

# End-to-End Fine-Grained Action Segmentation and Recognition Using Conditional Random Field Models and Discriminative Sparse Coding

Effrosyni Mavroudi<sup>†</sup>, Divya Bhaskara<sup>†‡</sup>, Shahin Sefati<sup>†‡</sup>, Haider Ali<sup>†</sup> and René Vidal<sup>†</sup>  
<sup>†</sup>Johns Hopkins University, <sup>‡</sup>University of Virginia, <sup>‡</sup>Comcast AI Research

## Abstract

*Fine-grained action segmentation and recognition is an important yet challenging task. Given a long, untrimmed sequence of kinematic data, the task is to classify the action at each time frame and segment the time series into the correct sequence of actions. In this paper, we propose a novel framework that combines a temporal Conditional Random Field (CRF) model with a powerful frame-level representation based on discriminative sparse coding. We introduce an end-to-end algorithm for jointly learning the weights of the CRF model, which include action classification and action transition costs, as well as an overcomplete dictionary of mid-level action primitives. This results in a CRF model that is driven by sparse coding features obtained using a discriminative dictionary that is shared among different actions and adapted to the task of structured output learning. We evaluate our method on three surgical tasks using kinematic data from the JIGSAWS dataset, as well as on a food preparation task using accelerometer data from the 50 Salads dataset. Our results show that the proposed method performs on par or better than state-of-the-art methods.*

## 1. Introduction

Temporal segmentation and recognition of complex activities in long continuous recordings is a useful, yet challenging task. Examples of complex activities comprised of fine-grained goal-driven actions that follow a grammar are surgical procedures [9], food preparation [31] and assembly tasks [35]. For instance, in the medical field there is a need to better train surgeons in performing surgical procedures using new technologies such as the daVinci robot. One possible approach is to use machine learning and computer vision techniques to automatically determine the skill level of the surgeon from kinematic data of the surgeon’s performance recorded by the robot [9]. Such an approach typically requires an accurate classification of the surgical gesture at each time frame [3] and a segmentation of the surgical task into the correct sequence of gestures [34]. Another

example of a complex activity with goal-driven fine-grained actions following a grammar is cooking. Although the actions performed while preparing a recipe and their relative ordering can vary, there are still temporal relations among them. For instance, the action *stir milk* usually happens after *pour milk*, or the action *fry egg* usually follows the action *crack egg*. Robots equipped with the ability to automatically recognize actions during food preparation could assist individuals with cognitive impairments in their daily activities by providing prompts and instructions. However, the task of fine-grained action segmentation and recognition is challenging due to the subtle differences between actions, the variability in the duration and style of execution among users and the variability in the relative ordering of actions.

Existing approaches to fine-grained action segmentation and recognition use a *temporal model* to capture the temporal evolution and ordering of actions, such as Hidden Markov Models (HMMs) [13, 32], Conditional Random Fields (CRF) [16, 17], Markov semi-Markov Conditional Random Fields (MsM-CRF) [34], Recurrent Neural Networks [8, 28] and Temporal Convolutional Networks (TCNs) [15]. However, such models cannot capture subtle differences between actions without a powerful, discriminative and robust *representation of frames* or short temporal segments. Sparse coding has emerged as a powerful signal representation in which the raw data in a certain time frame is represented as a linear combination of a small number of basis elements from an overcomplete dictionary. The coefficients of this linear combination are called *sparse codes* and are used as a new representation for temporal modeling. However, since the dictionary is typically learned in an unsupervised manner by minimizing a regularized reconstruction error [1], the resulting representation may not be discriminative for a given learning task. Task-driven discriminative dictionary learning addresses this issue by coupling dictionary and classifier learning [24]. For example, Sefati et al. [30] propose an approach to fine-grained action recognition called Shared Discriminative Sparse Dictionary Learning (SDSDL), where sparse codes are extracted at each time frame and a frame feature is computed by average pooling the sparse codes over a short temporal window

surrounding the frame. The dictionary is jointly learned with the per-frame classifier parameters, resulting in a discriminative mid-level representation that is shared across all actions/gestures. However, their approach lacks a temporal model, which is crucial for modeling temporal dependencies. Although prior work [38] has combined discriminative dictionary learning with CRFs for the purpose of saliency detection, such work is not directly applicable to fine-grained action recognition.

In this work we propose a joint model for fine-grained action recognition and segmentation that integrates a CRF for temporal modeling and discriminative sparse coding for frame-wise action representation. The proposed CRF models the temporal structure of long untrimmed activities via unary potentials that represent the cost of assigning an action label to a frame-wise representation of an action obtained via discriminative sparse coding, and pairwise potentials that capture the transitions between actions and encourage smoothness of the predicted label sequence. The parameters of the combined model are trained jointly in an end-to-end manner using a max-margin approach. Our experiments show competitive performance in the task of fine-grained action recognition, especially in the regime of limited training data. In summary, the contributions of this paper are three-fold:

1. We propose a novel framework for fine-grained action segmentation and recognition which uses a CRF model whose target variables (action labels per time step) are conditioned on sparse codes.
2. We introduce an algorithm for training our model in an end-to-end fashion. In particular, we jointly learn a task-specific discriminative dictionary and the CRF unary and pairwise weights by using Stochastic Gradient Descent (SGD).
3. We evaluate our model on two public datasets focused on goal-driven complex activities comprised of fine-grained actions. In particular, we use robot kinematic data from the JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) [9] dataset and evaluate our method on three surgical tasks. We also experiment with accelerometer data from the 50 Salads [31] dataset for recognizing actions that are labeled at two levels of granularity. Results show that our method performs on par with most state-of-the-art methods.

## 2. Related Work

The task of fine-grained action segmentation and recognition has recently received increased attention due to the release of datasets such as MPII Cooking [29], JIGSAWS [9] and 50 Salads [31]. In this section, we briefly review some of the main existing approaches for tackling this

problem. Besides, we briefly discuss existing work on discriminative dictionary learning. Note that since the focus of this paper is fine-grained action recognition from kinematic data, we do not discuss approaches for feature extraction or object parsing from video data.

**Fine-grained action recognition from kinematic data.** A straightforward approach to action segmentation and classification is the use of overlapping temporal windows in conjunction with temporal segment classifiers and non-maximum suppression (e.g., [29, 25]). However this approach does not exploit long-range temporal dependencies.

