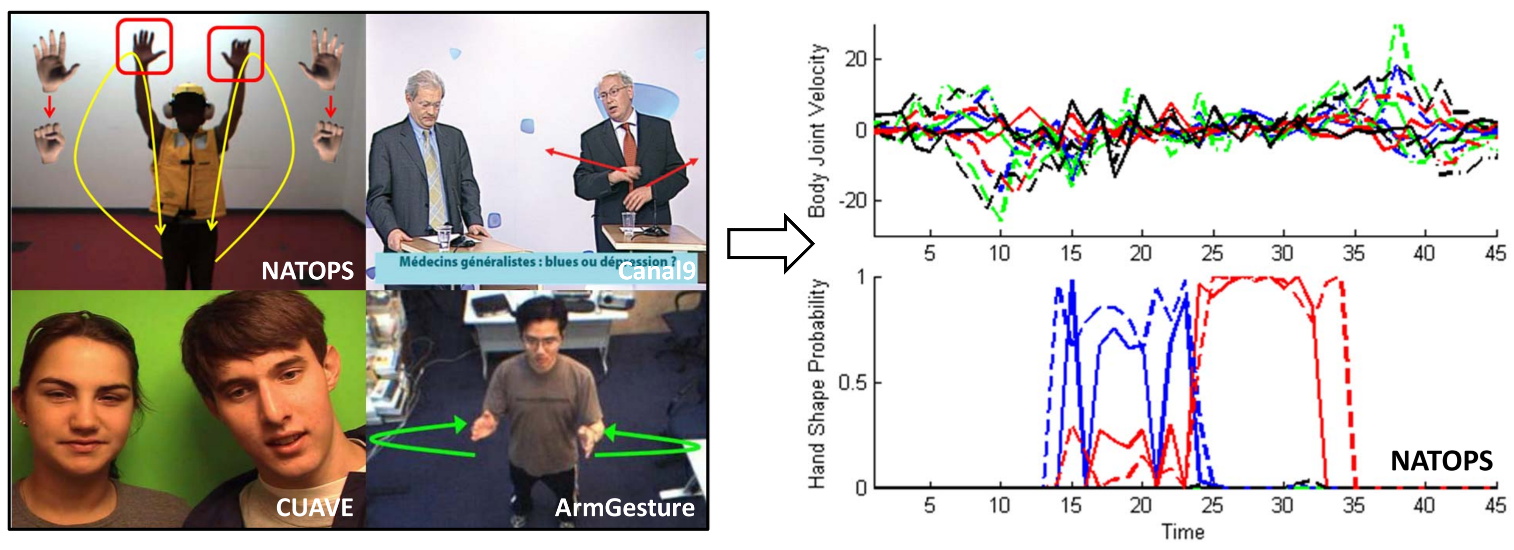
Multi-View Latent Variable Discriminative Models for Action Recognition

