JHU vision lab

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

Effrosyni Mavroudi[1], Divya Bhaskara[1,2], Shahin Sefati[1,3], Haider Ali[1], René Vidal[1]

[1]Johns Hopkins University, [2]University of Virginia

[3]Comcast AI Research

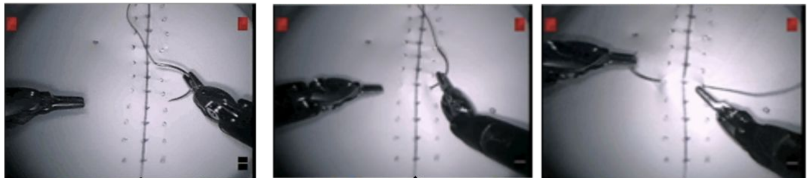
JOHNS HOPKINS MATHEMATICAL INSTITUTE for DATA SCIENCE

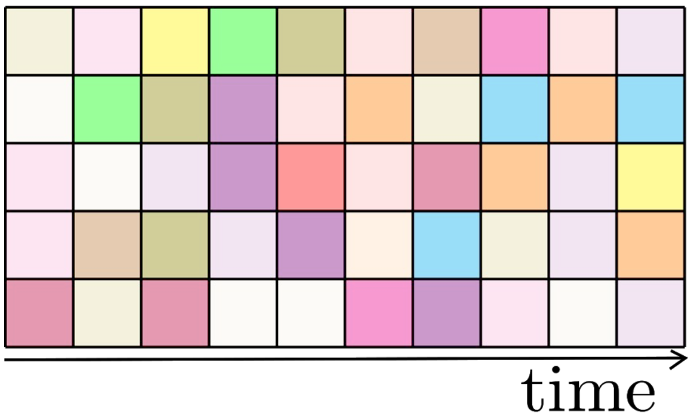
THE DEPARTMENT OF BIOMEDICAL ENGINEERING

The Whitaker Institute at Johns Hopkins

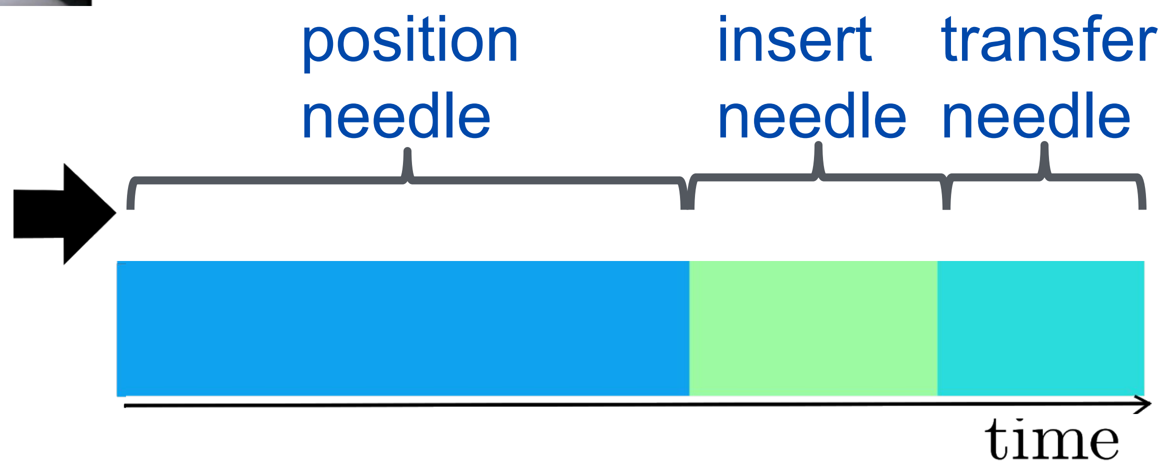
Center for IMAGING SCIENCE

Fine-grained Action Segmentation and Recognition





Input: Kinematic data time-series



Output: Action labels per frame

1) **Which** actions?
2) **When** does each action start/end?