Recently, deep learning approaches have started to emerge in the field. For instance, in [8] a recurrent neural network (Long Short Term Memory network - LSTM) is applied to kinematic data, while in [15] a Temporal Convolutional Network composed of 1D convolutions, nonlinearities and pooling/upsampling layers is introduced. Although these models yield promising results, they do not explicitly model correlations and dependencies among action labels.

Another line of work, including our proposed method, takes into account the fact that the action segmentation and classification problem is a structured output prediction problem due to the temporal structure of the sequence of action labels and thus employs structured temporal models such as HMMs and their extensions [32, 13, 14]. Among them, the work that is most related to this work is Sparse-HMMs [32], which combines dictionary learning with HMMs. However, a Sparse-HMM is a generative model in which a separate dictionary is learned for each action class. In this work we use a CRF, which is a discriminative model, and we learn a dictionary that is shared among all action classes. Discriminative models like CRFs [16, 17], semi-Markov CRFs [34] have gained popularity since they allow for flexible energy functions. Other types of temporal models include a duration model and language model recently proposed in [27] for modeling action durations and context. The input to these temporal models are either the kinematic data themselves or features extracted from them. For instance, in the Latent Convolutional Skip Chain CRF (LC-SC-CRF) [17] the responses to convolutional filters, which capture latent action primitives, are used as features.

**Discriminative Dictionary Learning.** Task-driven discriminative dictionary learning was introduced in the seminal work of Mairal et al. [24] and couples the process of dictionary learning and classifier training, thus incorporating supervised learning to sparse coding. Since then discriminative dictionary learning has enjoyed many successes in diverse areas such as handwritten digit classification [22, 39], face recognition [10, 39, 26], object category recognition [10, 26, 5], scene classification [5, 19, 26], and action classification [26].

The closest work to ours is the Shared Discriminative

Sparse Dictionary Learning (SDSDL) proposed by Sefati et al. [30], where sparse codes are used as frame features and a discriminative dictionary is jointly learned with per frame action classifiers for the task of surgical task segmentation. Our work builds on top of this model by replacing the per-frame classifiers, which compute independent predictions per frame, with a structured output temporal model (CRF), which takes into account the temporal dependencies between actions. While prior work has considered joint dictionary and CRF learning [33, 37, 38] for the tasks of semantic segmentation and saliency estimation, our work differs from these previous approaches in three key aspects. First, to the best of our knowledge, we are the first to apply joint dictionary and CRF learning to the task of action segmentation and classification. Second, we are learning unary CRF classifiers and pairwise transition scores, while in [33] only two scalar variables encoding the relative weight between the unary and pairwise potentials are learned. Third, we use local temporal average-pooling of sparse codes as a feature extraction process for capturing local temporal context instead of the raw sparse codes used in [37, 38].

### 3. Technical Approach

In this section, we introduce our temporal CRF model and frame-wise representation based on sparse coding and describe our algorithm for training our model. Figure 1 illustrates the key components of our model.

#### 3.1. Model

**Frame-wise representation.** Let  $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^T$  be a sequence of length  $T$ , with  $\mathbf{x}_t \in \mathbb{R}^p$  being the input at time  $t$  (e.g., the robot’s joint positions and velocities). Our goal is to compactly represent each  $\mathbf{x}_t$  as a linear combination of a small number of atomic motions using an overcomplete dictionary of representative atomic motions  $\Psi \in \mathbb{R}^{p \times m}$ , i.e.,  $\mathbf{x}_t \approx \Psi \mathbf{u}_t$ , where  $\mathbf{u}_t \in \mathbb{R}^m$  is the vector of sparse coefficients obtained for frame  $t$ . Such sparse codes can be obtained by considering the following optimization problem:

$$\min_{\{\mathbf{u}_t\}_{t=1}^T} \frac{1}{T} \sum_{t=1}^T \|\mathbf{x}_t - \Psi \mathbf{u}_t\|_2^2 + \lambda_u \|\mathbf{u}_t\|_1, \quad (1)$$

where  $\lambda_u$  is a regularization parameter controlling the trade-off between reconstruction error and sparsity of the coefficients. Problem (1) is a standard Lasso regression and can be efficiently solved using existing sparse coding algorithms [23]. After computing sparse codes  $\mathbf{u}_t$  for each time step of the input sequence, we follow the approach proposed in [30] to compute feature vectors  $\{\mathbf{z}_t\}_{t=1}^T$ . Namely, we initially split the positive and negative components of the sparse codes and stack them on top of each other. This

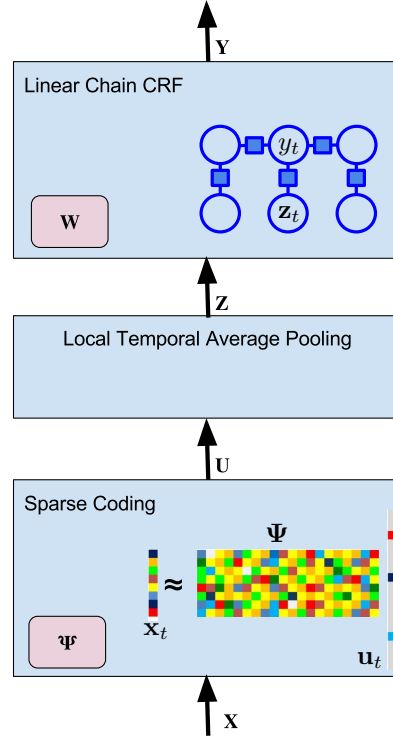


Figure 1: Overview of our framework. Given an input time series  $\mathbf{X}$ , we first extract sparse codes  $\mathbf{U}$  for each timestep using a dictionary  $\Psi$ . Sparse codes are then average pooled in short temporal windows yielding feature vectors  $\mathbf{Z}$  per timestep. These feature vectors are then given as inputs to a Linear Chain CRF with weights  $\mathbf{W}$ . Trainable parameters  $\Psi$  and  $\mathbf{W}$  are shown in light pink boxes.

step yields a vector  $\mathbf{a}_t \in \mathbb{R}^D$ ,  $D = 2m$ , which is given by:

$$\mathbf{a}_t = \begin{bmatrix} \max(0, \mathbf{u}_t) \\ \min(0, \mathbf{u}_t) \end{bmatrix}. \quad (2)$$

This is a common practice [6, 4], which allows the classification layer to assign different weights to positive and negative responses. Second, we compute a feature vector  $\mathbf{z}_t \in \mathbb{R}^D$  for each frame by average-pooling vectors  $\mathbf{a}_t$  in a temporal window  $T_t$  surrounding frame  $t$ , i.e.:

$$\mathbf{z}_t = \frac{1}{L} \sum_{j \in T_t} \mathbf{a}_j, \quad T_t \doteq \left\{ t - \left\lfloor \frac{L}{2} \right\rfloor, t + \left\lfloor \frac{L}{2} \right\rfloor \right\}, \quad (3)$$

where  $L$  is the length of the temporal window centered at frame  $t$ . This feature vector captures local temporal context.

