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

Ziheng Wang[1] Shangfei Wang[2] Qiang Ji[1]

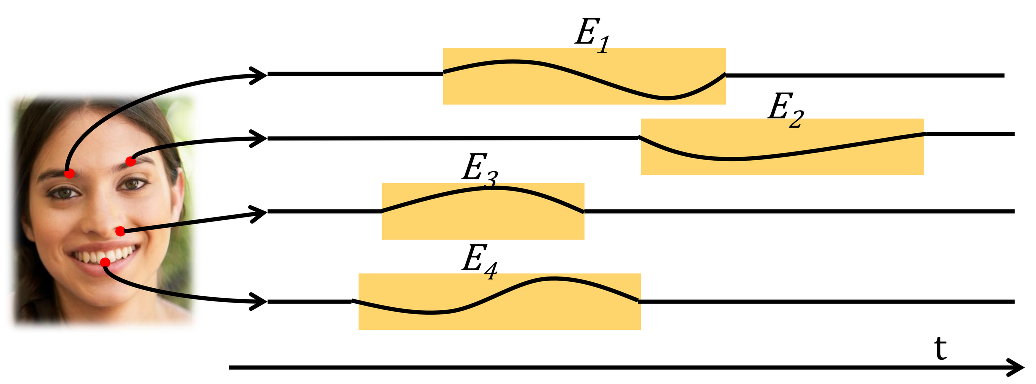
Rensselaer Polytechnic Institute[1] University of Science and Technology of China[2]

1. Problem

- Facial Expression Recognition

Happiness Surprise Disgust Sadness

2. Main Idea



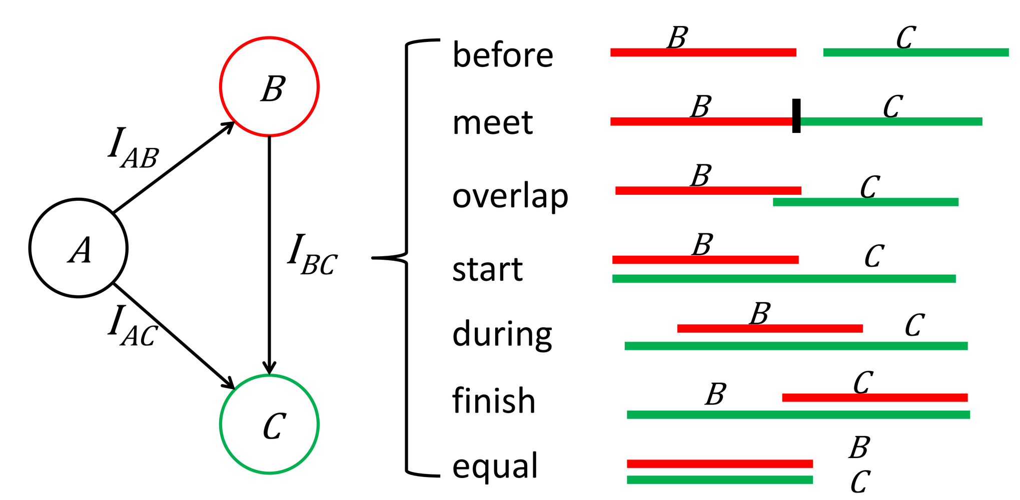
- Model Facial expression as a complex activity consisting of **sequential or overlapping** facial muscle events
- Propose an Interval Temporal Bayesian Network (ITBN) to explicitly capture **a larger variety of complex spatio-temporal relations** among facial muscle events for expression recognition

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 no reply on local assumptions and capture global relations
  - Capture a larger variety of complex relations
  - Remains fully probabilistic

Existing Dynamic Models	Proposed Model
Sequential events	**Sequential or overlapping events**
Local stationary relations	**Global relations**
3 relations (precede, follow, equal)	**13 complex relations**