Applications

Automatic Surgical Skill Assessment

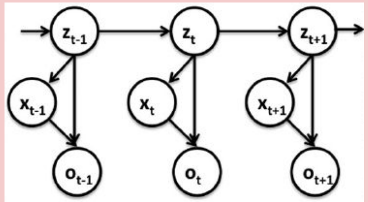
Assisted Living And Smart Home Environments

[1] https://biomedical.closeupengineering.it/wp-content/uploads/2015/07/immagine-evid-11.jpg
[2] http://cvip.computing.dundee.ac.uk/datasets/foodpreparation/50salads/

Prior Work

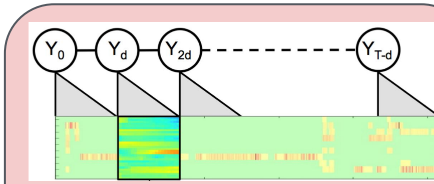
Frame Representation

- Sparse Coding
- Convolutional Filters
- Raw data

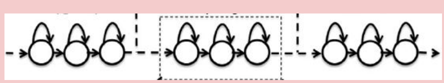


S-HMM

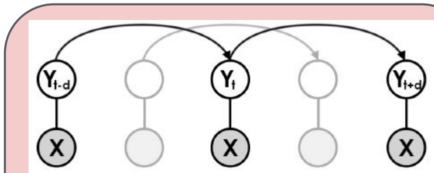
Ours: Discriminative Sparse Coding & Conditional Random Fields



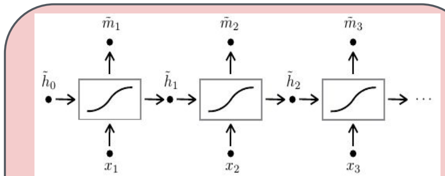
LC-SC-CRF



GMM-HMM



MsM-CRF, SC-CRF



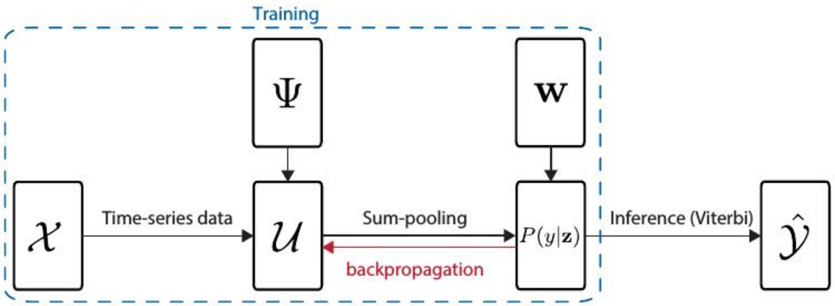
LSTM, TCN

Temporal Model

- Hidden Markov Models
- Conditional Random Fields
- Deep Temporal Models

Related Work

- Discriminative Sparse Dictionary Learning (SDSDL) [1]

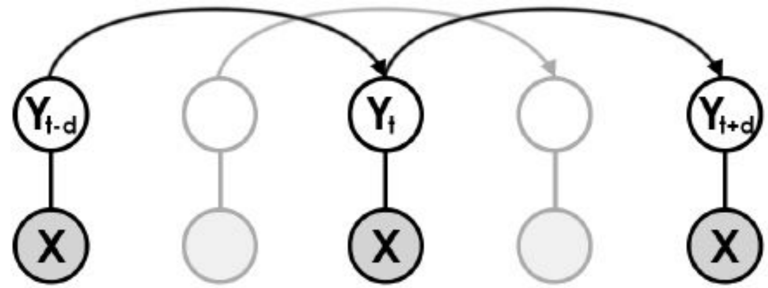


✔ Spatial model: Discriminative Sparse Coding

✘ Temporal model: Precomputed transition probabilities

Dictionary shared between actions and jointly trained with per-frame SVM classifier.

- Skip Chain Conditional Random Field (SC-CRF) [2]



