Capturing Complex Spatio-Temporal Relations among Facial Muscles for Facial Expression Recognition

Ziheng Wang¹, Shangfei Wang², Qiang Ji¹
Rensselaer Polytechnic Institute¹ University of Science and Technology of China²

1. Problem
Facial Expression Recognition

2. Main Idea
Model Facial expression as a complex activity consisting of sequential or overlapping facial muscle events
Propose an Interval Temporal Bayesian Network (ITBN) to explicitly capture a larger variety of complex spatio-temporal relations among facial muscle events for expression recognition

3. Related Work
Existing approaches
- Time-sliced graphical models such as hidden Markov models (HMMs) and dynamic Bayesian networks (DBNs)
- Syntactic and description-based approaches
Limitations
- Can only model a sequence of instantaneously occurring events
- Only offer three time point relations: precedes, follows and equals
- Typically assume first order Markov property and local stationary transition.
- Syntactic and description-based models lack the expressive power to capture uncertainties

The proposed model
- Model both sequential and overlapping events
- Do not rely on local assumptions and capture global relations
- Capture a larger variety of complex relations
- Remains fully probabilistic

<table>
<thead>
<tr>
<th>Existing Dynamic Models</th>
<th>Proposed Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequential events</td>
<td>Sequential or overlapping events</td>
</tr>
<tr>
<td>Local stationary relations</td>
<td>Global relations</td>
</tr>
<tr>
<td>3 relations (precede, follow, equal)</td>
<td>13 complex relations</td>
</tr>
</tbody>
</table>

4. Interval Temporal Bayesian Network

5. Implementation of ITBN
We propose to implement ITBN with a corresponding Bayesian Network (BN) to exploit the well developed BN mathematical machinery

6. Modeling Facial Expression with ITBN

7. Learning and Inference
Step 1: Temporal Relation Node Selection
- It is not necessary to consider the relations among all possible pairs of events
- A selection routine is performed to select those that maximize discrimination
- The following score is used to rank all the relation nodes. The top L nodes are selected and instantiated in the model

\[
S(I_{mk}) = \sum_{j} \left[ D_o(P_j | P_k) + D_o(P_k | P_j) \right]
\]

Step 2: Structure Learning
- Temporal nodes are linked to their corresponding event nodes
- Spatial links are learned by finding a network C that maximizes the BIC score on the training data D

\[
\max \left\{ \log P(D|C) - \frac{1}{2} |C| \log N \right\}
\]

Step 3: Parameter Estimation
- Parameters include the conditional probability distribution (CPD) \( P(k|j) \) and the CPD \( P(j|k) \)
- Tree structured CPD is introduced to reduce the number of parameters
- Parameters are learned by maximizing the log likelihood

\[
\theta^* = \arg \max \log P(D|\theta)
\]

8. Experimental Results

9. Conclusions
- Model a facial expression as a complex activity of temporally sequential or overlapping facial events
- Propose a novel ITBN to capture global and a larger variety of complex spatio-temporal relations among facial events
- The proposed model achieves comparable and even better results than existing methods, without using appearance information
- ITBN can be widely applied for analyzing other complex activities