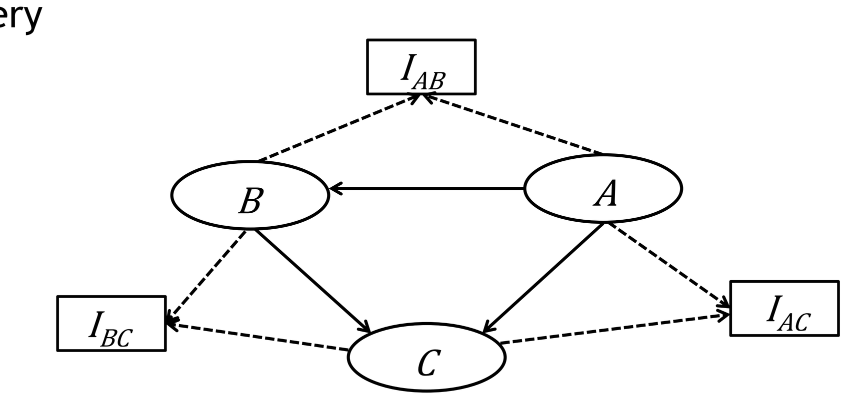
4. Interval Temporal Bayesian Network



- **Node: Temporal Entity (Event)**
  - A temporal entity is characterized by a pair $\langle \Sigma, \Omega \rangle$ where Σ is a set of all possible states for the temporal entity, and $\Omega = \{[a,b] \in \mathbb{R}: a < b\}$ is the period of time spanned by the temporal entity.
- **Link: Temporal Dependency**
  - A temporal dependency denoted as I_{XY} describes a temporal relation between two temporal entities $X = \langle \Sigma_X, \Omega_X \rangle$ and $Y = \langle \Sigma_Y, \Omega_Y \rangle$ with X as the reference.
  - The strength of I_{XY} is quantified by the conditional probability $P(I_{XY}|\Sigma_X, \Sigma_Y)$
  - 13 temporal dependencies are defined according to Allen's Interval Algebra

5. Implementation of ITBN

- We propose to implement ITBN with a corresponding Bayesian Network (BN) to exploit the well developed BN mathematical machinery

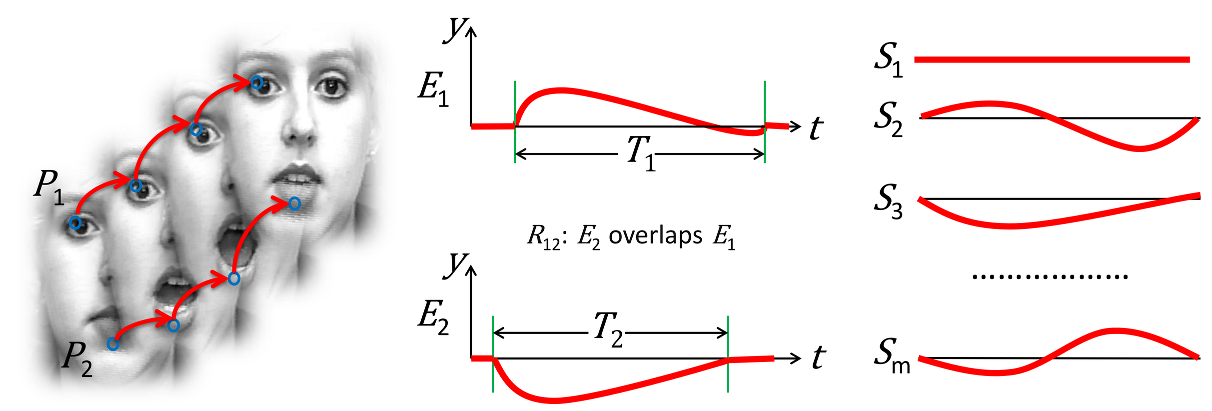


- *Circular node*: state of each temporal event
- *Square node*: type of temporal relation
- *Solid links*: capture the spatial dependencies among events
- *Dotted links*: connect the relation node with the corresponding temporal event nodes and capture the temporal dependencies

❖ Given a set of temporal events $\mathcal{E} = \{E_i\}_{i=1}^{n}$ and their pairwise relations $\mathcal{I} = \{I_k\}_{k=1}^{K}$, the joint distribution can be written as

$$P(\mathcal{E}, \mathcal{R}) = \underbrace{\prod_{i=1}^{n} P(E_i|\pi(E_i))}_{\text{Spatial Relations}} \underbrace{\prod_{k=1}^{K} P(I_k|\pi(I_k))}_{\text{Temporal Relations}}$$

6. Modeling Facial Expression with ITBN



- **Primitive Facial Muscle Event**
  - **Definition**: each primitive event is the movement of one facial feature point
  - **Duration**: the interval from the time the point starts to move till the time it finally comes back
  - **State**: one of the moving patterns
  - **Temporal relation**: based on the durations of the events, their temporal relations can be uniquely determined
- **Facial Expression Recognition with ITBN**
  - To recognize N expressions, we build N ITBN's $\{M_y: y = 1, \ldots, N\}$, each of which models one expression.
  - Given a query sample x, its expression is classified as:

$$y^* = \arg\max_y \frac{P(x|M_y)}{Complexity(M_y)}$$

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

$$S(I_{AB}) = \sum_{i>j} [D_{KL}(P_i \parallel P_j) + D_{KL}(P_j \parallel P_i)]$$

