Surgical Gesture Classification from Video Data

Benjamín Béjar Haro[1], Luca Zappella[1], and René Vidal

Center for Imaging Science, Johns Hopkins University

Abstract. Much of the existing work on automatic classification of gestures and skill in robotic surgery is based on kinematic and dynamic cues, such as time to completion, speed, forces, torque, or robot trajectories. In this paper we show that in a typical surgical training setup, video data can be equally discriminative. To that end, we propose and evaluate three approaches to surgical gesture classification from video. In the first one, we model each video clip from each surgical gesture as the output of a linear dynamical system (LDS) and use metrics in the space of LDSs to classify new video clips. In the second one, we use spatio-temporal features extracted from each video clip to learn a dictionary of spatio-temporal words and use a bag-of-features (BoF) approach to classify new video clips. In the third approach, we use multiple kernel learning to combine the LDS and BoF approaches. Our experiments show that methods based on video data perform equally well as the state-of-the-art approaches based on kinematic data.

Keywords: surgical gesture classification; time series classification; dynamical system classification; bag of features; multiple kernel learning.

1 Introduction

Recent technological advances have contributed to, and changed, the way in which surgery can be performed. One of them is Robotic Minimally Invasive Surgery (RMIS), which has several advantages over traditional surgery, such as better precision, smaller incisions and reduced recovery time. However, the steep learning curve together with the lack of fair and effective criteria for judging the skills acquired by a trainee, may reduce the benefits of this technology.

This has motivated a number of approaches for automatic RMIS skill assessment and gesture classification. One of the most natural approaches is to decompose a surgical task into a series of pre-defined 'atomic' gestures or *surgemes* [1–3], such as 'insert a needle', 'grab a needle', 'position a needle', etc. (Fig. 1 shows sample frames from three different surgemes). The problem then becomes how these surgemes can be segmented in time, recognized, and finally assessed.

Most of the prior work on surgical gesture recognition (see, e.g., [4–6]) uses hidden Markov models (HMMs) to analyze kinematic data stored by the robot, such as the position of the robot tools, angles between robot joints, velocity measurements and force/torque signatures. All these approaches model each surgeme

[1] Equal contribution

(a) S2: positioning needle (b) S4: Transferring needle from left to right (c) S5: moving to center with needle in grip

Fig. 1. Examples of three different surgemes in a suturing task.

as one or more states of an HMM. The main difference is in how these approaches model the observations within each surgeme. For example, [5] vector-quantizes the observations into discrete symbols, [7] uses a Gaussian model combined with linear discriminant analysis (LDA), [6] assumes that the observations are generated from a lower-dimensional latent space using Factor Analyzed HMMs (FA-HMMs) and Switched Linear Dynamical Systems (SLDSs), [8] uses a Gaussian mixture model (GMM), and [9] models the observations as a linear combination of atomic motions with sparse coefficients. All of these methods have significantly improved surgical gesture classification over a standard HMM.

In addition to kinematic measurements, RMIS systems are also typically equipped with cameras that record the entire procedure. The work in [10, 11] propose to recognize the different phases of a surgery (e.g. CO2 inflation, abdominal suturing, etc.) using laparoscopic videos. Other works on video data analysis [12, 13] focus on recognizing the phases of a surgery by also observing surgeons and nurses in the operating room. To the best of our knowledge, the only existing work on automatic skill and surgical gesture (rather than coarse phases) classification from video is [14], which uses basic visual cues based on optical flow and concludes that kinematic-based approaches are generally more accurate.

In this paper, we propose and evaluate three approaches to surgical gesture classification from video. The first approach uses linear dynamical systems (LDSs) to model each video clip from each surgeme. Distances between the parameters of the LDSs are then used to classify new video clips. The second approach is a bag-of-features (BoF) approach in which a dictionary of spatio-temporal words is learned from spatio-temporal features extracted from all video clips. Each video clip is then represented with a histogram of such words and distances between histograms are used to classify new video clips. The third approach combines the LDS and BoF approaches using multiple kernel learning (MKL). Our experiments on kinematic data from a typical surgical training setup show that methods based on LDSs already outperform state-of-the-art approaches based on HMMs [9]. For video data, the BoF approach performs better than the LDS approach, while the MKL approach performs equally well in terms of accuracy, but is typically more robust. Overall, our main conclusion is that methods based on video data perform equally well as methods based on kinematic data for a typical surgical training setup. This result should encourage

further investigation of video based techniques for surgical gesture classification as videos potentially carry more unexploited information than kinematic data.

