Knowledge Augmented Deep Neural Networks for Joint Facial Expression and Action Unit Recognition

Zijun Cui (cuiz3@rpi.edu)
Rensselaer Polytechnic Institute

Tengfei Song
Southeast University

Yuru Wang
Northeast Normal University

Qiang Ji (jiq@rpi.edu)
Rensselaer Polytechnic Institute

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

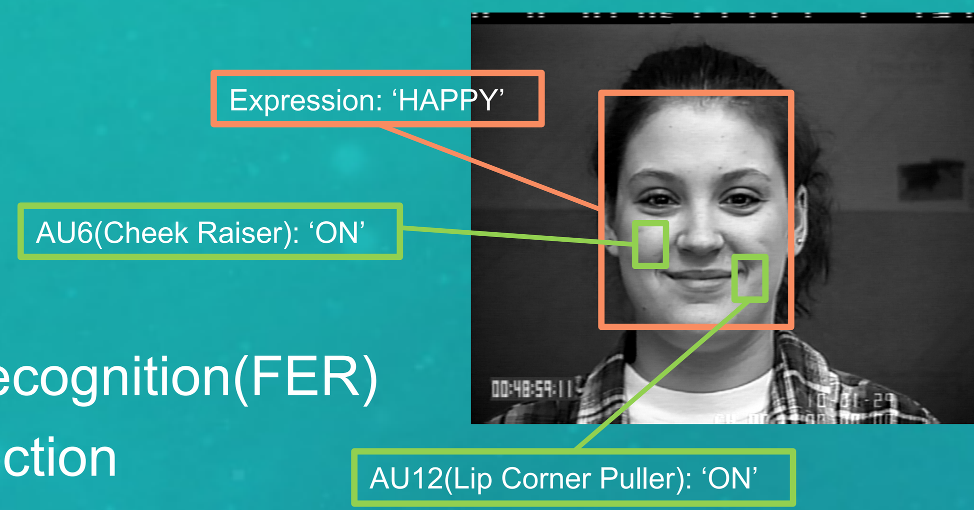

Figure: An example from CK+ dataset[1]

❑ Tasks:

- Facial Expression Recognition(FER)
- Action Unit(AU) Detection

❑ Motivations:

- Facial expression and AUs are strongly correlated
- Generic knowledge on expression-AUs relationships is available

❑ Contributions:

- A *knowledge model* encoding the generic knowledge systematically
- A deep learning framework for *joint* facial expression and AU recognition

Generic Knowledge as Probabilities

-- on expression-AUs probabilistic relationships

❑ Notation:

- Expression $X^e = \{1,2,\dots,E\}$
  E is the total number of expressions
- AUs $X_m^{au} = \{X_1^{au}, X_2^{au}, \dots, X_M^{au}\}$
  M is the total number of AUs and $X_m^{au} = \{0,1\}$

❑ *Expression-dependent single AU probabilities*

○ `AU4` is a primary AU given `Anger` expression

$$p(X_4^{au} = 1 | X^e = \text{Anger}) > p(X_4^{au} = 0 | X^e = \text{Anger})$$

❑ *Expression-dependent joint AU probabilities*

○ `AU6` and `AU12` are positively correlated given `Happy` expression

$$p(X_6^{au} = 1 | X_{12}^{au} = 1, X^e = \text{Happy}) > p(X_6^{au} = 0 | X_{12}^{au} = 1, X^e = \text{Happy})$$
$$p(X_6^{au} = 1 | X_{12}^{au} = 1, X^e = \text{Happy}) > p(X_6^{au} = 1 | X_{12}^{au} = 0, X^e = \text{Happy})$$

❑ *Expression-independent joint AU probabilities*

○ `AU1` and `AU2` are positively correlated

$$p(X_1^{au} = 1 | X_2^{au} = 1) > p(X_1^{au} = 0 | X_2^{au} = 1)$$
$$p(X_1^{au} = 1 | X_2^{au} = 1) > p(X_1^{au} = 1 | X_2^{au} = 0)$$