Yale Song, Louis-Philippe Morency, and Randall Davis

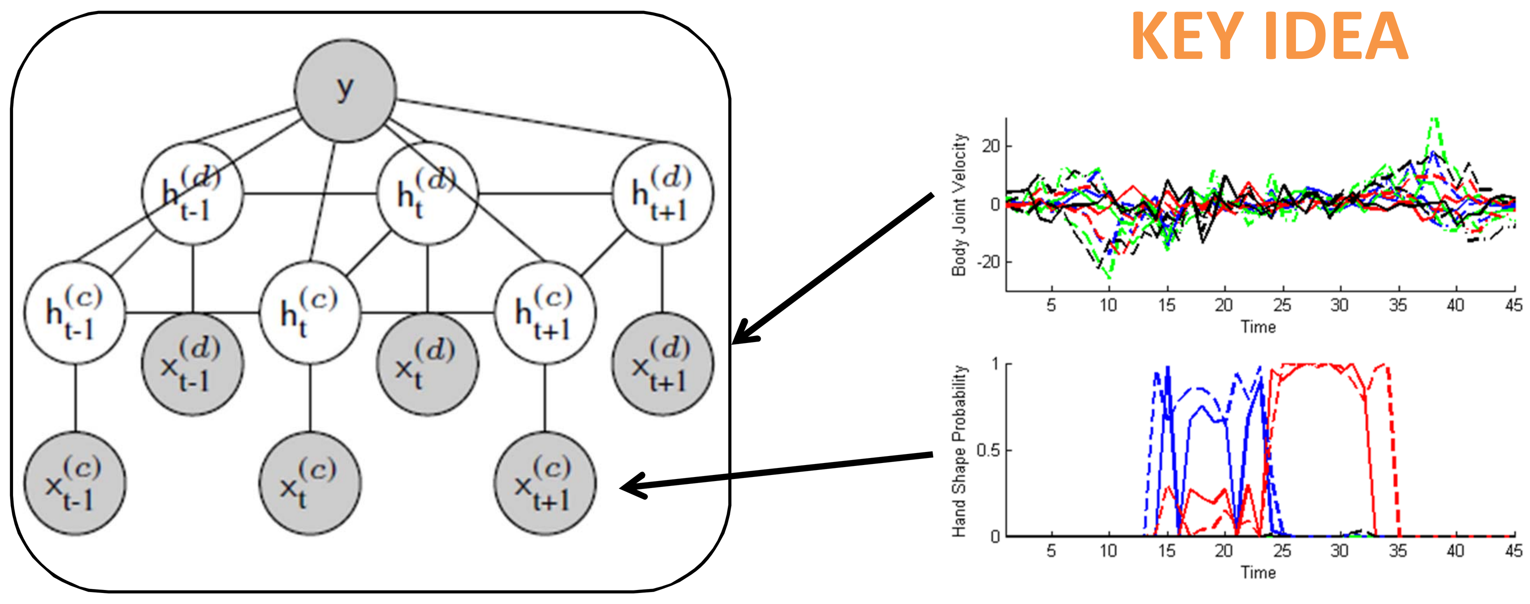
{yalesong,davis}@csail.mit.edu, morency@ict.usc.edu

Challenges in learning with multi-view action sequences



- Many real-world data contain observations from multiple views (e.g., audio-visual) with view-specific dynamics (e.g., variance, noise) and view-shared dynamics (e.g., peaks and valleys)
- Learning hidden dynamics in multi-view data is challenging, especially for learning both view-specific and view-shared interactions
- Standard latent temporal models (e.g., HMM, HCRF) typically have an exponential growth of latent space $O(\mathcal{H}^{\#views})$

KEY IDEA



- Our model jointly learns both *view-specific and view-shared* hidden dynamics using disjoint sets of latent variables
- Each view is treated independently at the observation level, thus our model captures different dynamics from each view more precisely
- Interaction between views are learned explicitly using disjoint sets of latent variables, thus complexity of latent space grows only linearly

Code & dataset available at http://people.csail.mit.edu/yalesong/cvpr12

Multi-View Latent Variable Discriminative Models

Goal: given a multi-view sequence $\hat{\mathbf{x}}_{1:t} = \{\mathbf{x}_{1:t}{}^{(1)}, \cdots, \mathbf{x}_{1:t}{}^{(C)}\}$, learn a temporal model $p(\cdot)$ to perform sequence labeling using disjoint sets of latent variables *specific to each view*, i.e., $\hat{\boldsymbol{h}}_{1:t} = \{\boldsymbol{h}_{1:t}{}^{(1)}, \cdots, \boldsymbol{h}_{1:t}{}^{(C)}\}$

Multi-View HCRF for Segmented Sequence Labeling

$$p(y \mid \hat{\mathbf{x}}; \boldsymbol{\Lambda}) = \sum_{\hat{\boldsymbol{h}}} p(y, \hat{\boldsymbol{h}} \mid \hat{\mathbf{x}}; \boldsymbol{\Lambda}) = \frac{1}{Z(\hat{\mathbf{x}}; \boldsymbol{\Lambda})} \sum_{\hat{\boldsymbol{h}}} \exp\{\Phi(y, \hat{\boldsymbol{h}}, \hat{\mathbf{x}}; \boldsymbol{\Lambda})\}$$

$$y^* = \arg\max_{y \in \mathcal{Y}} \log p(y \mid \hat{\mathbf{x}}; \boldsymbol{\Lambda}^*)$$