2 Video-Based Methods for Classifying Surgical Gestures

In this section we describe three techniques for surgical gesture classification based on video data. We assume that each video is segmented into *video surgemes*, i.e., video clips corresponding to a single execution of one out of a pre-defined set of surgemes. All three methods use labeled video surgemes to learn a model for each of them. We then show how these models can be compared and used for classifying gestures in new video surgemes.

2.1 Classification using Linear Dynamical Systems

In this approach, we model the raw pixel intensities of each frame in a video surgeme as the output of a Linear Dynamical System (LDS). More specifically, the raw pixel intensities at time instant k, $\mathbf{z}_k \in \mathbb{R}^p$, with $p \gg n$, are given by

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k, \tag{1}$$

$$\mathbf{z}_k = \mathbf{C}\mathbf{x}_k + \mathbf{n}_k, \tag{2}$$

where $\mathbf{x}_k \in \mathbb{R}^n$ is an unobserved (latent) continuous state, $\mathbf{u}_k$ is the state driving process (assumed to be Gaussian) with zero mean and identity covariance, i.e., $\mathbf{u}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\mathbf{n}_k$ represents the measurement noise, also Gaussian with $\mathbf{n}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$ and independent from $\mathbf{u}_k$. The matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ describe, respectively, the dynamics of the state variable, the correlation among the driving process samples and the mapping of the latent state to the observed signal.

Given a video surgeme, we identify the system's parameters $\mathbf{A}, \mathbf{B}, \mathbf{C}$ and $\mathbf{R}$ using a sub-optimal, but computationally efficient, method based on Principal Component Analysis proposed in [15]. Once, we have identified an LDS for each video surgeme, we need a distance to asses how close two given surgeme models are. A survey of different metrics that could be used can be found in [16]. We tried different distances on the space of LDSs based on subspace angles [17] (Finsler, Frobenius and Martin) and Binet-Cauchy kernels (Trace, Determinant and Max Singular Value) [17, 18]. Since the Martin and Frobenius distances performed best, we will present the results obtained with these two distances. More specifically, let $\theta_1, \ldots, \theta_{2n}$ be the subspace angles between the observability subspaces of two nth order LDS models $\mathcal{M}_1$ and $\mathcal{M}_2$. The (squared) Martin and Frobenius distances between the models $\mathcal{M}_1$ and $\mathcal{M}_2$ are, respectively, given by:

$$d_M^2(\mathcal{M}_1, \mathcal{M}_2) = -\log \prod_{i=1}^{2n} \cos^2(\theta_i) \quad \text{and} \quad d_F^2(\mathcal{M}_1, \mathcal{M}_2) = 2 \sum_{i=1}^{2n} \sin^2(\theta_i). \tag{3}$$

These distances can be used to classify new surgemes using a nearest neighbor approach. In our experiments we have used them to train a Support Vector Machine (SVM) classifier with a Radial Basis Function (RBF) kernel. That is $k(\mathcal{M}_i, \mathcal{M}_j) = e^{-\gamma d_X^2(\mathcal{M}_i, \mathcal{M}_j)}$, where $d_X = d_M$ or d_F and $\gamma > 0$ is a parameter.

2.2 Classification using Bag of Spatio-Temporal Features

The second approach is based on the Bag of Features (BoF) approach, a widely used technique for object recognition [19]. In the standard BoF approach, some salient features (e.g., SIFT features [20]) are first extracted from images of different objects. These features are then clustered to learn a dictionary of visual words given by the cluster centers. Each image is then represented in terms of the dictionary using a histogram, and classifiers are trained to recognize new images based on their histograms. The BoF approach can also be extended to action recognition tasks. The most direct way to do so is to build a histogram for each video, where the features are extracted from groups of frames rather than from a single image (see, e.g., [21–23]).

