Supplemental Material: A Dataset and Benchmarks for Segmentation and Recognition of Gestures in Robotic Surgery

Narges Ahmidi, Lingling Tao, Shahin Sefati, Yixin Gao, Colin Lea, Benjamín Béjar Haro, Luca Zappella, Sanjeev Khudanpur, René Vidal, *Fellow, IEEE,* Gregory D. Hager, *Fellow, IEEE*

I. Details of the methodology

The described methodology is strictly consistent in the way that the training, testing, and validation of the techniques are performed: (1) To avoid a biased selection of only trials with good results, all the models should be tested on all the provided test trials and their predictions (both good and bad) should get included in the validation results, (2) At each cross-validation fold, a set of particular training trials is provided for training the models. The models can be trained using any arbitrary subset of the training set. For example, it is accepted if a technology decides not to train a model for gesture G9. This will have the drawback for them that they can not correctly classify G9 at test time, (3) If a model requires a validation set (for hyper-parameter learning) then part of the training set should be used, and (4) No test trials should be used during the training or validation phase.

When reporting the results for a specific model, the "best" configuration of parameters was defined as the one that achieves the highest overall performance across different cross-validation settings (LOSO and LOUO) and for different tasks (suturing, needle-passing, and knot-tying).

II. Results

Tables I and II summarize the validation results of the reported algorithms in this paper when trained and tested with the same input features. The input features to all these algorithms are kinematics variables from both master and slave robot arms (all 76 dimensions). Note that the video-based techniques have already been trained and tested against identical input video sequences, generating their own model-specific intermediate features.

Tables III and IV summarize state-of-the-art results for skill assessment using the JIGSAWS dataset. Table III shows micro-average results of three techniques (KSVD-SHMM, MFA-HMM [1] and discrete-HMM [2]) for a 3-way skill classification problem. The skill classes are expert, intermediate, and expert, as defined in the JIGSAWS dataset.

NA, YG, CL, and GH are with the Department of Computer Science. LT, SS, LZ, BB and RV are with the Center for Imaging Science, Department of Biomedical Engineering. SK is with the Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, 21218 USA.

Correspondence e-mail: nahmidi1@jhu.edu.

Manuscript received September 2016; accepted December 2016.

Table IV reports results for the skill regression problem for two techniques: GMM-HMM and DCC [3]. Each technique produces a score for a given trial which is used to regress against the OSATS score provided for each trial in the JIGSAWS dataset. A leave-one-trial-out cross validation was used to compute the results. Note that the performances of these algorithms are not directly comparable as they were not assessed using the uniform validation methodology described in this paper, but they can provide a perspective on the accuracy of skill classification. In addition, other newer techniques [4]–[6] (using other larger and complex surgical datasets) have shown when gesture sequence and boundaries are known, they are able to predict the surgeon's skill level with up to 90% and 75% accuracy, under leave-one-trail-out and leave-one-user-out accuracy, respectively.

III. Computational Complexity

The complexity of the training phase is reported in Table V for training C classes using all m samples. The complexity of the decoding phase is reported for decoding one single test sample.

In this table, T is the number of frames in a given trial, t number of frames in a segment, C the number of gesture classes, m number of training samples, n number of iterations, D dimension of the features, d dimension of the hidden state, K the sparsity level, x number of dense features extracted from one frame, L number of the atoms in the dictionary, v the number of SVM support vectors, and S the total number of hidden states.

IV. Surgical Gesture Classification

A. Bag of Spatio-Temporal Features: BoF

Classification: The three different types of kernels used in section V-A for classification are defined as bellow:

$$K_I(h_i, h_j) = \min(h_i, h_j) \tag{1}$$

$$K_{\mathcal{X}}(h_i, h_j) = \sum_k 2\frac{h_i(k)h_j(k)}{h_i(k) + h_j(k)} \tag{2}$$

$$K_{\mathrm{RBF}}(h_i, h_j) = exp(-\gamma \sum_{q=1}^{Q} \frac{1}{\mu_q} d(h_i^q, h_j^q)) \tag{3}$$

where $\gamma \in \mathbb{R}^+$ is a parameter. When we have Q different types of features extracted from the data (e.g. in BoF we have

TABLE I: Performance of gestures joint segmentation and classification techniques validated on the JIGSAWS using the same input features (76 dimensions of kinematics). GMM-HMM ($S = 3$, $M = 1$, $d = 1$, f_1), KSVD-S-HMM ($K = 3$, 300-word dictionary), MsM-CRF (kinematic m_2), and SC-CRF ($\delta = 30$).

