‘Exemplar Hidden Markov Models for Classification of Facial Expressions in Videos’

Workshop on Analysis and Modeling of Face and Gesture
CVPR 2015

Karan Sikka
Machine Perception Lab
UCSD

Joint work with Dr. Abhinav Dhall and Dr. Marian Bartlett
Automatic Facial Expression Recognition

- Classify underlying expressions in a video.
- Emotions, Pain, engagement level.

Smile
Disgust
Surprise

Smile
Disgust
Surprise
Previous Art

Image Based Approaches

- Spatial features + Classifier.

- Issues
  - Require key-shots (apex frames).
  - No explicit dynamics.

- Gabor, LBP, SIFT.

- Image -> Video based approaches.
Previous Art
Video-based approaches

• **Space-Time**
  – Extract localized S-T features across entire video.
  – Feature pooling + Classifier.
  – LBPTOP, BoW, facial point time-series.

• **Issues**
  1. Pooling from multiple expressions.
     • Loss of discriminative power (unsegmented videos).
  2. Loss of temporal information.
     • No temporal correspondence between facial states.
Previous Art for AFER
Video based approaches

• Sequential
  – Analyze an expression as a sequence of features.
  – Explicitly model spatio-temporal aspects.

• Focus on HMMs.
  – Desirable properties for modeling expressions.
Why HMM

- HMMs model expression dynamics.
  
  ![Diagram showing HMM states and transitions](image)

- **Hidden states**: Temporal Segmentation (variable length)
- **Model per state**: Model behavior for each facial state (local states).
- **Transition probabilities**: Temporal dynamics
**HMMs -> Exemplar HMMs**

- **In PRACTICE** $\text{Accuracy(HMM)} < \text{Accuracy (Discriminative)}$
  - Generative model.
  - Modeling decision boundary is easier than modeling classes.

- **Solution Proposed**
  - *Structural* advantages of HMM + *discriminative* ability SVMs.
  - Probabilistic kernels.

- Probabilistic kernels.
  - Recognition of dynamics textures, handwritten text, shapes.
  - Jaakkola et al., Jebara et al., Vasconcelos et al.
Kernels and implicit space

• Dot products in SVM can be replaced by Kernel functions (Kernel SVM)
  \[ K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle \]

• Possible to compute Dot products indirectly for points in non-Euclidian (implicit) space
  – \( \Phi \) maps HMM models to a vector space.
  – \( K(p_i, p_j) = \langle \Phi(p_i), \Phi(p_j) \rangle \)
Exemplar-HMMs

Videos

$p(x_1)$

$p(x_j)$

$p(x_f)$

HMM

Implicit Projection function via Kernel

Decision Boundary via kernel SVM
Probabilistic Product Kernel (PPK)

- PPK (Jebara et al.) to compute distance between two HMMs:
  \[ k(p_1, p_2) = \int p_1(x) p_2(x) dx \]

- Closed form solution of HMM (Exponential family).

- Intuitive: Compares all states from two HMMs while using transition probabilities.
## Experiments
### Basic Emotions

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Type</th>
<th>Videos/subjects</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>CK+</td>
<td>Posed</td>
<td>327 (118 subj)</td>
<td>7 emotions (leave-one-subject)</td>
</tr>
<tr>
<td>Oulu-CASIA VIS</td>
<td>Posed</td>
<td>480 (80 sub)</td>
<td>6 emotions (10 fold)</td>
</tr>
<tr>
<td>FEEDTUM</td>
<td>Spontaneous</td>
<td>320 (19 subj)</td>
<td>6 emotions (leave-one-subject)</td>
</tr>
</tbody>
</table>

- Time-series of facial landmarks points (49*2 dim).
- PCA.
- Metric: Average accuracy across all classes.
Competing algorithms

1. Global S-T
   - Landmark features + pooling +SVM
   - LBPTOP: Local Binary Patterns (texture) from XYT planes.
     • S-T Histograms + Pooling

2. Baseline generative model
   - HMM generative classifier

   - Explicit temporal info inside texture features.
   - Universal GMM (UGMM) learned.
   - Video-> Align localized S-T features with UGMM.
## Experiments