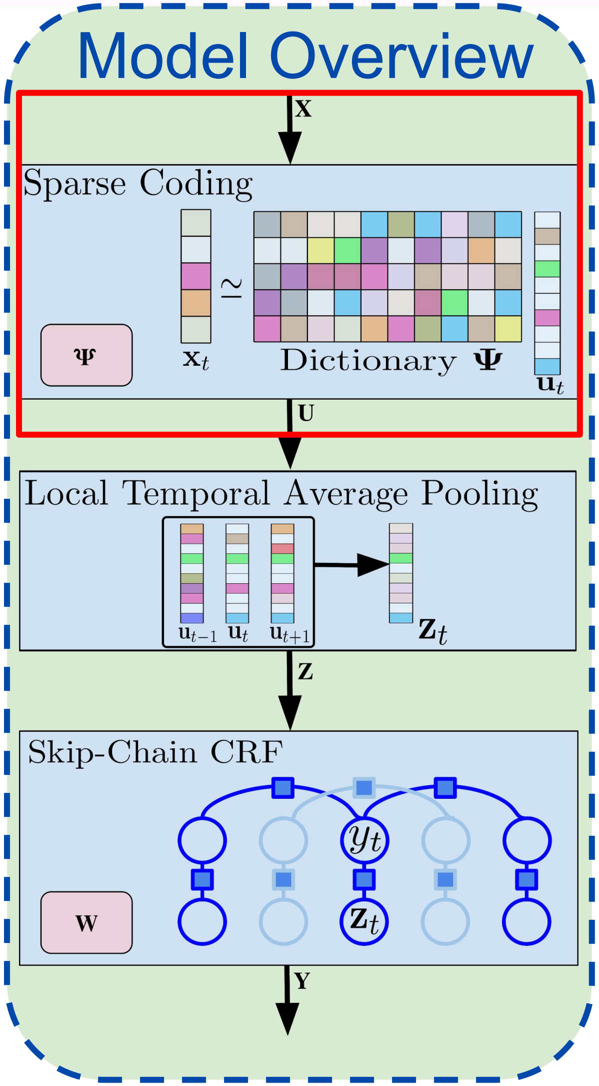
✘ Spatial model: Raw Kinematic Data

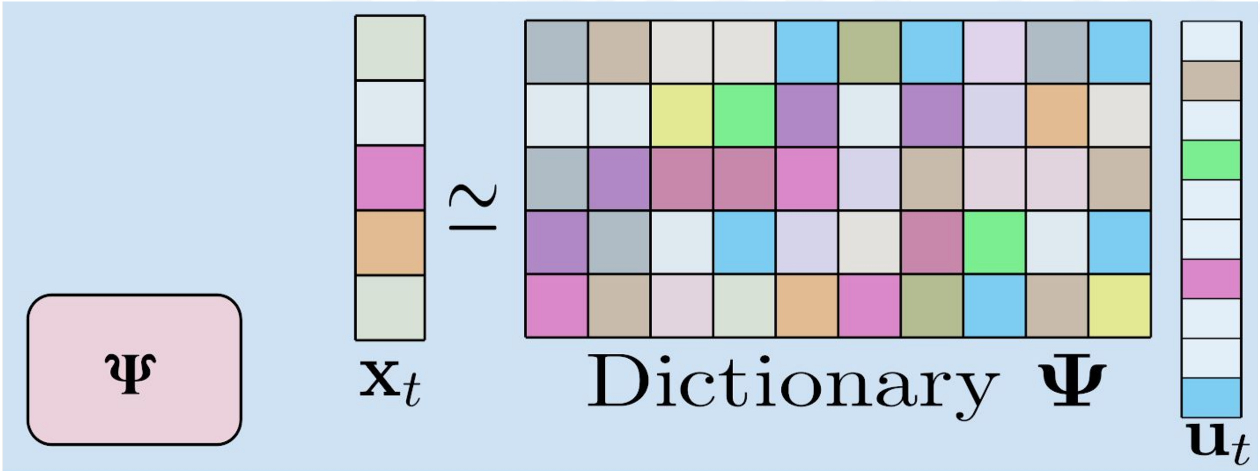
✔ Temporal model: Skip-Chain CRF

SC-CRF can model action to action transitions over large periods of time.

