An Improved Model for Segmentation and Recognition of Fine-grained Activities with Application to Surgical Training Tasks

Colin Lea Gregory D. Hager René Vidal
Johns Hopkins University
clea1@jhu.edu, hager@cs.jhu.edu, rvidal@cis.jhu.edu

Abstract

Automated segmentation and recognition of fine-grained activities is important for enabling new applications in industrial automation, human-robot collaboration, and surgical training. Many existing approaches to activity recognition assume that a video has already been segmented and perform classification using an abstract representation based on spatio-temporal features. While some approaches perform joint activity segmentation and recognition, they typically suffer from a poor modeling of the transitions between actions and a representation that does not incorporate contextual information about the scene. In this paper, we propose a model for action segmentation and recognition that improves upon existing work in two directions. First, we develop a variation of the Skip-Chain Conditional Random Field that captures long-range state transitions between actions by using higher-order temporal relationships. Second, we argue that in constrained environments, where the relevant set of objects is known, it is better to develop features using high-level object relationships that have semantic meaning instead of relying on abstract features. We apply our approach to a set of tasks common for training in robotic surgery: suturing, knot tying, and needle passing, and show that our method increases micro and macro accuracy by 18.46% and 44.13% relative to the state of the art on a widely used robotic surgery dataset.

1. Introduction

In this work we approach the problem of automated segmentation and recognition of actions in constrained environments. We define a fine-grained activity as a sequence of action primitives that take place in a specific environment with a finite set of relevant objects. Some examples with this structure include cooking [19, 10], sports [12], and many robotic manipulation tasks [33]. In this paper we target three robotic surgical training tasks: suturing, knot tying, and needle passing

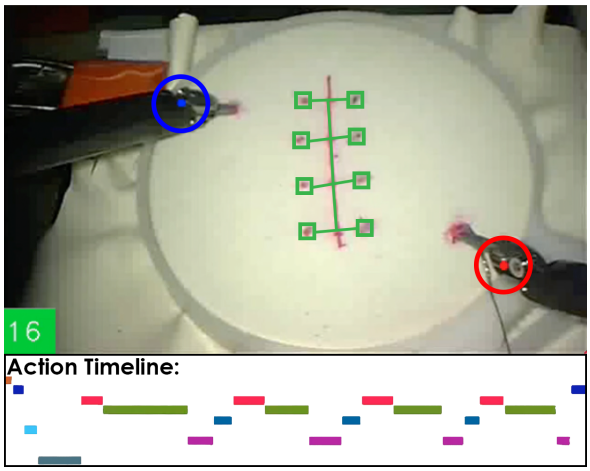

Figure 1. (Top) Image from the suturing video in the JIGSAWS dataset. Positions of the robot tools are highlighted in blue and red. Insertion points extracted using a Deformable Parts Model are highlighted in green. (Bottom) The timeline depicting which actions occur when. Each of the 10 colors is a unique action primitive.

We focus on constrained environments in efforts to develop high-level features based on object relationships. Numerous papers in recent years have evaluated large scale video datasets that contain a large number of objects. The scope of these datasets is too large to reliably recognize each of the objects using current algorithms. In constrained environments it is reasonable to assume that most objects can be recognized. In our surgical environment there are three dominant objects: a structured set of insertion points for suturing, a set of nodes for needle passing, and a rod with suture knots for knot tying.

There have been many calls for improving the quality and efficacy of training for robotic surgery in recent years [2]. Current methods for skill evaluation in surgical training tasks tends to be either too subjective or too time-consuming. By automatically and quantitatively evaluating users it is hypothesized that the current inter-reviewer

variability and bias can be reduced. By performing activity recognition on this training data we believe that we can more robustly evaluate the skill of a user. Furthermore, recognizing actions in real time enables the creation of smart assistants that notify a user of an error or, in a robotic setting, physically assist the user with a task.

It has become common in recent years for fine-grained tasks to contain multi-modal data. For example, datasets in a home setting often contain human skeleton positions extracted from RGBD images (e.g. using a Kinect). In the surgical data we have video from each task as well as all of the joint positions and rotations from a medical robot. Given that in both cases we have pose information for the human/robot, it behooves us to relate this to known object-level information.