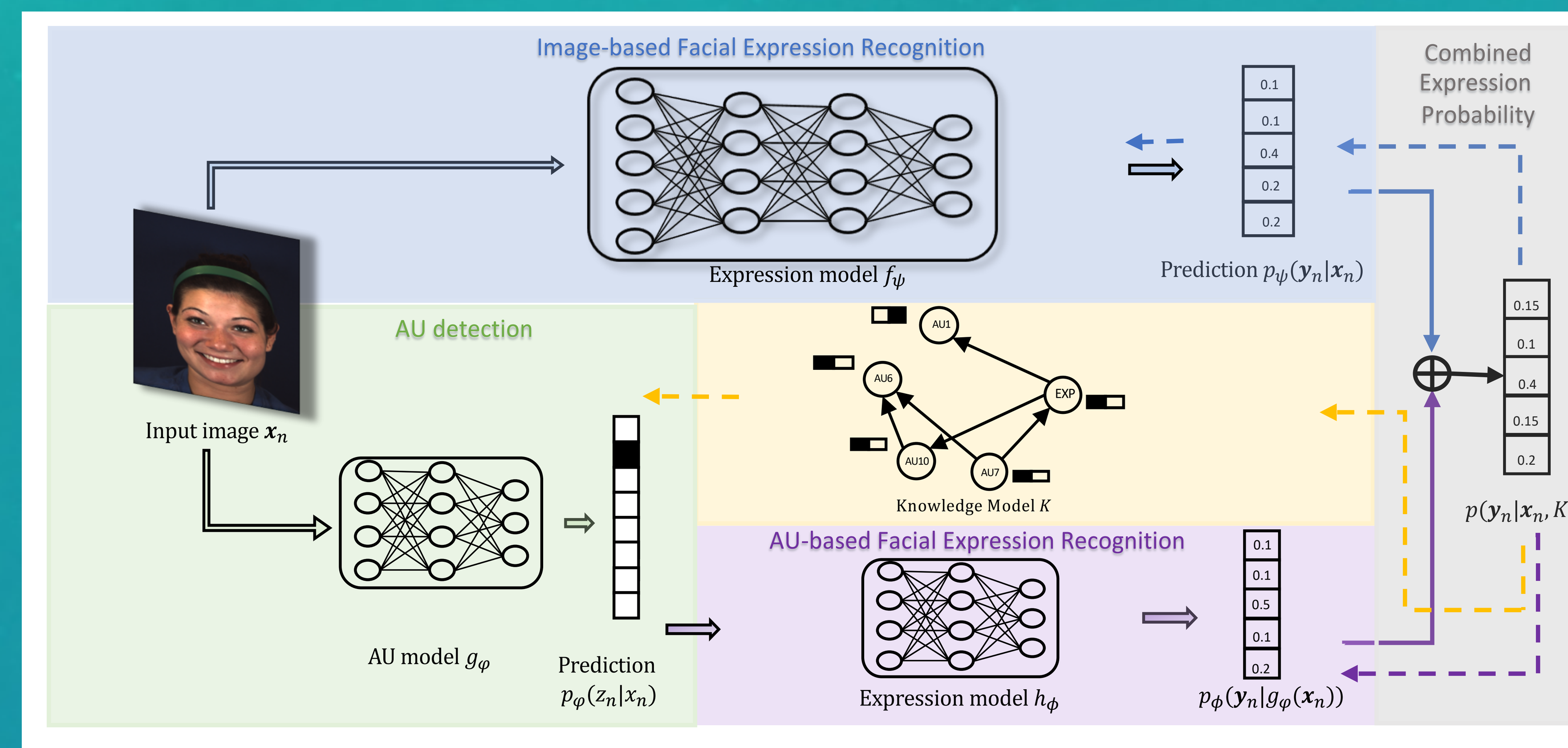

Figure: Overview of the proposed framework

Encoding of the Generic Knowledge

-- Bayesian Network(BN) Learning with Probability Constraints

❑ Definitions of the Bayesian Network

- Conditional probabilities are parameterized with the regression equations
  $$p(X_i = k | \pi(X_i)) = \sigma_M\left(\sum_{j=1}^{J} w_{ijk}\pi_j(X_i) + b_{ik}\right)$$
  where weights $\boldsymbol{w} = \{w_{ijk}\}$ and bias $\boldsymbol{b} = \{b_{ik}\}$ are to be learned
- $A(\boldsymbol{w})$ is the weighted adjacency matrix defining the structure[2]: $A_{ij} = \sum_{k=1}^{K} ||w_{ijk}||_2^2$
- The constraint of Directed Acyclic Graph(DAG)[3]: $\mathrm{tr}\left(e^{A(\boldsymbol{w})\circ A(\boldsymbol{w})}\right) - N = 0$

