is implicitly dependent on the dictionary Ψ through the feature vectors $\mathbf{z}_i$s, which are constructed by average pooling of the sparse codes. Therefore, the optimization in (7) can be written as a function of both $\mathbf{w}$ and Ψ as follows:

$$J(\mathbf{w}, \Psi) = \frac{\lambda_w}{2} \|\mathbf{w}\|_F^2 + \frac{1}{N_t} \sum_{i=1}^{N_t} \ell\big(\mathbf{w}; (\mathbf{z}_i(\Psi), y_i)\big). \tag{10}$$

To jointly solve for the classifier parameters and the dictionary, we propose to use a stochastic sub-gradient descent algorithm [18], which requires computing the sub-gradient of the cost function with respect to both $\mathbf{w}$ and Ψ.

The sub-gradient with respect to the classifier parameters, $\mathbf{w}$, is given by:

$$\frac{\partial J}{\partial \mathbf{w}} = \lambda_w \mathbf{w} + \frac{1}{N_t} \sum_{i=1}^{N_t} \frac{\partial \ell(\mathbf{w}; (\mathbf{z}_i, y_i))}{\partial \mathbf{w}}. \tag{11}$$

If $\ell(\mathbf{w}; (\mathbf{z}_i, y_i))$ is equal to zero then $\partial \ell(\mathbf{w}; (\mathbf{z}_i, y_i))/\partial \mathbf{w} = \mathbf{0}$, but if the loss is greater than zero, the l^{th} column of $\partial \ell(\mathbf{w}; (\mathbf{z}_i, y_i))/\partial \mathbf{w}$ is computed as follows:

$$(\partial \ell(\mathbf{w}; (\mathbf{z}_i, y_i))/\partial \mathbf{w})^l = \begin{cases} -\mathbf{z}_i & \text{if } l = y_i \\ \mathbf{z}_i & \text{if } l = y_i' \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

On the other hand, the cost in (10) is implicitly dependent on the dictionary through the sparse codes. The sub-gradient of the loss function with respect to the dictionary can be computed by the chain rule as shown in [3]—similar to the "backpropagation" technique in neural networks [10]. Using the chain rule, we have:

$$\frac{\partial J}{\partial \Psi_{gh}} = \frac{1}{N_t} \sum_{i=1}^{N_t} \frac{\partial \ell(\mathbf{w}; (\mathbf{z}_i, y_i))}{\partial \Psi_{gh}} = \frac{1}{N_t} \sum_{i=1}^{N_t} \frac{\partial \ell(\mathbf{w}; (\mathbf{z}_i, y_i))}{\partial \mathbf{z}_i}^\top \frac{\partial \mathbf{z}_i}{\partial \Psi_{gh}}, \tag{13}$$

where Ψ_{gh} is the entry in the gth row and hth column of the dictionary Ψ, $\partial \ell(\mathbf{w}; (\mathbf{z}_i, y_i))/\partial \mathbf{z}_i = w_{(y_i')} - w_{(y_i)}$, for $\ell(\mathbf{w}; (\mathbf{z}_i, y_i)) > 0$. Then, recalling (6), we need to compute the derivative of the sparse codes with respect to the dictionary

$$\frac{\partial \mathbf{z}_i}{\partial \Psi_{gh}} = \frac{1}{\mathcal{I}} \sum_{j \in I_i} \frac{\partial \boldsymbol{\alpha}_i}{\partial \Psi_{gh}}. \tag{14}$$

For simplicity, let us drop the subscript i and write $\mathbf{z}$ for $\mathbf{z}_i$ and $\mathbf{u}$ for $\mathbf{u}_i$. Since $\mathbf{u}$ is the solution to the optimization problem in (2), it must satisfy the equation:

$$\Psi^\top \Psi \mathbf{u} - \Psi^\top \mathbf{x} = -\lambda_u \text{sign}(\mathbf{u}). \tag{15}$$

