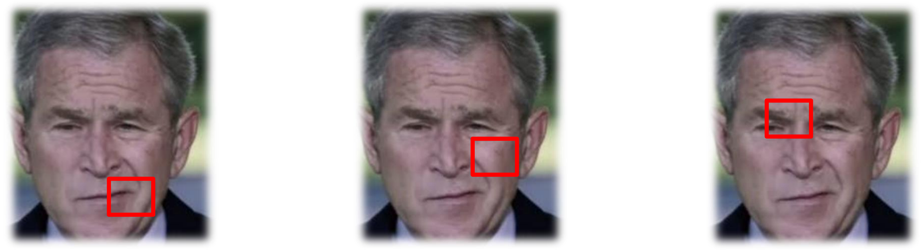
Capturing Global Semantic Relationships for Facial Action Unit Recognition

Ziheng Wang[1] Yongqiang Li[3] Shangfei Wang[2] Qiang Ji[1]

Rensselaer Polytechnic Institute[1] University of Science and Technology of China[2] Harbin Institute of Technology[3]

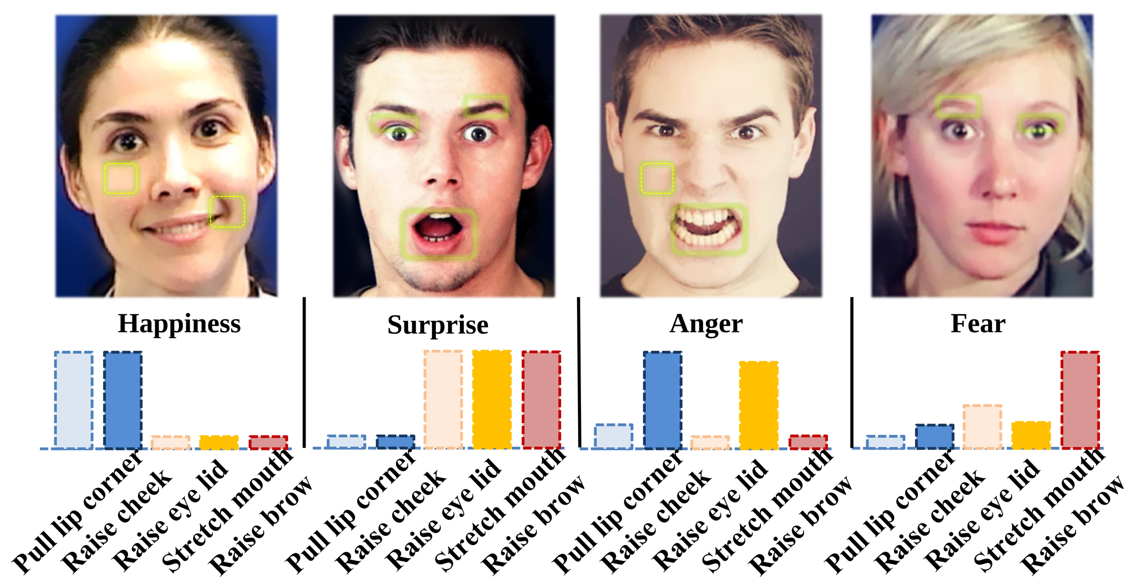
1. Problem

- Facial Action Unit Recognition



Pull Lip Corner Raise Cheek Lower Eyebrow

2. Main Idea

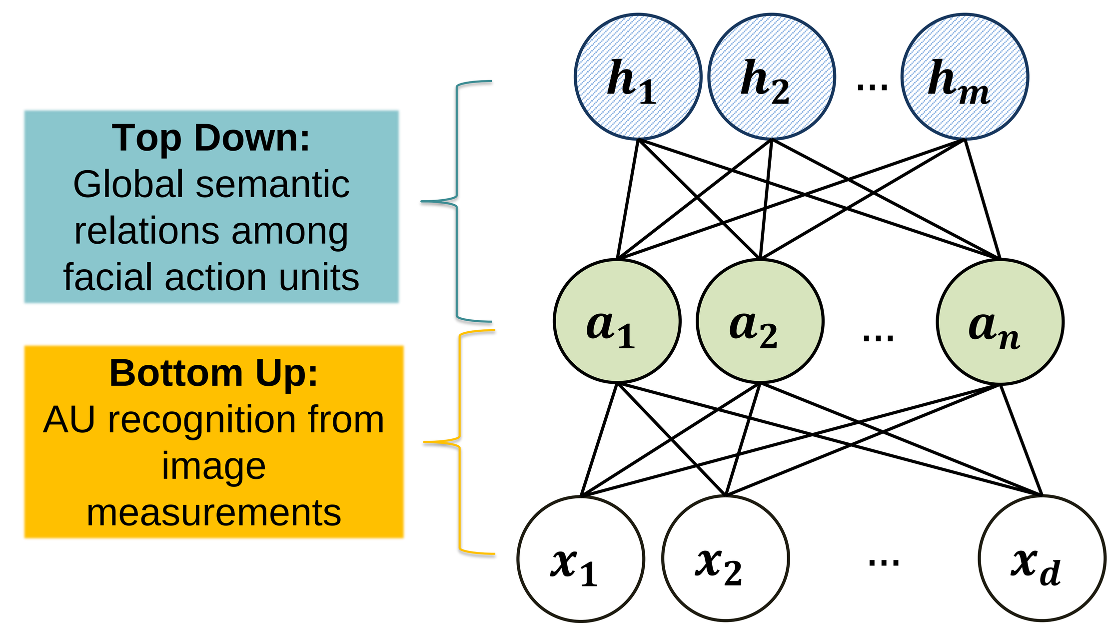


- Facial action units are NOT independent
  - AUs do not occur alone and some combinations of action units are frequently observed
  - Some AUs must or must not be present at the same time due to the limitations of facial anatomy
- Relationships among AUs are influenced by facial expression
  - "Stretch mouth" and "raise brow" are likely to be both absent during happiness, both present during surprise, and mutually exclusive during anger or fear
- Propose to use restricted Boltzmann machine (RBM) to capture global complex relationships among AUs for AU recognition
- Propose to use 3-way RBM to capture facial expression to more accurately characterize AU relationships

