Group emotion recognition in the wild by combining deep neural networks for facial expression classification and scene-context analysis

Conference Paper · November 2017

DOI: 10.1145/3136755.3143010

2 authors:

Asad Abbas
Austal
9 PUBLICATIONS **123** CITATIONS
SEE PROFILE

Stephan Chalup
The University of Newcastle, Callaghan, Australia
167 PUBLICATIONS **1,473** CITATIONS
SEE PROFILE

Group Emotion Recognition in the Wild by Combining Deep Neural Networks for Facial Expression Classification and Scene-Context Analysis

Asad Abbas
School of Electrical Engineering and Computing
The University of Newcastle
Callaghan NSW 2308, Australia
asad.abbas@uon.edu.au

Stephan K. Chalup
School of Electrical Engineering and Computing
The University of Newcastle
Callaghan NSW 2308, Australia
stephan.chalup@newcastle.edu.au

ABSTRACT

This paper presents the implementation details of a proposed solution to the Emotion Recognition in the Wild 2017 Challenge, in the category of group-level emotion recognition. The objective of this sub-challenge is to classify a group's emotion as Positive, Neutral or Negative. Our proposed approach incorporates both image context and facial information extracted from an image for classification. We use Convolutional Neural Networks (CNNs) to predict facial emotions from detected faces present in an image. Predicted facial emotions are combined with scene-context information extracted by another CNN using fully connected neural network layers. Various techniques are explored by combining and training these two Deep Neural Network models in order to perform group-level emotion recognition. We evaluate our approach on the Group Affective Database 2.0 provided with the challenge. Experimental evaluations show promising performance improvements, resulting in approximately 37% improvement over the competition's baseline model on the validation dataset.

CCS CONCEPTS

• **Computing methodologies → Neural networks**; *Computer vision*; Supervised learning by classification;

KEYWORDS

Affective Computing; Group Emotion Recognition; Deep Neural Networks; Facial Emotion Recognition

ACM Reference Format:
Asad Abbas and Stephan K. Chalup. 2017. Group Emotion Recognition in the Wild by Combining Deep Neural Networks for Facial Expression Classification and Scene-Context Analysis. In *Proceedings of 19th ACM International Conference on Multimodal Interaction (ICMI'17).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3136755.3143010

1 INTRODUCTION

Automatically recognizing and perceiving human emotions has attracted considerable attention in the artificial intelligence domain during the past decade. It has many potential applications, ranging from medical applications such as pain detection [1], monitoring of depression [5] and helping individuals with autism spectrum disorder [9] to commercial use cases such as call-center systems, marketing, and advertisement [23]. Its associated research is inherently a multidisciplinary enterprise spread across a variety of related fields, including human-computer interaction, linguistics, learning theory, neuroscience, cognitive psychology, behavioral robotics and affect-aware game development [35].

Emotion recognition systems can depend on different types of input signals, such as visual signals (image/video), audio, text and physiological signals. But facial expressions are one of the most widely acknowledged forms of sentic modulation [25]. Most present day attempts to recognize facial expression are based on the Facial Action Coding System (FACS) of psychologist Paul Ekman which provides mappings between measurable muscle actions and an emotion space [8]. Automated facial expression recognition systems usually attempt to classify an image or a sequence of images as one of six basic emotions.

Most of the existing work on facial expression analysis focused on the classification of emotions of individual faces. More recently, the social nature of emotions led to research that addressed analysis of emotions of groups of people as a whole. The success of social network websites played an important role in research on images containing multiple participants. A large amount of images containing groups of people are uploaded every day. Automatically analyzing such image data has direct applications in image retrieval, early event prediction, surveillance, personal photo albums, crowd affective analysis and security.

Unlike emotions observed on a single facial expression, the group-as-whole approach is one in which researchers try to capture the dimensions of the elusive feelings arising from group dynamics and how the group as a holistic entity influences the feelings and behaviors of the individuals within it [2]. Literature in social psychology shows that emotions observed from a group of people arises from the combinations of its “bottom-up" components and its “top-down" components [2]. Bottom-up components result from the combinations of individual-level affective factors that group members possess such as facial expressions. Top-down components result from group or contextual-level factors that define or shape

ICMI'17, November 13–17, 2017, Glasgow, UK

ACM ISBN 978-1-4503-5543-8/17/11…$15.00
https://doi.org/10.1145/3136755.3143010

affective experiences such as environment, dressing or arrangement of the people.

Recently group level emotion prediction has been addressed by the Emotion Recognition in the Wild (EmotiW) challenge series. During the EmotiW 2016 challenge, the group-level task was to estimate the intensity of happiness of a group of people in images [6]. The EmotiW 2017 sub-challenge *group-level emotion recognition* aims to classify a group's perceived emotion as Positive, Neutral or Negative on Group Affective Database [7]. The present paper accompanies the EmotiW'17 sub-challenge on Group-level Emotion Recognition. We propose a Deep Neural Network (DNN) architecture to predict group-level emotion. The architecture models group emotion by using image context on one hand, and facial expressions on the other hand. Our proposed method outperforms the baseline approach, which comprises the CENTRIST descriptor [33] and Support Vector Machines (SVM), and achieves an accuracy of 52.97% on the validation set.