**Temporal model.** Let  $\mathbf{Z} = \{\mathbf{z}_t\}_{t=1}^T$  be a sequence of length  $T$  with  $\mathbf{z}_t$  being the feature vector representing the input at time  $t$ , and  $\mathbf{Y} = \{y_t\}_{t=1}^T$  be the corresponding sequence of action labels per frame,  $y_t \in \{1, \dots, N_c\}$ , with  $N_c$  being the number of action classes. Let  $G = \{\mathcal{V}, \mathcal{E}\}$  be the graph

whose nodes correspond to different frames ( $|\mathcal{V}| = T$ ) and whose edges connect every  $d$  frames (with  $d = 1$  corresponding to consecutive frames). Our CRF models the conditional distribution of labels given the input features with a Gibbs distribution of the form  $P(\mathbf{Y} | \mathbf{Z}) \propto \exp E(\mathbf{Z}, \mathbf{Y})$ , where the energy  $E(\mathbf{Z}, \mathbf{Y})$  is factorized into a sum of potential functions defined on cliques of order less than or equal to two. Formally, the energy function can be written as:

$$E(\mathbf{Z}, \mathbf{Y}) = \sum_{t=1}^T \mathbf{U}_{y_t}^\top \mathbf{z}_t + \sum_{t=1}^{T-d} \mathbf{P}_{y_t, y_{t+d}}, \quad (4)$$

where the first term is the unary potential which models the score of assigning label  $y_t$  to frame  $t$  described by feature  $\mathbf{z}_t$ , while the second term is called pairwise potential and models the score of assigning labels  $y_t$  and  $y_{t+d}$  to frames  $t$  and  $t + d$  respectively ( $d$  is a parameter called the skip length and a CRF with  $d > 1$  is called Skip-Chain CRF (SC-CRF) [16, 17]).  $\mathbf{U}_{y_t} \in \mathbb{R}^D$  is a linear unary classifier corresponding to action class  $y_t$  and  $\mathbf{P} \in \mathbb{R}^{N_c \times N_c}$  is the pairwise transition matrix. Note that there exist different variants to this model. For instance, one can use precomputed unary and pairwise potentials and learn two scalar coefficients that encode the relative weights of the two terms [33].

We now show how this energy can be written as a linear function with respect to a parameter vector  $\mathbf{W} \in \mathbb{R}^{N_c D + N_c^2}$ . The unary term can be rewritten as follows:

$$\sum_{t=1}^T \mathbf{U}_{y_t}^\top \mathbf{z}_t = [\mathbf{U}_1^\top, \dots, \mathbf{U}_{N_c}^\top] \begin{bmatrix} \sum_{t=1}^T \mathbf{z}_t \delta(y_t = 1) \\ \vdots \\ \sum_{t=1}^T \mathbf{z}_t \delta(y_t = N_c) \end{bmatrix} \quad (5)$$

$$= \mathbf{W}_U^\top \Phi_U(\mathbf{Z}, \mathbf{Y}),$$

where  $\mathbf{W}_U$  and  $\Phi_U(\mathbf{Z}, \mathbf{Y}) \in \mathbb{R}^{N_c D}$  are, respectively, the unary CRF weights and the unary joint feature. Similarly the pairwise term can be written as:

$$[\mathbf{P}_{11}, \dots, \mathbf{P}_{N_c N_c}] \begin{bmatrix} \sum_{t=1}^{T-d} \delta(y_t = 1) \delta(y_{t+d} = 1) \\ \vdots \\ \sum_{t=1}^{T-d} \delta(y_t = N_c) \delta(y_{t+d} = N_c) \end{bmatrix} \quad (6)$$

$$= \mathbf{W}_P^\top \Phi_P(\mathbf{Y}),$$

where  $\mathbf{W}_P$ ,  $\Phi_P(\mathbf{Y}) \in \mathbb{R}^{N_c^2}$  are the pairwise CRF weights and pairwise joint feature. Therefore, the overall energy function can be written as:

$$E(\mathbf{Z}, \mathbf{Y}) = \begin{bmatrix} \mathbf{W}_U \\ \mathbf{W}_P \end{bmatrix}^\top \begin{bmatrix} \Phi_U(\mathbf{Z}, \mathbf{Y}) \\ \Phi_P(\mathbf{Y}) \end{bmatrix} = \mathbf{W}^\top \Phi(\mathbf{Z}, \mathbf{Y}), \quad (7)$$

where  $\mathbf{W}$  is the vector of CRF weights and  $\Phi(\mathbf{Z}, \mathbf{Y})$  the joint feature [11]. At this point, we should emphasize that feature vectors  $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_T]$  are constructed by local average pooling of the sparse codes and are therefore implicitly dependent of the input data  $\mathbf{X}$  and the dictionary  $\Psi$ . For the rest of this manuscript, we will denote this dependency by substituting  $\mathbf{Z}$  with the notation  $\mathbf{Z}(\mathbf{X}, \Psi)$ . So our energy can be rewritten as:

$$E(\mathbf{Z}(\mathbf{X}, \Psi), \mathbf{Y}) = \mathbf{W}^\top \Phi(\mathbf{Z}(\mathbf{X}, \Psi), \mathbf{Y}). \quad (8)$$

It should be now clear that if  $\Psi$  is fixed, then the energy is linear with respect to the parameter vector  $\mathbf{W}$ , like in a standard CRF model. However, if  $\Psi$  is a parameter that needs to be learned, then the energy function is nonlinear with respect to  $(\mathbf{W}, \Psi)$  and thus training is not straightforward. The training problem is addressed next.