3. Related Work

- **Existing approaches**
  - Treat AUs are uncorrelated entities
  - Use Bayesian network (BN) to model AU relationships
- **Limitations**
  - Models such as BN are based on the first-order Markov assumption and therefore can only capture local, i.e. *pairwise* relationships between action units
  - Finding the optimal structure of a large AU network is difficult
  - Modeling AU relationships without considering the influence of facial expression, which could lead to incorrect estimation of AU dependencies
- **Proposed Method**
  - Model *global* AU relationships and consider the influence of *expression*

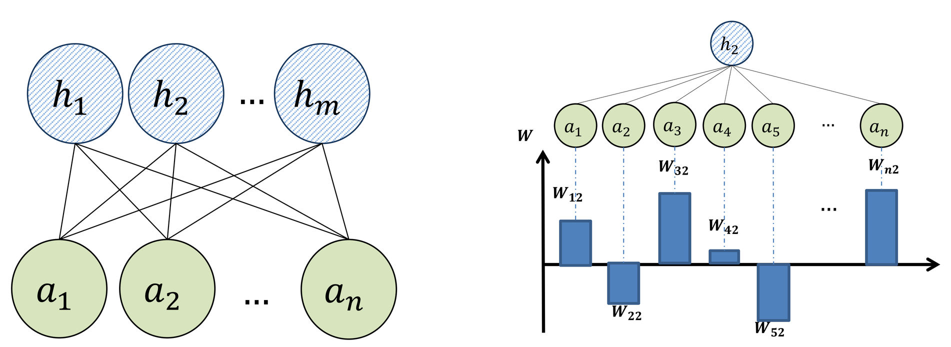
4. Hierarchical Model for AU Recognition



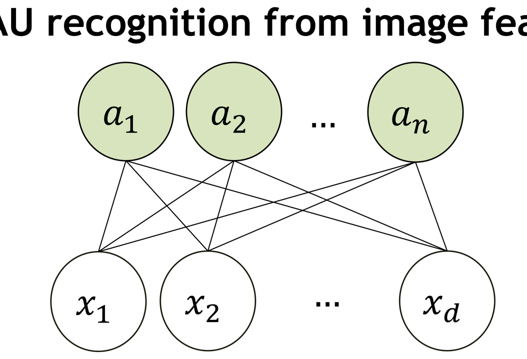
- **Middle layer a_1 to a_n:** binary state of AU_1 to AU_n
- **Bottom layer x_1 to x_d:** image features
- **Top layer h_1 to h_m:** latent nodes modeling AU relationships
- **Total Energy:**

$$E(\boldsymbol{x},\boldsymbol{a},\boldsymbol{h};\theta) = -\sum_i\sum_j a_i W_{ij}^1 h_j - \sum_j c_j h_j - \sum_i b_i a_i - \sum_i\sum_t W_{it}^2 a_i x_t$$

