Group-Level Emotion Recognition using Deep Models with A Four-stream Hybrid Network
Ahmed Shehab Khan1, Zhiyuan Li1, Jie Cai1, Zibo Meng2, James O’Reilly1, and Yan Tong1
1 Department of Computer Science, University of South Carolina, Columbia, United States
2 InnoPeak Technology Inc, Palo Alto, California, United States
This work was supported by National Science Foundation under CAREER Award IIS-1149787

Introduction
• Group Emotion Recognition (GER) is a challenging task due to scene variations and interactions among multiple persons in group
• Lots of data is available in public domain thanks to easy access to camera and large number of social network users
• GER has importance in various applications such as image retrieval, early event prediction, surveillance etc.

Contribution
A hybrid framework for GER consists of
• A global stream
• A novel face-location aware global stream to incorporate face-related information in the global stream
• A multi-scale face stream to handle large variations in input image size and face size
• A global blurred stream to extract scene-only features from the whole image

Overall Framework

Multi-Scale Face Stream:
• Employing two networks, one for small faces (SFN) and the other for larger faces (BFN)
• Fusing predictions of these two networks by weighted average
• Calculating weights based on face area using the following formula

\[ w_i = \frac{1}{\sqrt{A_i}} \]

\[ w_{gs-fs} = \sum w_i \]

The Four Streams

Global Stream:
• Recognizing group-level emotion from whole image
• Capturing both scene-related and face-related features
• Modeling spatial relationships and interactions among faces
• CNN Architecture: VGG16 with batch normalization

Global Blurred Stream:
• Scene-features are important
• Blurring faces in the image forcing CNN to learn only scene-related features
• CNN Architecture: VGG16 with batch normalization

Global-Blurred Stream:
• Scene-features are important
• Blurring faces in the image forcing CNN to learn only scene-related features
• CNN Architecture: VGG16 with batch normalization

Face-location Aware Global Stream:
• Face locations and sizes are utilized as an attention heatmap

\[ H_i(l + \epsilon, j + \delta) = \begin{cases} 1 & \text{if } |l(j)| \leq 3r \text{ otherwise.} \\ 0 & \end{cases} \]

• CNN Architecture: VGG16 with batch normalization

Multi-Scale Face Stream:
• Employing two networks, one for small faces (SFN) and the other for larger faces (BFN)
• Fusing predictions of these two networks by weighted average
• Calculating weights based on face area using the following formula

\[ w_i = \frac{1}{\sqrt{A_i}} \]

\[ w_{gs-fs} = \sum w_i \]

Experimental Results

Fusion
• Weighted average of four streams’ predictions
• Weights were estimated by grid search on the validation set

\[ \mathbf{P} = w_{gs-fs} \cdot \mathbf{P}_{gs-fs} + w_{gs} \cdot \mathbf{P}_{gs} + w_{ms-ffs} \cdot \mathbf{P}_{ms-ffs} + w_{gb} \cdot \mathbf{P}_{gb} \]

Overall Accuracy: 65.59
Positive Class: 73.93
Neutral Class: 55.34
Negative Class: 64.17

Conclusion and Future work
• Face plays a major role in recognizing strong emotion expressions
• Scene is important in recognizing the neutral expression
• Developing methodologies dealing with hard class and hard samples in future work