In this paper we address two questions. First, how can we best capture transitions between actions? Common time series models like Hidden Markov Models and Linear Chain Conditional Random Fields assume each frame is only linked to the frame before it. However, since a video is typically recorded at 30 frames per second and each action can last several seconds, frame-to-frame transition probabilities can be very different from action-to-action transition probabilities. Moreover, different videos can have different frame rates, and thus having models based on frame-to-frame transition probabilities could lead to erroneous segmentation. We posit that using models that capture long-range temporal interactions provides superior performance than using frame-to-frame interactions.

Second, does using high-level object information improve our recognition performance? Many recent methods rely on abstract feature representations that extract texture patches in an image. While these may be appropriate for large-scale datasets with numerous, diverse events, it is unclear that they are the correct tool for fine-grained activities. We posit that using these high-level relationships provides superior recognition performance than abstract feature representations.

Our contributions are as follows:

- We develop of a variation on the Skip-Chain Conditional Random Field to better capture transitions between actions.
- We propose use of a Deformable Part Model to capture high-level features relating the robot and object parts in an image.
- We evaluate on suturing, needle passing, and knot tying data from a widely used dataset for robot surgery training.

Our contributions increase the average Leave One User Out micro and macro accuracy on the dataset by 18.46% and 44.13% respectively relative to the state of the art [26].

2. Related Work

Action Analysis

Recent work on activity recognition in the computer vision community tends to gravitate towards large-scale recognition of diverse actions. In these papers, humans are either performing actions in isolation [22] or interacting with a large assortment of objects that cannot easily be identified [24]. Work on fine-grained activity analysis has modeled activities like cooking and sports. Lei *et al.* [10] use hand and object detectors to extract semantic information from a kitchen activity. They then extract abstract trajectory features to recognize a set of seven actions such as *placing*, *chopping*, and *mixing* ingredients. Rohrbach *et al.* [19] evaluate human pose-based models and "holistic" models that are composed of features like Histogram of Oriented Gradients and Motion Boundary Histograms. They then use a Support Vector Machine to classify actions.

There has been other work focusing on modeling sequences of actions at a higher level. For example [12] Morariu and Davis use Markov Logic Networks with Allen Temporal Logic to recognize events happening on two-player basketball games. Graphical models like Hidden Markov Models and Conditional Random Fields have been a mainstay in modeling of complex activities. Other kinds of Markov models have also been developed for activity analysis including Hidden Semi Markov Models [32], Coupled Hidden Semi Markov Models [15], and Max-Margin Hidden Conditional Random Fields [13].

Activity Recognition for Surgical Training Tasks

There have been many models developed to recognize activities in surgical data [25, 26, 33, 7, 29]. Much of this work has been directed towards variations on Hidden Markov Models [25, 20, 17, 11], Conditional Random Fields [26], and Linear Dynamical Systems [29, 33, 3]. The typical approach is to take the input data including features such as positions, rotations, joint configurations, and joint velocities and feed it into a temporal model. Few approaches take advantage of video captured by the robot. In [26] they use the video in a data-driven approach using Bag of Words with Dense Trajectory features. This is the most similar to our work, except their image features do not elicit semantic meaning.

Object Models

Part-based models have recently become popular for decomposing objects into a set of parts. In [1] they introduce an efficient technique for performing inference based on the original work of [5] on Pictorial Structures. Similar techniques has substantially improved human pose estimation [31, 21] and object recognition and localization

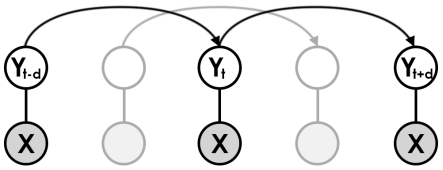

Figure 2. A Skip-Chain Conditional Random Field is used to capture state transitions over large periods of time. In this figure we show a skip-length of 2.

[4, 1]. While originally these approaches only worked on images, recent work has extended the idea to the video domain [27, 16, 30]. Results from these papers show promise and have influenced our idea of using high-level object representations in activity recognition. However, they do not explicitly model interactions between the objects in the environment. Instead they use textural features computed in the area surrounding the part locations.

3. Methodology

In this section we discuss our models for predicting actions in time series data using a Skip-Chain CRF and for detecting object part locations using a Deformable Part Model. In addition we show how we can compute additional features for the SC-CRF that are based on relationships between objects in the scene.