The rest of the paper is organized as follows. Section 2 reviews previous work which is related to the task of the challenge. Section 3 provides an overview of our approach and the dataset used in this work. Section 4 discusses different experiments conducted in this study. Section 5 presents results and related discussion. Finally, Section 6 concludes the paper with a discussion of possible future work.

2 RELATED WORK

In past three decades, facial expression recognition (FER) has been widely studied [28]. There are numerous publicly available labelled facial expression databases which have helped to accelerate research in the domain of automated facial analysis. Databases commonly used for facial action unit and expression recognition include: Cohn-Kanade (also its extended edition known as CK+) [20], Multipie [11], AM-FED [24], Genki-4k [32] and UNBC-McMaster Pain archive [21]. Most of the datasets were captured in controlled environments which do not reflect the type of conditions seen in real life applications. There is a rich literature on hand-crafted features extracted from images and videos for encoding facial expressions. Studies that applied traditional FER approaches to classify images that were captured in a naturalistic, spontaneous or uncontrolled setting, or when traditional FER approaches were evaluated across databases, they failed to give satisfactory results [22].

Recent advances in emotion recognition focused on recognizing spontaneous facial expressions. Since 2012, when AlexNet [18] was used for image classification of ImageNet [26], deep neural networks became state-of-the-art for many vision tasks. Yu and Zhang used an ensemble of Convolutional Neural Networks (CNNs) with five convolutional layers to achieve state-of-the-art results in classifying emotions in the wild in EmotiW in 2015 [34]. Kim et al. [16] used a hierarchical decision tree and an exponential rule to combine decisions of varying networks and parameters in EmotiW 2015.

Using only facial emotion information to infer the overall emotion of a group is a difficult problem, as accurate face detection cannot be performed on images captured in an uncontrolled environment. There is some existing work on group-level emotion analysis, most of which was presented during the EmotiW challenges. Dhall et al. [7] computed group happiness as weighted average of predicted individual happiness of a group using CENTRIST descriptor [33]. Authors classify group affect into three categories (positive, negative, neutral) by combining scene descriptors with face-centered descriptors. Vonikakis et al. [31] used a predominately bottom-up approach based on geometric features and dataset balancing to achieve group-level happiness prediction. Li et al. [19] used both holistic features and face-level features for a group happiness prediction task by employing residual nets to extract deep face feature representations and Long Short Term Memory (LTSM) networks [27] to aggregate the scene features and face features in a sequential manner. Aleksandra Cerekovic [3] used both image context and individual face information by proposing a method based on LTSM networks encoding face happiness intensity and the spatial distribution of faces forming a group to predict group happiness.

3 PROPOSED APPROACH

Following social psychology literature we argue that group emotion in an image is a combination of individual face expressions and overall scene context [2]. In our approach deep neural networks are trained to classify facial emotions of detected faces in an image as well as the overall scene context. First faces are detected using a mixtures of trees method as discussed in [37]. Then cropped faces are fed into a Convolutional Neural Network (CNN) which predicts probabilities of faces belonging to a class. These probabilities are combined with context information extracted by another CNN trained on image context, thus providing overall group emotion. The overall approach is depicted in Figure 1. The method proposed in [37] is used to automatically detect faces in an image. The authors of [37] assumed that the face shape is a tree structure (for fast inference), and used a part-based model for face detection, pose estimation, and facial feature detection. For face emotion classification, we experimentally found that fine tuning a CNN model trained on the FER2013 dataset [10] provides good results. As images are captured in an unconstrained environment, face detection models suffer from false detections as well as no detections at all in some cases. In such cases scene context information plays an important role for group emotion prediction. We classified image scene context by fine-tuning pre-trained VGG16 [29] and InceptionV3 [30] models. Probabilites of face classification and group context were then combined in the last step using three fully connected neural network layers, providing an overall group-level emotion recognition system.

3.1 Dataset

The dataset for the EmotiW 2017 group-level emotion recognition challenge is crawled from Flickr.com and Google Images, and is related to keywords[1]. Images in the database have high intra-class variance and are captured "in the wild" (i.e. in an uncontrolled environment). As a result images contain a variety of scenes and subjects, and may include changes of illumination, partial or full

[1] The sample keywords were: Tahrir Square, London Protest, Brazil Football Fans, Excited People, Happy People, Humanitarian Aid, Delhi Protest, Gaza Protest, Party Friends, Police Brutality, Celebration etc.