  - $a_i W_{ij}^1 h_j$: compatibility between AU_i and latent node h_j
  - $c_j h_j$: bias for each latent node h_j
  - $b_i a_i$: bias for each AU node a_i
  - $W_{it}^2 a_i x_t$: compatibility between AU_i and feature x_t
- **Top down - capturing global relations among AUs**



  - Each latent node is connected to all AU nodes and therefore modeling their higher-order relationships
  - The captured AU relationships can be implicitly inferred from the model parameters W_{ij}^1
  - Vector $[w_{im}]_{i=1}^n$ captures a specific presence and absence pattern of all the action units
  - W_{im}^1 large → AU_i more likely to occur in pattern m
  - W_{im}^1 small → AU_i less likely to occur in pattern m
- **Bottom up: AU recognition from image features**



  - Each AU node is connected to all the image features with the energy $E(a_i, x) = -\sum_t W_{it}^2 a_i x_t$
  - Equivalent to a set of linear AU classification models

5. Learning and Inference

- **Discriminative Learning**
  - Given training data $\{(\boldsymbol{x}_i, \boldsymbol{a}_i)\}_{i=1}^N$

$$\boldsymbol{\theta}^* = \arg\max_{\theta} \mathcal{L}(\boldsymbol{\theta}) = \arg\max \frac{1}{N}\sum_{i=1}^N \log P(\boldsymbol{a}_i|\boldsymbol{x}_i;\boldsymbol{\theta})$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left\langle \frac{\partial E}{\partial \boldsymbol{\theta}} \right\rangle_{P(\boldsymbol{h}|\boldsymbol{a},\boldsymbol{x};\boldsymbol{\theta})} - \left\langle \frac{\partial E}{\partial \boldsymbol{\theta}} \right\rangle_{P(\boldsymbol{h},\boldsymbol{a}|\boldsymbol{x};\boldsymbol{\theta})}$$

  - Calculating the gradient requires $P(\boldsymbol{h}|\boldsymbol{a},\boldsymbol{x};\boldsymbol{\theta})$ and $P(\boldsymbol{h},\boldsymbol{a}|\boldsymbol{x};\boldsymbol{\theta})$
  - $P(\boldsymbol{h}|\boldsymbol{a},\boldsymbol{x};\boldsymbol{\theta})$ can be analytically computed

$$P(h_j|\boldsymbol{a},\boldsymbol{x};\boldsymbol{\theta}) = P(h_j|\boldsymbol{a};\boldsymbol{\theta}) = \sigma\left(-c_j - \sum_i W_{ij}^1 a_i\right) \quad (1)$$

  - $P(\boldsymbol{h},\boldsymbol{a}|\boldsymbol{x};\boldsymbol{\theta})$ is intractable, we revised contrastive divergence (CD) algorithm to compute it
  - The basic idea is to approximate $P(\boldsymbol{h},\boldsymbol{a}|\boldsymbol{x};\boldsymbol{\theta})$ by sampling h with Equation 1 and then sampling a with Equation 2

$$P(a_i|\boldsymbol{h},\boldsymbol{x};\theta) = \sigma\left(-b_i - \sum_j W_{ij}^1 h_j - \sum_t W_{it}^2 x_t\right) \quad (2)$$

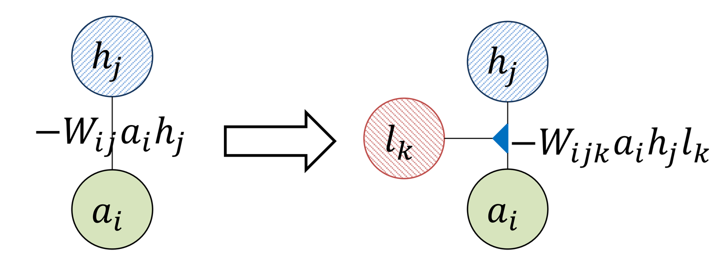
- **Inference**
  - Given a query sample $\boldsymbol{x}$, each action unit can be inferred by