In the case of surgical gesture recognition, we extract Space-Time Interest Points (STIP) [21] from each video surgeme. STIP are salient points where the video has significant variations both in space and in time (as opposed to uniform regions). Hence, STIP can be seen as an extension of space corners to the space-time domain. Moreover, STIP are always detected in correspondence of motion, thus most of the information contained in the static background is automatically discarded. A 3D cuboid is then centered around each of the detected STIP and the local information contained in the cuboid is used to build a 72-bin histogram of oriented gradients (HOG) and a 90-bin histogram of optical flow (HOF), as described in [24]. Therefore, each STIP is described with a vector of size 162 that contains gradient and motion information. The HOG-HOF features extracted from a training set of videos are then clustered by K-means to form a dictionary of N words and histograms of words are built for each video surgeme. Given these histograms, we compute the χ^2-kernel and train an SVM classifier for each surgeme.

2.3 Classification using Multiple Kernel Learning

Both the LDS and BoF techniques previously described use visual data. However, while the LDS approach tries to capture the dynamics of the scene, the BoF approach is based on sparse (due to feature detection) local structures of the frame (captured by HOG) and very small and sparse motion (captured by HOF). Hence, it seems natural to think about a strategy that integrates these complementary techniques.

One way of combining the LDS and BoF approaches is to exploit the fact that both techniques use a kernel to train an SVM classifier. Therefore, we can combine the kernels using a Multiple Kernel Learning (MKL) framework [25]. In this framework, the SVM optimization problem is solved with respect to a new kernel obtained as a weighted linear combination of a set of given kernels. Thus, the principle behind MKL is to simultaneously solve for the classifier parameters and the kernel weights. Specifically, given a training set of features $\{\mathbf{x}_i\}$ and their labels $\{y_i\}$, the objective is to learn a classification function of the form $f(\mathbf{x}) = \mathbf{w}^t\phi(\mathbf{x}) + b$, where the kernel is given by $\phi^t(\mathbf{x}_i)\phi(\mathbf{x}_j) = \sum_k d_k\phi_k(\mathbf{x}_i)^t\phi_k(\mathbf{x}_j)$,

with d_k being the weight of each kernel k. The problem, therefore, is:

$$\min_{\mathbf{w},b,\mathbf{d}} \frac{1}{2}\mathbf{w}^\top\mathbf{w} + C\sum_i l(y_i, f(\mathbf{x}_i)) + r(\mathbf{d}), \quad \text{subject to } d_k \geq 0, \tag{4}$$

where $r(\cdot)$ is a regularizer (ℓ_1 or ℓ_2 norm), $l(y_i, f(\mathbf{x}_i)) = \max(0, 1 - y_i f(\mathbf{x}_i))$ is the loss function and $C > 0$ is a parameter that sets the trade-off between maximizing the margin and minimizing the loss [25].

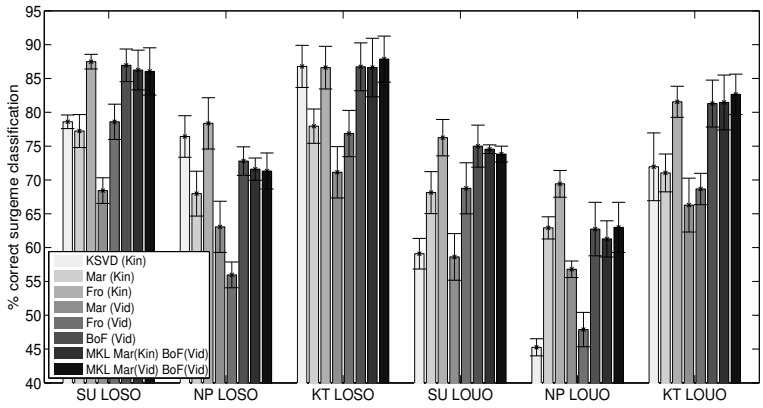
3 Experiments

Surgical data. For our tests we used the California dataset [3]. The dataset consists of three different tasks: suturing (SU, 39 trials), needle passing (NP, 26 trials) and knot tying (KT, 36 trials). Each task is performed by 8 surgeons with different skill levels. Typically each surgeon performed around 3 to 5 trials for each task. 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 consists of JPEG images of size 320×240.

