From Hard to Soft: Towards more Human-like Emotion Recognition by Modelling the Perception Uncertainty

Jing Han[1], Zixing Zhang[2], Maximilian Schmitt[1], Maja Pantic[2] and Björn Schuller[1,2]

[1] Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2] Department of Computing, Imperial College London, UK

Introduction

- **Motivations:**
  - Automatic Emotion Recognition (AER) is highly **subjective**, which differs from many other pattern recognition tasks that have a ground truth
  - conventional methods – 'hard' prediction: emotion prediction = emotional state, i.e., a unique category or value is provided for AER
- **Major Contributions:**
  - Propose a 'soft' prediction strategy: emotion prediction = emotional state + **perception uncertainty**

Human-like AER

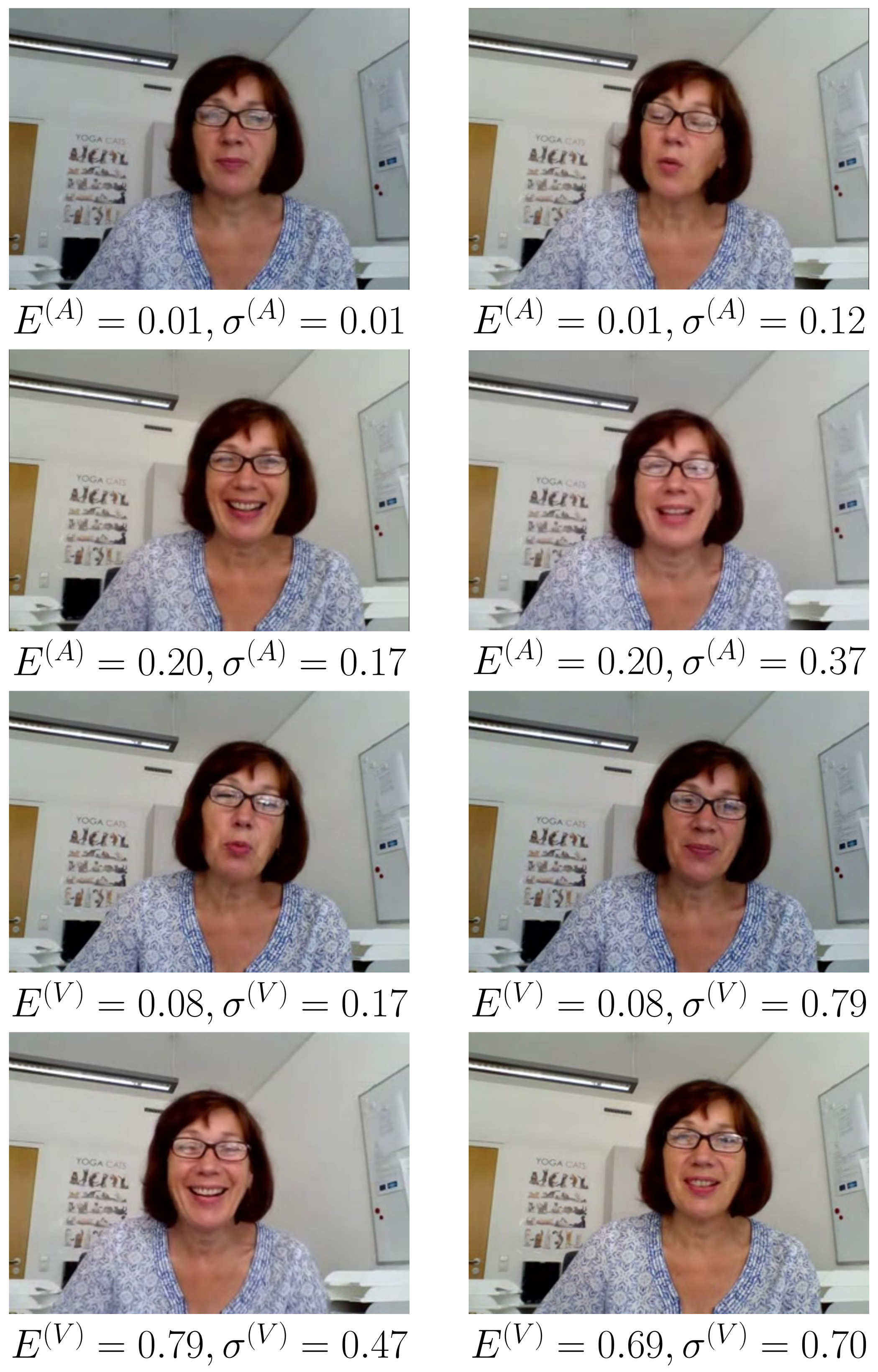
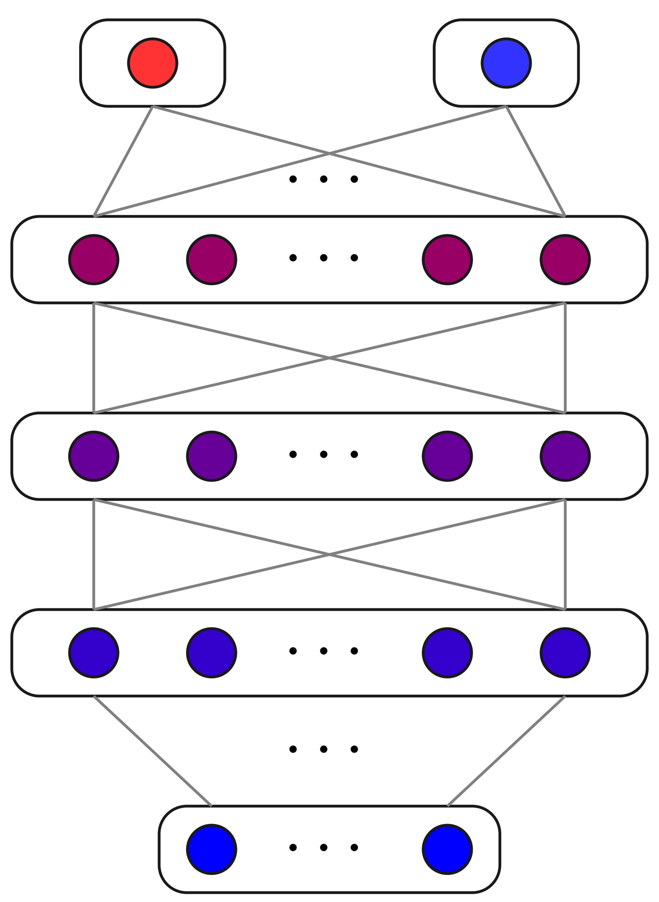