$$a_i^* = \arg\max_{a_i} P(a_i|\boldsymbol{x})$$

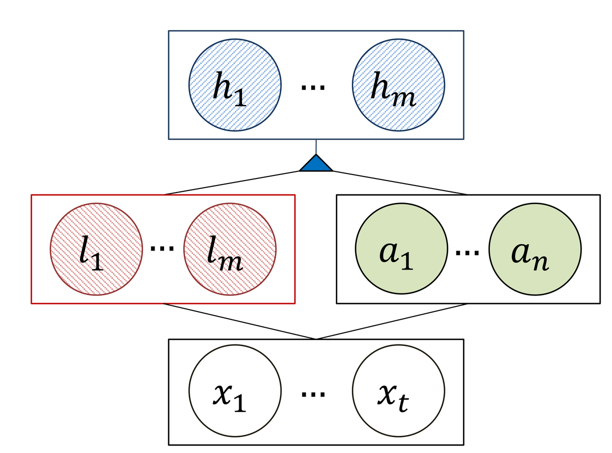
  - Can be efficiently performed with Gibbs sampling by iteratively sampling $\boldsymbol{h}$ from $P(\boldsymbol{h}|\boldsymbol{x},\boldsymbol{a})$ and sampling a from $P(\boldsymbol{a}|\boldsymbol{h},\boldsymbol{x})$

6. Incorporating Expression to Model AU Relations

- **Relations among the AUs depend on facial expressions**
  - Expression is known during training but known during testing
  - Basic idea: modulate the connection between each pair of action unit and latent unit (a_i, h_j) with an expression variable l



  - Model



$$E(\boldsymbol{x},\boldsymbol{a},\boldsymbol{l},\boldsymbol{h};\theta) = -\sum_i\sum_j\sum_k W_{ijk}^1 a_i h_j l_k - \sum_i\sum_t W_{it}^2 a_i x_t - \sum_k\sum_t W_{kt}^3 l_k x_t - \sum_j c_j h_j - \sum_i b_i a_i - \sum_k d_k l_k$$

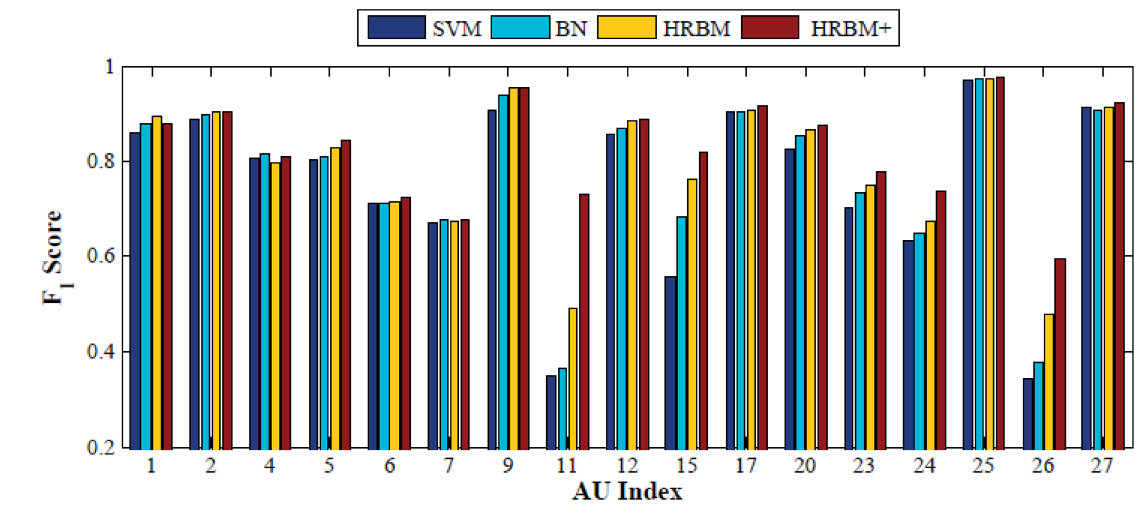
$$\boldsymbol{\theta}^* = \arg\max \sum_{i=1}^N \log P(\boldsymbol{a}_i, \boldsymbol{l}_i|\boldsymbol{x}_i;\boldsymbol{\theta})$$

$$a_i^* = \arg\max P(a_i|\boldsymbol{x}) = \arg\max \sum_{\boldsymbol{l}} P(a_i, \boldsymbol{l}|\boldsymbol{x})$$

