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

Effrosyni Mavroudi† Divya Bhaskara†‡ Shahin Sefati†♮ Haider Ali† René Vidal†

†Johns Hopkins University, ‡University of Virginia, ♮Comcast AI Research

## Motivation

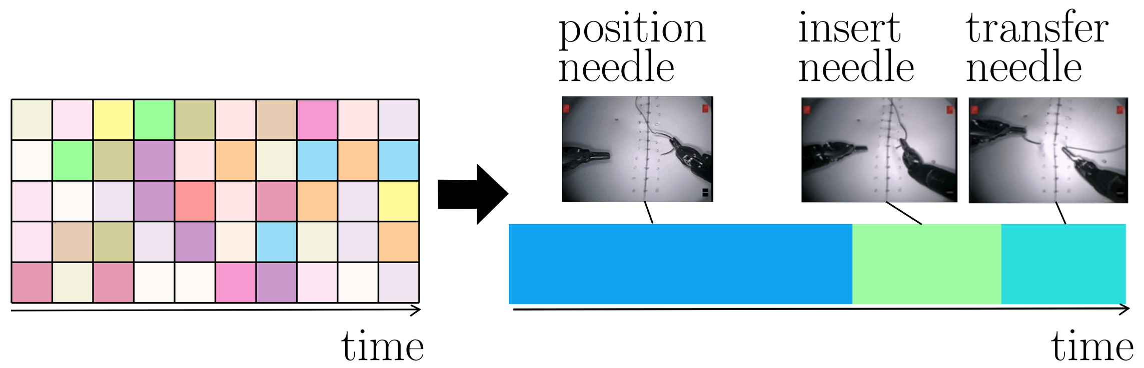
Applications of fine-grained action segmentation and recognition.

(a) Automatic Surgical Skill Evaluation (b) Assisted Living and Smart Home Environments

## Contributions

- Propose a novel spatio-temporal model for fine-grained action segmentation and recognition.
  - Frame representation: Discriminative Sparse Coding.
  - Temporal model: Conditional Random Field (CRF).
- Propose an algorithm for training our model in an end-to-end fashion.
  - Jointly learn a task-specific discriminative dictionary and the CRF unary and pairwise parameters using Stochastic Gradient Descent (SGD).

## Data & Prior Work




Input: Kinematic data time-series Output: Labels per frame

- JIGSAWS:
  - 76-dimensional surgical robot kinematic data.
  - 3 tasks: Suturing (SU), Knot Tying (KT), Needle Passing (NP).
  - 2 experimental setups: leave-one-supertrial-out (LOSO), leave-one-user-out (LOUO).
- 50 Salads:
  - Data recorded by 10 accelerometers attached to kitchen tools.
  - 2 levels of granularity for annotations: *eval* and *mid*.
  - 5 train/test splits [7].
- Prior work:

| Frame Feature \ Temporal Model | Precomputed transition probabilities | HMM | CRF | Deep temporal |
|---|---|---|---|---|
| Raw kinematic | - | GMM-HMM | MsM-CRF, SC-CRF | LSTM, BiLSTM, TCN |
| Convolutional Filters | - | - | LC-SC-CRF | - |
| Sparse Coding | SDSDL | S-HMM | **Ours** | - |

This work was supported by NIH grant R01HD87133.