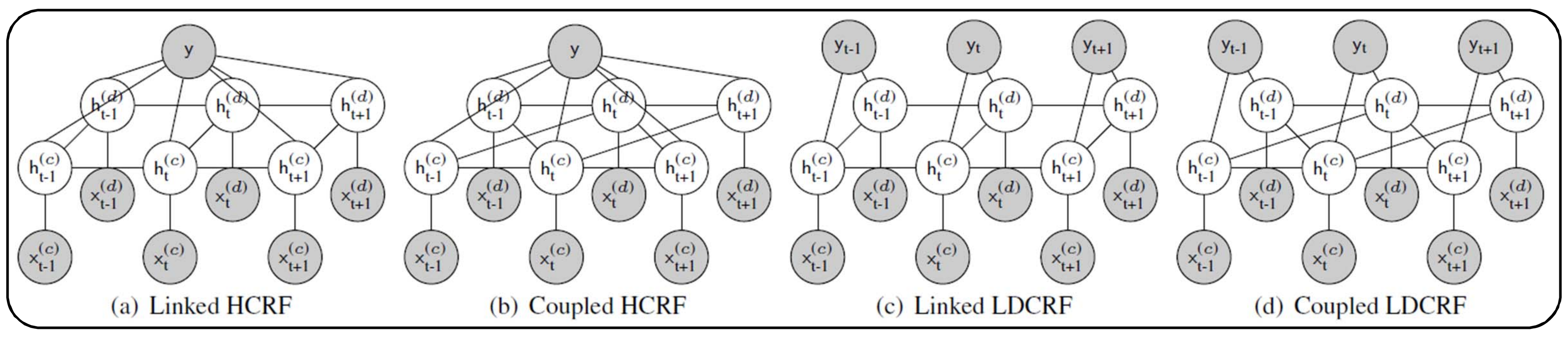
Multi-View LDCRF for Unsegmented Sequence Labeling

$$p(\boldsymbol{y} \mid \hat{\mathbf{x}}; \boldsymbol{\Lambda}) = \sum_{\hat{\boldsymbol{h}}: \forall h_t \in \mathcal{H}_y} p(\hat{\boldsymbol{h}} \mid \hat{\mathbf{x}}; \boldsymbol{\Lambda}) = \frac{1}{Z(\hat{\mathbf{x}}; \boldsymbol{\Lambda})} \sum_{\hat{\boldsymbol{h}}: \forall h_t \in \mathcal{H}_y} \exp\{\Phi(\boldsymbol{y}, \hat{\boldsymbol{h}}, \hat{\mathbf{x}}; \boldsymbol{\Lambda})\}$$

$$\boldsymbol{y}^* = \arg\max_{y \in \mathcal{Y}} \sum_{c=1}^{C} \sum_{\hat{\boldsymbol{h}}: \forall h_t \in \mathcal{H}_y} \alpha_c \cdot \log p(\hat{\boldsymbol{h}}^{(c)} \mid \hat{\mathbf{x}}; \boldsymbol{\Lambda}^*)$$

- Following [2], we assume that each class $y \in \mathcal{Y}$ and each view $c \in \mathcal{C}$ has a disjoint set of latent variables $\mathcal{H}_y{}^{(c)}$, i.e.,

$$p(\boldsymbol{y} \mid \hat{\boldsymbol{h}}, \hat{\mathbf{x}}) = 0, \qquad h_t{}^{(c)} \notin \mathcal{H}_y{}^{(c)}$$



Potential Function and Graph Topology

$$\Phi(\cdot) = \underbrace{\sum_{s,c,k} \lambda^1{}_k \cdot f^1{}_k(h_s{}^{(c)}, \mathbf{x}^{(c)})}_{\text{Observation}} + \underbrace{\sum_{s,c,k} \lambda^2{}_k \cdot f^2{}_k(y, h_s{}^{(c)})}_{\text{Label}} + \underbrace{\sum_{s,t,c,d,k} \lambda^3{}_k \cdot f^3{}_k(y, h_s{}^{(c)}, h_t{}^{(d)})}_{\text{Edge}}$$

$$f^3{}_k(y, h_s{}^{(c)}, h_t{}^{(d)}) = 1 \Leftrightarrow \begin{cases} (s+1 = t \wedge c = d) \vee (s = t \wedge c \neq d) & \text{(linked)} \\ (s+1 = t) & \text{(coupled)} \\ (s+1 = t) \vee (s = t \wedge c \neq d) & \text{(linked-coupled)} \end{cases}$$