7. Experimental Results

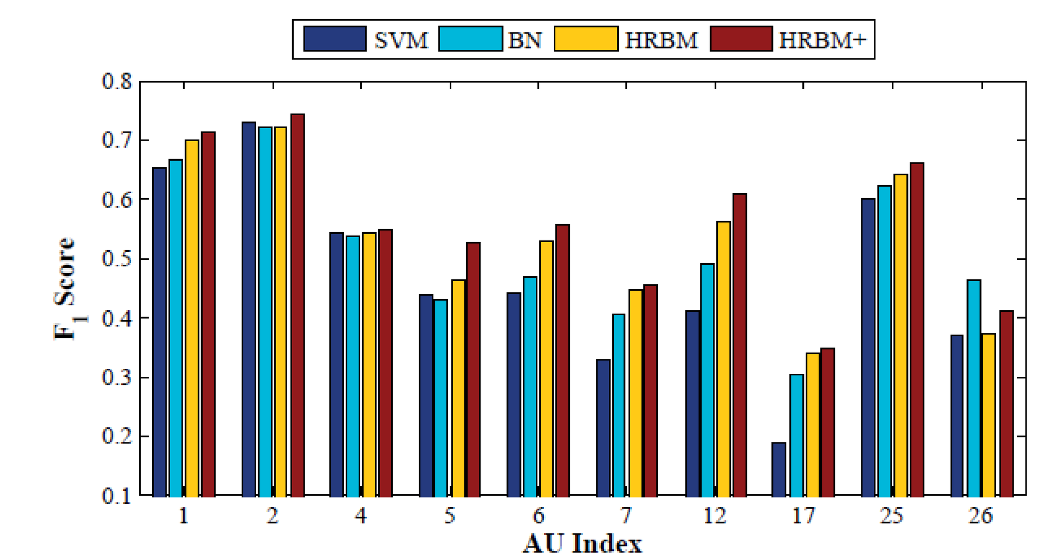
- **Methods**
  - SVM, BN [Tong *et al.* 2007], HRBM, HRBM+
- **Posed Facial Action Units Recognition**
  - CK+: 593 peak images, 17 action units

Method	SVM	BN	HRBM	HRBM+
Average F_1-score	74.70%	76.70%	**79.21%**	**82.44%**

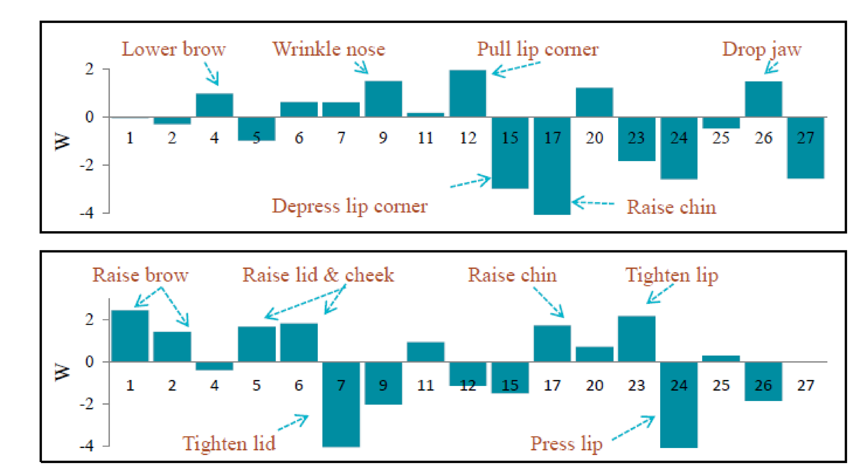


- **Non-posed Facial Action Unit Recognition**
  - SEMAIN: 180 image frames, 10 action units
  - Train on CK+, test on SEMAIN

Method	SVM	BN	HRBM	HRBM+
Average F_1-score	47.70%	51.09%	**54.76%**	**56.14%**



- **Semantic Relationship Analysis**



8. Conclusions

- Proposed a hierarchical model for AU recognition
- Capture higher-order AU interactions
- Consider the influence of facial expression on AU relations
- Experimental results demonstrate the effectiveness of the proposed approach