## Spatio-temporal Representation

**Frame Representation**

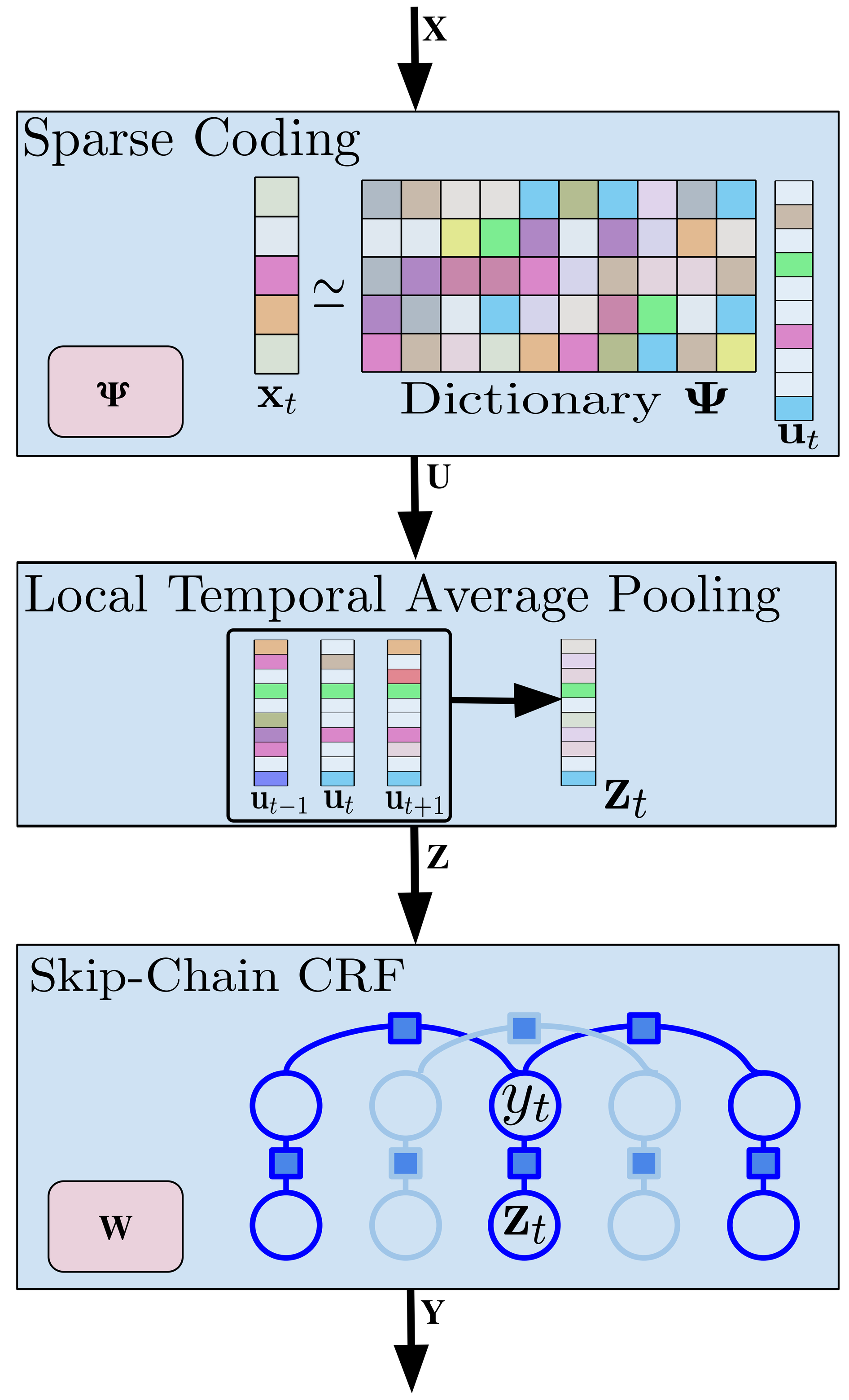
Represent kinematic data $\mathbf{x}_t$ at time $t$ as a linear combination of a small number of basis elements from an overcomplete dictionary $\mathbf{\Psi}$. *Sparse codes* $\mathbf{u}_t$ are the coefficients of this linear combination.

**Local Temporal Representation**

Perform average pooling of sparse codes in a short temporal window to obtain frame feature $\mathbf{z}_t$.

**Global Temporal Representation**

CRF unary potentials represent cost of assigning a label to a frame and are obtained by applying a linear classifier ($\mathbf{W}^U$) to the local temporal representation. Pairwise weights ($\mathbf{W}^P$) capture the transitions between actions and encourage smoothness of the predicted label sequence. ($\mathbf{W} = [\mathbf{W}^U\ \mathbf{W}^P]$).