[1] S. Sefati, N. J. Cowan, and R. Vidal. Learning shared, discriminative dictionaries for surgical gesture segmentation and classification, M2CAI 2015
[2] 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, WACV15

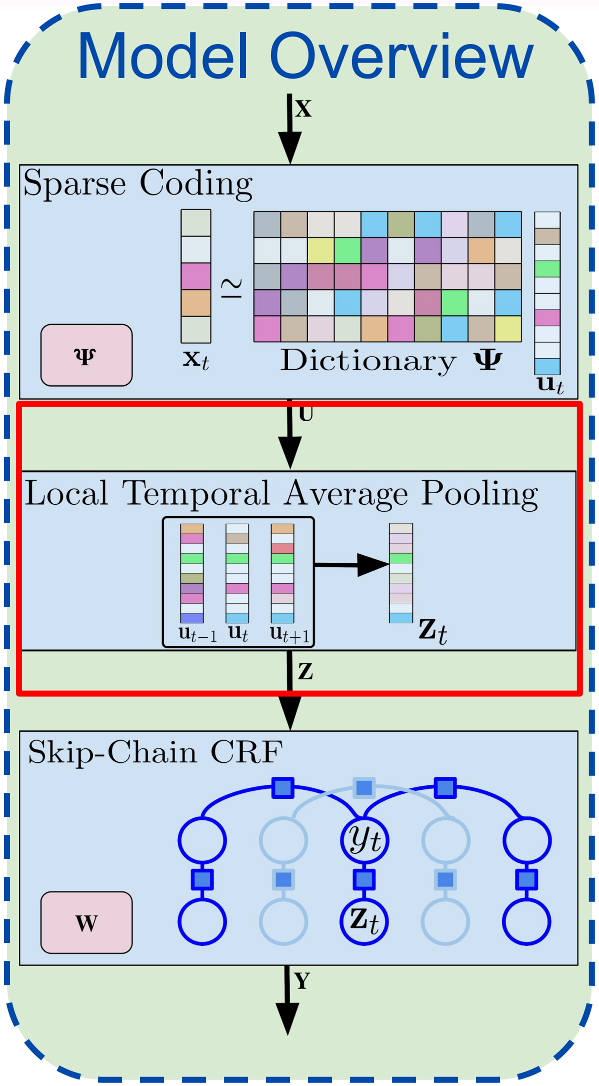
Our model: Frame Representation

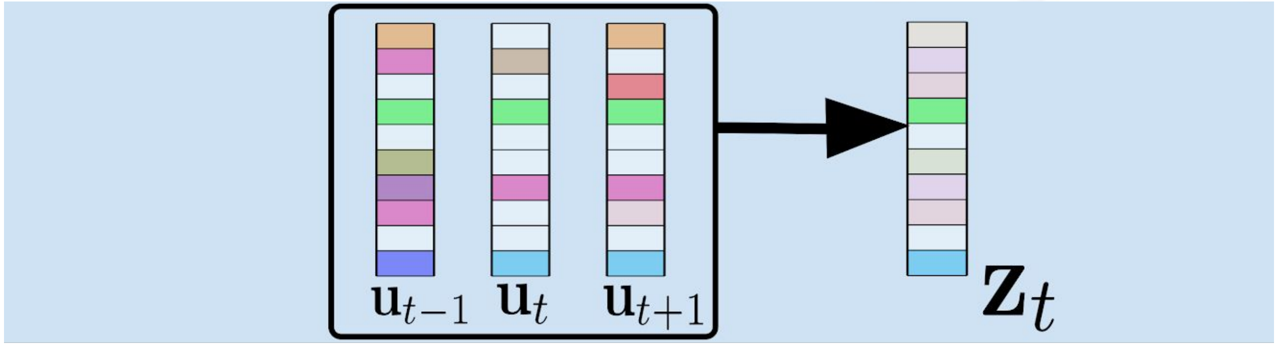




Sparse coding: represent input kinematic data at time t as a combination of a small number of atoms from dictionary $\mathbf{\Psi}$.

