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.

AER with Perception Uncertainty

Emotional perception state: $E$ uncertainty: $\sigma$

- input: audio/visual feature vectors $x$
- outputs: $(E^{(A)}, \sigma^{(A)})$ for arousal, $(E^{(V)}, \sigma^{(V)})$ for valence
- $E$ is calculated by EWE over all raters by given instance $n$:
  $$E^{(i)}_n = \frac{1}{K-1} \sum_{k=1}^{K} r^{(i)}_k \epsilon^{n,k}_i,$$
  where $r^{(i)}_k$ is a rater-dependent weight for rater $k$.
- $\sigma$ is calculated by inter-rater disagreement level:
  $$\sigma^{(i)}_n = \frac{1}{K-1} \sum_{k=1}^{K} \epsilon^{n,k}_i - c_{MLE} \epsilon^{n}_i^2.$$
- loss in multi-task learning:
  $$J(\theta) = w_E \cdot MSE_E + w_v \cdot MSE_v,$$
  with $w_E + w_v = 2$
- audiovisual late fusion:
  $$y = \epsilon + \gamma \cdot y_i$$

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 w/ two hidden layers
  - 240 LSTM cells per layer
  - hyperparameter and post-processing parameters are optimised based on the development set.

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

<table>
<thead>
<tr>
<th>CCC modality</th>
<th>task</th>
<th>dev.</th>
<th>test</th>
<th>dev.</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>video</td>
<td>single</td>
<td>$E^{(A)}$</td>
<td>$\sigma^{(A)}$</td>
<td>$E^{(V)}$</td>
<td>$\sigma^{(V)}$</td>
</tr>
<tr>
<td></td>
<td>audio</td>
<td>281</td>
<td>103</td>
<td>234</td>
<td>183</td>
</tr>
<tr>
<td></td>
<td>single</td>
<td>298</td>
<td>075</td>
<td>267</td>
<td>015</td>
</tr>
<tr>
<td></td>
<td>multi</td>
<td>296</td>
<td>180</td>
<td>292</td>
<td>089</td>
</tr>
<tr>
<td>video</td>
<td>single</td>
<td>386</td>
<td>204</td>
<td>295</td>
<td>174</td>
</tr>
<tr>
<td></td>
<td>audio</td>
<td>456</td>
<td>266</td>
<td>402</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td>single</td>
<td>502</td>
<td>261</td>
<td>478</td>
<td>111</td>
</tr>
<tr>
<td></td>
<td>multi</td>
<td>575</td>
<td>235</td>
<td>515</td>
<td>110</td>
</tr>
</tbody>
</table>

Figure: Four pairs of frames with comparable emotional states ($E$) but distinct perception uncertainties ($\sigma$) in (A) arousal and (V) valence, respectively.

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. 338364).