❑ Probability constraints derived from the generic knowledge

- Strictly inequality constraints: $\{g_i(\boldsymbol{w}, \boldsymbol{b}) < 0\}_{i=1}^{G}$
  To better handle g_i, we define positive margin with additional variable s_i
  The strictly inequality constraints become equality constraints:
  $g_i(\boldsymbol{w}, \boldsymbol{b}) + e^{s_i} = 0, i = 1, \dots, G$
- Inequality constraints: $\{l_j(\boldsymbol{w}, \boldsymbol{b}) \le 0\}_{j=1}^{L}$
- Equality constraints: $\{h_k(\boldsymbol{w}, \boldsymbol{b}) = 0\}_{k=1}^{H}$
- For example:
  $$g_i(\boldsymbol{w}, \boldsymbol{b}) = p(X_1^{au} = 0 | X_2^{au} = 1; \boldsymbol{w}, \boldsymbol{b}) - p(X_1^{au} = 1 | X_2^{au} = 1; \boldsymbol{w}, \boldsymbol{b}) < 0$$

❑ A penalty function $f(\boldsymbol{w}, \boldsymbol{b}; \boldsymbol{s})$ measures the violation of constraint

$$f(\boldsymbol{w}, \boldsymbol{b}; \boldsymbol{s}) = \frac{1}{G}\sum_{i=1}^{G} \log((g_i(\boldsymbol{w}, \boldsymbol{b}) + e^{s_i})^2 + 1)$$
$$+ \frac{1}{L}\sum_{j=1}^{L} \log\left(\left(l_j^+(\boldsymbol{w}, \boldsymbol{b})\right)^2 + 1\right) + \frac{1}{K}\sum_{k=1}^{K} \log\left(\left(h_i(\boldsymbol{w}, \boldsymbol{b})\right)^2 + 1\right)$$

with weights $\boldsymbol{w}$, bias $\boldsymbol{b}$ and current margins $e^{\boldsymbol{s}}$. And $l_j^+ = \max\{0, l_j\}$

- $f(\boldsymbol{w}, \boldsymbol{b}; \boldsymbol{s}) = \boldsymbol{0}$ if and only if all the constraints are satisfied

❑ A *Constrained Optimization Approach* for BN learning

$$\boldsymbol{w}^*, \boldsymbol{b}^*, \boldsymbol{s}^* = \arg\min_{\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{s}} f(\boldsymbol{w}, \boldsymbol{b}; \boldsymbol{s}) + \gamma||\boldsymbol{w}||_1 - \mu||\boldsymbol{s}||_2^2$$
$$\text{s.t.}\ \mathrm{tr}\left(e^{A(\boldsymbol{w})\circ A(\boldsymbol{w})}\right) - N = 0$$

where $||\boldsymbol{w}||_1$ penalizes the density of the structure, and $||\boldsymbol{s}||_2^2$ encourages the bigger positive margins

❑ The learned Bayesian Network serves as our knowledge model K

AU detection model and FER models

❑ We learn AU detection model and FER models with:

- The training images $\boldsymbol{x}_n, n = 1, \dots, N$
- The GT expression labels $y_n^{GT}, n = 1, \dots, N$
- The knowledge model K

* N is the total number of training samples

❑ Phase 1: Initialization of AU detection and FER models

➢ Weakly supervised AU detection model g_φ

$$\varphi^* = \operatorname{argmin}_\varphi \frac{1}{N}\sum_{n=1}^{N} E_{p(z_n|y_n^{GT}, K)}\, l(z_n, g_\varphi(\boldsymbol{x}_n))$$

$p(z_n|y_n^{GT}, K)$ is the probability of AU configuration z_n computed from the BN model and the y_n^{GT}