For a very small perturbation of the dictionary atoms, we can assume that the support (set of non-zero entries) of $\mathbf{u}$, denoted by S, does not change. Under this assumption, we can compute the gradient of the kth non-zero entry of the sparse coefficient with respect to the active columns of the dictionary as follows:

$$\frac{\partial \mathbf{u}_S(k)}{\partial \Psi_S} = (\mathbf{x} - \Psi_S \mathbf{u}_S) A_{[k]} - (\Psi_S A)_{<k>} \mathbf{u}_S^\top, \tag{16}$$

where $\mathbf{u}_S$ represents the vector containing the non-zero entries of $\mathbf{u}$, Ψ_S contains the active columns of Ψ, $A = (\Psi_S^\top \Psi_S)^{-1}$, the subscript $[k]$ represents the kth row of the matrix, and the subscript $< k >$ represents the kth column of the matrix.

We use stochastic sub-gradient descent to minimize the cost in (10), where at each iteration we compute the sub-gradients using a subset of the training data.

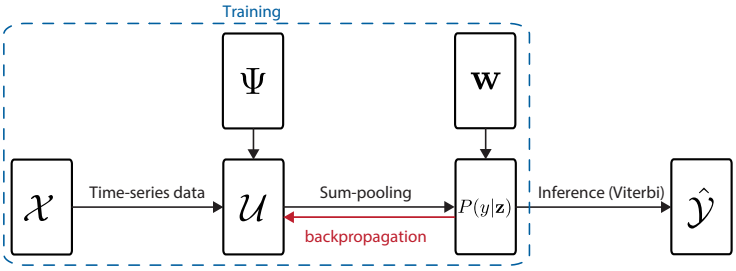

Fig. 2. Two stages of the training method shown in the blue dashed line box. For testing, inference is done by Viterbi algorithm. $\mathcal{X}$: Input data; Ψ: dictionary; $\mathcal{U}$: concatenation of sparse coefficients; $\mathbf{w}$: multi-class linear SVM classifier parameters; $\hat{\mathcal{Y}}$: output sequence of predicted class labels

After learning the parameters of the classifier and the dictionary, the model can predict a gesture label for the data in the testing set in a frame by frame manner. This approach ignores the fact that a surgical task is composed of multiple surgical gestures that are executed in a particular order and does not capture the temporal coherence in the predicted labels. Statistical methods such as Hidden Markov Model (HMM) and Conditional Random Field (CRF) methods are effective frameworks for modeling the time evolution of the surgical gestures [17, 11, 15, 21, 19, 20, 9]. To capture the temporal coherence of the predicted gesture labels, we integrate the output of the learned discriminative classifier into an HMM-like framework. Specifically, we use the soft-max function

$$P(y = l \mid \mathbf{z}) = \frac{e^{w_{(l)}^{\top}\mathbf{z}}}{\sum_{k=1}^{L} e^{w_{(k)}^{\top}\mathbf{z}}} \tag{17}$$

to convert the classifier scores for feature vector $\mathbf{z}$ into the probability of predicting gesture label l for feature $\mathbf{z}$. The HMM transition probabilities can be directly computed from the frequency of surgical gestures' transitions. The surgeme labels of the testing data can be inferred by the Viterbi algorithm [6].

3 Results

3.1 Dataset description

We evaluate the performance of the proposed SDSDL method for joint segmentation and classification of surgical gestures on the JHU-ISI Gesture and Skill Assessment Dataset (JIGSAWS) described in [7]. The JIGSAWS dataset includes eight subjects with three different skill levels (novice, intermediate and expert) performing 3–5 trials of three tasks (suturing, knot tying and needle passing). Each trial lasts about 2 minutes and the kinematic data of both master and slave manipulators of the da Vinci robotic surgical system is recorded at a constant rate of 30 Hz. Kinematic data consists of 76 motion variables including positions and velocities of both master (38 variables) and slave (38 variables) manipulators. All trials in the JIGSAWS dataset are manually segmented to 15

surgical gestures [15, 7]. Fig. 3 shows a manually labeled suturing trial with the corresponding surgical gestures listed in the caption. For each trial, we apply a mean-variance normalization to all kinematic variables. Specifically, let $\mathbf{x}_k(j)$ denote the jth variable of the kinematic data at frame k. The corresponding normalized feature, $\hat{\mathbf{x}}_k(j)$, is computed as follows:

$$\hat{\mathbf{x}}_k(j) = \frac{\mathbf{x}_k(j) - \mu_j}{\sigma_j}, \tag{18}$$

where μ_j and σ_j are, respectively, the mean and standard deviation of the jth kinematic variable.