## Joint Dictionary and CRF Learning

We use a max-margin formulation and SGD to jointly learn the dictionary $\mathbf{\Psi}$ and the Conditional Random Field weights $\mathbf{W}$ by minimizing:

$$\frac{C}{N_s}\sum_{n=1}^{N_s}\max_{\mathbf{Y}}[\underbrace{\Delta(\mathbf{Y}^n,\mathbf{Y})}_{\text{Hamming Loss}}+\langle\mathbf{W},\underbrace{\mathbf{\Phi}(\mathbf{Z}^n(\mathbf{X}^n,\mathbf{\Psi}),\mathbf{Y})}_{\text{CRF Joint Feature}}\rangle] - \langle\mathbf{W},\mathbf{\Phi}(\mathbf{Z}^n(\mathbf{X}^n,\mathbf{\Psi}),\mathbf{Y}^n)\rangle + \frac{1}{2}\|\mathbf{W}\|_F^2$$

## Quantitative Results

| Method | LOSO | | LOUO | |
|---|---|---|---|---|
| | SU | NP | SU | NP |
| GMM-HMM [1] | 82.22 | 70.55 | 73.95 | 64.13 |
| SHMM [2, 1] | 83.40 | 73.09 | 73.45 | 62.78 |
| MsM-CRF [3, 1] | 81.99 | 72.44 | 67.84 | 63.28 |
| SC-CRF-SL [4, 1] | 85.18 | *75.09* | 81.74 | **74.77** |
| SDSDL [5] | **86.32** | 74.88 | 78.68 | 66.01 |
| LSTM [6] | - | - | 78.38 | - |
| BiLSTM [6] | - | - | 80.15 | - |
| TCN [7] | - | - | 79.6 | - |
| LC-SC-CRF [8] | - | - | **83.4** | - |
| Ours | *86.21* | **75.19** | 78.16 | *66.25* |

Comparison with state-of-the-art on JIGSAWS dataset.

| Method | 50 Salads | |
|---|---|---|
| | *eval* | *mid* |
| LC-SC-CRF [8] | 77.8 | *55.05* |
| LSTM [7] | 73.3 | - |
| TCN [7] | **82.0** | - |
| Ours | *80.04* | **56.72** |

Comparison with state-of-the-art on 50 Salads dataset.