### 3.2. Training

Let  $\{\mathbf{X}^n\}_{n=1}^{N_s}$  be  $N_s$  training sequences with associated label sequences  $\{\mathbf{Y}^n\}_{n=1}^{N_s}$ . We formulate the training problem as one of minimizing the following regularized loss:

$$J(\mathbf{W}, \Psi) = \frac{1}{2} \|\mathbf{W}\|_F^2 + \frac{C}{N_s} \sum_{n=1}^{N_s} \max_{\mathbf{Y}} [\Delta(\mathbf{Y}^n, \mathbf{Y}) + \langle \mathbf{W}, \Phi(\mathbf{Z}^n(\mathbf{X}^n, \Psi), \mathbf{Y}) \rangle] - \langle \mathbf{W}, \Phi(\mathbf{Z}^n(\mathbf{X}^n, \Psi), \mathbf{Y}^n) \rangle, \quad (9)$$

where  $C$  is a regularization parameter controlling the regularization of the CRF weights,  $\Delta(\hat{\mathbf{Y}}, \mathbf{Y}) = \sum_{t=1}^T \delta(\hat{y}_t \neq y_t)$  is the Hamming loss between two sequences of labels  $\hat{\mathbf{Y}}$  and  $\mathbf{Y}$ , and  $\mathbf{Z}^n$  is the matrix of feature vectors extracted from the frames of input sequence  $\mathbf{X}^n$ , i.e.,  $\mathbf{Z}^n = [\mathbf{z}_1^n, \dots, \mathbf{z}_T^n]$ . This max-margin formulation performs regularized empirical risk minimization and bounds the hamming loss from above. We use a Stochastic Gradient Descent algorithm for minimizing the objective function in Eq. (9). Our algorithm is based on the task-driven dictionary learning approach developed by Mairal et al. [24]. Notice that, although the sparse coefficients are computed by minimizing a non-differentiable objective function (Eq. 1),  $J(\mathbf{W}, \Psi)$  is differentiable and its gradient can be computed [22]. In particular, the function relating the sparse codes  $\mathbf{u}_t$  and the dictionary is differentiable almost everywhere, except at the points where the set of non-zero elements of  $\mathbf{u}_t$  (called the support set and denoted by  $S_t$ ) changes. Assuming that the perturbations of the dictionary atoms are small so that the support set stays the same, we can compute the gradient of the non-zero coefficients with respect to the columns of  $\Psi$  indexed by  $S_t$ , denoted as  $\Psi_{S_t}$ ,



as follows [33]:

$$\frac{\partial \mathbf{u}_t(k)}{\partial \Psi_{S_t}} = (\mathbf{x}_t - \Psi_{S_t}(\mathbf{u}_t)_{S_t})(\mathbf{A}_t^{-1})_{[k]} - (\Psi_{S_t} \mathbf{A}_t^{-\top})_{\langle k \rangle} (\mathbf{u}_t)_{S_t}^\top \quad (10)$$

where  $k \in S_t$ ,  $(\mathbf{u}_t)_{S_t}$  denotes the sub-vector of  $\mathbf{u}_t$  with entries in  $S_t$ ,  $\mathbf{A}_t = \Psi_{S_t}^\top \Psi_{S_t}$ , and the subscripts  $[k]$  and  $\langle k \rangle$  denote, respectively, the  $k$ -th row and column of the corresponding matrix.

Given the dictionary and CRF weights computed at the  $(i-1)$ -th iteration, the main steps of our iterative algorithm at the  $i$ -th iteration are:

1. Randomly select a training sequence  $(\mathbf{X}^i, \mathbf{Y}^i)$ .
2. Compute sparse codes  $\mathbf{u}_t$  with Eq. 1 and feature vectors  $\mathbf{z}_t$  with Eq. 3 using dictionary  $\Psi^{(i-1)}$ .
3. Find the sequence  $\hat{\mathbf{Y}}$  that yields the most violated constraint by solving the loss augmented inference problem:

$$\hat{\mathbf{Y}} = \underset{\mathbf{Y}}{\operatorname{argmax}} \Delta(\mathbf{Y}^i, \mathbf{Y}) + \left\langle \mathbf{W}^{(i-1)}, \Phi(\mathbf{Z}^i(\mathbf{X}^i, \Psi^{(i-1)}), \mathbf{Y}) \right\rangle \quad (11)$$

using the Viterbi algorithm (see [17] for details regarding inference when using a SC-CRF ( $d > 1$ )).

4. Compute gradient with respect to the CRF parameters  $\mathbf{W}$ :

$$\frac{\partial J}{\partial \mathbf{W}} = \mathbf{W}^{(i-1)} + C(\Phi(\mathbf{Z}^i(\mathbf{X}^i, \Psi^{(i-1)}), \hat{\mathbf{Y}}) - \Phi(\mathbf{Z}^i(\mathbf{X}^i, \Psi^{(i-1)}), \mathbf{Y}^i)). \quad (12)$$

5. Compute gradients with respect to the dictionary  $\Psi$  using the chain rule:

$$\begin{aligned} \frac{\partial J}{\partial \Psi} &= \sum_{t=1}^T \left( \frac{\partial J}{\partial \mathbf{z}_t} \right)^\top \frac{\partial \mathbf{z}_t}{\partial \Psi} = \sum_{t=1}^T \left( \frac{\partial J}{\partial \mathbf{z}_t} \right)^\top \frac{1}{L} \sum_{j \in T_t} \frac{\partial \mathbf{a}_j}{\partial \Psi}, \\ &= \frac{1}{L} \sum_{t=1}^T \sum_{j \in T_t} (\mathbf{x}_j - \Psi_{S_j}^{(i-1)}(\mathbf{u}_j)_{S_j}) \left( \mathbf{A}_j^{-1} \left( \frac{\partial J}{\partial \mathbf{z}_t} \right)_{\tilde{S}_j} \right)^\top \\ &\quad - \Psi_{S_j}^{(i-1)} \mathbf{A}_j^{-\top}(\mathbf{u}_j)_{S_j} \left( \frac{\partial J}{\partial \mathbf{z}_t} \right)_{\tilde{S}_j}^\top, \end{aligned} \quad (13)$$

where  $\frac{\partial J}{\partial \mathbf{z}_t} = \mathbf{U}_{\hat{\mathbf{y}}_t} - \mathbf{U}_{\mathbf{y}_t} \in \mathbb{R}^D$ ,  $S_j$  is the set of indices corresponding to the non-zero entries of the vector  $\mathbf{u}_j$ ,  $\tilde{S}_j$  is the set of indices corresponding to the non-zero entries of the vector  $\mathbf{a}_j$ ,  $A_j = \Psi_{S_j}^\top \Psi_{S_j}$ ,  $\Psi_{S_j}$  denotes the active columns of the dictionary indexed by  $S_j$ ,  $(\mathbf{u}_j)_{S_j}$  denotes the non-zero entries of vector  $\mathbf{u}_j$  and  $\left( \frac{\partial J}{\partial \mathbf{z}_t} \right)_{\tilde{S}_j}$  denotes the entries of the partial derivative corresponding to non-zero entries of vector  $\mathbf{a}_j$ .