Parameter Estimation and Optimization

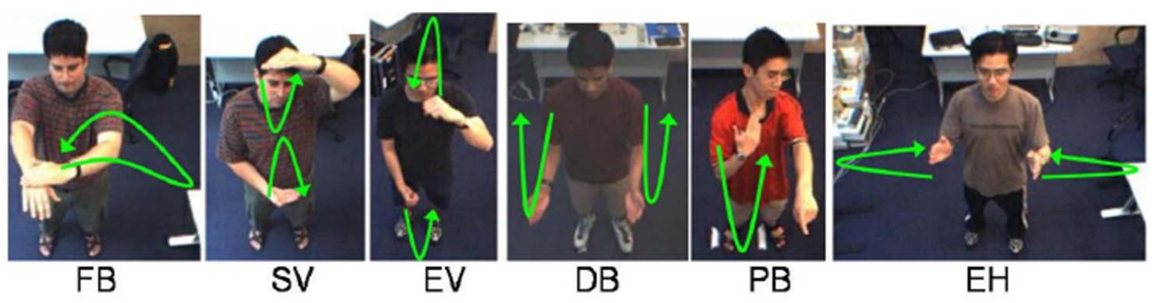
- Given a training dataset $\mathcal{D} = \{y_i, \hat{\mathbf{x}}_i\}, i = 1 \cdots N$, find $\boldsymbol{\Lambda}^*$ by solving

$$\boldsymbol{\Lambda}^* = \arg\min_{\boldsymbol{\Lambda}} -\sum_{i=1}^{N} \log p(y_i \mid \hat{\mathbf{x}}_i; \boldsymbol{\Lambda}) + \frac{\gamma}{2} \|\boldsymbol{\Lambda}\|_2$$

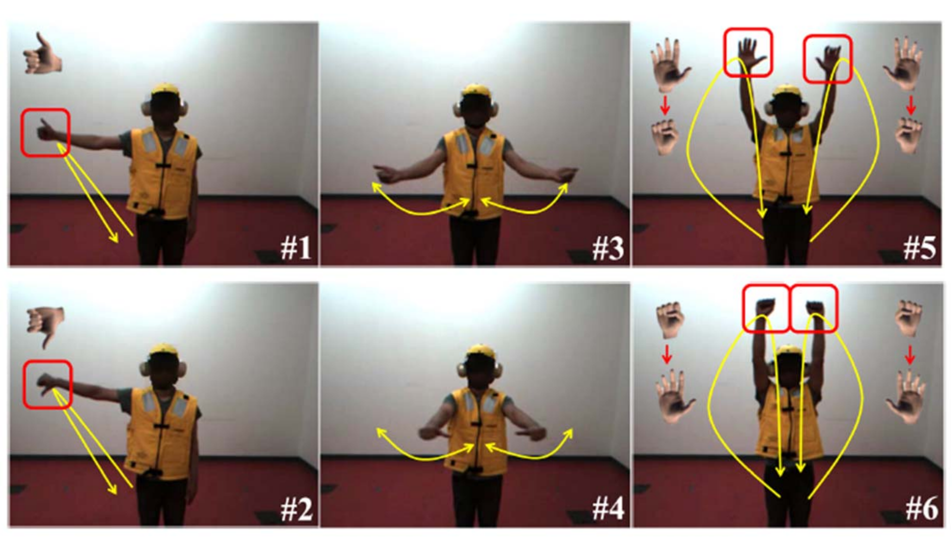
- Optimization is done by using non-convex regularized bundle method
- Inference is done by using Junction Tree or Loopy Belief Propagation

Experiments

Datasets



ArmGesture (left + right arms)



NATOPS (body + hand)