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

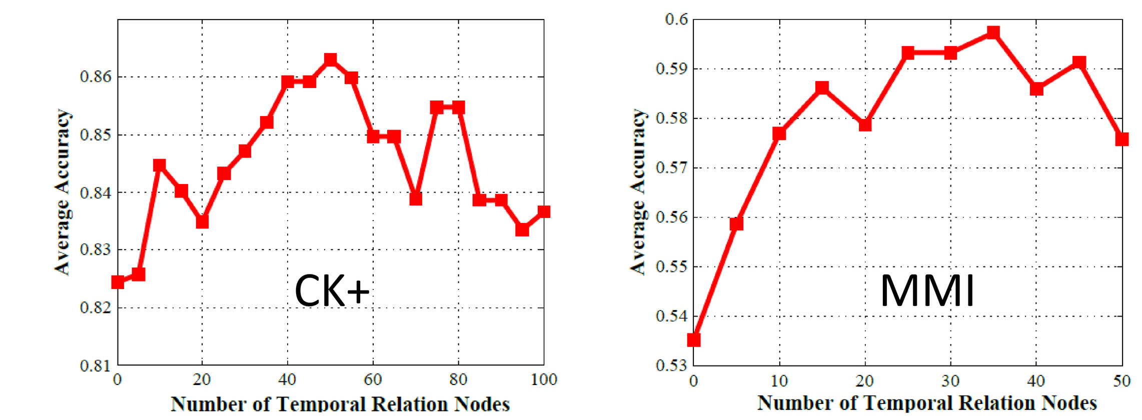
$$\max_G \left[\log P(D|G,\Theta) - \frac{|\Theta| \log N}{2}\right]$$

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

$$\Theta^* = \arg\max_\Theta \log P(D|\Theta)$$

8. Experimental Results

- **Data Sets**
  - CK+: 7 expressions, 327 video sequences
  - MMI: 6 expressions, 205 video sequences
- **Performance Vs Number of Relation Nodes**

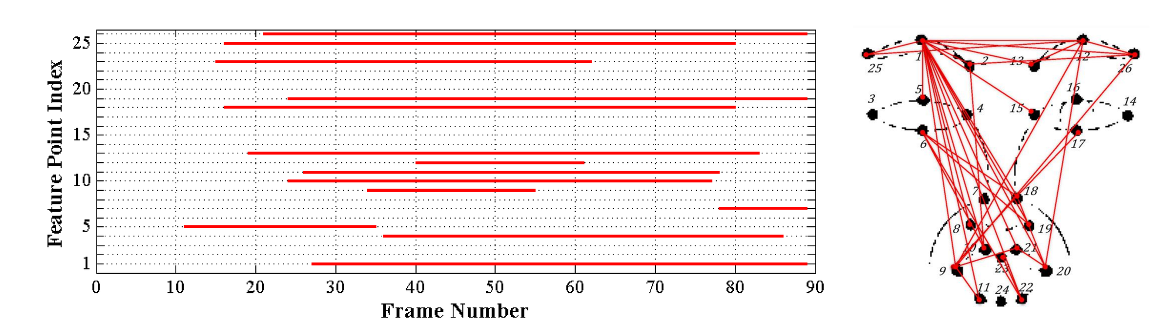


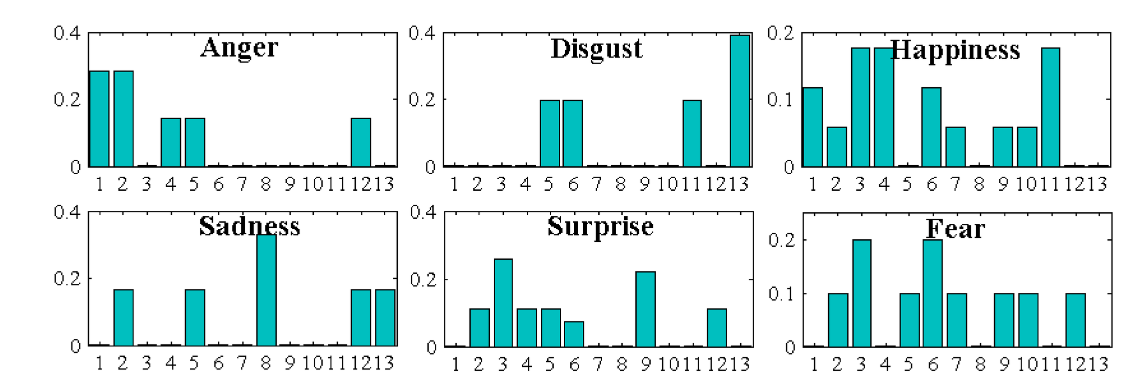
- **Performance Vs Related works**

	ITBN	HMM	Lucey *et al.*
CK+	86.3%	83.5%	83.3%

	ITBN	HMM	Zhong *et al.*		
			CPL	CSPL	ADL
MMI	59.7%	51.5%	49.4%	73.5%	47.8%

  - ITBN outperforms time-slice based models such as HMM
  - ITBN achieves comparable and even better performance than the related works, even though it is based purely on the tracking results
- **Relationship Analysis**





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