6. Update  $\mathbf{W}$ ,  $\Psi$  using stochastic gradient descent.

7. Normalize the dictionary atoms to have unit  $l_2$  norm. This step prevents the columns of  $\Psi$  from becoming arbitrarily large, which would result in arbitrarily small sparse coefficients.

## 4. Experiments

We evaluate our method on two public datasets for fine-grained action segmentation and recognition: JIGSAWS [9] and 50 Salads [31]. First, we report our results on each dataset and compare them with the state of the art. Next, we examine the effect of different model components.

### 4.1. Datasets

#### JHU-ISI Gesture and Skill Assessment (JIGSAWS) [9].

This dataset provides kinematic data of the right and left manipulators of the master and slave da Vinci surgical robot recorded at 30 Hz during the execution of three surgical tasks (Suturing (SU), Knot-tying (KT) and Needle-passing (NP)) by surgeons with varying skill levels. In particular, kinematic data include positions, orientations, velocities etc. (76 variables in total), and there are 8 surgeons performing a total of 39, 36 and 26 trials for the Suturing, Knot-tying and Needle-passing surgical tasks, respectively. This dataset is challenging due to the significant variability in the execution of tasks by surgeons of different skill levels and the subtle differences between fine-grained actions. There are 10, 6 and 8 different action classes for the Suturing, Knot-tying and Needle-passing tasks, respectively. Examples of action classes are *orienting needle*, *reaching for needle with right hand*, *pulling suture with left hand*, and *making C loop*. We evaluate our method using the standard Leave-One-User-Out (LOUO) and Leave-One-Supertrial-Out (LOSO) cross-validation setups [2].

**50 Salads [31].** This dataset provides data recorded by 10 accelerometers attached to kitchen tools, such as knife, peeler, oil bottle etc., during the preparation of a salad by 25 users. This dataset features annotations at four levels of granularity, out of which we use the *eval* and *mid* granularities. The former consists of 10 actions that can be reasonably recognized based on the utilization of accelerometer-equipped objects, such as *add oil*, *cut*, *peel* etc., while the latter consists of 18 mid-level actions, such as *cut tomato*, *peel cucumber*. Both granularities include a background class. We evaluate our method using the ground truth labels and the 5-fold cross-validation setup proposed by the authors of [18, 15].

In summary, these two datasets provide kinematic/sensor data recorded during the execution of long goal-driven complex activities, which are comprised of multiple fine-grained action instances following a grammar. Hence, they are suitable for evaluating our method, which was designed

Method	LOSO			LOUO		
	SU	KT	NP	SU	KT	NP
GMM-HMM [2]	82.22	80.95	70.55	73.95	72.47	64.13
KSVD-SHMM [32, 2]	83.40	83.54	73.09	73.45	74.89	62.78
MsM-CRF [34, 2]	81.99	79.26	72.44	67.84	44.68	63.28
SC-CRF-SL [16, 2]	85.18	<b>84.03</b>	75.09	81.74	<b>78.95</b>	<b>74.77</b>
SDSDL [30]	<b>86.32</b>	82.54	74.88	78.68	75.11	66.01
LSTM (5Hz) [8]*	-	-	-	80.5	-	-
LSTM (30Hz) [8]*	-	-	-	78.38	-	-
BiLSTM (5Hz) [8]*	-	-	-	83.3	-	-
BiLSTM (30Hz) [8]*	-	-	-	80.15	-	-
TCN [18]	-	-	-	79.6	-	-
LC-SC-CRF [17]**	-	-	-	<b>83.4</b>	-	-
Ours	<i>86.21 (0.34)</i>	<i>83.89 (0.08)</i>	<b>75.19 (0.12)</b>	<i>78.16 (0.42)</i>	<i>76.68 (1.20)</i>	<i>66.25 (0.06)</i>

Table 1: Average per-frame action recognition accuracy for surgical task segmentation and recognition on the JIGSAWS dataset [9]. The results are averaged over three random runs, with the standard deviation reported in parentheses. Best results are shown in bold, while second best results are denoted in italics.\* Our results are not directly comparable with those of [8], since they were using data downsampled in time (5Hz). For a fair comparison, results for LSTM, BiLSTM on non-downsampled data (30Hz) were obtained using the code and default parameters publicly available from the authors [8].\*\* Our results are not directly comparable with those of LC-SC-CRF [17], where authors were using both kinematic data as well as the distance from the tools to the closest object in the scene from the video.

for kinematic data and features a temporal model that is able to capture action transitions. Other datasets collected for action segmentation with available skeleton data, such as CAD-120 [12], Composable Activities [20], Watch-n-Patch [36] and OAD [7], have a mean number of 3 to 12 action instances per sequence [21], while for example the Suturing task in the JIGSAWS dataset features an average of 20 action instances per sequence, ranging from 17 to 37. It is therefore more challenging for comparing temporal models. Recently, the PKU-MMD dataset [21] was proposed, which is of larger scale and also contains around 20 action instances per sequence. However, the actions in this dataset are not fine-grained (e.g., *hand waving*, *hugging* etc.).