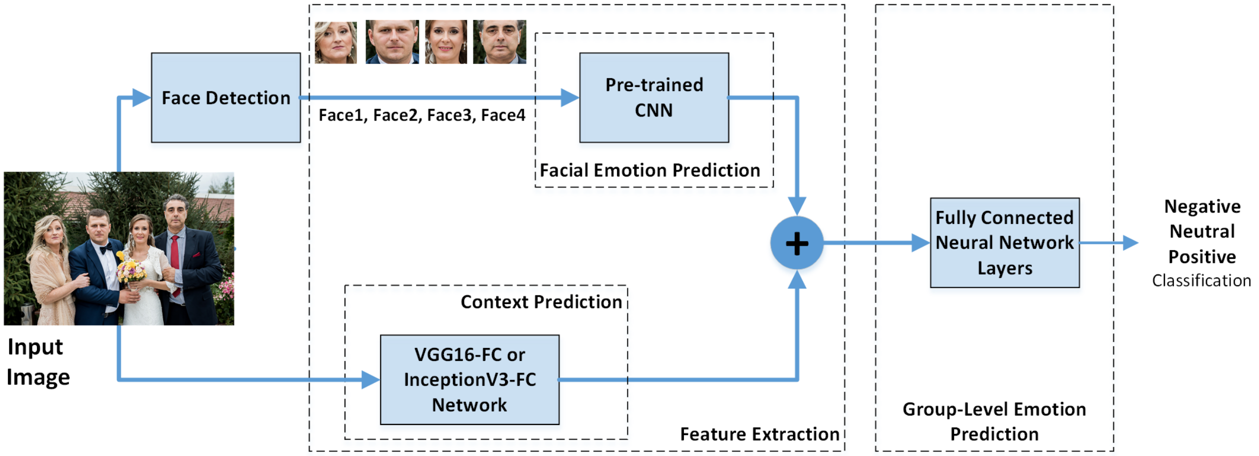
Face Detection
Face1, Face2, Face3, Face4
Pre-trained CNN
Facial Emotion Prediction
Input Image
Context Prediction
VGG16-FC or InceptionV3-FC Network
Feature Extraction
+
Fully Connected Neural Network Layers
Group-Level Emotion Prediction
Negative
Neutral
Positive
Classification

Figure 1: Our Approach: Information extracted from each face in the image (top arc) and context (bottom arc) are combined by a third neural network module to classify the overall group emotion. [Input image Flickr ID:121092974@N05]

face occlusions, head pose variations, background noise as well as a range of emotions, ages and ethnicities. The database consists of 5695 images for the combined Training and Validation database. Initially, we planned to collect additional training samples but couldn't collect any substantial amount due to time limitations and therefore the additional training data consisted of only 3,462 images that were collected from flickr and Google Images. The dataset distribution is shown in Table 1

Table 1: Dataset distribution including additional training samples

Database	Negative	Neutral	Positive
Training Samples	1159	1199	1272
Validation Samples	564	728	773
Additional Training Samples	1281	1043	1138

4 EXPERIMENTS

In this section, we describe our experiments for group-level emotion recognition. Firstly, we describe the training of CNNs used for facial emotion classification. Secondly, different configurations of CNNs used for image context classification are discussed. Finally information extracted from the faces and from the image context is used to predict the overall group emotion in test images. All experiments were conducted on our university GPU servers equipped with Tesla K80 and K20 GPUs. CNNs were implemented using Keras[2] as API with Tensorflow[3] as backend. Different hyperparameters were tested at the time of training and CNN models took several hours to train. In our case all CNN models showed similar training behaviour over different hyperparameters and initialization conditions with very low standard deviation. We were not able to perform exhaustive grid search over hyperparameters due to limited computing resources and thus only best performing CNN models are reported in the results.

4.1 Facial Emotion Prediction

To build the training set for this stage, we used [37] for face detection on the Group Affect Database 2.0 provided by the EmotiW 2017 challenge. False face detections in the training data were manually removed resulting in 1,852 Negative, 2,465 Neutral and 2,440 Positive face images. Similarly face detection was performed on the validation set resulting in 1,486 Negative, 1,622 Neutral and 2,274 Positive face images but no false face detections were removed in this case.

4.1.1 Face-CNN from Scratch. In the first experiment, we trained a CNN from scratch, with a configuration similar to the CNNs used to detect basic facial expressions. The CNN comprised 10 convolutional layers with a stride of 1. The input was padded so that the output had the same length as the original input. Average pooling with stride 2 and rectified linear units (ReLU) as activation function were used. Network details are shown in Table 2. Further, the training set was augmented during training using the image data generator API from Keras. For augmentation, we used rotation, shifting, zooming and horizontal flip. Input images were resized to $48 \times 48 \times 3$, normalized in the range [0,1], randomly shuffled and grouped in batches of 64. To prevent over-fitting a dropout of 0.25 was applied throughout the network. The CNNs were trained for 100 epochs using the Adam optimizer [17] with a fixed learning rate of 1e-05. This model contains approximately 630,000 trainable parameters and achieved an accuracy of 56.12%. Learning curves are shown in Figure 2.

