Style Extractor for Facial Expression recognition in the Presence of Speech

Ali N. Salman And Carlos Busso
Facial Emotion Recognition (FER) is a hard problem
- Prediction are not reliable during speech

Wide application domains
- Human-Computer/Robot Interaction
- Driver distraction
- Medical monitoring
Related Work

- **Dynamic FER**
  - Emotions perceived from isolated frames is different from emotions perceived from watching corresponding video [Salman and Busso 2020]

Table: Compares the perceptual evaluation between videos (different annotators) or videos compared to frames.

<table>
<thead>
<tr>
<th>Label</th>
<th>Set</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happiness</td>
<td>Video/Video</td>
<td>0.91</td>
<td>0.84</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>Video/Frame</td>
<td>0.67</td>
<td>0.97</td>
<td>0.79</td>
</tr>
<tr>
<td>Anger</td>
<td>Video/Video</td>
<td>0.73</td>
<td>0.67</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>Video/Frame</td>
<td>0.55</td>
<td>0.14</td>
<td>0.22</td>
</tr>
<tr>
<td>Sadness</td>
<td>Video/Video</td>
<td>0.77</td>
<td>0.79</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>Video/Frame</td>
<td>0.66</td>
<td>0.57</td>
<td>0.61</td>
</tr>
<tr>
<td>Neutral</td>
<td>Video/Video</td>
<td>0.72</td>
<td>0.72</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>Video/Frame</td>
<td>0.54</td>
<td>0.77</td>
<td>0.63</td>
</tr>
<tr>
<td>Average</td>
<td>Video/Video</td>
<td>0.78</td>
<td>0.76</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>Video/Frame</td>
<td>0.61</td>
<td>0.61</td>
<td>0.56</td>
</tr>
</tbody>
</table>

No Audio
Related Work

- **Facial Regions**
  - Lower facial regions are greatly affected by speech articulation
    - Lower regions contain valuable features for emotion classification
    - Some emotions are better perceived in the lower regions [Hoffmann et. al. 2013, Busso and Narayanan 2006]
  - Lower facial regions contain both emotional and lexical information
    - Separating emotion facial features from speech articulations is challenging

Activeness of different facial regions during speech
Related Work

- **Lexical Dependent FER**
  - A phone or viseme dependent classifier can be used to increase the reliability of emotion recognition [Mariooryad and Busso 2013, Kim and Provost 2015]
  - Another approach is to treat the lower and upper area and use a phoneme dependent classifier on the lower region [Kim and Provost 2019]
  - Transcriptions can be costly (manual), unreliable (ASR), or not feasible to attain (no audio) in real world application

- **Blind-Lexical FER**
  - Separating the facial region into many area can improve accuracy [Kim and Provost 2015]
  - Using an asymmetric bilinear factorization model to extract emotion information without knowing the phonetic labels [Mariooryad and Busso 2015].
Goal

▪ Dynamic FER
  ▪ Static FER have restriction in real-world application
  ▪ Aggregating features collected from static FER systems is not enough to capture temporal information

▪ Blind-Lexical compensation
  ▪ Transcriptions can be costly and might not be available during real world application
  ▪ The use of transcription during training is valid
  ▪ Separate the emotional and lexical attribute

▪ End-to-End Image FER
  ▪ Using just image sequences as input and not relying on special hardware to capture (i.e., motion capture, depth sensing)
Proposed Model

- **Feature Extraction**
  - Extracts facial features using a CNN model

- **Style Extractor**
  - Separates the emotional facial information (i.e., style) from the lexical facial information (i.e., content)

- **Fusion Model**
  - Combines the features extractor and style extractor features to predict the emotion
Feature Extraction

- **VGG16 architecture**
  - Initialize model using VGG-Face weights [Parkhi et. al. 2015]
  - Train the model for the emotional classes using categorical cross-entropy.

- **Static Features**
  - The first fully connected layer (red arrow) to represent the features
Style Extraction

- **Model**
  - FC (blue) and LSTM (orange) model to transform the input sequence from emotional to neutral
  - The model also predicts the phoneme for each
  - The model takes a facial mesh as input and normalize the emotional features
  - We use the difference between the input mesh and output mesh to represent the style
  - Additionally, we predict the phoneme for each mesh to assist in learning phoneme dependent features
Data

- Manually align emotional and neutral videos that contain the same lexical contents but different emotions
  - Alignment at the phone level
- Z-Face [Jeni et. al. 2015] to extract the 3D facial mesh
- Use the 3D mesh of the aligned pairs (emotional to neutral) to train the model

Style Extraction

Neutral

Emotional

No Audio
Feature Extraction
- Extracts facial features using a CNN model

Style Extractor
- Separates the emotional facial information (i.e., style) from the lexical facial information (i.e., content)

Fusion Model
- Combines the features extractor and style extractor features to predict the emotion
AffectNet Database