## 4.2. Implementation Details

Input data are normalized to have zero mean and unit standard deviation. We apply PCA on the robot kinematic data of the JIGSAWS dataset to reduce their dimension from 76 to 35 following the setup of [30]. The dictionary is initialized using the SPAMS dictionary learning toolbox [23] and the CRF parameters are initialized to 0. We use Stochastic Gradient Descent with a batch size of 1 and momentum of 0.9. We also reduce the learning rate by one half every 20 epochs and train our models for 100 epochs. Parameters such as the regularization cost  $C$ , initial learning rate  $\eta$ , temporal window size for average-pooling  $L$ , Lasso regularizer parameter  $\lambda_u$ , skip chain length  $d$  and dictionary size  $m$  vary with each dataset, surgical task or granularity. The window size was fixed to 71 for JIGSAWS and 51 for 50 Salads, the dictionary size  $M$  was chosen

via cross-validation from the values  $\{50, 100, 150, 200\}$ ,  $\lambda_u$  from values  $\{0.1, 0.5\}$ ,  $C$  from  $\{0.001, 0.01, 0.1, 1\}$ ,  $\eta$  from  $\{0.0001, 0.001, 0.01\}$  and  $d$  from  $\{21, 51, 81\}$ . To perform cross-validation we generate five random splits of the available sequences of each dataset task/granularity. Note that since both datasets have a fixed test setup, with all users appearing in the test set exactly once, it is not clear how to use them for hyperparameter selection without inadvertently training on the test set. Here we randomly crop a temporal segment from each of the videos instead of using the whole sequences for cross-validation, in order to avoid using the exact same video sequences which will be used for evaluating our method. The length of these segments is 80% of the original sequence length. Furthermore, we select  $m$ ,  $\lambda_u$  and  $d$  by using the initialized dictionary  $\Psi^0$  and learning the weights of a SC-CRF, while we choose  $C$  and  $\eta$  by jointly learning the dictionary and the SC-CRF weights.

## 4.3. Results

**Overall performance.** We first compare our method with state-of-the-art methods on the JIGSAWS and 50 Salads datasets. The per-frame action recognition accuracies of all the compared methods on JIGSAWS are summarized in Table 1. It can be seen that our method yields the best or second best performance for all tasks on both the LOSO and LOUO setups, except for Suturing LOUO, where LC-SC-CRF achieves per-frame action recognition accuracies up to 83%. However, their result is not directly comparable to ours, since they employ additional video-based features. Also note that in [16] they use a SC-CRF with an addi-

Method	50 Salads	
	<i>eval</i>	<i>mid</i>
LC-SC-CRF [17]	77.8	55.05*
LSTM [18]	73.3	-
TCN [18]	<b>82.0</b>	-
Ours	80.04 (0.11)	<b>56.72</b> (0.72)

Table 2: Results for action segmentation and recognition on the 50 Salads dataset using granularities *eval* and *mid*. Results are averaged over three random runs, with the standard deviation reported in parentheses. Best results are shown in bold, while second best results are denoted in italics.\* LC-SC-CRF [17] was evaluated on the *mid* granularity with smoothed out short interstitial background segments [18].

tional pairwise term (skip-length data potentials), which is not incorporated in our model and could potentially improve our results. However, it is worth noting that our method achieves comparable performance to deep recurrent models such as LSTMs [8] and the newly proposed TCN [18], which possibly captures complex temporal patterns, such as action compositions, action durations, and long-range temporal dependencies. Furthermore, our method consistently improves over SDSDL [30], which was based on joint sparse dictionary and linear SVM learning, as well as a temporal smoothing of results using the Viterbi algorithm with precomputed action transition probabilities.

Table 2 summarizes our results on the 50 Salads dataset under two granularities. Although the modality used in this dataset is different (accelerometer data), it can be seen that our method is very competitive among all the compared methods, even with respect to methods relying on powerful deep temporal models such as LSTMs.

**Ablative analysis.** In Tables 4, 3 we analyze the contribution of the key components of our method, namely the contribution of a) using sparse features (Eq. 3) obtained from an unsupervised dictionary in conjunction with a Linear Chain CRF, b) substituting the Linear Chain CRF with a Skip Chain CRF (SC-CRF) and c) jointly learning the dictionary used in sparse coding and the CRF unary and pairwise weights. As expected, using sparse features instead of the raw kinematic features consistently boosts performance across all tasks on JIGSAWS. Similarly, sparse coding of accelerometer data improves performance on 50 Salads and notably this improvement is larger in the case of fine-grained activities (*mid* granularity). Furthermore, using a SC-CRF further boosts performance as expected, since it is more suitable for capturing action-to-action transition probabilities in contrast to the Linear Chain CRF which captures frame-to-frame action transition probabilities.

It is however surprising that learning a discriminative dictionary jointly with the CRF weights does not significantly improve performance, yielding an improvement of at most  $\sim 1\%$ . Further investigating this result, we computed

Method	50 Salads	
	<i>eval</i>	<i>mid</i>
raw + CRF	71.81 (0.55)	44.83 (0.73)
SF + CRF	76.65 (0.19)	52.63 (0.23)
SF + SC-CRF	80.24 (0.20)	<b>56.73</b> (0.08)
SDL + SC-CRF	<b>80.54</b> (0.11)	56.72 (0.72)

Table 3: Analysis of contribution to recognition performance from each model component in the 50 Salads dataset. Results are averaged over three random runs, with the standard deviation reported in parentheses. raw+CRF: use kinematic data as input to a CRF, SF + CRF: use sparse features  $z$  as input to a CRF, SF + SC-CRF: use sparse features  $z$  as input to a SC-CRF, SDL + SC-CRF: joint dictionary and SC-CRF learning.

additional metrics for evaluating the segmentation quality on the JIGSAWS dataset. In particular, we report the edit score [17], a metric measuring how well the model predictions the ordering of action segments, and segmental-f1@10 score as defined in [15]. As it can be seen in Table 5, performance is similar across all metrics for both unsupervised and discriminative dictionary, except for a consistent improvement in Needle Passing. One possible explanation could be that the computation of features based on average pooling of sparse codes in a temporal window might reduce the impact of the discriminatively trained dictionary. However, repeating the experiment on JIGSAWS (Suturing LOSO) without average temporal pooling leads to the same behavior, i.e. using a dictionary learned via unsupervised training with a SC-CRF yields a per-frame accuracy of 86.64%, while using a dictionary jointly trained with the SC-CRF yields 86.11%. Our findings could be attributed to the limited training data. They also seem to corroborate the conclusions drawn by Coates et al. [6], who have experimentally observed that the superior performance of sparse coding, especially when training samples are limited, arises from its non-linear encoding scheme and not from the basis functions that it uses.

**Qualitative results.** In Fig. 2 we show examples of ground truth segmentations and predictions for selected testing sequences from JIGSAWS Suturing. As it can be seen, the LOUO setup is more challenging since the model is asked to recognize actions performed by a user it has not seen before and in addition to that there is great variability in experience and styles between surgeons. In all cases our model outputs smooth predictions, without significant over-segmentations.