3.2 Experimental setup

We consider two different test setups for surgical gesture classification. Setup 1 is the *leave-one-supertrial-out* (LOSO) setup, where we leave one trial of each subject out for testing, and use the remaining trials for training. Setup 2 is the *leave-one-user-out* (LOUO) setup, where we leave all trials corresponding to one subject out for testing and use all the trials from the remaining users for training.

3.3 Implementation details

To initialize the dictionary and the sparse codes, we solve the optimization problem in (2) using the dictionary learning and sparse coding toolbox from the SPAMS software [13]. Our results show that a choice for the LASSO regularizer λ_u in (2) in the interval $(0.05, 0.1)$ gives a good initialization. The dictionary size, m, and window size, $\mathcal{I}$, are chosen from $m = \{40, 60, 80, 100, 150, 200\}$ and $\mathcal{I} = \{1 : 10 : 101, 151, 201\}$ by cross-validation. The regularizer for the multi-class linear SVM, λ_w, is varied from 10^{-4} to 1 and is chosen by cross-validation.

3.4 Segmentation and classification results

We apply the proposed SDSDL method to the following combinations of kinematic variables: (1) all 38-dimensional kinematic data from the master robot, (2) all 38-dimensional kinematic data from the slave robot, (3) 76-dimensional kinematic data (combined slave and master data), (4) principal component analysis (PCA) projections of (3) to $n = \{5, 15, 25, 35\}$. While SDSDL had a similar performance for each combination, it performed slightly better with the data projected to $n = 35$ dimensions by PCA. The method achieves its best performance for the window sizes in the range of 51 to 81, and the dictionary sizes of 100 and 150. The prediction accuracy of the method is stable for the classifier regularizer in the range of 0.01 to 0.06.

Table 1 reports the accuracy of SDSDL on three surgical tasks for the two different setups explained in Section 3.2. The results of SDSDL outperform those of S-HMM in joint segmentation and classification of surgical gestures using only kinematic data. In particular, SDSDL gives an $8 - 10\%$ improvement in the more challenging LOUO setup for all three tasks. SDSDL also outperforms other state-of-the-art statistical methods including LDA-HMM (with 3 states for each gesture and one Gaussian per state) and MsM-CRF. However, notice that SDSDL performs slightly worse than Skip-Chain CRF (SC-CRF) [9] in the

LOUO setup. Notice also that, in general, all prior methods perform considerably better in the LOSO setup than in the LOUO setup. The reason is that surgeons have their own styles in surgery, and when the model is not trained with the surgeon's style (LOUO), it performs worse. By skipping some frames, the SC-CRF model captures higher-order temporal relationships between the gestures, achieves a more robust performance in LOUO, and thus becomes more invariant to surgeon style. These results suggest that the performance of SDSDL could be improved by integrating SDSDL with a SC-CRF, rather than an HMM.

