Feature-level and Model-level Audiovisual Fusion for Emotion Recognition in the Wild

Jie Cai\textsuperscript{1}, Zibo Meng\textsuperscript{2}, Ahmed Shehab Khan\textsuperscript{1}, Zhiyuan Li\textsuperscript{1}, James O’Reilly\textsuperscript{1}, Shizhong Han\textsuperscript{3}, Ping Liu\textsuperscript{4}, Min Chen\textsuperscript{5} and Yan Tong\textsuperscript{1}

\textsuperscript{1}University of South Carolina
\textsuperscript{2}Innopeak Technology Inc.
\textsuperscript{3}12 Sigma Technologies
\textsuperscript{4}JD. com, Inc.
\textsuperscript{5}University of Washington Bothell
Introduction

- Posed dataset:
  Neutral  Peak (Angry)  Neutral

- Emotion Recognition in the Wild:
  Illumination changes
Introduction

- **Posed dataset:**

  Neutral
  .................................................................
  Peak (Angry)
  Neutral

- **Emotion Recognition in the Wild:**

  Occlusions
Introduction

- **Posed dataset:**
  
  Neutral  ..........  Peak (Angry)  ..........  Neutral

- **Emotion Recognition in the Wild:**
  
  Head pose variations
Introduction

- Posed dataset:

Neutral  ..........  Peak (Angry)  ..........  Neutral

- Emotion Recognition in the Wild:

Solution:
Exploiting information from audio and visual channels
Related Work

- **Feature-level fusion**
  Combining audio and visual features
  Training a classifier for emotion recognition

- **Decision-level fusion**
  Combining recognition results from audio and visual signals
  o majority voting
  o averaging prediction scores
  o weighted prediction scores

- **Model-level fusion**
  Exploiting correlation between audio and visual channels
  o Bayesian Network (BN)
  o Coupled, tripled, or multistream fused HMMs
  o Artificial Neural Network (ANN)
  o Multiple Kernel Learning (MKL)
A Flowchart of The Proposed Fusion Frameworks

Feature-level fusion
A Flowchart of The Proposed Fusion Frameworks

Video

Face Preprocessing → Visual Feature Extraction → LBP-TOP → Feature Level Fusion → SVM Classifier → Emotion Label

Audio

Audio Feature Extraction → LBP-TOP SVM Classifier → CNN SVM Classifier → CNN-BLSTM Classifier → Audio SVM Classifier → Probabilistic Graphical Model → Emotion Label

Model-level fusion
Audiovisual Feature Extraction

• Audio features (1582 dimension):
  o 34 spectral related low-level feature descriptors with corresponding delta coefficients × 21 functional (1428 dimension)
  o 4 voicing related low-level feature descriptors with corresponding delta coefficients × 19 functional (152 dimension)
  o the number of pitch onsets + the total duration of the input (2 dimension)

• Visual features:
  o Human-crafted features, i.e., LBP-TOP
  o Deep learning features, i.e., aggregated CNNs and CNN-BLSTM
Aggregated CNNs

- k-average temporal pooling
- k was set to 7 empirically
- per-video-clip CNN features are 49 dimensions
Aggregated CNNs

\[ L_{IL} = \frac{1}{2} \sum_{i=1}^{m} \left\| x_i - c_{y_i} \right\|_2^2 + \lambda_1 \sum_{c_j \in \mathcal{N}} \sum_{\substack{c_k \in \mathcal{N} \ni c_k \neq c_j}} (1 + \frac{c_k \cdot c_j}{\|c_k\|_2 \|c_j\|_2}) \]
Aggregated CNNs

- An ensemble of CNNs:
  - 10 VGG-Face CNNs + 10 shallow CNNs + 10 VGG-Face CNNs ($L_{IL}$) + 10 shallow CNNs ($L_{IL}$)
- Average scores of the softmax layer outputs of the 40 CNNs as the per-frame feature
Aggregated CNNs

- Two backbone CNNs: VGG-Face and a shallow CNN
- VGG-Face fine-tuned (1) CK+, MMI, Oulu-CASIA, RAF-DB and ExpW; (2) AFEW training set
- Shallow CNN fine-tuned (1) FER-2013 dataset; (2) AFEW training set
- CNN features of the fine-tuned VGG-Face model
- BLSTM features taken from dense layer (512 dimensions)
- Training: a series of 20 images were randomly chosen from each video
- Testing: a series of 20 evenly spaced images were chosen from each video
Audiovisual Feature-Level Fusion

Four types of features:
- The first 20 principal components of the audio features
- The first 150 principal components of the LBP-TOP
- The 49-dimensional (7×7 bins) aggregated CNN features
- The first 50 principal components of the BLSTM features