## 5. Conclusion

We have presented a novel end-to-end learning framework for fine-grained action segmentation and recognition, which combines features based on sparse coding with a Linear Chain CRF model. We also proposed a max-margin

Method	LOSO			LOUO		
	SU	KT	NP	SU	KT	NP
raw + CRF	79.57 (0.04)	76.39 (0.09)	66.24 (0.10)	71.77 (0.05)	69.63 (0.06)	59.47 (0.18)
SF + CRF	85.70 (0.01)	82.06 (0.03)	71.72 (0.07)	76.64 (0.05)	73.58 (0.07)	60.59 (0.19)
SF + SC-CRF	<b>87.60</b> (0.03)	83.71 (0.03)	74.63 (0.02)	<b>79.95</b> (0.05)	<b>76.88</b> (0.14)	65.75 (0.12)
SDL + SC-CRF	86.21 (0.34)	<b>83.89</b> (0.07)	<b>75.19</b> (0.12)	78.16 (0.42)	76.68 (1.20)	<b>66.25</b> (0.06)

Table 4: Analysis of contribution to recognition performance from each model component in the JIGSAWS dataset. Results are averaged over three random runs, with the standard deviation reported in parentheses. raw+CRF: use kinematic data as input to a Linear Chain CRF, SF + CRF: use sparse features  $z$  as input to a CRF, SF + SC-CRF: use sparse features  $z$  as input to a SC-CRF, SDL + SC-CRF: joint dictionary and SC-CRF learning.

Method	LOSO			LOUO		
	SU	KT	NP	SU	KT	NP
SF + SC-CRF	<b>87.57/82.92/88.59</b>	83.08/82.87/87.46	74.62/73.05/76.01	<b>79.92/63.39/75.00</b>	76.93/63.61/71.38	65.81/55.45/62.30
SDL + SC-CRF	85.90/75.45/83.47	<b>83.97/82.82/87.94</b>	<b>75.33/76.63/79.85</b>	78.42/58.02/69.22	76.39/65.55/72.87	<b>66.29/60.85/64.43</b>

Table 5: Comparison of unsupervised and supervised dictionary used for sparse coding on JIGSAWS dataset. Metrics reported are: accuracy/edit score/segmental f1 score. Results are from a single random run. SF + SC-CRF: use sparse features  $z$  obtained from unsupervised dictionary as input to a SC-CRF, SDL + SC-CRF: use sparse features  $z$  from discriminative dictionary learned jointly with a SC-CRF.

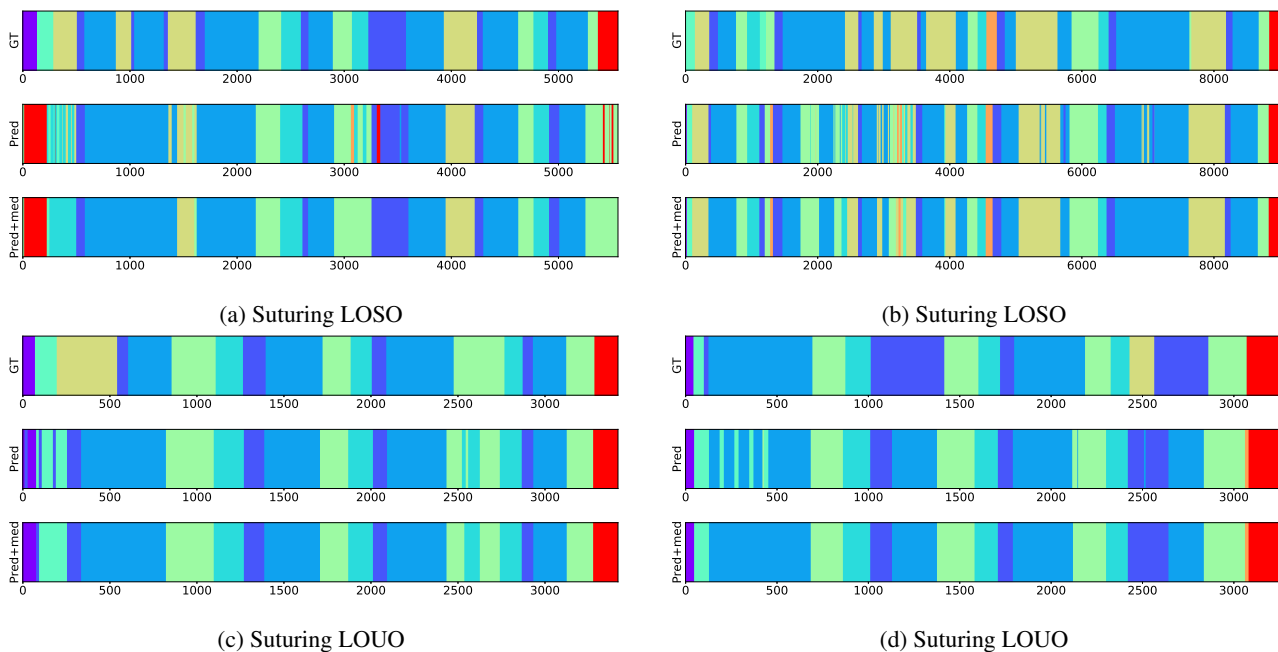


Figure 2: Qualitative examples of ground truth and predicted temporal segmentations (before and after median filtering) on JIGSAWS data. Each color denotes a different action class. (Best viewed in color.)

approach for jointly learning the sparse dictionary and the CRF weights, resulting in a dictionary adapted to the task of action segmentation and recognition. Experimental evaluation of our method on two datasets showed that our method performs on par or outperforms most of the state-of-the-art methods. Given the recent success of deep convolutional networks (CNNs), future work will explore using deep fea-

tures as inputs to the temporal model and jointly learning the CNN and CRF parameters in a unified framework.

**Acknowledgements.** We would like to thank Colin Lea and Lingling Tao for their insightful comments and for their help with the JIGSAWS dataset, and Vicente Ordóñez for his useful feedback during this research collaboration. This work was supported by NIH grant R01HD87133.