Figure: Four pairs of frames with comparable emotional states (E) but distinct perception uncertainties (σ) in (A)rousal and (V)alence, respectively

AER with Perception Uncertainty

emotional state: E perception uncertainty: σ



feature vector: $\mathbf{x}$

- input: audio/visual feature vectors $\mathbf{x}$
- outputs:
  $(E^{(A)}, \sigma^{(A)})$ for arousal
  $(E^{(V)}, \sigma^{(V)})$ for valence
- E is calculated by EWE over all raters by given instance n:

$$E_n^{(i)} = \frac{1}{\sum_{k=1}^{K} r_k^{(i)}} \sum_{k=1}^{K} r_k^{(i)} e_{n,k}^{(i)},$$

  where $r_k^{(i)}$ is a rater-dependent weight for rater k
- σ is calculated by inter-rater disagreement level:

$$\sigma_n^{(i)} = \sqrt{\frac{1}{K-1} \sum_{k=1}^{K} (e_{n,k}^{(i)} - e_n^{\mathrm{MLE},(i)})^2}.$$

- loss in multi-task learning:

$$\mathcal{J}(\theta) = w_E \cdot MSE_E + w_\sigma \cdot MSE_\sigma,$$

  with $w_E + w_\sigma = 2$
- audiovisual late fusion: $y = \epsilon + \sum \gamma_i \cdot y_i$

Experiments and Results

- features:
  - audio: mean and variance of 65 LLDs from ComParE13
  - video: 49-point facial landmark locations
- on-line standardisation
- annotation delay compensation: 4s
- network architecture:
  - BLSTM-RNN with two hidden layers
  - 240 LSTM cells per layer
- hyper-parameter and post-processing parameters are optimised based on the *development* set