4.1.2 Face-Pre-Trained CNN. We trained a CNN model[4] inspired by the Xception [4] architecture. This architecture provides more efficient use of model parameters as compared to VGG nets and contains residual modules [13] and depth-wise separable convolutions [14]. Firstly, this network was trained on the FER2013

[2] https://github.com/fchollet/keras
[3] https://www.tensorflow.org/
[4] See mini-Xception model available at https://github.com/oarriaga/face_classification

Table 2: Configuration of CNN trained from scratch to predict facial emotions

Layers	Output Size	Patch Size
Conv1	48×48×16	7×7
Conv2	48×48×16	7×7
Average Pooling	24×24×16	2×2
Conv3	24×24×32	5×5
Conv4	24×24×32	5×5
Average Pooling	12×12×32	2×2
Conv5	12×12×64	3×3
Conv6	12×12×64	3×3
Average Pooling	6×6×64	2×2
Conv7	6×6×128	3×3
Conv8	6×6×128	3×3
Average Pooling	3×3×128	2×2
Conv9	3×3×256	3×3
Conv10	3×3×3	2×2
Softmax Activation	3	

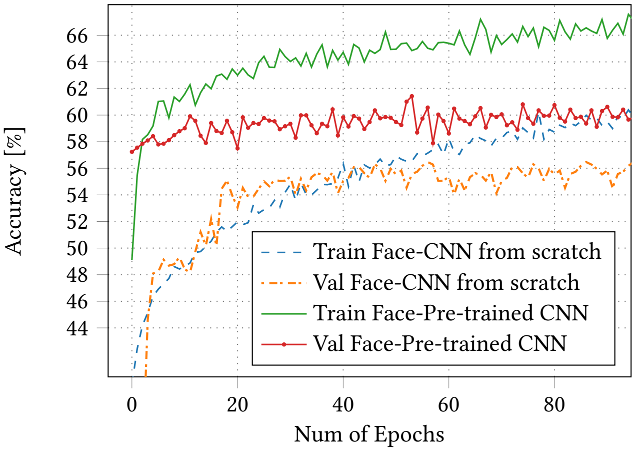

Figure 2: Training and Validation accuracy as a function of number of epochs for Face-CNN trained from scratch and Face-Pre-trained CNN

dataset [10] and achieved an accuracy of 66% for the emotion classification task on the FER2013 dataset. Secondly, we used 2048-dimensional top level features from a model trained on the FER2013 dataset and connected 3 fully connected (FC) layers, and fine-tuned the new model by applying back-propagation to the whole network. Sizes of the fully connected layers were 48, 48 and 3. The fully connected layers were randomly initialized, had ReLU activation functions and a dropout of 0.5 was applied for regularization. The training set was augmented during training using the image data generator API from Keras. For augmentation, we used rotation, shifting, zooming and horizontal flip. Due to the pre-trained Xception architecture's [4] input size requirements, augmented face images were re-sized to 64×64 and converted to gray-scale. A batch size of 32 was used and the network was trained for 100 epochs. The Adam optimizer [17] was used, starting with a learning rate of 0.001. This model contained approximately 66,000 trainable parameters. The training accuracy and the validation accuracy reached a maximum of 66.11% and 60.00% on 33rd and 21st epoch, respectively. Learning curves are shown in Figure 2. This result is understandable, given that the validation data contains many false face detections as well as blended emotions that can be even hard for humans to classify.

4.2 Context Prediction

This stage aims to learn top-down components present in images that show, for example, environment, dressing, arrangement of people and the overall context of the image. VGG16 [29] and InceptionV3 [30] networks pre-trained on ImageNet [26] data were fine-tuned on the EmotiW 2017 dataset to predict the overall emotion present in the image. The training dataset was augmented using rotation, shifting, zooming and horizontal flip. Input images were resized to 256×256×3, randomly shuffled and grouped into batches consisting of 64 images. Adam optimizer [17] was used with starting learning rate of 0.001 and learning rate was reduced on reaching plateau by a factor of 0.1. After initial pilot tests early stopping with a patience of 50 epochs was applied.

4.2.1 VGG16-FC. We used the VGG16 model [29], with weights pre-trained on ImageNet. 5 fully connected layers were stacked on top of a 512-dimensional feature vector from the VGG16 model, and the new model was fine-tuned by applying back-propagation to the whole network. Sizes of the fully connected layers were 512, 256, 128, 64 and 3, respectively. Fully connected layers were randomly initialized, used ReLU activation and a dropout of 0.5 was applied for regularization. A softmax activation was used in the last layer to obtain output prediction probabilities. This model contained about 435,000 parameters.

4.2.2 InceptionV3-FC. We used an InceptionV3 model [30], pre-trained on ImageNet. Similar to the VGG16-FC model, 5 fully connected layers were put on top of a 2048-dimensional feature vector and fine-tuned using the EmotiW 2017 dataset. Sizes of fully connected layers were 512, 256, 128, 64 and 3, respectively. Similar to the VGG16-FC network, fully connected layers were randomly initialized, had ReLU activation and a dropout of 0.5 was applied for regularization. Softmax activation in the last layer provided output prediction probabilities. This model contained about 1,220,000 trainable parameters.