ArmGesture

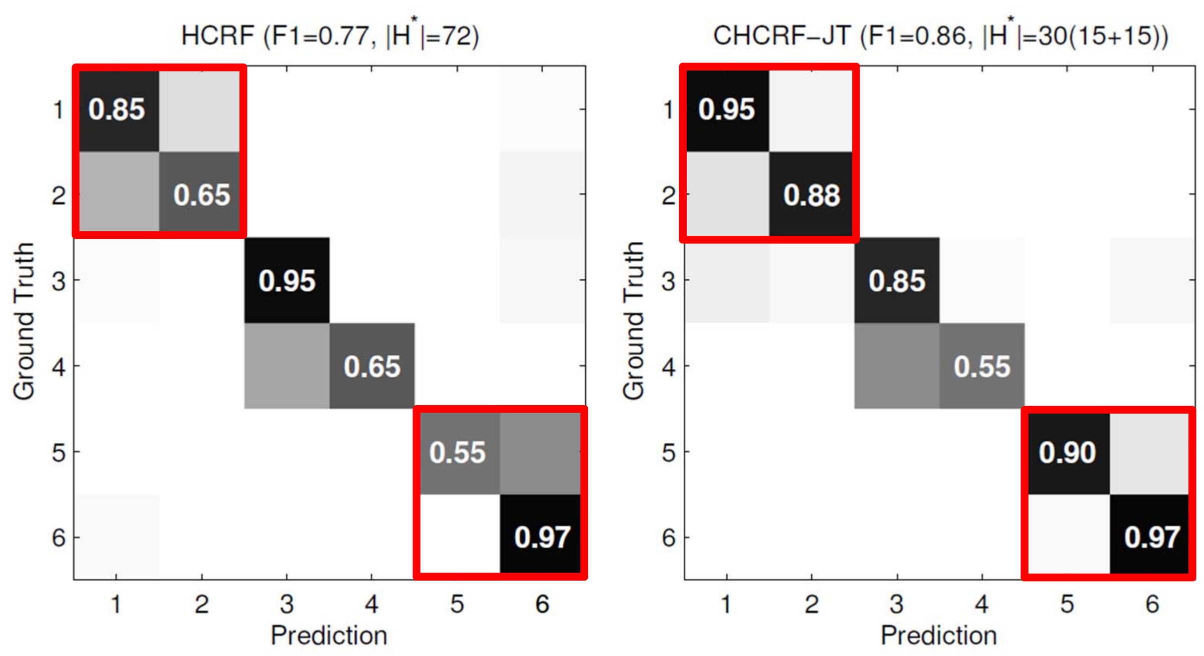
Models	Accuracy (%)
HMM [18]	84.22
CRF [18]	86.03
HCRF [18]	91.64
MM-HCRF	93.79
S-KDR-SVM [20]	95.30
Linked HCRF	**97.65**
Coupled HCRF	**97.24**

ArmGesture-Continuous

Models	Accuracy (%)
CRF	90.80
LDCRF	91.02
Linked LDCRF	**92.51**
Coupled LDCRF	**92.44**

NATOPS

Models	Accuracy (%)
HMM	77.67
CRF	53.30
HCRF	78.00
Linked HCRF	**87.00**
Coupled HCRF	**86.00**



Key Findings

- Learning hidden dynamics in multi-view data using ***view-specific latent variables*** significantly improved recognition accuracy
- Improvement is especially clear ***when each view has important information***, e.g., for the first and third gesture pair in NATOPS, CHCRF achieved 92% and 94%, while for HCRF they were 75% and 76%.

Acknowledgements:

This work was funded by ONR #N000140910625, NSF #IIS-1018055, and the U.S. Army Research, Development, and Engineering Command (RDE-COM).

Corresponding Author:

Yale Song, yalesong@csail.mit.edu

References:

1. A. Quattoni, S. B. Wang, L.-P. Morency, M. Collins, and T. Darrell. *Hidden conditional random fields*. TPAMI 29(10), 2007.
2. L.-P. Morency, A. Quattoni, and T. Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In CVPR 2007.
3. Y. Song, D. Demirdjian, and R. Davis. *Tracking body and hands for gesture recognition: NATOPS aircraft handling signals database*. In FG 2011.