Table: Concordance Correlation Coefficient (CCC) of the soft predictions

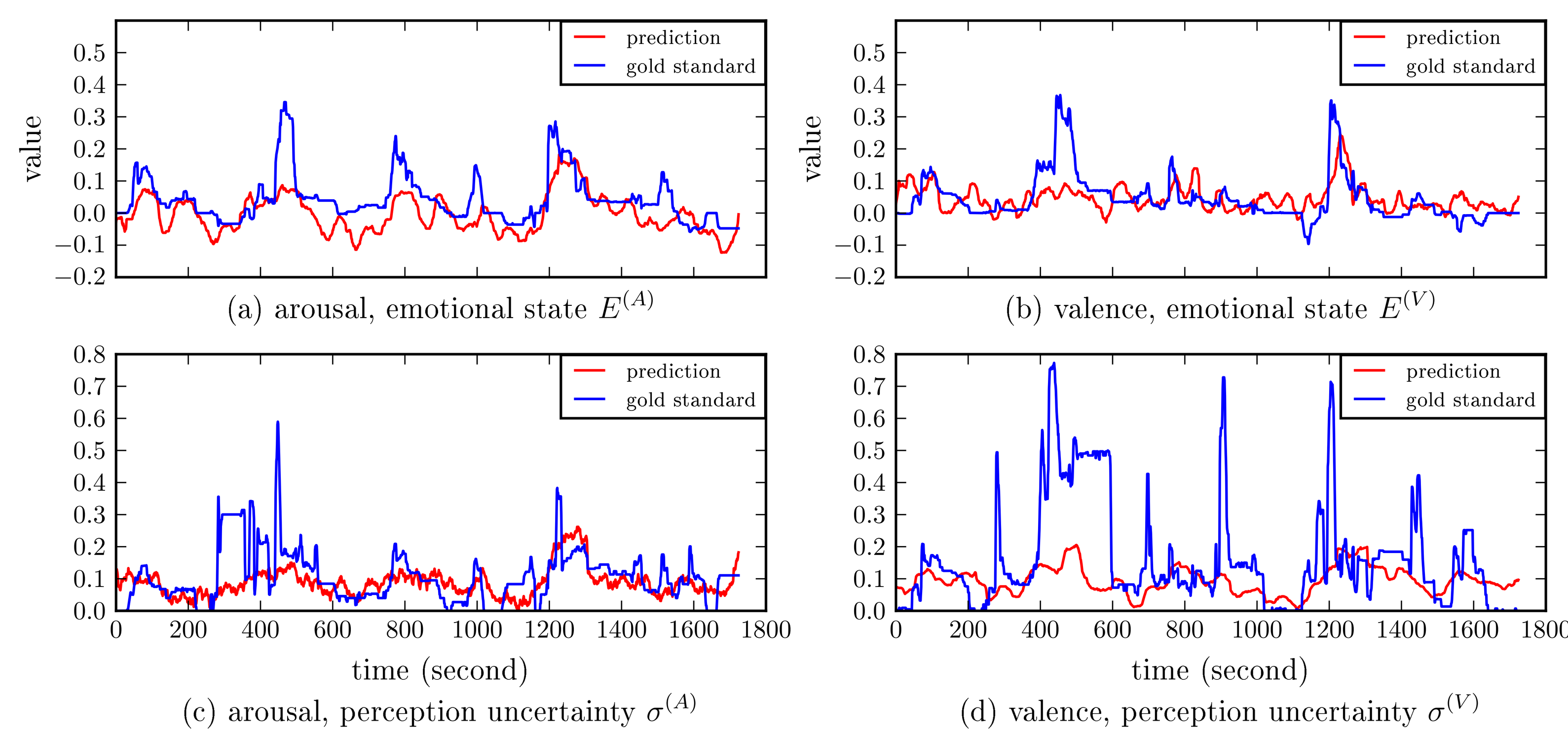
CCC		*dev.*		*test*		*dev.*		*test*	
modality	*task*	$E^{(A)}$	$\sigma^{(A)}$	$E^{(A)}$	$\sigma^{(A)}$	$E^{(V)}$	$\sigma^{(V)}$	$E^{(V)}$	$\sigma^{(V)}$
audio	*single*	.281	.103	.234	.185	.298	.075	.267	.015
	multi	.356	.181	.275	**.246**	.396	.180	.292	.089
video	*single*	.386	.204	.295	.171	.456	.266	.402	.120
	multi	.477	**.276**	.373	.167	**.588**	**.317**	.505	**.153**
audio+video	*single*	.505	.195	.386	.193	.502	.261	.478	.111
	multi	**.559**	.273	**.450**	.200	.575	.235	**.515**	.110

Dataset

SEWA German Video-chat Database:

- # pairs of *spontaneous* chats: 32 (# audio-visual recordings: 64)
- # frames in train/ valid/ test sets: 55 072/ 22 307/ 27 597
- # raters for arousal and valence: 6

Performance Illustration



Conclusion

- provide two indicators for AER, i.e., the perception uncertainty together with the emotional state
- soft prediction with multi-task learning performs better
- performance is further enhanced when combining audio and video information
- future work:
  - evaluate on more emotion datasets
  - address other subjective tasks
  - consider other deep learning frameworks

Acknowledgements

This work was supported by the EU's H2020 Programme SEWA (No. 645094) and the EU's 7th Framework Programme iHEARu (No. 338164).