Figure 3 shows the training and validation accuracies for the VGG16-FC and InceptionV3-FC networks. The VGG16-FC and InceptionV3 models achieved top validation accuracies of 66.30% and 63.19%, respectively. This result is understandable as both networks were pre-trained on the Imagenet dataset, which contained very few images similar to the EmotiW'17 dataset. Moreover, the EmotiW dataset has a limited size and contains a lot of intra- and inter-class variance, therefore neural networks struggle to construct the underlying distribution of the dataset they are being trained on. Interestingly, the accuracy of the larger InceptionV3-FC model was lower than that of the smaller VGG16-FC in our case. Different hyperparameters were tested at the time of training, but in all cases VGG16-FC performed better than InceptionV3-FC. Some possible

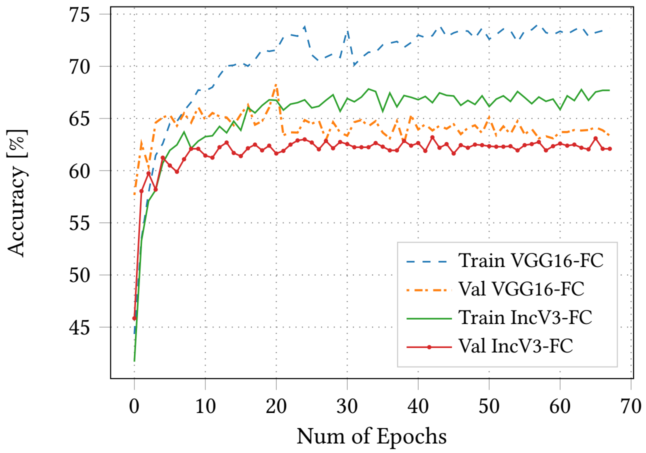

Figure 3: Training and Validation accuracy as a function of number of epochs for VGG16-FC and InceptionV3-FC model for scene context analysis

explanations are that this might be due to the specific selection of hyperparameters and network initialization or the effective capacity of the VGG16 architecture is more suited to the training dataset.

4.3 Overall Group-level Emotion Recognition

This stage combines the outputs of the facial emotion prediction and context prediction stages and feeds the combined information into a densely connected neural network as shown in Figure 1. The output of the facial emotion prediction stage was first converted into a 1 × 3 vector, representing the probabilities of faces classified as negative, neutral and positive in an image. Similarly a 1 × 3 vector containing softmax probabilities of an image belonging to a negative, neutral or positive class was obtained from the context prediction stage. Both vectors were combined to form a 1×6 vector, which served as input to the densely connected neural network. The sizes of the fully connected layers were 128, 128, 128 and 3, and were initialized with the initializers of He et al. [12]. ReLU activation, a dropout of 0.75 and a L2 regularizer were applied for regularization. A softmax activation was used in the last layer to obtain output predictions. This model contained approximately 35,000 trainable parameters and training lasted for only 5 minutes.

Experiments were performed using the VGG16-FC as well as the InceptionV3-FC networks trained in context-model stage. Figure 4 shows the training and validation accuracy for overall group-level emotion recognition when using the VGG16-FC model and the InceptionV3-FC model. We obtained best validation accuracy for group-level emotion recognition at 72.38% with VGG16-FC as context prediction model. By using InceptionV3-FC as context prediction model we achieved a validation accuracy of 70.09%.

5 RESULTS AND DISCUSSION

Comparative results are shown in Table 3. Our proposed models are compared with the given baseline model on the EmotiW'17 challenge. All reported models are evaluated on the Group Affect Database 2.0 provided by challenge. Our best performing overall group-level emotion prediction model is a combination of a Xception architecture based fine-tuned CNN network and a fine-tuned VGG16 network pre-trained on the FER2013 and ImageNet databases, respectively. Our proposed model shows an improvement of approximately 37% over the baseline model on the validation dataset. Our algorithm leverages both "Top-down" and "Bottom-up" components present in an image to predict overall group-level emotion.

To further analyze our approach, confusion matrices for two of the best performing models are shown in Figure 5 and Figure 6. For the Face - Pretrained CNN + VGG16-FC Network the accuracies were: Positive= 607/(26 + 140 + 607) = 78.52%, Neutral= 472/(150 + 101 + 472) = 65.28% and Negative= 410/(53 + 101 + 410) = 76.78% (Figure 5). Similarly, for the Face - Pretrained CNN + InceptionV3-FC Network the accuracies were: Positive = 76.19%, Neutral = 64.59% and Negative = 68.79% (Figure 6). I.e.in both models, the positive class has the best accuracy and the neutral class has the worst accuracy. On visual inspection of the classified images, we found that there are three main factors for low overall accuracy. Firstly, false faces were detected in most of the misclassified images, which results in misleading facial emotion predictions. Secondly, the reason for low accuracies are wrong facial emotion classifications. As discussed earlier, most of the faces are detected in an uncontrolled environment, and include changes of illumination, partial or full face occlusions and head pose variations. We also

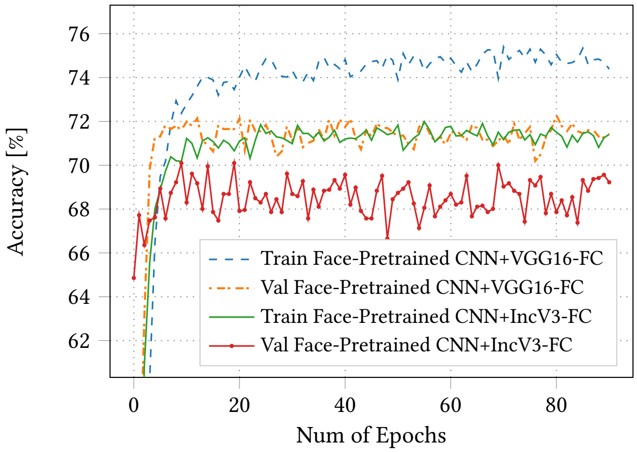

Figure 4: Training and Validation accuracy for overall group-level emotion recognition