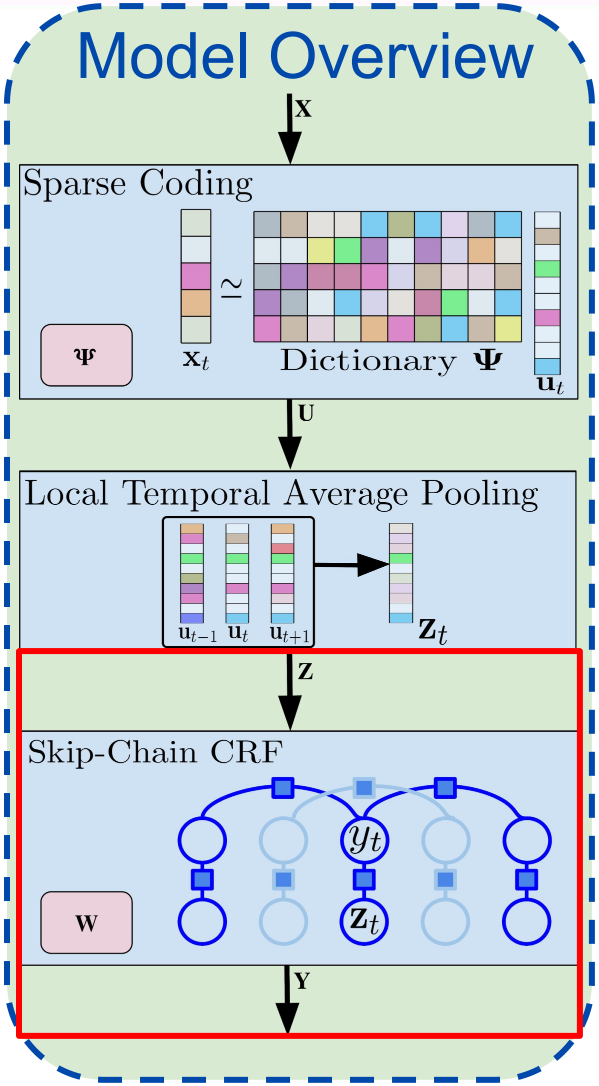
Our model: Local Temporal Representation

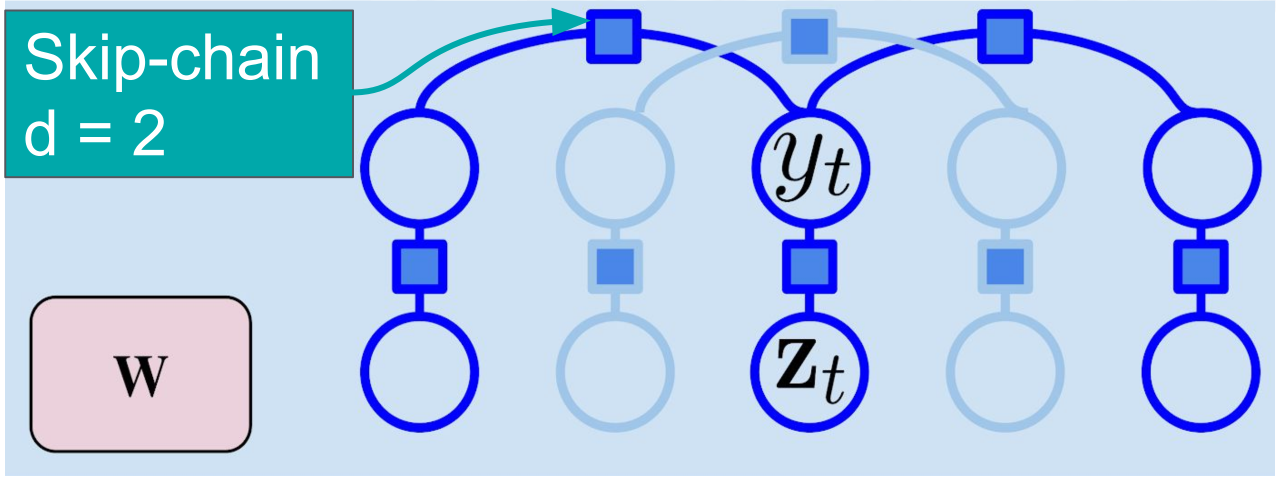




Obtain frame feature by average pooling sparse codes in a short temporal window.