➢ Facial Expression Recognition(FER) Models

- Image-based FER model f_ψ: $\psi^* = \operatorname{argmin}_\psi \frac{1}{N}\sum_{n=1}^{N} l(y_n^{GT}, f_\psi(\boldsymbol{x}_n))$
- AU-based FER model h_ϕ: $\phi^* = \operatorname{argmin}_\phi \frac{1}{N}\sum_{n=1}^{N} l\left(y_n^{GT}, h_\phi\left(g_\varphi(\boldsymbol{x}_n)\right)\right)$

  where $g_\varphi(\boldsymbol{x}_n)$ is the output of the AU model g_φ

* l is the cross-entropy loss

❑ Phase 2: Integration among AU and Expression Models

➢ The combined expression probability

$$p(\boldsymbol{y}_n|\boldsymbol{x}_n, K) = w_1 p_\psi(\boldsymbol{y}_n|\boldsymbol{x}_n) + w_2 p_\phi(\boldsymbol{y}_n|g_\varphi(\boldsymbol{x}_n), K)$$

$p_\psi(\boldsymbol{y}_n|\boldsymbol{x}_n)$ is the output of f_ψ and p_ϕ is the output of h_ϕ. w_1, w_2 are the weights

➢ Expression-augmented AU detection model

$$\varphi^* = \operatorname{argmin}_\varphi \frac{1}{N}\sum_{n=1}^{N} E_{p(z_n|y_n^{GT}, K)}\, l\left(z_n, g_\varphi(\boldsymbol{x}_n)\right) + \lambda_1 E_{p(\boldsymbol{y}_n|\boldsymbol{x}_n, K)} E_{p(z_n|\boldsymbol{y}_n, K)}\, l\left(z_n, g_\varphi(\boldsymbol{x}_n)\right)$$

➢ Knowledge-augmented image-based FER model

$$\psi^* = \operatorname{argmin}_\psi \frac{1}{N}\sum_{n=1}^{N} l(y_n^{GT}, f_\psi(\boldsymbol{x}_n)) + \lambda_2 E_{p(\boldsymbol{y}_n|\boldsymbol{x}_n, K)}\, l\left(\boldsymbol{y}_n, f_\psi(\boldsymbol{x}_n)\right)$$

* l is the cross-entropy loss. λ_1, λ_2 are the hyper-parameters to be tuned

Experiments

-- comparisons with state-of-the-art models

- Action Unit Detection

Table 6: Comparison to the SoAs on AU detection.

Supervision	Method	BP4D	CK+	MMI
Supervised	HRBM[47]	.67	.79	.56
	MC-LVM[8]	-	**.80***	-
	JPML[56]	**.68***	.78*	-
	AU R-CNN[30]	.63*	-	-
Weakly-supervised	HTL[40]	.50	.66	.42
	LP-SM[54]	.55	.72*	.50
	TCAE[22]	.56*	-	-
	AUD-BN(baseline)	.56	.69	.47
	AUD-EA(gBN)	**.57**	**.74**	**.58**

- Facial Expression Recognition(FER)

Table 8: Comparison with SoA FER methods

Methods	BP4D	CK+	MMI	EmotioNet
STM-Explet[27]	-	94.19*	75.12*	-
DTAGN(Joint)[12]	-	97.25*	70.24*	-
DeRL[50]	-	97.30*	73.23*	-
ILCNN[3]	-	94.35*	70.67*	-
DAM-CNN[49]	-	95.88*	-	-
FMPN-FER[4]	60.16	96.53	82.74*	84.88
DeepEmotion[32]	79.54	95.23	72.66	81.51
FER-I(baseline)	61.68	94.29	67.35	80.85
FER-IK(gBN)	**83.82**	**97.59**	**84.90**	**95.55**

[1] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanadedataset (ck+): A complete dataset for action unit and emotion-specified expression. CVPR 2010
[2] Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing Learning sparse nonparametricdags. InInternational Conference on Artificial Intelligence and Statistics, 2020
[3] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing Dags with no tears: Continuousoptimization for structure learning. InAdvances in Neural Information Processing Systems, 2018.