Cross validation	Method (Data type)	Evaluation	**Suturing**	**Needle-passing**	**Knot-tying**
LOSO	**GMM-HMM (kin)**	Micro	78.76	64.28	75.03
		Macro±std	70.34±26.36	61.00±12.08	70.37±13.71
		Precision±std	71.30±27.72	61.81±17.85	75.52±11.91
		β ($\mu \pm \sigma$)	79.46±23.59	64.39±8.724	74.53±40.12
		β 95%CI	69.18–88.11	58.51–70.07	61.20–85.861
	KSVD-SHMM (kin)	Micro	83.94	70.69	77.83
		Macro±std	74.27±27.02	65.29±20.43	74.57±9.32
		Precision±std	84.55±10.52	65.21±21.16	83.64±10.95
		β ($\mu \pm \sigma$)	84.63±0.11	71.12±0.26	77.78±0.15
		β 95%CI	77.38–90.69	60.64–80.56	69.58–85.02
	MsM-CRF (kin)	Micro	81.99	72.44	79.26
		Macro±std	72.56±26.70	67.73±16.93	79.05±7.62
		Precision±std	72.23±27.69	67.54±18.97	82.12±7.27
		β ($\mu \pm \sigma$)	83.04±0.24	73.24±0.29	79.19±0.14
		β 95%CI	72.29–91.57	62.075–83.09	71.15–86.22
LOUO	**GMM-HMM (kin)**	Micro	65.21	45.88	64.49
		Macro±std	51.62±28.57	39.19±12.64	58.71±12.62
		Precision±std	54.01±32.33	49.32±27.13	64.31±10.90
		β ($\mu \pm \sigma$)	64.62±72.27	45.92±73.06	62.22±1.413
		β 95%CI	47.20–80.24	29.50–62.79	37.82–83.66
	KSVD-SHMM (kin)	Micro	70.81	55.02	67.89
		Macro±std	51.48±27.08	41.16±19.92	64.60±11.29
		Precision±std	68.59±19.01	50.02±23.91	72.76±13.58
		β ($\mu \pm \sigma$)	70.63±0.19	56.52±0.57	69.44±1.00
		β 95%CI	61.54–78.97	41.43–71.02	48.24–86.99
	MsM-CRF (kin)	Micro	67.84	44.68	63.28
		Macro±std	51.05±28.62	37.58±12.63	53.05±26.31
		Precision±std	54.27±32.20	47.43±27.12	57.95±31.66
		β ($\mu \pm \sigma$)	68.24±1.34	46.83±2.14	64.40±1.13
		β 95%CI	43.67–88.26	19.36–75.43	42.38–83.58
	SC-CRF (kin)	Micro	78.22	69.16	67.69
		Macro±std	59.10±34.72	59.74±20.10	60.68±13.94
		Precision±std	62.71±32.21	62.23±17.87	61.18±17.29
		β ($\mu \pm \sigma$)	78.32±30.93	69.49±90.13	68.50±75.81
		β 95%CI	66.50–88.15	49.48–86.22	50.34–84.12

both HOG and HOF descriptors, i.e. $Q = 2$), we can either concatenate them and construct only one histogram, or we can have Q separate histogram representations (h^q for $q = \{1, 2, \ldots, Q\}$). In the latter case, each representation can been considered as a *channel*, and we can compute a kernel using a multi-channel approach. In the definition of RBF kernel above, the function $d(h_i^q, h_j^q)$ returns the $\mathcal{X}^2$ distance between the two histograms h_i^q and h_j^q from channel q, and μ_q is the average empirical distance between all pairs of training histograms for channel q [7].

B. Linear Dynamical System

Comparing LDSs: To assess the similarity or the distance between two LDS models, we used three different types of distance metrics in section V-B that we briefly describe below:

TABLE II: Performance of gesture classification techniques validated on the JIGSAWS using the same input features (76 dimensions of kinematics). LDS($n = 15$, SVM classifier, BC metric for kinematics), and GMM-HMM ($S = 3$, $M = 1$, $d = 1$).