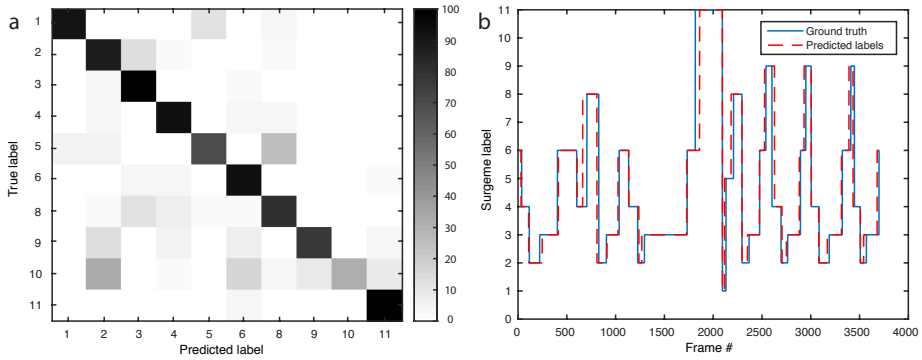

Fig. 3. (**a**) Confusion matrix corresponding to the LOSO setup for the suturing task. (**b**) Sample time series in suturing task. List of surgemes: 0. Idle motion, 1. Reach needle, 2. Position needle, 3. Insert/push needle through tissue, 4. Transfer needle, 5. Move to center with needle (right hand), 6. Pull suture with left hand, 7. Pull suture with right hand, 8. Orienting needle, 9. Right hand assisting left in tightening suture, 10. Loosen more suture, 11. Drop suture (end of trial).

Effect of temporal windowing: Our experiments show that constructing a histogram by sum-pooling the sparse codes in a temporal window around each frame improves the gesture classification results. Testing this effect over a wide range of values for the window size reveals that a window sizes around $61-81$ frames are the optimal values for capturing the local statistics of the surgical motions in the JIGSAWS dataset.

Effect of duplicating sparse codes to positive and negative: Duplicating the entries of sparse codes to positive and negative components before applying sum- or max-pooling have been shown to result in improved classification performance [4, 14]. In our experiments on the JIGSAWS dataset, splitting the positive and negative components of the sparse code (see equation (4)) improves the prediction accuracy by at least $1-2\%$.

4 Conclusion

We have proposed a sparse-representation-based algorithm for the segmentation and classification of atomic gestures in robotic surgery. The architecture of the

Table 1. Average percentage of correctly classified frames using the kinematic data of the JIGSAWS dataset [7].

Task	Setup	LDA-HMM [1]	MsM-CRF [1]	SC-CRF [9]	S-HMM [1]	SDSDL
Suturing	LOSO	82.21%	81.99%	85.18%	83.94 %	**86.32**%
	LOUO	73.95%	67.84%	**80.29**%	70.81%	78.68%
Needle Passing	LOSO	70.54 %	72.44%	**77.30**%	70.69%	74.88%
	LOUO	64.12 %	44.68%	**74.77**%	55.02%	66.01%
Knot Tying	LOSO	80.95%	79.26%	80.72%	77.83%	**82.54**%
	LOUO	72.47%	63.28%	**78.95**%	67.89%	75.11%

proposed model is simple. By learning a shared dictionary of atomic surgical motions for all the data in the training set, the model requires fewer parameters compared to similar modeling approaches such as S-HMM [19], where a separate dictionary is learned for each surgical gesture. The proposed model, can be integrated with other statistical models such as Conditional Random Fields (CRF) models [20, 9] to capture higher-order temporal information. More specifically, the proposed sparse representation in this paper can be used to model the unary potentials in the CRF model.

Acknowledgement. This material is based upon work supported by the NSF under grant 1335035 (R.V.), ONR under grant N000141310116 (R.V.), and a James S. McDonnell Complex Systems Scholar award (S.S., N.J.C. and R.V.). The authors would like to thank Dr. N. Ahmidi, Colin Lea, Lingling Tao, and Prof. G. Hager for insightful discussions and comments.