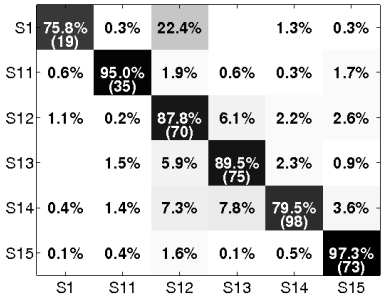
The data was manually segmented based on the surgeme's definition of [3]. Specifically, the vocabulary of possible atomic actions consisted of 14 surgemes: 1) reaching for needle with right hand, 2) positioning needle, 3) pushing needle through tissue, 4) transferring needle from left to right, 5) moving to center with needle in grip, 6) pulling suture with left hand, 7) pulling suture with right hand, 8) orienting needle, 9) using right hand to help tighten suture, 10) loosening more suture, 11) dropping suture at end and moving to end points, 12) reaching for needle with left hand, 13) making 'C' loop around right hand, 14) right hand reaches for suture and 15) both hands pull.

Results. In order to compare the accuracy of the surgeme recognition task using kinematic versus visual data, we created two different test setups. The first setup is the *leave-one-super-trial-out* (LOSO), where we leave one trial for each one of the users out for testing. The second setup is the *leave-one-user-out* (LOUO), where we leave all the trials from one user out for testing. For each task we performed a training and a test phase using only the surgemes that appeared in that task.

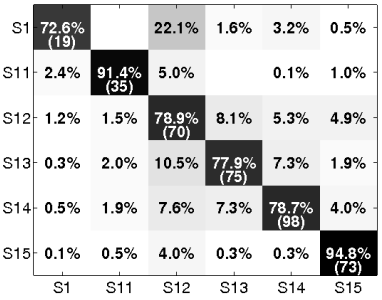
Note that the LDS approach is not restricted to video data, in fact we also present here the results of LDS with kinematic data. For kinematic data, an additional approach based on sparse dictionary learning (KSVD) [9] is evaluated. With the exception of [9], all other techniques use the SVM classifier (one-versus-one multi-class classification) [26]. The SVM penalty parameter C is estimated using 3-fold cross validation. We empirically set $\gamma = 10^{-3}$ for the RBF kernel, $n = 15$ for the order of the LDS, and $N = 300$ for the size of the BoF dictionary. For MKL, we use ℓ_2 norm regularization on the kernel weights. In order to avoid over-fitting in favor of the most frequent surgemes, we randomly sample no more than 40 surgemes per class and average the results over 20 repetitions.



(a) Average surgeme classification rates



(b) Confusion matrix KT task MKL Mar (Vid) BoF (Vid) - LOSO



(c) Confusion matrix KT task MKL Mar (Vid) BoF (Vid) - LOUO

Fig. 2. Results of Kinematic- (Kin) and Video- (Vid) based techniques.

The performance is measured as the percentage of correctly identified surgemes averaged over all tests and repetitions for each setup (see Fig. 2). The intervals at the top of the bars of Fig. 2(a) correspond to the average standard deviation for that experiment. Fig. 2(b) and 2(c) show the confusion matrices for Knot Tying (with video data) for the LOSO and LOUO setups, respectively. The numbers in parentheses along the main diagonal represent the number of times that the corresponding surgeme appeared in the dataset.

Among the kinematic-based algorithms the LDS technique with the Frobenius distance outperforms the LDS with Martin distance and the KSVD approach of [9] in almost all of the cases. The combination of gradient and optical flow features extracted by BoF leads to higher accuracy than LDS on video in all of the cases. When merging LDS and BoF using the MKL framework, the average accuracy seems to be slightly improved although not in all of the tested cases. However, as shown in Fig. 2(b) and 2(c), the errors become almost equally spread among classes.

Particularly interesting is the LOUO test, which provides an insight into the ability of the algorithms to generalize and recognize actions performed by users

that were unseen during the training phase. The results show that kinematic- and video-based algorithms are able to generalize equally well in this setting. Overall, we observe a decrease in performance of around 10 percentage points for all approaches, with KSVD being the most sensitive.

4 Conclusion

We have proposed three methods for surgical gesture classification from video data. The results showed that video data can be as discriminative as kinematic data. However, in this paper we used fairly low-level visual features, such as image intensities, image gradients and optical flow. Future work includes using more advanced visual features, such as detection and tracking of surgical tools.

Acknowledgments. This work was funded by NSF grants 0931805 and 0941362, and by the Talentia Fellowships Programme of the Andalusian Regional Ministry of Economy, Innovation and Science. The authors thank Intuitive Surgical and Carol Reiley for providing the dataset, and Greg Hager and Nicolas Padoy for discussions about the use of dynamical models for surgical gesture recognition.