Table 3: Comparision of validation accuracies of the proposed models

Model	Val Accuracy
SVR + CENTRIST (baseline)	52.97%
Face-Pretrained CNN	60.00%
InceptionV3-FC (ours)	63.19%
VGG16-FC (ours)	66.30%
Face - Pretrained CNN + InceptionV3-FC (ours)	70.09%
Face - Pretrained CNN + VGG16-FC (ours)	**72.38%**

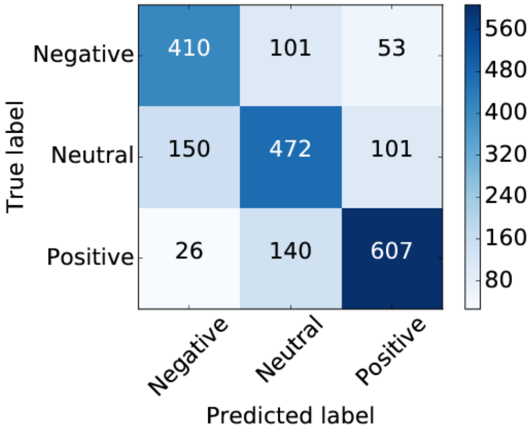

Figure 5: Confusion Matrix for Face - Pretrained CNN + VGG16-FC Network

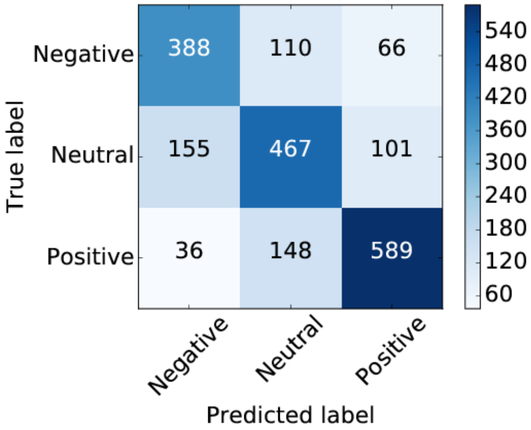

Figure 6: Confusion Matrix for Face - Pretrained CNN + InceptionV3-FC Network

observed that facial emotions in many negatively and positively classified images provided by the validation dataset, look 'Neutral' to human observers. These factors led to poor 'Neutral' class classification which impacted on the final overall group-level emotion prediction. Lastly, in many negative or neutral cases, the algorithm failed to detect any face at all. In such scenarios, our approach relies on image scene-context information extracted by the VGG16-FC or InceptionV3-FC networks alone. Both networks were pre-trained on the ImageNet database, which does not contain enough images related to the target categories and thus may perform poorly in such scenarios.

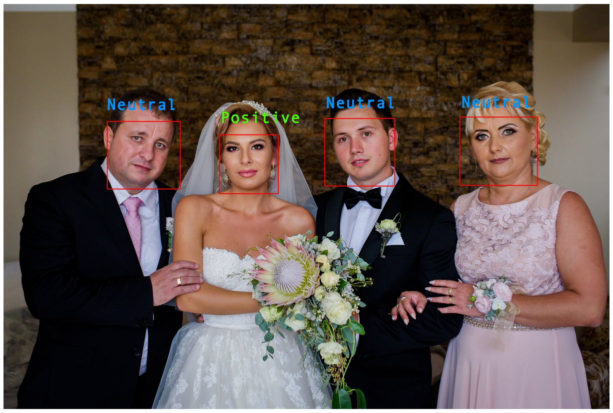

Figure 7: An example image from Flickr correctly classified as Positive in presence of Neutral Facial Emotions. [Flickr ID: 121092974@N05]

Figures 7, 8 and 9 show illustrative example images taken from Flickr public domain[5]. Images like the example in Figure 7 were correctly classified by the 'Faces-Pretrained CNN + VGG16-FC' model as 'Positive' even if they included more 'Neutral' than 'Positive' faces. Interestingly, also human observers agree that this image contains more neutral faces and that they were correctly detected and classified by the facial emotion prediction model. In such cases image context plays an important role. If enough training data related to such context is present then the image context model can rectify and overcome misleading facial expression classifications. Figure 8 shows another illustrative example image taken from the Flickr public domain, and represents examples that have been correctly classified by the 'Faces-Pretrained CNN + VGG16-FC' model as 'Negative'. In our opinion this image contains 'Neutral' and 'Positive' faces and they were correctly classified by the facial expression prediction model, but the context is negative. A possible explanation of this phenomenon is that the image context model (VGG16-FC) plays an important role in correct classification, which we think is due to the fact that the training data contains many similar negative image samples containing 'anti-Trump slogans'. Figure 9 shows an example image with no face detection at all. Images like this will be classified only based on context information.

