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 \( \mathcal{O}(N^p) \)

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

**Multi-View HCRF for Segmented Sequence Labeling**

\[
p(y | x; \lambda) = \frac{1}{Z(\lambda)} \sum_{h \in \mathcal{H}} \exp \Phi(y, h, x; \lambda)
\]

\[
y^* = \arg \max_{y \in \mathcal{Y}} \log p(y | x; \lambda)
\]

**Multi-View LDCRF for Unsegmented Sequence Labeling**

\[
p(y | h; \lambda) = \frac{1}{Z(\lambda)} \sum_{h \in \mathcal{H}} \exp \Phi(h, y; \lambda)
\]

\[
y^* = \arg \max_{y \in \mathcal{Y}} \log p(y | h; \lambda)
\]

- 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.,

\[
p(y | h, x) = 0, \quad h \notin \mathcal{H}_{y,c}
\]

**Potential Function and Graph Topology**

\[
\Phi(c) = \sum_{c \in \mathcal{C}} \sum_{c \in \mathcal{C}} \Phi_c(\overline{c}; \lambda)
\]

- Given a training dataset \( D = \{y_i, \mathbf{x}_i\}, i = 1 \ldots N \), find \( \lambda^* \) by solving

\[
\lambda^* = \arg \min_{\lambda} - \sum_{i=1}^{N} \log p(y_i | \mathbf{x}_i; \lambda) + \frac{1}{2} \|\lambda\|_2
\]

- 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
    - HMM: 84.22
    - CRF \([18]\): 86.03
    - HCRF \([13]\): 91.64
    - MM-HCRF: 93.79
    - S-KDR-SVM \([20]\): 95.30
    - Linked HCRF: 97.65
    - Coupled HCRF: 97.24

- **ArmGesture-Continuous**
  - Models
    - CRF: 90.80
    - LDCRF: 91.02
    - Linked LDCRF: 92.51
    - Coupled LDCRF: 92.44

- **NATOPS**
  - Models
    - 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, HCRF achieved 92% and 94%, while for HCRF they were 75% and 76%.