Multimodal Human Behavior Analysis: Learning Correlation and Interaction Across Modalities

Yale Song
MIT CSAIL

Speaker

Louis-Philippe Morency
USC ICT

Randall Davis
MIT CSAIL
Human Communication is Multimodal

Speech + Body gesture + Facial Expression + …
Multimodal data is abundant

Canal9 – Political debates
A. Vinciarelli et al. ACII 2011

ICT Youtube
L.-P. Morency et al. ICMI 2011

AVEC 2012

D-META 2012

Challenges: ML perspective

- Modality-specific characteristic
  e.g., range, mean, variance, noise, frame rate

- Modality-specific dynamic
  e.g., spatio-temporal dynamics, prosodic dynamics

- Complex correlation/interaction
  across modalities
  e.g., loud voice + exaggerated gesture

Can we build an algorithm that considers all these?
(Typical) multimodal learning

- Early fusion

  - Audio
  - Video
  
  Fusion
  
  $f(x_A, x_V)$

  Strong modality can dominate (e.g., high variance)

- Late fusion

  - Audio
  - Video
  
  $f(x_A)$
  
  $f(x_V)$
  
  Fusion

  Cross-modality pattern is lost

Previous work

- Co-training, co-regularization, and variants
  - Blum and Mitchell, COLT 98; Nigam and Ghani, CIKM 00; Balcan et al., NIPS 04; Zhou and Li, IJCAI 05; Kumar and Daume III, ICML 11; Guo and Xiao, ICML 12
  - Strict assumptions (e.g., sufficiency, compatibility, conditional independence)

- Sparse coding / dictionary learning
  - Mairal et al., NIPS 08; Jia et al., NIPS 10; Yang and Yang, CVPR 12
  - Hard to generalize to multimodal sequence learning

- Component / correlation analysis (PCA, CCA, ICA,...)
  - Hardoon et al., Neural Comp. 04; Cui et al., ICDM 07; Chaudhuri et al., ICML 09
  - Usually data-driven (not task-specific), no sequence learning

- Factorized graphical models (DBN,...)
  - Brand et al., CVPR 97; Chen et al., NIPS 10; Jia et al., ICCV 11; Song et al., CVPR 12
  - Restricted to log-linear model; non-linear relationship is not captured
Key Idea: KCCA + MV-HCRF

**Kernel CCA**

- Transform $x_A, x_V$ s.t. correlation is max-ed

**Multi-view HCRF**

- Learn the *interaction* btn. $x'_A, x'_V$ using multi-view CRF w/ latent vars
Key Idea: KCCA + MV-HCRF

Kernel CCA

Transform $x_A, x_V$ s.t. correlation is max-ed

Learn the interaction btn. $x_A', x_V'$ using multi-view CRF w/ latent vars

Fusion $f(x_A', x_V')$
Canonical Correlation Analysis

• Given a set of paired samples (e.g., audio-visual)

\[ \{(a_i, v_i) \mid a_i \in \mathbb{R}^{d_A \times T}, v_i \in \mathbb{R}^{d_V \times T}\}_{i=1}^{N} \]

\[ A = [a_1, \ldots a_N], V = [v_1 \ldots v_N] \]

• Find a pair of transformations \((w_a, w_v)\) s.t.

\[ \max_{w_a, w_v} \rho(w_a \cdot A, w_v \cdot V) \]

• CCA finds \((w_a, w_v)\) that are linear in feature space; thus may not reveal nonlinear relationship in the data

Kernalize!
The “Kernel Trick”

\[ \rho(\cdot, \cdot) = \max_{w_a, w_v} \frac{\mathbb{E}[w_a^T a v^T w_v]}{\sqrt{\mathbb{E}[w_a^T a a^T w_a]} \sqrt{\mathbb{E}[w_v^T v v^T w_v]}} \]

\[ = \max_{w_a, w_v} \frac{w_a^T A V^T w_v}{\sqrt{w_a^T A A^T w_a} w_v^T V V^T w_v} \]

\[ = \max_{\alpha, \beta} \frac{\alpha A A^T V V^T \beta}{\sqrt{\alpha A A^T A A^T \alpha \cdot \beta V V^T V V^T \beta}} \]

\[ = \max_{\alpha, \beta} \frac{\alpha^T K_a K_v \beta}{\sqrt{\alpha^T K_a^2 \alpha \cdot \beta^T K_v^2 \beta}}. \]
Kernel CCA on Audio-Visual Data

from **Canal9 Database**, A. Vinciarelli *et al*. ACII 2011

Emphasizing the amplitude of ‘*head shake*’ and ‘*shoulder shrug*’ maximized the correlation between audio and visual channels
Key Idea: KCCA + MV-HCRF

Transform $x_A, x_V$ s.t. correlation is max-ed

Learn the interaction btn. $x'_A, x'_V$ using multi-view CRF w/ latent vars
Multi-view HCRF [Song et al., CVPR 12]

Multi-view HCRF [Song et al., CVPR 12]

Multi-view HCRF [Song et al., CVPR 12]