Finally, for the EmotiW'17 competition, our best performing model 'Faces-Pretrained CNN + VGG16-FC' produced an accuracy of 63.43% which shows an improvement of approximately 19% over the EmotiW'17 baseline.

6 CONCLUSION

In this paper, we proposed a Deep Neural Network model to extract "Top-down" and "Bottom-up" components and predict the overall group-level emotion in an image. Our model combines individual face-level information present in an image and the overall image

[5] The actual sample images from the dataset used for the experiments could not be displayed in the paper due to copyright restrictions. Instead we employed images from Flickr's public domain as illustrative examples. All Flickr images in this paper were made available under the Creative Commons CC0 1.0 Universal Public Domain Dedication.

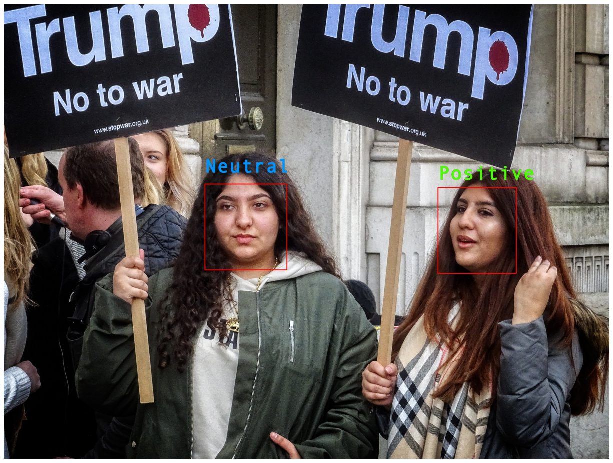

Figure 8: An example image from Flickr correctly classified as Negative in presence of Neutral and Positive Facial Emotions. [Flickr ID: garryknight]

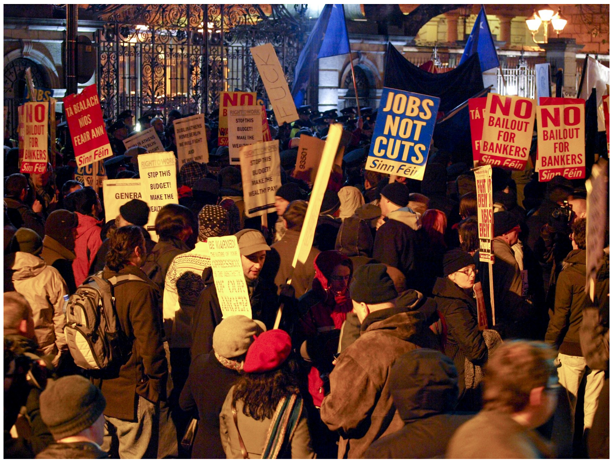
NO BAILOUT FOR BANKERS Sinn Féin
JOBS NOT CUTS Sinn Féin
NO BAILOUT FOR BANKERS Sinn Féin
NO BAILOUT FOR BANKERS Sinn Féin
STOP THIS BUDGET! MAKE THE WEALTHY PAY

Figure 9: An example image from Flickr correctly classified as Negative in absence of any face detection. [Flickr ID: sebastiandooris]

context to classify an input image as 'Negative', 'Neutral' or 'Positive'. We present our model in EmotiW'17 challenge on Group-level emotion recognition. Due to the relatively small size of the provided dataset, our best performing CNN models are obtained by fine-tuning pretrained CNN models. Our resulting model achieved an improvement of approximately 37% over the baseline model.

As the accuracy of the overall model depends on face detection, we plan in future research to experiment with different face detection models available in literature. We are particularly interested in fine tuning CNN based object detection models such as [15] for face detection. For scene-context prediction, we are interested in fine tuning a CNN model pretrained on the Places365 dataset [36]. It has been reported in literature that a CNN (AlexNet, GoogLeNet and VGG) trained on scene-centric data (Places) outperforms CNN (AlexNet, GoogLeNet and VGG) trained on the ImageNet dataset on scene-related datasets. On the other hand CNNs trained on ImageNet datasets outperform CNNs trained on the Places dataset on object-related image datasets [36]. Unfortunately, we could not perform such experiments on the EmotiW'17 dataset due to time constraints. And finally, collecting more data related to misclassified samples would be helpful to boost the overall performance of our approach. Manual collection and labeling of data is a laborious task and would be proposed as a last resort to achieve a further increase in validation accuracy.

ACKNOWLEDGMENT

AA was supported by a UNRSC50:50 scholarship. The authors would like to thank Aaron Scott from the UON ARCS team for technical support with the GPU servers.

REFERENCES

[1] Ahmed Bilal Ashraf, Simon Lucey, Jeffrey F Cohn, Tsuhan Chen, Zara Ambadar, Kenneth M Prkachin, and Patricia E Solomon. 2009. The Painful Face–pain Expression Recognition Using Active Appearance Models. *Image and Vision Computing* 27, 12 (2009), 1788–1796.