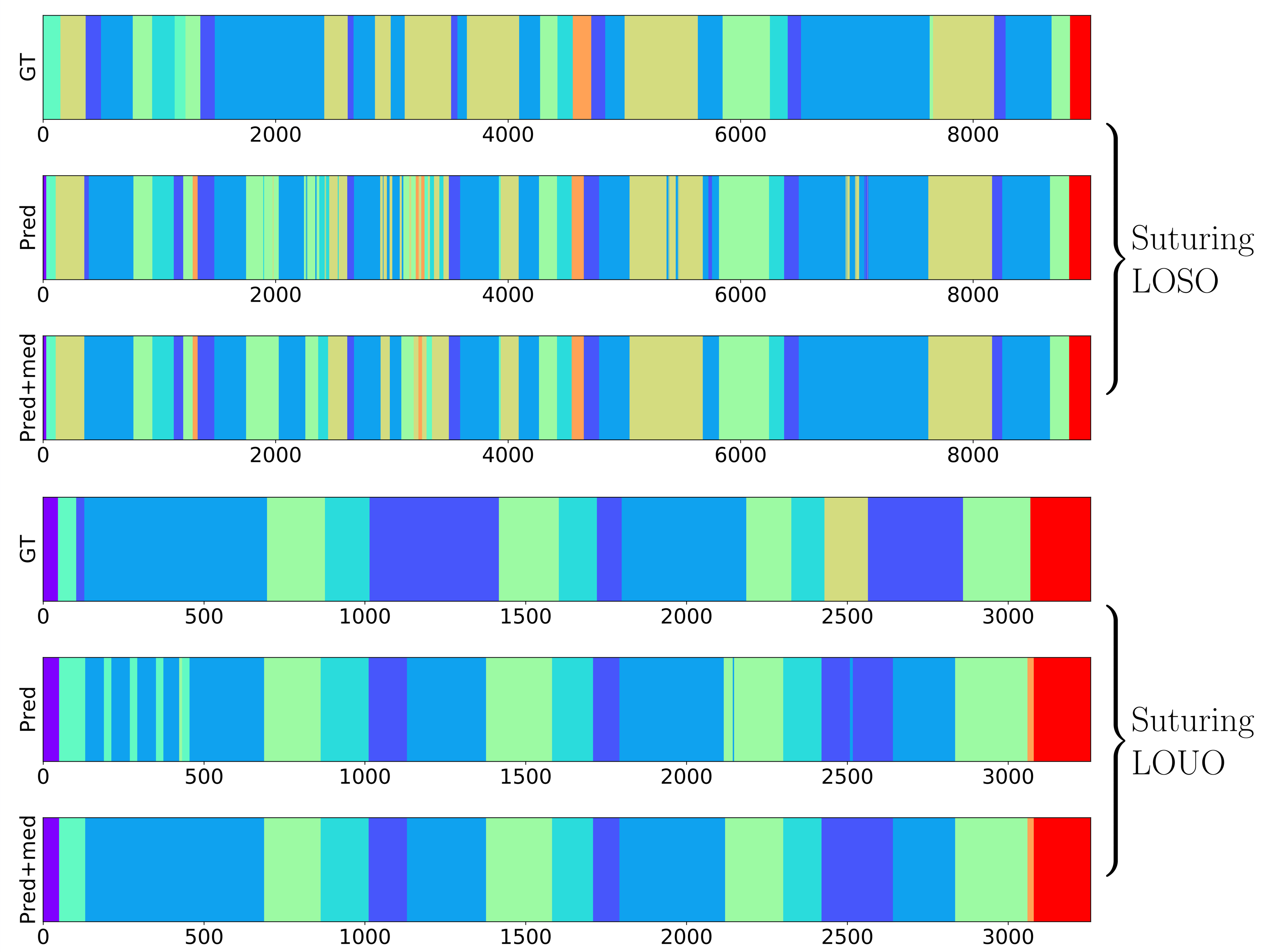
## Ablation Analysis

| Method | JIGSAWS | |
|---|---|---|
| | NP LOSO | NP LOUO |
| raw + CRF | 66.24 (0.10) | 59.47 (0.18) |
| SF + CRF | 71.72 (0.07) | 60.59 (0.19) |
| SF + SC-CRF | 74.63 (0.02) | 65.75 (0.12) |
| SDL + SC-CRF | **75.19** (0.12) | **66.25** (0.06) |

| Method | 50 Salads | |
|---|---|---|
| | *eval* | *mid* |
| 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) | **56.73** (0.08) |
| SDL + SC-CRF | **80.54** (0.11) | 56.72 (0.72) |

- ✓ Sparse coding features (SF + CRF) improve over raw kinematic features.
  - Dictionary learned in an unsupervised manner from training data.
- ✓ Skip-Chain CRF (SC-CRF) improves over Linear Chain CRF.
- ✓ Joint learning of Dictionary and CRF (SDL+CRF) generally boosts performance.

## Qualitative Results




Qualitative examples of ground truth temporal segmentations (GT), predicted temporal segmentations (Pred) and predictions postprocessed by median filtering (Pred+med).

## References

[1] N. Ahmidi et al., A Dataset and Benchmarks for Segmentation and Recognition of Gestures in Robotic Surgery. TBME'17.
[2] L. Tao et al., Sparse Hidden Markov Models for Surgical Gesture Classification and Skill Evaluation. IPCAI'12.
[3] L. Tao et al., Segmentation and Recognition of Surgical Gestures from Kinematic and Video Data. MICCAI'13.
[4] C. Lea et al., An Improved Model for Segmentation and Recognition of Fine-Grained Activities with Application to Surgical Training Tasks. WACV'15.
[5] S. Sefati et al., Learning Shared, Discriminative Dictionaries for Surgical Gesture Segmentation and Classification. M2CAI'15.
[6] R. DiPietro et al., Recognizing surgical activities with recurrent neural networks. MICCAI'16.
[7] C. Lea et al., Temporal Convolutional Networks: A Unified Approach to Action Segmentation. ECCV16-WBNIMR.
[8] C. Lea et al., Learning Convolutional Action Primitives for Fine-grained Action Recognition. ICRA'16.