Our model: Global Temporal Representation

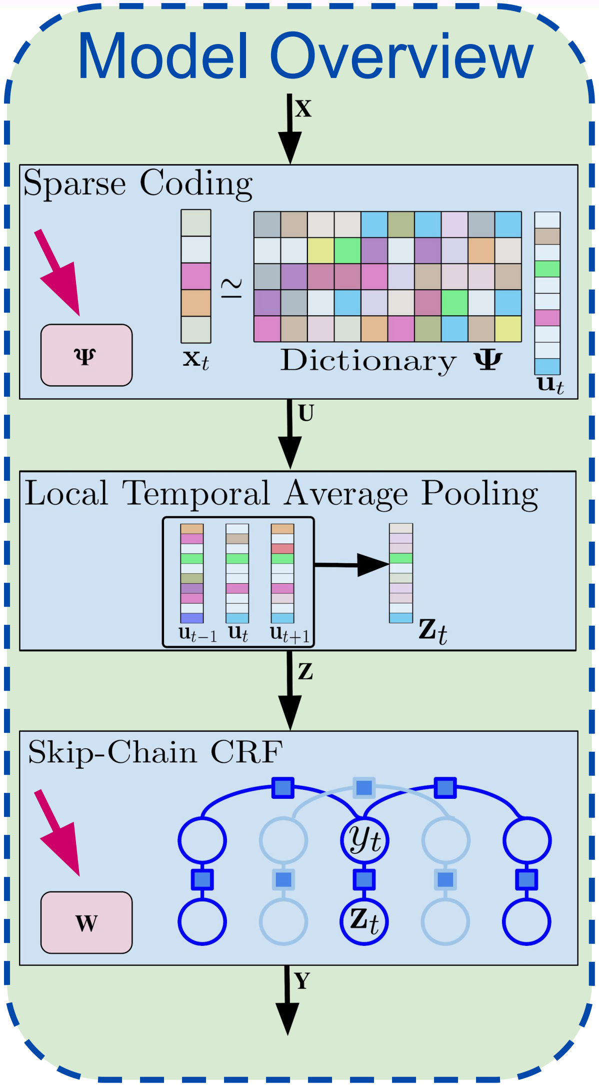




Unary potentials: cost of assigning a label y_t to frame t.

Pairwise potentials: capture transitions between actions and encourage smoothness of labels.

Our model: Training



- End-to-end training of task-driven discriminative dictionary $\mathbf{\Psi}$ and CRF parameters $\mathbf{W}$.
- Use max-margin formulation for structured prediction and optimize using Stochastic Gradient Descent.
- Key challenge: Computing gradient w.r.t dictionary $\mathbf{\Psi}$.