- Hence, a joint feature vector consisting of 269 features for each video clip.
Audiovisual Model-Level Fusion

\[
\text{Emotion}^* = \arg \max P(\text{Emotion}|M_{\text{audio}}, M_{\text{LBP-TOP}}, M_{\text{CNN}}, M_{\text{BLSTM}})
\]
**Dataset**

- Emotion Recognition on the “Acted Facial Expression in the Wild dataset” – **AFEW dataset**

- Classification into 7 emotion categories: Anger, Disgust, Fear, Happiness, Neutral, Sadness and Surprise

- Video clips collected from “close-to-real-world” environments, i.e., Hollywood movies, reality TV shows and sitcom

- Training set (773), Validation set (383), Test set (653)
## Experimental Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Channel</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hu et. al [19]</td>
<td>audio, visual</td>
<td>59.01</td>
<td>60.34</td>
</tr>
<tr>
<td>Fan et. al [12]</td>
<td>audio, visual</td>
<td>–</td>
<td>59.02</td>
</tr>
<tr>
<td>Vielzeuf et. al [38]</td>
<td>audio, visual</td>
<td>–</td>
<td>58.81</td>
</tr>
<tr>
<td>Yao et. al [42]</td>
<td>audio, visual</td>
<td>51.96</td>
<td>57.84</td>
</tr>
<tr>
<td>Ouyang et. al [30]</td>
<td>audio, visual</td>
<td>–</td>
<td>57.20</td>
</tr>
<tr>
<td>Kim et. al [23]</td>
<td>audio, visual</td>
<td>50.39</td>
<td>57.12</td>
</tr>
<tr>
<td>Yan et. al [41]</td>
<td>audio, visual</td>
<td>–</td>
<td>56.66</td>
</tr>
<tr>
<td>Wu et. al [40]</td>
<td>audio, visual</td>
<td>–</td>
<td>55.31</td>
</tr>
<tr>
<td>Kaya et. al [21]</td>
<td>audio, visual</td>
<td>57.02</td>
<td>54.55</td>
</tr>
<tr>
<td>Ding et. al [9]</td>
<td>audio, visual</td>
<td>51.20</td>
<td>53.96</td>
</tr>
<tr>
<td>Yao et. al [43]</td>
<td>audio, visual</td>
<td>49.09</td>
<td>53.80</td>
</tr>
<tr>
<td>Kaya et. al [20]</td>
<td>audio, visual</td>
<td>52.30</td>
<td>53.62</td>
</tr>
<tr>
<td>Kahou et. al [10]</td>
<td>audio, visual</td>
<td>–</td>
<td>52.88</td>
</tr>
<tr>
<td>Sun et. al [37]</td>
<td>audio, visual</td>
<td>–</td>
<td>51.43</td>
</tr>
<tr>
<td>Pini et. al [33]</td>
<td>audio, visual</td>
<td>49.92</td>
<td>50.39</td>
</tr>
<tr>
<td>Li et. al [26]</td>
<td>audio, visual</td>
<td>–</td>
<td>50.46</td>
</tr>
<tr>
<td>Gideon et. al [14]</td>
<td>audio, visual</td>
<td>38.81</td>
<td>46.88</td>
</tr>
<tr>
<td>Bargal et. al [2]</td>
<td>visual</td>
<td>59.42</td>
<td>56.66</td>
</tr>
<tr>
<td>Sun et. al [36]</td>
<td>visual</td>
<td>50.67</td>
<td>50.14</td>
</tr>
<tr>
<td>Audio (baseline)</td>
<td>audio</td>
<td>35.51</td>
<td>–</td>
</tr>
<tr>
<td>LBP-TOP (baseline)</td>
<td>visual</td>
<td>38.90</td>
<td>–</td>
</tr>
<tr>
<td>CNN (baseline)</td>
<td>visual</td>
<td>47.00</td>
<td>–</td>
</tr>
<tr>
<td>CNN-BLSTM (baseline)</td>
<td>visual</td>
<td>49.09</td>
<td>–</td>
</tr>
<tr>
<td>Feature-level fusion</td>
<td>audio, visual</td>
<td>53.79</td>
<td>56.81</td>
</tr>
<tr>
<td>Model-level fusion</td>
<td>audio, visual</td>
<td>54.83</td>
<td>54.06</td>
</tr>
</tbody>
</table>
Conclusion

• Two novel audiovisual fusion methods by exploiting audio features, LBP-TOP-based features, aggregated CNN features, and CNN-BLSTM features.
• Both the proposed fusion methods outperform the baseline methods that employ a single type of feature.
• More advanced techniques will be developed to improve the recognition performance.
Acknowledgement

This work was supported by National Science Foundation under CAREER award IIS-1149787.

The Titan Xp used for this research was donated by the NVIDIA Corporation.
Thank you!