3.1. Skip-Chain Conditional Random Field

We propose a variation on the Skip-Chain Conditional Random Field (SC-CRF) [23], shown in Figure 2, that is more capable of capturing transitions in actions over many frames than typical linear chain CRF models. We describe this model as follows. Let X_t be a feature extracted at time t from both kinematic and video data. Let Y_t be the label (action primitive) being performed at time t. We model $Y = (Y_1, ..., Y_T)$ given $X = (X_1, ..., X_T)$ with a Gibbs distribution $P(Y|X) \propto \exp(-E(X,Y))$ using the following energy:

$$E_C(X,Y) = \sum_{t=1}^{T} w_{cu}^T \phi_c(X_t, Y_t) + \sum_{t=d}^{T} w_{cp}^T \psi_c(Y_t, Y_{t-d}) + \sum_{t=d}^{T} w_{cs}^T \gamma_c(X_t, X_{t-d}, Y_t) \quad (1)$$

where ϕ_c, ψ_c, and γ_c are, respectively, unary, pairwise and skip-length potentials to be described below, and w_{cu}, w_{cp}, w_{cs} are their corresponding parameters.

Observe that this energy can be rewritten as a combination of weights $w = [w_{cu}, w_{cp}, w_{cs}]^T$ and a function $\Psi(X,Y)$ representing the unary, pairwise, and skip-length features:

$$E_c(X,Y) = w^T \Psi(X,Y) \quad (2)$$

where Ψ is a concatenation of the sum of each potential over time:

$$\Psi(X,Y) = \begin{bmatrix} \sum_{t=1}^{T} \phi_c(X_t, Y_t) \\ \sum_{t=d}^{T} \psi_c(Y_t, Y_{t-d}) \\ \sum_{t=d}^{T} \gamma_c(X_t, X_{t-d}, , Y_t) \end{bmatrix} \quad (3)$$

Unary Potentials: The unary potential models the cost of assigning label Y_t to frame t, given feature X_t. Here X_t is a subset of the kinematic data and features extracted from the image-based object relationship data. Our unary term is simply a linear combination of each of these features for each class. Thus w_{cu} is a vector of size F_u, the number of features.

Pairwise Potentials: The pairwise potential is the cost of transitioning from label Y_{t-d} to Y_t where d is the skip-length. The vector $\psi_c(Y_t, Y_{t-d})$ is 1 for the index (Y_t, Y_{t-d}) and 0 elsewhere. The pairwise parameter w_{cp} is of size L^2 where L is the number of action primitives.

Skip-length Data Potentials: The skip-length data potential is the cost of assigning label Y_t to data X_t and X_{t-d}. In particular this is used with binary features such as gripper state and part occlusion state to model transitions over time. This is modeled with a Dirac delta function $\gamma_c(X_t^f, X_{t-d}^f) = \delta(X_t^f - X_{t-d}^f)$ where f is the index of a binary feature. There are $L \cdot F_s$ parameters in w_{cs} where F_s is the number of skip-length features.

Inference

Typically, inference in higher order models has a high computationally complexity. However, given the structure of our skip-chains we can efficiently perform inference using a modified version of the Viterbi algorithm. As shown in Figure 2, a sequence can be viewed as a set of d independent chains. Thus, we can compute Viterbi on every chain and then merge the results into a single output by interlacing.

Learning

We learn the parameters of the SC-CRF using the Structural Suport Vector Machine formulation proposed by [28] where we minimize:

$$\min_{w,\xi} ||w||_2 + C \sum_{n=1}^{N} \xi_n \quad \text{s.t.}$$
$$w^T(\Psi(X^n, Y^n) - \Psi(X^n, Y)) \geq \Delta(Y^n, Y) - \xi^n$$
$$n = 1...N, \forall Y \in \mathcal{Y} \quad (4)$$

X^n and Y^n are sample sequences in our training set, Δ is the Hamming loss function, and ξ is a slack variable. Recall

that each sequence can be split into d independent chains. For training, we separate each training example into d parts $X^n = \{X^{n_i}\}_{i=1}^d$.

We use Block Coordinate Frank Wolfe [9], as implemented in pyStruct [14], to jointly learn all parameters. In total there are L^2 pairwise transition parameters, $L \cdot F_u$ unary parameters, and $L \cdot F_s$ skip-length parameters.