Cross validation	Method (Data type)	Evaluation	**Suturing**	**Needle-passing**	**Knot-tying**
LOSO	**LDS (kin)**	Micro	84.61	59.76	81.67
		Macro±std	63.87±30.82	46.55±25.81	74.51±23.73
		Precision±std	73.30±28.41	52.91±17.31	76.07±18.72
		β ($\mu \pm \sigma$)	84.77±0.06	59.78±0.01	81.67±0.016
		β 95%CI	79.36–89.49	57.42–62.12	79.12–84.09
	GMM-HMM (kin)	Micro	87.76	68.14	82.48
		Macro±std	75.13±28.63	64.52±16.51	78.86±14.08
		Precision±std	75.13±28.55	66.14±19.24	81.85±9.75
		β ($\mu \pm \sigma$)	87.76±0.01	68.12±0.01	82.35±0.34
		β 95%CI	85.8–89.60	65.71–70.47	69.44–92.27
LOUO	**LDS (kin)**	Micro	73.64	47.96	71.42
		Macro±std	51.75±32.91	32.59±29.74	63.99±24.51
		Precision±std	53.39±32.01	32.01±27.76	65.74±21.54
		β ($\mu \pm \sigma$)	73.80±0.30	45.68±1.45	71.44±0.15
		β 95%CI	62.31–83.84	22.92–69.44	63.31–78.93
	GMM-HMM (kin)	Micro	66.58	48.64	67.65
		Macro±std	50.40±30.33	46.61±13.00	62.34±17.14
		Precision±std	56.89±34.06	56.70±28.23	65.62±16.82
		β ($\mu \pm \sigma$)	66.98±0.68	48.42±0.90	67.10±1.17
		β 95%CI	49.89–82.03	30.08–66.99	44.35–86.21

TABLE III: Skill classification results (micro averages) reported in [1] and [2]. Three classes are self-claimed expertise level: expert, intermediate, and novice.

Cross validation	Method (Data type)	**Suturing**	**Needle-Passing**	**Knot-Tying**
LOSO	KSVD-SHMM (kin) [1]	97.4	96.2	94.4
LOSO	MFA-HMM (kin) [1]	92.3	76.9	86.1
LOSO	discrete-HMM (kin) [2]	72	Not available	Not available
LOUO	KSVD-SHMM (kin) [1]	59.0	26.9	58.3
LOUO	MFA-HMM (kin) [1]	38.5	46.2	44.4

(1) Subspace angles: Metrics based on subspace angles ([8], [9]) measure the dissimilarity of two LDS models (of order n represented by $M_1 = (A_1, C_1)$ and $M_2 = (A_2, C_2)$) using the subspace angles $(\theta_1, \ldots \theta_{2n})$ between the range spaces of their infinite observability matrices of the dynamical models. The LDS model's infinite observability matrix is defined as follows:

$$\mathcal{O}_i = [C_i^T, (C_i A_i)^T, (C_i A_i^2)^T, \ldots], \qquad i = 1, 2 \qquad (4)$$

The subspace angles are invariant with respect to a change

TABLE IV: Skill regression results (similarity, OSATS mean and standard deviation of error, Area Under the Curve, and 95% confidence interval) using the techniques reported in [3]. A leave-one-trial-out LOTO cross validation was performed.

Cross validation	Method (Data type))	Evaluation	**Suturing**
LOTO	GMM-HMM (kin) [3]	Sim Error μ, σ AUC 95% CI	82.68 17.3±15.10 0.97 0.92–1.0
LOTO	DCC (kin) [3]	Sim Error μ, σ AUC 95% CI	85.79 0.14±0.15 0.98 0.93–1.0

TABLE V: Computational complexity.

Technique	Training Phase m samples	Decoding Phase 1 sample
HMM	$O(mTD^2)$: LDA $O(mTS^2)$: Baum Welch	$O(TS^2)$
S-HMM	$O(mT(K^2+D)L)$	$O(TC^2)$
MsM-CRF	$O(mt)$: feature extraction $O(mtnTC^2)$: parameter learning	$O(mtnTC^2)$
SC-CRF	$O(mTC^2)$: inference $O(nmTC^2)$: optimization (Block Coordinate Frank Wolfe)	$O(TC^2)$
BoF	$O(mt)$: feature extraction $O(mtxn)$: dictionary learning $O(mtx)$: histogram generation $O(Lm^2)$: kernel generation $O(m^2C)$: SVM kernel learning	$O(tmx + v^2LC)$
LDS	$O(mdD^2)$: PCA $O(m^2d^3)$: LDS kernel $O(m^2d^3)$: Sylvester eq. for all pairs $O(m^2C)$: SVM kernel	$O(dD^2 + v^2d^3C)$

of basis in the state spaces and thus one can define different metric distances based on the subspace angles [8]. For a stable system ($\|A_i\| < 1$), the subspace angle is defined as $\theta_i = cos^{-1}(\sqrt{\lambda_i})$ where λ_i is the i^{th} eigenvalue of $P_{11}^{-1}P_{12}P_{22}^{-1}P_{21}$ and P_{ij} is the solution to the Sylvester equation with constant $\rho = 1$:

$$P_{ij} = \rho A_i^T P_{ij} A_j + C_i^T C_j \tag{5}$$

After measuring the subspace angles, we then can compute the dissimilarity between the two given models using the Martin distance metric:

$$d_M^2(M_1, M_2) = -\log \prod_{i=1}^{2n} \cos^2(\theta_i) \tag{6}$$

or the Frobenius distance metric:

$$d_F^2(M_1, M_2) = 2\sum_{i=1}^{2n} \sin^2(\theta_i) \tag{7}$$

(2) Binet-Cauchy (BC) kernels: The distance metrics based on BC-kernels [10] depends not only on the parameters (A, C), but also on the initial condition x_0. For our particular application of classification, we use a special case of BC kernel, called the normalized determinant kernel that is independent of the initial conditions and is invariant to basis transformation. The normalized determinant kernel is defined as follows:

$$k_D(M_1, M_2) = \frac{det(P_{12})^2}{det(P_{11})det(P_{22})} \tag{8}$$

where P_{ij} is the solution to the Sylvester equation with $0 < \rho < 1$ being a parameter (not a constant as for subspace angles). We then compute the distance between the two given models as:

$$d_D^2(M_1, M_2) = k_D(M_1, M_1) + k_D(M_2, M_2) - 2k_D(M_1, M_2) \tag{9}$$

(3) Action-induced distances: This approach aims to find the "closest" representation between two LDS models through a basis transformation [11]. The non-singular matrix transformation Q is restricted to be from the orthogonal group $O(n)$ for a more tractable computation. A Frobenius norm is used to measure the squared Align metric between the two model parameters as follows:

$$\begin{aligned} d_A^2(M_1, M_2) = \min_{Q \in O(n)} \{\lambda_A \|Q^T A_1 Q - A_2\|^2 + \\ \lambda_c \|C_1 Q - C_2\|^2 + \lambda_B \|Q^T B_1 - B_2\|\} \end{aligned} \tag{10}$$

with weight parameters $\lambda_A \geq 0$, $\lambda_B \geq 0$ and $\lambda_C \geq 0$.

V. Surgical Gesture Segmentation and Classification

A. Sparse Hidden Markov Model: S-HMM

Inference: As discussed in section IV-B, when modeling the surgical gestures with S-HMM, we can infer the sequence of gesture labels $\{z_t\}_{t=1}^T$ using a dynamic programming method similar to the Viterbi algorithm [12].

However, in our model, the marginal probability $p(o_t|z_t)$ cannot be computed in closed form because x_t has a Laplace distribution. Thus instead of marginalizing over x_t and computing $p(s_{1:T}, o_{1:T})$, we choose the best x_t^* for each z_t and only maximize over z_t with fixed corresponding x_t. More specifically, we can write the following recursion

$$\begin{aligned} \alpha_t(z, \boldsymbol{x}) \triangleq & p(o_t|x_t = \boldsymbol{x}, z_t = z) \cdot p(x_t = \boldsymbol{x}|z_t = z) \\ & \cdot \max_{z', \boldsymbol{x}'} \{q_{z',z} \cdot \alpha_{t-1}(z', \boldsymbol{x}')\} \end{aligned} \tag{11}$$

From the last equality, one can see that the value of x_t only affects the first two probabilities and has no influence on the last term. Now, since the number of states Z is finite, for each z we can find the $\hat{\boldsymbol{x}}_z$ that maximizes $p(o_t|\boldsymbol{x}, z)p(\boldsymbol{x}|z)$. That is, $\hat{\boldsymbol{x}}_z = \arg\min_{\boldsymbol{x}} \ \lambda_z \|\boldsymbol{x}\|_1 + \frac{1}{2\sigma_z^2}\|o_t - \boldsymbol{D}_z \boldsymbol{x}\|^2$, which can be found using Basis Pursuit [13] or Orthogonal Matching Pursuit (OMP) [14]. Since the learning algorithm uses K-SVD method which uses OMP for sparse coding, we also use OMP here.

B. Semantic Image Model and Skip-Chain Conditional Random Field

Semantic Image Models and Features:

In section IV-D, we developed a deformable part model [15] to detect and localize the positions of the objects in the video.

As the first step in training the semantic model, we manually label the position of the objects in the first (non-occluded) frame of each video file. Having multiple samples of the location of each object q_i, we can model the distance between the pairs (q_i, q_j) using a Gaussian function with mean and standard deviation (μ_{ij}, Σ_{ij}).

The next step is to localize the objects in each image. We use a template matching algorithm to find these objects. To reduce the false positive rate of the template matching, we learn the incorrect patches by training a classifier (SVM) and using its output when deciding whether to accept or reject a candidate object.