### Basic Emotions - Posed

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (CK +)</th>
<th>Accuracy (Oulu)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geom. + Mean-pooling</td>
<td>93.00 (±1.55)</td>
<td>70.83 (±2.84)</td>
</tr>
<tr>
<td>Geom. + Max-pooling</td>
<td>92.85 (±1.67)</td>
<td>69.16 (±1.80)</td>
</tr>
<tr>
<td>LBPTOP</td>
<td>91.30 (±1.79)</td>
<td>72.08 (±2.22)</td>
</tr>
<tr>
<td>HMM</td>
<td>85.35 (±2.16)</td>
<td>63.54 (±3.10)</td>
</tr>
<tr>
<td>STM-ExpLet</td>
<td>94.19 (N/A)</td>
<td>74.59 (N/A)</td>
</tr>
<tr>
<td>ITBN</td>
<td>86.3 (±N/A)</td>
<td>NA</td>
</tr>
<tr>
<td>Exemplar-HMMs</td>
<td><strong>94.60 (±1.55)</strong></td>
<td><strong>75.00 (±2.12)</strong></td>
</tr>
</tbody>
</table>

- Significant improvement compared to S-T approaches.
# Experiments

## Basic Emotions - Posed

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (CK +)</th>
<th>Accuracy (Oulu)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geom. + Mean-pooling</td>
<td>93.00 (±1.55)</td>
<td>70.83 (±2.84)</td>
</tr>
<tr>
<td>Geom. + Max-pooling</td>
<td>92.85 (±1.67)</td>
<td>69.16 (±1.80)</td>
</tr>
<tr>
<td>LBPTOP</td>
<td>91.30 (±1.79)</td>
<td>72.08 (±2.22)</td>
</tr>
<tr>
<td>HMM</td>
<td>85.35 (±2.16)</td>
<td>63.54 (±3.10)</td>
</tr>
<tr>
<td>STM-ExpLet</td>
<td>94.19 (N/A)</td>
<td>74.59 (N/A)</td>
</tr>
<tr>
<td>ITBN</td>
<td>86.3 (±N/A)</td>
<td>NA</td>
</tr>
<tr>
<td>Exemplar-HMMs</td>
<td>94.60 (±1.55)</td>
<td>75.00 (±2.12)</td>
</tr>
</tbody>
</table>

- Advantages of discriminative modeling over generative modeling.
### Experiments

**Basic Emotions - Spontaneous**

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (FEEDTUM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geom. + Mean-pooling</td>
<td>48.91 (±3.70)</td>
</tr>
<tr>
<td>Geom. + max-pooling</td>
<td>53.87 (±2.59)</td>
</tr>
<tr>
<td>LBPTOP</td>
<td>48.17 (±3.31)</td>
</tr>
<tr>
<td>HMM</td>
<td>48.23 (±3.88)</td>
</tr>
<tr>
<td>Exemplar-HMMs</td>
<td><strong>54.14 (±3.72)</strong></td>
</tr>
</tbody>
</table>
AMFED Dataset

• Videos of participants watching 3 superbowl commercials.

• Video responses collected over the internet along with self-ratings describing:
  – Like/not-like
  – Watch again or not

• Public Dataset
  – 242 videos
  – Expert annotations for: AU 2, 4, 5 9, 12, 14, 15, 17, 18, 26 + Smile + Expressability.
  – Annotations in form of agreement between annotators.
AMFED

• 2 Binary self-report prediction tasks
  – Predict whether a video is rated liked/not-liked.
  – Predict whether a video will be watched again or not.

• Using time-series of AU annotations.
  – Threshold to 0 (<50%) and 1 (>=50%) based on agreement.

• 3 Fold
  – AUC
## Results

### AMFED

<table>
<thead>
<tr>
<th>Method</th>
<th>Like/Don’t Like</th>
<th>Watch-again/Don’t Watch-again</th>
</tr>
</thead>
<tbody>
<tr>
<td>AU + Mean-pooling</td>
<td>.66</td>
<td>.87</td>
</tr>
<tr>
<td>AU + max-pooling</td>
<td>.61</td>
<td>.89</td>
</tr>
<tr>
<td>HMM</td>
<td>.58</td>
<td>.84</td>
</tr>
<tr>
<td>Exemplar-HMMs</td>
<td><strong>.84</strong></td>
<td><strong>.92</strong></td>
</tr>
</tbody>
</table>
Devil in the details

- Bayesian HMMs avoid overfitting and lead to better results.
- Cross-validation necessary to select the kernel parameters.
- Gaussian assumption limits dimensionality.
  - To be extended for texture (high dim.) features.
Summary

• Explored approach for using HMMs within a discriminative framework for AFER.

• Exemplar-HMMs for temporal modeling
  – Temporal segmentation + model expression states
  – Model dynamics
  – Maintains specificity of each example

• PPK for model-based similarity
  – Comprehensively compares states from two HMM.
  – Takes into account temporal information.
Questions?

Karan Sikka  Abhinav Dhall  Dr. Marian S. Bartlett

Machine Perception Lab, UCSD

Thanks