JIGSAWS Dataset

76-dim robot kinematic data

3 surgical tasks

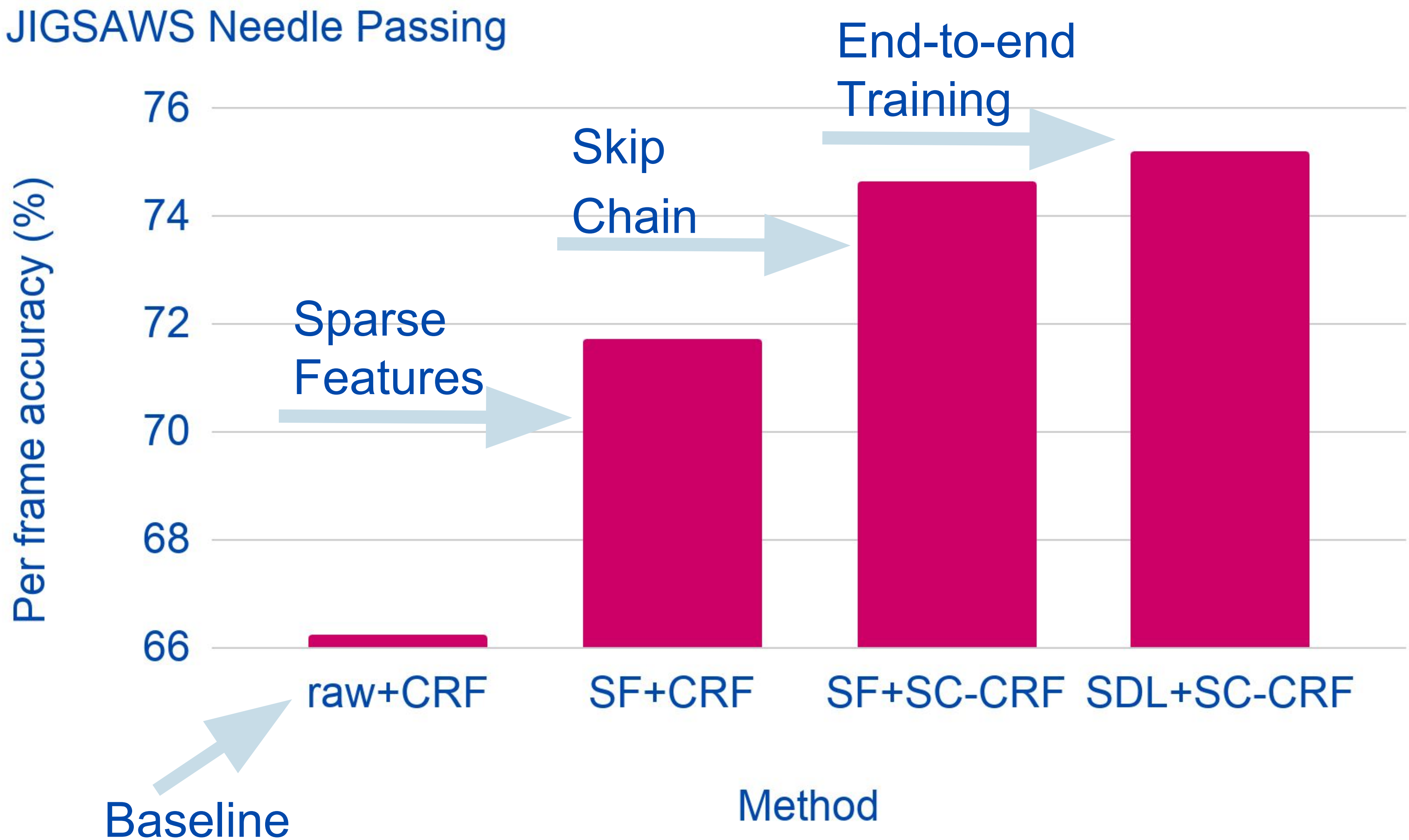
8 surgeons

2-5 min trials (30 Hz)

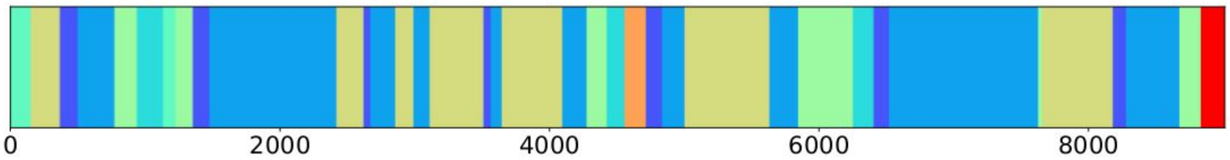
6-10 action classes for each task

[1] JIGSAWS dataset https://cirl.lcsr.jhu.edu/research/hmm/datasets/jigsaws_release/