Finally, our goal is to find a most likely configuration for all objects $Q = \{q_1, \ldots, q_n\}$ in a given image I by optimizing $P(Q|I) \propto exp(-E_V(I, Q))$, where E_V is the following energy function:

$$E_V(I, Q) = \sum_{i \in nodes} w^{vu} \phi_V(I, q_i) + \sum_{i \in edges} w^{ve} \psi_V(I, q_i, q_j) \tag{12}$$

where ϕ_V is the unary function which returns the score from the template-matching step. The pairwise term ψ_V returns a dissimilarity score for the edge length between two nodes using the Gaussian model learned for that particular edge:

$$\psi_V(q_i, q_j) = (q_i - q_j - \mu_{ij})^T \Sigma_{ij}^{-1} (q_i - q_j - \mu_{ij}) \tag{13}$$

For inference, we use a computationally efficient variation of belief propagation proposed by [15].

After determining the object locations in the image, we compute two new semantic-driven features. To do so, we first project the tool Cartesian position (p_K known from the kinematics) into the current image frame (p_I). Then we measures the distance between p_I and the closest object in the image:

$$f_d(p_I, Q) = \min_i \|p_I - q_i\|_2. \tag{14}$$

The second feature measures the relative position between the projection of the tool and the closest object:

$$f_o(p_I, Q) = p_I - q_{i^*} \tag{15}$$

where $i^* = \arg\min_i \|p_I - q_i\|_2$.

REFERENCES

[1] L. Tao, E. Elhamifar, S. Khudanpur, G. D. Hager, and R. Vidal, "Sparse hidden markov models for surgical gesture classification and skill evaluation," *Information Processing in Computer-Assisted Interventions*, vol. 7330, pp. 167–177, 2012.

[2] C. E. Reiley and G. D. Hager, "Task versus subtask surgical skill evaluation of robotic minimally invasive surgery," *Medical Image Computing and Computer-Assisted Intervention*, pp. 435–442, 2009.

[3] N. Ahmidi, G. D. Hager, L. Ishii, G. L. Gallia, and M. Ishii, "Robotic path planning for surgeon skill evaluation in minimally-invasive sinus surgery." *Medical Image Computing and Computer-Assisted Intervention*, vol. 7510, pp. 471–478, 2012.

[4] S. S. Vedula, A. O. Malpani, L. Tao, G. Chen, Y. Gao, P. Poddar, N. Ahmidi, C. Paxton, R. Vidal, S. Khudanpur, G. D. Hager, and C. C. G. Chen, "Analysis of the structure of surgical activity for a suturing and knot-tying task," *PLOS one*, 2016.

[5] N. Ahmidi, Y. Gao, B. Bejar, S. S. Vedula, S. Khudanpur, R. Vidal, and G. D. Hager, "String motif-based description of tool motion for detecting skill and gestures in robotic surgery," *Medical Image Computing and Computer-Assisted Intervention*, 2013.

[6] N. Ahmidi, P. Poddar, J. D. Jones, S. S. Vedula, L. Ishii, G. D. Hager, and M. Ishii, "Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty," *International Journal of Computer Assisted Radiology and Surgery*, pp. 981–991, 2015.

[7] L. Zappella, B. Béjar, G. D. Hager, and R. Vidal, "Surgical gesture classification from video and kinematic data," *Medical Image Analysis*, vol. 17, pp. 732 – 745, 2013.

[8] K. Cock and B. Moor, "Subspace angles and distances between ARMA models," *System and Control Letters*, vol. 46, pp. 265–270, 2002.

[9] A. Martin, "A metric for ARMA processes," *IEEE Trans. on Signal Processing*, vol. 48, pp. 1164–1170, 2000.

[10] S. Vishwanathan, A. Smola, and R. Vidal, "Binet-Cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes," *International Journal of Computer Vision*, vol. 73, pp. 95–119, 2007.

[11] B. Afsari, R. Chaudhry, A. Ravichandran, and R. Vidal, "Group action induced distances for averaging and clustering linear dynamical systems with applications to the analysis of dynamic visual scenes," *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[12] G. Forney-Jr, "The Viterbi algorithm," *Proceedings of the IEEE*, vol. 61 (3), 1973.

[13] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, pp. 33–61, 1998.

[14] J. Tropp and A. Gilbert, "Multi-scale hybrid linear models for lossy image representation," *IEEE Trans on Information Theory*, vol. 15 (12), pp. 3655–3671, 2006.

[15] P. Felzenszwalb and D. Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision*, vol. 61(1), pp. 55–79, 2005.