3.2. Object Model

We propose to use a Deformable Part Model (DPM) [1] to detect and localize objects in videos. We model this as a graph where nodes are object parts and edges act as springs that regulate the distance between nodes as shown in Figure 4.1. The unary term is the output of template-matching performed at each location in the image. Pairwise scores are a function of the distance between nodes connected in the graph. Let Z_i be the part index for node i and I be an image. Our goal is to find the most likely configuration, given an image, by modeling $P(Z|I) \propto \exp(-E_V(I, Z))$ where

$$E_v(I, Z) = \sum_{i \in Nodes} w_{vu}^T \phi_v(I, Z_i) + \sum_{e \in Edges(i)} w_{ve}^T \psi_v(I, Z_i, Z_j) \quad (5)$$

Potentials: The unary term $\phi_v(I, Z_i)$ is the output of a cross correlation-based template-matching technique with template Z_i^t at image location Z_i^{loc}. The pairwise scores are computed as the distance between a given configuration location and the expected location with

$$\psi_v(Z_i, Z_j) = (|Z_j^{loc} - Z_i^{loc}| - \mu_{ij})^T \Sigma_{ij}^{-1} (|Z_j^{loc} - Z_i^{loc}| - \mu_{ij}) \quad (6)$$

where i and j are the current and parents nodes and $\Sigma_{Z_{t_i}, Z_{t_j}}$ is the covariance of the offset between part types Z_{t_i} and Z_{t_j}.

Learning

There are two components that must be learned for the unary potential and pairwise potentials. First, for the unary potential, we need a set of templates that correspond to each object part. A single template is learned per part by averaging each labeled image per pixel for that part. We also need to learn the weights w_{vu} for a linear classifier on the unary potential. These are learned using a Support Vector Machine where the training examples are the outputs of our template matching function. For each object part, positive training examples was computed from a set of labeled training images and a large number of negative training examples was extracted from random other locations in the training images.

The parameters for the pairwise potential, which measure deformations between object parts, are computed from the mean, μ, and covariance, Σ between respective parts. In addition, during the training process we compute a minimum spanning tree on the set of labels to define which parts connect with which other parts. Using this tree allows us to perform inference in a much more tractable manner.

Inference

We use the variation of belief propagation proposed by [1]. This is an efficiently technique for inference in tree graphs for applications where we are optimizing over locations in an image. This involves a two-step procedure that computes a likelihood score for each potential part location. There is a forward pass where messages go from leaf nodes to the root and a backward pass going from the root to the leaves. The part configurations can be at any location in the image thus inference would normally be very expensive. [1] caches pairwise distance scores using the generalized distance function to efficiently optimize over the part locations. The best set of part configurations can then be found by finding the part-model with the best score in the image using the energy in equation (5).

3.3. Kinematic and Image Features

The set of features at time t, X_t, consists of both kinematic and video features. The kinematic features include the positions of the left and right robot end effector P_k, the end effector velocities V_k, binary opened/closed gripper states for each end effector G_k, and skip-length features S_k that look at the change in each gripper state between timesteps.

In addition, using the object model we develop a set of features relating the object part information to the robot kinematic data and use them as additional terms in the unary function of the SC-CRF. In particular compute two scene-based features. The first feature measures distance to the closest object part from each tool: $f_d(P_i, Z) = \min_i ||P_i - Z_i^{loc}||_2$ where P_i as the projection of a tool position on the image image and Z_i^{loc} as the position of the i^{th} object part. In the suturing videos the value should be very small when the needle is being inserted (thus occluding the insertion point) and large when the tool is far from any points. The second feature measures the relative position of each tool to the closest object part: $f_o(P_i, Z) = P_i - Z_i^{loc}$ for $i = \arg\min_i ||P_i - Z_i^{loc}||_2$. This offers a more rich representation of the tool relative to the insertion points.

For the SC-CRF results in Section 4.1 we use just the kinematic features and for the object models results in Section 4.2 we use both kinematic and image features.

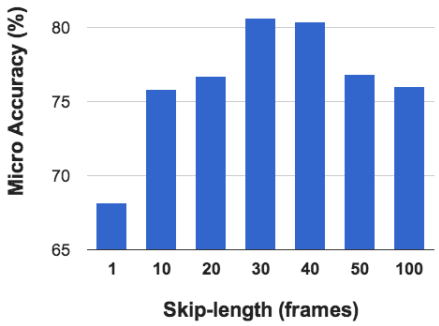

Figure 3. Results of varying the skip-length parameter in our Skip-Chain Conditional Random Field model.