## References

- [1] M. Aharon, M. Elad, and A. M. Bruckstein. K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006. 1
- [2] N. Ahmidi, L. Tao, S. Sefati, Y. Gao, C. Lea, B. Béjar, L. Zappella, S. Khudanpur, R. Vidal, and G. D. Hager. A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. *IEEE Transactions on Biomedical Engineering*, 2017. 5, 6
- [3] B. Béjar, L. Zappella, and R. Vidal. Surgical gesture classification from video data. In *Medical Image Computing and Computer Assisted Intervention*, pages 34–41, 2012. 1
- [4] L. Bo, X. Ren, and D. Fox. Multipath sparse coding using hierarchical matching pursuit. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 660–667, 2013. 3
- [5] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2559–2566. IEEE, 2010. 2
- [6] A. Coates and A. Y. Ng. The importance of encoding versus training with sparse coding and vector quantization. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 921–928, 2011. 3, 7
- [7] R. De Geest, E. Gavves, A. Ghodrati, Z. Li, C. Snoek, and T. Tuytelaars. Online action detection. In *European Conference on Computer Vision*, pages 269–284. Springer, 2016. 6
- [8] R. DiPietro, C. Lea, A. Malpani, N. Ahmidi, S. S. Vedula, G. I. Lee, M. R. Lee, and G. D. Hager. Recognizing surgical activities with recurrent neural networks. In *Medical Image Computing and Computer Assisted Intervention*, pages 551–558. Springer, 2016. 1, 2, 6, 7
- [9] Y. Gao, S. S. Vedula, C. E. Reiley, N. Ahmidi, B. Varadarajan, H. C. Lin, L. Tao, L. Zappella, B. Béjar, D. D. Yuh, C. Chiung, G. Chen, R. Vidal, S. Khudanpur, and G. D. Hager. JHU-ISI gesture and skill assessment working set (JIGSAWS): a surgical activity dataset for human motion modeling. In *Fifth Workshop on Modeling and Monitoring of Computer Assisted Interventions M2CAI*, 2014. 1, 2, 5, 6
- [10] Z. Jiang, Z. Lin, and L. S. Davis. Learning a discriminative dictionary for sparse coding via label consistent K-SVD. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1697–1704, 2011. 2
- [11] T. Joachims, T. Finley, and C.-N. J. Yu. Cutting-plane training of structural SVMs. *Machine Learning*, 77(1):27–59, 2009. 4
- [12] H. S. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from RGB-D videos. In *International Journal of Robotics Research*, 2013. 6
- [13] H. Kuehne, A. Arslan, and T. Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 780–787, 2014. 1, 2
- [14] H. Kuehne, J. Gall, and T. Serre. An end-to-end generative framework for video segmentation and recognition. In *IEEE Winter Applications of Computer Vision Conference*, Lake Placid, Mar 2016. 2
- [15] C. Lea, M. Flynn, R. Vidal, A. Reiter, and G. Hager. Temporal convolutional networks for action segmentation and detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 5, 7
- [16] C. Lea, G. D. Hager, and R. Vidal. 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*, pages 1123–1129, 2015. 1, 2, 4, 6
- [17] C. Lea, R. Vidal, and G. D. Hager. Learning convolutional action primitives for fine-grained action recognition. In *IEEE International Conference on Robotics and Automation*, 2016. 1, 2, 4, 5, 6, 7
- [18] C. Lea, R. Vidal, A. Reiter, and G. D. Hager. Temporal convolutional networks: A unified approach to action segmentation. In *Workshop on Brave New Ideas on Motion Representation*, 2016. 5, 6, 7
- [19] X.-C. Lian, Z. Li, B.-L. Lu, and L. Zhang. Max-margin dictionary learning for multiclass image categorization. *European Conference on Computer Vision*, pages 157–170, 2010. 2
- [20] I. Lillo, A. Soto, and J. C. Niebles. Discriminative hierarchical modeling of spatio-temporally composable human activities. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 6
- [21] C. Liu, Y. Hu, Y. Li, S. Song, and J. Liu. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *CoRR*, 2017. 6
- [22] J. Mairal, F. Bach, and J. Ponce. Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):791–804, 2012. 2, 4
- [23] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 11:19–60, 2010. 3, 6
- [24] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. R. Bach. Supervised dictionary learning. In *Neural Information Processing Systems*, pages 1033–1040, 2009. 1, 2, 4
- [25] D. Oneata, J. Verbeek, and C. Schmid. Action and event recognition with Fisher vectors on a compact feature set. In *IEEE International Conference on Computer Vision*, pages 1817–1824, 2013. 2
- [26] Y. Quan, Y. Xu, Y. Sun, Y. Huang, and H. Ji. Sparse coding for classification via discrimination ensemble. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5839–5847, 2016. 2
- [27] A. Richard and J. Gall. Temporal action detection using a statistical language model. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016. 2
- [28] A. Richard, H. Kuehne, and J. Gall. Weakly supervised action learning with RNN based fine-to-coarse modeling. *CoRR*, abs/1703.08132, 2017. 1
- [29] 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*, 2012. 2

- [30] S. Sefati, N. J. Cowan, and R. Vidal. Learning shared, discriminative dictionaries for surgical gesture segmentation and classification. In *MICCAI 6th Workshop on Modeling and Monitoring of Computer Assisted Interventions (M2CAI)*, Munich, Germany, 2015. 1, 3, 6, 7
- [31] S. Stein and S. J. McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 729–738. ACM, 2013. 1, 2, 5
- [32] L. Tao, E. Elhamifar, S. Khudanpur, G. Hager, and R. Vidal. Sparse hidden Markov models for surgical gesture classification and skill evaluation. In *Information Processing in Computed Assisted Interventions*, 2012. 1, 2, 6
- [33] L. Tao, F. Porikli, and R. Vidal. Sparse dictionaries for semantic segmentation. In *European Conference on Computer Vision*, 2014. 3, 4, 5
- [34] L. Tao, L. Zappella, G. Hager, and R. Vidal. Segmentation and recognition of surgical gestures from kinematic and video data. In *Medical Image Computing and Computer Assisted Intervention*, 2013. 1, 2, 6
- [35] N. N. Vo and A. F. Bobick. From stochastic grammar to bayes network: Probabilistic parsing of complex activity. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2641–2648, 2014. 1
- [36] C. Wu, J. Zhang, S. Savarese, and A. Saxena. Watch-n-patch: Unsupervised understanding of actions and relations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4362–4370, 2015. 6
- [37] J. Yang and M.-H. Yang. Top-down visual saliency via joint CRF and dictionary learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 3
- [38] J. Yang and M.-H. Yang. Top-down visual saliency via joint CRF and dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(3):576–588, 2017. 2, 3
- [39] J. Yang, K. Yu, and T. Huang. Supervised translation-invariant sparse coding. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3517–3524. IEEE, 2010. 2