References

1. Ahmidi, N.: Activity Detection and Skill Assessment for Dexterous Motions in Robotic and Minimally-Invasive Surgery. Ph.D. thesis, Johns Hopkins University (2015)
2. Ahmidi, N., Gao, Y., Béjar, B., Vedula, S.S., Khudanpur, S., Vidal, R., Hager, G.D.: String motif-based description of tool motion for detecting skill and gestures in robotic surgery. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 26–33. Springer (2013)
3. Boureau, Y.L., Bach, F., LeCun, Y., Ponce, J.: Learning mid-level features for recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2559–2566 (2010)
4. Coates, A., Ng, A.Y.: The importance of encoding versus training with sparse coding and vector quantization. In: Proceedings of the 28th International Conference on Machine Learning. pp. 921–928 (2011)
5. Datta, V., Mackay, S., Mandalia, M., Darzi, A.: The use of electromagnetic motion tracking analysis to objectively measure open surgical skill in the laboratory-based model. Journal of the American College of Surgeons 193(5), 479–485 (2001)
6. Forney Jr, G.D.: The Viterbi algorithm. Proceedings of the IEEE 61(3), 268–278 (1973)
7. Gao, Y., Vedula, S.S., Reiley, C.E., Ahmidi, N., Varadarajan, B., Lin, H.C., Tao, L., Zappella, L., Béjarl, B., Yuh, D.D., Chen, C.C.G., Vidal, R., Khudanpur, S., Hager, G.D.: The JHU-ISI gesture and skill assessment dataset (JIGSAWS): A surgical activity working set for human motion modeling. In: Medical Image Computing and Computer-Assisted Intervention M2CAI Workshop (2014)

8. Judkins, T.N., Oleynikov, D., Stergiou, N.: Objective evaluation of expert and novice performance during robotic surgical training tasks. Surgical endoscopy 23(3), 590–597 (2009)
9. Lea, C., Hager, G.D., Vidal, R.: An improved model for segmentation and recognition of fine-grained activities with application to surgical training tasks. In: IEEE Winter Conference on Applications of Computer Vision (WACV) (2015)
10. LeCun, Y., Bottou, L., Orr, G.B., Muller, K.R.: Efficient bachprop. Lecture Notes in Computer Science 1524, 0009 (1998)
11. Leong, J.J., Nicolaou, M., Atallah, L., Mylonas, G.P., Darzi, A.W., Yang, G.Z.: HMM assessment of quality of movement trajectory in laparoscopic surgery. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 752–759. Springer (2006)
12. Lobel, H., Vidal, R., Soto, A.: Learning shared, discriminative, and compact representations for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2015)
13. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online dictionary learning for sparse coding. In: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 689–696. ACM (2009)
14. Mairal, J., Bach, F., Ponce, J., et al.: Sparse modeling for image and vision processing. Foundations and Trends® in Computer Graphics and Vision 8(2-3), 85–283 (2014)
15. Reiley, C.E., Hager, G.D.: Task versus subtask surgical skill evaluation of robotic minimally invasive surgery. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 435–442. Springer (2009)
16. Richards, C., Rosen, J., Hannaford, B., Pellegrini, C., Sinanan, M.: Skills evaluation in minimally invasive surgery using force/torque signatures. Surgical Endoscopy 14(9), 791–798 (2000)
17. 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. IEEE Transactions on Biomedical Engineering 48(5), 579–591 (2001)
18. Shalev-Shwartz, S., Singer, Y., Srebro, N., Cotter, A.: Pegasos: Primal estimated sub-gradient solver for SVM. Mathematical programming 127(1), 3–30 (2011)
19. Tao, L., Elhamifar, E., Khudanpur, S., Hager, G.D., Vidal, R.: Sparse hidden markov models for surgical gesture classification and skill evaluation. In: Information Processing in Computer-Assisted Interventions, pp. 167–177. Springer (2012)
20. Tao, L., Zappella, L., Hager, G.D., Vidal, R.: Surgical gesture segmentation and recognition. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 339–346. Springer (2013)
21. Varadarajan, B.: Learning and inference algorithms for dynamical system models of dextrous motion. Ph.D. thesis, Johns Hopkins University (2011)
22. 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), pp. 426–434. Springer (2009)
23. Yamauchi, Y., Yamashita, J., Morikawa, O., Hashimoto, R., Mochimaru, M., Fukui, Y., Uno, H., Yokoyama, K.: Surgical skill evaluation by force data for endoscopic sinus surgery training system. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 44–51. Springer (2002)
24. Zappella, L., Béjar, B., Hager, G., Vidal, R.: Surgical gesture classification from video and kinematic data. Medical image analysis 17(7), 732–745 (2013)