4. Experiments

All experiments are performed on the JIGSAWS dataset [6], which includes eight subjects performing *suturing*, *needle passing*, and *knot tying* five times each. These tasks are decomposed into about 10 unique action primitives such as "inserting needle into skin" and "tying knot." Each performance is around two minutes long and contains 15 to 20 action primitives per video. We show the accuracy for each experiment using Leave One Super-trial Out (LOSO) and Leave One User Out (LOUO). LOSO trains on four (of five) instances from all eight users and tests on each left out instance. LOUO trains on seven users and tests on the eighth. The accuracy for each is averaged over each left out set. Micro is defined as $m_i(y, y^*) = \frac{1}{N}\sum_{i=1}^{N}\delta(y_i, y_i^*)$ where δ is 1 if y_i and y_i^* are the same label and 0 otherwise. Macro is an average of averages and is defined as $m_a(y, y^*) = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{N_c}\sum_{i=1}^{N}(\delta(y_i, y_i^*)\cdot\delta(y_i^*, c))$ where N_c is the number of true labels of class c.

4.1. Skip-Chain CRF

We vary the skip-length parameter in the SC-CRF from 1 to 100. $d = 1$ is the special case of a linear chain CRF and $d = 100$ implies that each node is connected to another node 100 frames earlier. Figure 3 shows how the micro accuracy in our system varies with parameter d. Note that because each node is independent of its neighbors for all $d > 1$ the output can be noisy. Thus for all results we apply a median filter with width d to smooth out the result.

Table 3 shows our final results compared to prior work. Each of our test setups uses the following parameters: skip-length of $d = 30$ frames, a feature vector of kinematic positions, velocities, and gripper states. All features are demeaned and normalized on a per-video per-feature basis.

We compare our method against [26] which uses a Markov Semi-Markov CRF with kinematics- and video-based features. They show their results are superior to other methods, like a baseline linear chain CRF. Due to differing test setups we are unable to directly compare our results to their baselines. One key difference is that in their approach they use abstract texture features and in our approach we use semantically meaningful object information. They use a set of 78 kinematic features that includes robot joint angles as well as master and slave position and velocity information. Our kinematic feature set is much smaller and also more discriminative.

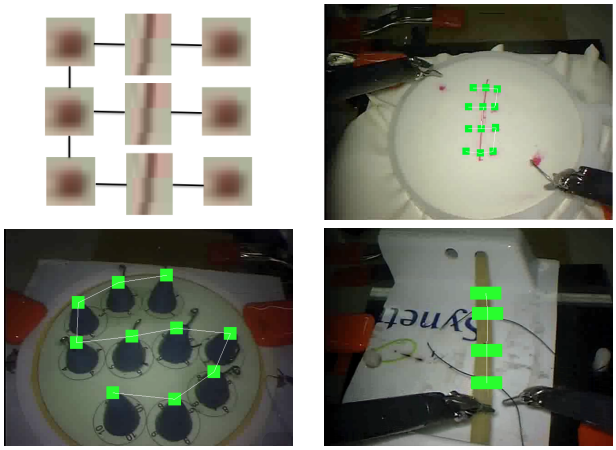

Figure 4. Deformable Part Model for three surgical tasks: (top left) cartoon diagram of the suturing model (top right) Suturing (botom left) Needle Passing (bottom right) Knot Tying

4.2. Object Model

We calculate the accuracy of the predicted positions of the object parts using LOUO testing with the Deformable Part Model. The part locations in one frame of every video were hand-annotated. We claim a part is correct if it is within half of the width of the template. The complete set of results is shown in Table 1.

Table 1. Accuracy and mean error for the deformable parts model on the suturing, needle passing, and knot tying videos.

Task	Accuracy	Error (px)
Suturing	100%	1.08
Needle Passing	92.9%	3.92
Knot Tying	91.9%	8.73

In Table 2 we evaluate the SC-CRF with various subsets of video and kinematic features. Image positions are projected from the kinematic data onto the image. For a fair comparison we also show results assuming only X and Y coordinates of the position and velocity data are available. Note that the worse results when projecting positions onto the image versus the raw X/Y kinematic data is due to an imperfect calibration of the camera model.

Table 3. Final Results using the kinematic data skip-length $d = 30$. Bold text implies the best score for that surgical task.