- **AffectNet [Mollahosseini et. al. 2019]**
  - Collected from the internet using major search engines
    - 1250 emotional keywords in 6 different languages
  - Over 1 million images
    - Around 440 thousand are manually annotated with seven discrete emotional labels
    - Valence and arousal annotation (not used in this study)
    - 425x425 average resolution
  - We consider 4 classes (happiness, anger, sadness, and neutral state)
    - Downsample to 24,882 images per class (training set)
      - Random split 80/20 for training/validation
      - Validation set as testing set
  - This dataset is used to train the feature extractor

http://mohammadmahoor.com/affectnet/
MSP-IMPROV Dataset

- **MSP-IMPROV [Busso et. al. 2017]**
  - Multimodal emotional database
  - 12 subjects (six males, six females)
  - 1,440 x 1,080 resolution
  - Same sentences are spoken with different target emotions
    - Improvisations are used before/after to the target sentences to capture naturalistic data
  - Target sentences are manually annotated in different modalities
    - 652 speaking turns
    - We only consider video-only annotations (happiness, anger, sadness, and neutral state)
  - This dataset is used to train the style extractor

https://ecs.utdallas.edu/research/researchlabs/msp-lab/MSP-Improv.html
CREMA-D Dataset

- CREMA-D [Cao et. al. 2014]
  - Multimodal emotional dataset
    - 91 subjects (six males, six females)
    - 960 x 720 resolution
    - Same sentences are spoken with different target emotions
    - Target sentences are manually annotated in different modalities
      - 7,442 annotated clips
      - We only consider 5,093 video-only labeled clips (happiness, anger, sadness, and neutral state)
    - 81/4/7 actors for train/validate/test
  - After training the style/feature extractor models we use this dataset to train the fusion model

https://github.com/CheyneyComputerScience/CREMA-D
Results - Feature Extractor

- **Feature Extractor**
  - Trained on a subset of the AffectNet database
  - Down sampled to match the minimum number of samples in a class
  - Results are reported on the validation set, which we use as our testing set

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Precision [%]</th>
<th>Recall [%]</th>
<th>F1-score [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happiness</td>
<td>89.8</td>
<td>91.0</td>
<td>90.5</td>
</tr>
<tr>
<td>Anger</td>
<td>76.7</td>
<td>71.2</td>
<td>73.9</td>
</tr>
<tr>
<td>Sadness</td>
<td>75.8</td>
<td>71.6</td>
<td>73.7</td>
</tr>
<tr>
<td>Neutral</td>
<td>63.7</td>
<td>70.1</td>
<td>67.0</td>
</tr>
<tr>
<td>Average</td>
<td>76.5</td>
<td>76.2</td>
<td>76.3</td>
</tr>
</tbody>
</table>

Performance of the static FER system in the feature extractor model. The reported values are on the AffectNet corpus.
- **Style Extractor**
  - We expect that the mesh from the style extractor looks more neutral than the original input
  - We trained a vanilla 3D mesh classifier on MSP-IMPROV
    - model achieved 60% F1-score
  - Testing the 3D mesh emotion classifier on CREMA-D
    - Around 4% of the original meshes are classified as neutral
    - 31% of the normalized meshes are classified as neutral

<table>
<thead>
<tr>
<th>Emotion\Mesh</th>
<th>Original</th>
<th>Normalized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happiness</td>
<td>15,886</td>
<td>3,332</td>
</tr>
<tr>
<td>Anger</td>
<td>191,309</td>
<td>122,075</td>
</tr>
<tr>
<td>Sadness</td>
<td>332,825</td>
<td>260,350</td>
</tr>
<tr>
<td>Neutral</td>
<td>25,060</td>
<td>179,323</td>
</tr>
</tbody>
</table>
Proposed Model

- To assess the effectiveness of the proposed approach we train two models
  - With the Style Extractor (Model [A])
  - Without the Style Extractor (Model [B])

- The model with the Style Extractor performs 7% better (absolute)
- The Style Extractor helps with generalization
  - Similar performance on train set (A vs B)
  - Smaller gap between train and validation

Results - Proposed Model

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A [%]</td>
<td>B [%]</td>
<td>A [%]</td>
</tr>
<tr>
<td>Happiness</td>
<td>87.8</td>
<td>81.1</td>
<td>83.0</td>
</tr>
<tr>
<td>Anger</td>
<td>89.2</td>
<td>51.0</td>
<td>50.9</td>
</tr>
<tr>
<td>Sadness</td>
<td>78.6</td>
<td>83.0</td>
<td>60.5</td>
</tr>
<tr>
<td>Neutral</td>
<td>68.8</td>
<td>65.0</td>
<td>89.9</td>
</tr>
<tr>
<td>Average</td>
<td>81.1</td>
<td>70.0</td>
<td>71.0</td>
</tr>
</tbody>
</table>

Performance of the proposed FER system for videos on the test set of the CREMA-D corpus.
Conclusion

- **Proposed Approach**
  - FER system that does not require transcription during inference
  - Style extractor that extracts the emotional features, not speech articulations

- **Future Research**
  - Find ways to align/pair data for training
  - Improve the feature extractor by extracting spatial-temporal features
  - Improve the style extractor by using images instead of 3D mesh
This work was funded by NEC Foundation and NSF under Grant IIS-1718944