Multi-view HCRF [Song et al., CVPR 12]

Capture the interaction across modalities more precisely (e.g., modality-specific and modality-shared patterns)

Multi-view HCRF (cont’d)

\[
p(y | \hat{x}; \Lambda) = \sum_{\hat{h}} p(y, \hat{h} | \hat{x}; \Lambda) = \frac{1}{Z(\hat{x}; \Lambda)} \sum_{\hat{h}} \exp\{\Phi(y, \hat{h}, \hat{x}; \Lambda)\}
\]

\[
\Phi(\cdot) = \sum_{s,c,k} \lambda^1_k \cdot f^1_k(h_s^{(c)}, x^{(c)}) + \sum_{s,c,k} \lambda^2_k \cdot f^2_k(y, h_s^{(c)}) + \sum_{s,t,c,d,k} \lambda^3_k \cdot f^3_k(y, h_s^{(c)}, h_t^{(d)})
\]

\[
f^3_k(y, h_s^{(c)}, h_t^{(d)}) = 1 \iff \begin{cases} 
(s + 1 = t \land c = d) \lor (s = t \land c \neq d) & \text{(linked)} \\
(s + 1 = t) & \text{(coupled)} \\
(s + 1 = t) \lor (s = t \land c \neq d) & \text{(linked-coupled)} 
\end{cases}
\]

Given a training dataset \( D = \{y_i, \hat{x}_i\}, i = 1 \cdots N \), find \( \Lambda^* \) by solving

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

- **Canal9 Database** [Bousmalis et al. 2011]

  **Goal:** binary sequence classification \{agree | disagree\}

- **Observational features**

  2D audio features (F0 and Energy)  
  8D visual features (binary, 8 gestures)

- **5-fold Cross Validation**

- **Three questions:**
  
  - Does combining audio-visual features help?  
  - Does MV-HCRF improve performance?  
  - Does KCCA + MV-HCRF improve performance?


Experiments (cont’d)

<table>
<thead>
<tr>
<th>Models</th>
<th>Audio</th>
<th>Video</th>
<th>Audio+Video</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM</td>
<td>.54 (.08)</td>
<td>.58 (.11)</td>
<td>.59 (.09)</td>
</tr>
<tr>
<td>CRF</td>
<td>.48 (.05)</td>
<td>.58 (.15)</td>
<td>.61 (.04)</td>
</tr>
<tr>
<td>HCRF</td>
<td>.52 (.09)</td>
<td>.60 (.09)</td>
<td>.64 (.13)</td>
</tr>
</tbody>
</table>

- Yes!
Experiments (cont’d)

<table>
<thead>
<tr>
<th>Models</th>
<th>Audio</th>
<th>Video</th>
<th>Audio+Video</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM</td>
<td>.54 (.08)</td>
<td>.58 (.11)</td>
<td>.59 (.09)</td>
</tr>
<tr>
<td>CRF</td>
<td>.48 (.05)</td>
<td>.58 (.15)</td>
<td>.61 (.04)</td>
</tr>
<tr>
<td>HCRF</td>
<td>.52 (.09)</td>
<td>.60 (.09)</td>
<td>.64 (.13)</td>
</tr>
<tr>
<td>MV-HCRF</td>
<td>•</td>
<td>•</td>
<td>.68 (.13)</td>
</tr>
</tbody>
</table>

- Yes!
- Yes! by learning *interaction* across modalities
Experiments (cont’d)

<table>
<thead>
<tr>
<th>Models</th>
<th>Audio</th>
<th>Video</th>
<th>Audio+Video</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM</td>
<td>.54 (.08)</td>
<td>.58 (.11)</td>
<td>.59 (.09)</td>
</tr>
<tr>
<td>CRF</td>
<td>.48 (.05)</td>
<td>.58 (.15)</td>
<td>.61 (.04)</td>
</tr>
<tr>
<td>HCRF</td>
<td>.52 (.09)</td>
<td>.60 (.09)</td>
<td>.64 (.13)</td>
</tr>
<tr>
<td>MV-HCRF</td>
<td>·</td>
<td>·</td>
<td>.68 (.13)</td>
</tr>
<tr>
<td><strong>KCCA + MV-HCRF</strong></td>
<td>·</td>
<td>·</td>
<td><strong>.72 (.07)</strong></td>
</tr>
</tbody>
</table>

- Yes!
- Yes! by Learning *interaction* across modalities
- Yes! by learning *correlation+interaction* across modalities
Contributions and Future Work

We showed that

- Learning *correlation & interaction* across modalities using KCCA + MV-HCRF improves performance

In the future,

- Evaluate our model using other multimodal datasets
- Combine the two-step process into a unified framework e.g., CCA + SVM = SVM-2K [Farquhar et al. NIPS ’06]
Thank You!

KCCA

\[ X_A, X'_A, X_V, X'_V \]

Multi-view HCRF

Fusion \[ f(x'_A, x'_V) \]

C++ code for MV-HCRF is available at
http://people.csail.mit.edu/yalesong/cvpr12