		Leave Super Trial Out			Leave User Out		
Method	**Metric**	**Suturing**	**Needle Passing**	**Knot Tying**	**Suturing**	**Needle Passing**	**Knot Tying**
[26]	Micro	81.35	73.02	**80.95**	71.02	63.60	63.69
	Macro	62.12	71.90	79.07	41.75	54.87	52.60
Ours	Micro	**85.24**	**77.47**	80.64	**80.29**	**75.33**	**78.91**
	Macro	**74.24**	**73.63**	**79.70**	**65.25**	**70.48**	**77.67**

Table 2. Results on Suturing using the SC-CRF with $d = 30$. *P=position, V=velocity, G=gripper,* $(\cdot)_k$*=kinematics-based,* $(\cdot)_i$*=image-based,* f_**=image features.*

	Features	**Axes**	**Micro**	**Macro**
Kine	All kinematics	-	74.97	54.06
	$P_kP_kV_kG_k$	xyz	80.29	65.25
	P_kV_k	xyz	71.52	53.21
	P_kV_k	xy	63.35	45.27
Video	P_iV_i	xy	62.60	44.44
	P_iV_i f_d	xy	63.24	46.70
	P_iV_i f_o	xy	63.24	46.70
Mix	$P_kV_kG_k$ f_d	xyz	**81.53**	**67.38**
	$P_kV_kG_k$ f_o	xyz	78.12	61.07
[26]	kinematics + vision	-	71.02	41.75

5. Discussion

Relative to the state of the art [26][1] we perform 18.46% better in micro LOUO and 44.13% better in macro LOUO. We also increase the LOSO accuracy by 3.50% micro and 11.89% macro.

In other work it is typically the case that LOSO performs substantially better than LOUO. The intuition is that each user has their own unique style and that when the model is trained with that style it is able to perform with much better accuracy; LOUO does not contain that user's style and thus would perform worse. Given that our accuracies for LOUO and LOSO have a smaller disparity than in other papers, we interpret that our model is more invariant to user style.

As depicted in Figure 3, our increase in accuracy is mostly part due to the Skip-Chain CRF. The intuition for this increase is as follows. Each node in the SC-CRF is looking back at the state d steps earlier. If d is large enough then it is likely for the state to be different than the current state. Furthermore this captures higher level structural information; instead of the pairwise term prompting the node to keep its own label it tries to push it into a different label. In this work we determined the optimal skip-length based on cross validation. Future work will look at developing an analytical solution based on the duration of each of the actions. Additional work may look at the effect of using one skip-length parameter per class as opposed to a single global parameter.

While the combined vision and kinematics results provide only a modest increase in accuracy we still believe that using this object-based information is advantageous. It would likely have a larger impact on accuracy if there was greater diversity in the dataset. It may also be possible to use the deformable parts model to develop more sophisticated features such as a measure of the deformation over time.

By assessing the results in table 2 we see that gap between the best vision results and the best combined results are due to the lack of depth (z-axis) information and the lack of gripper state. In order to use this work in the non-robotic laparoscopic setting it would be necessary to use tool tracking models [18, 8] with stereo video. It may be possible to detect the gripper state by extending one of these models.

Our method is very efficient compared to other methods like [26] which compute complex, time-consuming features such as Dense Trajectories. The Deformable Part Model runs at about 2 frames per second and inference in the SC-CRF runs at 50 frames per second. Future work may investigate how to use temporal information to compute the part model more efficiently using video. The ability to recognize actions in realtime enables new applications in human-robot collaboration and robot skill analysis.

[1] Updated results obtained from the authors

Acknowledgements

This work was funded in part by the National Science Foundation under grants CPS CNS-0931805, DGE-1232825, and ONR-N000141310116.

References

[1] Pictorial Structures for Object Recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005.

[2] C. B. Barden, M. C. Specht, M. D. McCarter, J. M. Daly, and T. J. Fahey. Effects of limited work hours on surgical training. *J. Am. Coll. Surg.*, 195(4):531–538, Oct 2002.

[3] B. B. Bejar. Classification methods for linear dynamical systems with application to surgical gesture recognition. Master's thesis, Department of Biomedical Engineering, Johns Hopkins University, 2013.

[4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analalysis and Machine Intelligence (PAMI)*, 32(9):1627–1645, Sept. 2010.

[5] M. Fischler and R. Elschlager. The Representation and Matching of Pictorial Structures. *IEEE Transactions on Computers*, C-22, 1973.