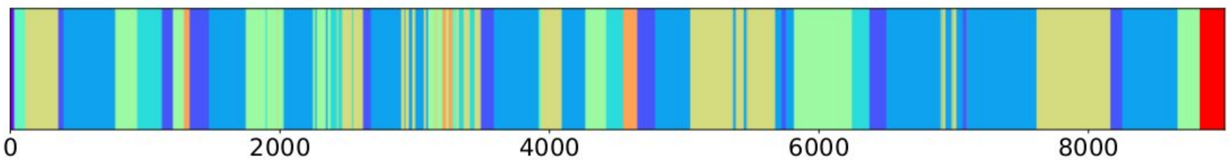
Experimental Results



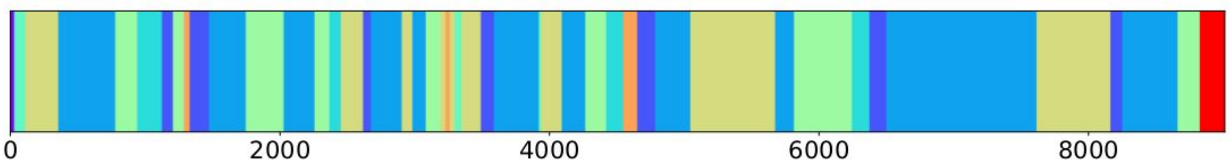
Qualitative Results



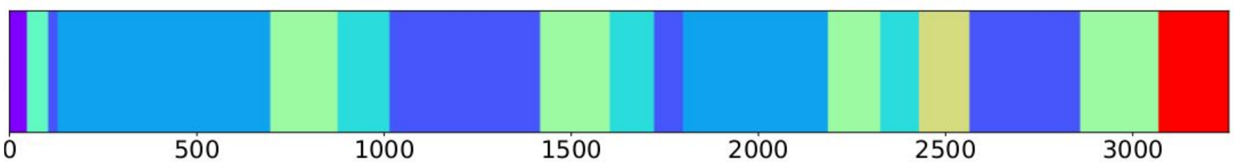
Ground Truth Labels



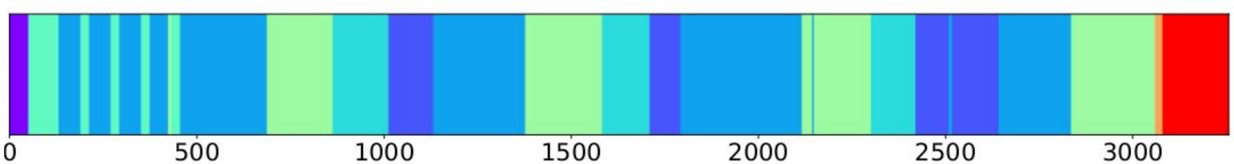
Predicted Labels



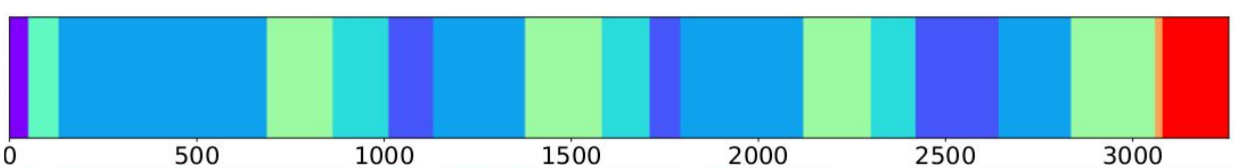
Predicted Labels + median filtering



Ground Truth Labels



Predicted Labels



Predicted Labels + median filtering

Quantitative Results

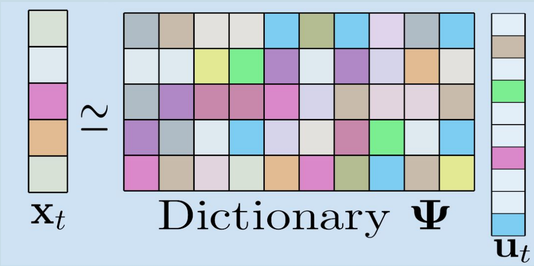
	LOSO			LOUO		
	SU	**KT**	**NP**	**SU**	**KT**	**NP**
GMM-HMM	82.22	80.95	70.55	73.95	72.47	64.13
KSVD-SHMM	83.40	83.54	73.09	73.45	74.89	62.78
MsM-CRF	81.99	79.26	72.44	67.84	44.68	63.28
SC-CRF-SL	85.18	**84.03**	75.09	81.74	**78.95**	**74.77**
LC-SC-CRF				**83.40**		
LSTM				78.38		
BiLSTM				80.15		
TCN				79.6		
SDSDL	**86.32**	82.54	74.88	78.68	75.11	66.01
Ours	*86.21*	*83.89*	**75.19**	78.16	*76.68*	*66.25*

Competitive performance: among 2 best methods for almost all tasks

Conclusions

- A novel spatio-temporal model for fine-grained action segmentation and recognition

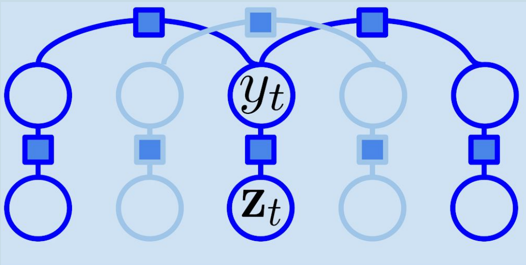
Frame Representation



Discriminative Dictionary Learning

\+

Temporal Model



Skip-Chain CRF

- A novel end-to-end max-margin learning method

For more details visit poster 3B-2 !

More information,

This work was supported by NIH grant R01HD087133

Vision Lab @ Johns Hopkins University

http://www.vision.jhu.edu

Center for Imaging Science @ Johns Hopkins University

http://www.cis.jhu.edu

Johns Hopkins Mathematical Institute for Data Science

http://www.minds.jhu.edu

Thank You!