[2] Sigal G Barsade and Donald E Gibson. 1998. Group Emotion: A View From Top and Bottom. *Research on Managing Groups And Teams* 1, 4 (1998), 81–102.

[3] Aleksandra Cerekovic. 2016. A Deep Look into Group Happiness Prediction from Images. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 437–444.

[4] François Chollet. 2016. Xception: Deep Learning with Depthwise Separable Convolutions. *arXiv preprint arXiv:1610.02357* (2016).

[5] Jeffrey F Cohn, Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai Nguyen, Margara Tejera Padilla, Feng Zhou, and Fernando De la Torre. 2009. Detecting Depression From Facial Actions and Vocal Prosody. In *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009.* IEEE, 1–7.

[6] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. 2016. Emotiw 2016: Video and Group-Level Emotion Recognition Challenges. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 427–432.

[7] Abhinav Dhall, Jyoti Joshi, Karan Sikka, Roland Goecke, and Nicu Sebe. 2015. The More the Merrier: Analysing the Affect of a Group of People in Images. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, Vol. 1. IEEE, 1–8.

[8] Paul Ekman and Wallace V Friesen. 1976. Measuring Facial Movement. *Environmental Psychology and Nonverbal Behavior* 1, 1 (1976), 56–75.

[9] Rana El Kaliouby and Peter Robinson. 2005. Real-Time Inference of Complex Mental States From Facial Expressions and Head Gestures. In *Real-Time Vision for Human-Computer Interaction*. Springer, 181–200.

[10] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. 2013. Challenges in Representation Learning: A Report on Three Machine Learning Contests. In *International Conference on Neural Information Processing*. Springer, 117–124.

[11] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. 2010. Multi-pie. *Image and Vision Computing* 28, 5 (2010), 807–813.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *CoRR* abs/1502.01852 (2015).

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[14] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv preprint arXiv:1704.04861* (2017).

[15] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. 2016. Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors. *arXiv preprint arXiv:1611.10012* (2016).

[16] Bo-Kyeong Kim, Hwaran Lee, Jihyeon Roh, and Soo-Young Lee. 2015. Hierarchical Committee of Deep Cnns with Exponentially-Weighted Decision Fusion for Static Facial Expression Recognition. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 427–434.

[17] Diederik Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980* (2014).

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet Classification With Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NIPS 2012)*, Vol. 25. 1097–1105.

[19] Jianshu Li, Sujoy Roy, Jiashi Feng, and Terence Sim. 2016. Happiness Level Prediction with Sequential Inputs via Multiple Regressions. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 487–493.

[20] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. The Extended Cohn-Kanade Dataset (ck+): A Complete Dataset for Action Unit and Emotion-Specified Expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 94–101.

[21] Patrick Lucey, Jeffrey F Cohn, Kenneth M Prkachin, Patricia E Solomon, and Iain Matthews. 2011. Painful data: The UNBC-McMaster shoulder pain expression archive database. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 57–64.

[22] C Mayer, M Eggers, and B Radig. 2014. Cross-Database Evaluation for Facial Expression Recognition. *Pattern Recognition and Image Analysis* 24, 1 (2014), 124–132.

[23] Daniel McDuff, Rana El Kaliouby, Karim Kassam, and Rosalind Picard. 2010. Affect Valence Inference From Facial Action Unit Spectrograms. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 17–24.

[24] D. McDuff, R. el Kaliouby, T. Senechal, M. Amr, J. F. Cohn, and R. Picard. 2013. Affectiva-MIT Facial Expression Dataset (AM-FED): Naturalistic and Spontaneous Facial Expressions Collected In-the-Wild. In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 881–888.

[25] Rosalind W Picard and Roalind Picard. 1997. *Affective Computing*. Vol. 252. MIT Press, Cambridge.

[26] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.

[27] Haşim Sak, Andrew Senior, and Françoise Beaufays. 2014. Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling. In *Fifteenth Annual Conference of the International Speech Communication Association*.

[28] Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. 2015. Automatic Analysis of Facial Affect: A Survey of Registration, Representation, and Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 6 (2015), 1113–1133.

[29] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014).

[30] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the Inception Architecture for Computer Vision. *CoRR* abs/1512.00567 (2015).

[31] Vassilios Vonikakis, Yasin Yazici, Viet Dung Nguyen, and Stefan Winkler. 2016. Group Happiness Assessment Using Geometric Features and Dataset Balancing. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 479–486.

[32] Jacob Whitehill, Gwen Littlewort, Ian Fasel, Marian Bartlett, and Javier Movellan. 2009. Toward Practical Smile Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 11 (2009), 2106–2111.

[33] Jianxin Wu and Jim M Rehg. 2011. CENTRIST: A Visual Descriptor for Scene Categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 8 (2011), 1489–1501.

[34] Zhiding Yu and Cha Zhang. 2015. Image Based Static Facial Expression Recognition with Multiple Deep Network Learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 435–442.

[35] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. 2009. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 1 (2009), 39–58.

[36] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).

[37] Xiangxin Zhu and Deva Ramanan. 2012. Face Detection, Pose Estimation, and Landmark Localization in the Wild. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2879–2886.