[6] Y. Gao, S. S. Vedula, C. E. Reiley, N. Ahmidi, B. Varadarajan, H. C. Lin, L. Tao, L. Zappella, B. Bejar, D. D. Yuh, C. C. G. Chen, R. Vidal, S. Khudanpur, and G. D. Hager. 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 - MICCAI Workshop*, 2014.

[7] B. B. Haro, L. Zappella, and R. Vidal. Surgical gesture classification from video data. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 15:34–41, 2012.

[8] S. Kumar, M. S. Narayanan, P. Singhal, J. J. Corso, and V. Krovi. Product of tracking experts for visual tracking of surgical tools. *IEEE International Conference on Automation Science and Engineering (CASE)*, 2013.

[9] S. Lacoste-julien, M. Jaggi, M. Schmidt, and P. Pletscher. Block-coordinate frank-wolfe optimization for structural svms. In S. Dasgupta and D. Mcallester, editors, *International Conference on Machine Learning (ICML)*, volume 28, pages 53–61, 2013.

[10] J. Lei, X. Ren, and D. Fox. Fine-grained kitchen activity recognition using rgb-d. In *ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 208–211, New York, NY, USA, 2012. ACM.

[11] C. Loukas and E. Georgiou. Surgical workflow analysis with gaussian mixture multivariate autoregressive (gmmar) models: a simulation study. *Comput Aided Surg*, 2013.

[12] V. Morariu and L. Davis. Multi-agent event recognition in structured scenarios. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2011.

[13] G. Mori. Max-margin hidden conditional random fields for human action recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 872–879, June 2009.

[14] A. C. Müller and S. Behnke. pystruct - learning structured prediction in python. *Journal of Machine Learning Research*, 15:2055–2060, 2014.

[15] P. Natarajan and R. Nevatia. Coupled Hidden Semi Markov Models for Activity Recognition. *IEEE Workshop on Motion and Video Computing (WMVC)*, 2007.

[16] M. Raptis and L. Sigal. Poselet Key-Framing: A Model for Human Activity Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2650–2657, June 2013.

[17] 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 (MICCAI)*, pages 435–442, 2009.

[18] A. Reiter, P. Allen, and T. Zhao. Feature classification for tracking articulated surgical tools. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 592–600, 2012.

[19] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele. A database for fine grained activity detection of cooking activities. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, United States, June 2012.

[20] J. Rosen, B. Hannaford, C. G. Richards, and M. N. Sinanan. Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills. *IEEE Trans. Biomed. Engineering*, 48(5):579–591, 2001.

[21] B. Sapp, A. Toshev, and B. Taskar. Cascaded Models for Articulated Pose Estimation. In *IEEE European Conference on Computer Vision (ECCV)*, volume 6312. 2010.

[22] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. *International Conference on Pattern Recognition (ICPR)*, 2004.

[23] C. Sutton and A. Mccallum. *Introduction to Conditional Random Fields for Relational Learning*. MIT Press, 2006.

[24] K. Tang, L. Fei-Fei, and D. Koller. Learning latent temporal structure for complex event detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[25] L. Tao, E. Elhamifar, S. Khudanpur, G. D. Hager, and R. Vidal. Sparse hidden markov models for surgical gesture classification and skill evaluation. In *International Conference on Information Processing in Computer-Assisted Interventions (IPCAI)*, volume 7330 of *Lecture Notes in Computer Science*, pages 167–177. Springer, 2012.

[26] L. Tao, L. Zappella, G. Hager, and R. Vidal. Surgical Gesture Segmentation and Recognition. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2013.

[27] Y. Tian, R. Sukthankar, and M. Shah. Spatiotemporal Deformable Part Models for Action Detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2642–2649, June 2013.

[28] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, Dec. 2005.

[29] B. Varadarajan. *Learning and inference algorithms for dynamical system models of dextrous motion*. PhD thesis, Johns Hopkins University, 2011.

[30] C. Wang, Y. Wang, and A. L. Yuille. An Approach to Pose-Based Action Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 915–922, June 2013.

[31] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1385–1392. Dept. of Computer Science, University of California, Irvine, IEEE, 2011.

[32] S.-Z. Yu. Hidden semi-Markov models. *Artificial Intelligence*, 174(2):215–243, Feb. 2010.

[33] L. Zappella, B. Béjar, G. Hager, and R. Vidal. Surgical gesture classification from video and kinematic data. *Medical image analysis*, 17(7):732–45, Oct. 2013.