References

1. 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) (2002) 49–61
2. McKenzie, C., Ibbotson, J., Cao, C., Lomax, A.: Hierarchical decomposition of laparoscopic surgery: A human factors approach to investigating the operating room environment. Journal of Minimally Invasive Therapy and Allied Technologies **10(3)** (2001) 121–127
3. 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. (2008) 396–401
4. Dosis, A., Bello, F., Gillies, D., Undre, S., Aggarwal, R., Darzi, A.: Laparoscopic task recognition using hidden Markov models. Studies in Health Technology and Informatics **111** (2005) 115–122
5. Reiley, C.E., Hager, G.D.: Task versus subtask surgical skill evaluation of robotic minimally invasive surgery. In: MICCAI. (2009) 435–442
6. Varadarajan, B.: Learning and inference algorithms for dynamical system models of dextrous motion. PhD thesis, Johns Hopkins University (2011)
7. Varadarajan, B., Reiley, C., Lin, H., Khudanpur, S., Hager, G.: Data-derived models for segmentation with application to surgical assessment and training. In: MICCAI. (2009) 426–434
8. Leong, J., Nicolaou, M., Atallah, L., Mylonas, G., Darzi, A., Yang, G.: HMM assessment of quality of movement trajectory in laparoscopic surgery. In: MICCAI. (2006) 752–759
9. Tao, L., Elhamifar, E., Khudanpur, S., Hager, G., Vidal, R.: Sparse hidden Markov models for surgical gesture classification and skill evaluation. In: Conference on Information Processing in Computer-Assisted Interventions. (2012)

10. Blum, T., Feussner, H., Navab, N.: Modeling and segmentation of surgical workflow from laparoscopic video. In: MICCAI. (2010) 400–407
11. Padoy, N., Blum, T., Ahmadi, S., Feussner, H., Berger, M., Navab, N.: Statistical modeling and recognition of surgical workflow. Medical Image Analysis **16**(3) (2012) 632 – 641
12. Lalys, F., Riffaud, L., Bouget, D., Jannin, P.: An application-dependent framework for the recognition of high-level surgical tasks in the OR. In: MICCAI. (2011) 331–338
13. Miyawaki, F., Masamune, K., Suzuki, S., Yoshimitsu, K., Vain, J.: Scrub nurse robot system - intraoperative motion analysis of a scrub nurse and timed-automata-based model for surgery. Transactions on Industrial Electronics **52**(5) (2005) 1227–1235
14. Lin, H.: Structure in surgical motion. PhD thesis, Johns Hopkins University (2010)
15. Doretto, G., Chiuso, A., Wu, Y., Soatto, S.: Dynamic textures. Int. Journal of Computer Vision **51**(2) (2003) 91–109
16. Chaudhry, R., Vidal, R.: Recognition of visual dynamical processes: Theory, kernels and experimental evaluation. Technical Report 09-01, Department of Computer Science, Johns Hopkins University (2009)
17. Cock, K.D., Moor, B.D.: Subspace angles and distances between ARMA models. System and Control Letters **46**(4) (2002) 265–270
18. Martin, A.: A metric for ARMA processes. IEEE Trans. on Signal Processing **48**(4) (2000) 1164–1170
19. Dance, C., Willamowski, J., Fan, L., Bray, C., Csurka, G.: Visual categorization with bags of keypoints. In: European Conference on Computer Vision. (2004)
20. Lowe, D.G.: Object recognition from local scale-invariant features. In: IEEE Conf. on Computer Vision and Pattern Recognition. (1999) 1150–1157
21. Laptev, I.: On space-time interest points. Int. Journal of Computer Vision **64(2-3)** (2005) 107–123
22. Willems, G., Tuytelaars, T., Gool, L.J.V.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: European Conference on Computer Vision. (2008)
23. Chaudhry, R., Ravichandran, A., Hager, G., Vidal, R.: Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In: IEEE Conference on Computer Vision and Pattern Recognition. (2009)
24. Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: British Machine Vision Conference. (2009) 1–11
25. Varma, M., Babu, R.: More generality in efficient multiple kernel learning. In: International Conference on Machine Learning. (2009) 1065–1072
26. Chang, C.C., Lin, C.J.: LIBSVM : a library for support vector machines. (2001) Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.