Group-Level Emotion Recognition using Hybrid Deep Models based on Faces, Scenes, Skeletons and Visual Attentions

Xin Guo
Department of Electrical and Computer Engineering, University of Delaware
Newark, DE, USA
guoxin@udel.edu

Bin Zhu
Department of Electrical and Computer Engineering, University of Delaware
Newark, DE, USA
zhubin@udel.edu

Luisa F. Polanía
American Family Mutual Insurance Company
Madison, WI, USA
polania@amfam.com

Charles Boncelet
Department of Electrical and Computer Engineering, University of Delaware
Newark, DE, USA
boncelet@udel.edu

Kenneth E. Barner
Department of Electrical and Computer Engineering, University of Delaware
Newark, DE, USA
barner@udel.edu

ABSTRACT

This paper presents a hybrid deep learning network submitted to the 6th Emotion Recognition in the Wild (EmotiW 2018) Grand Challenge [9], in the category of group-level emotion recognition. Advanced deep learning models trained individually on faces, scenes, skeletons and salient regions using visual attention mechanisms are fused to classify the emotion of a group of people in an image as positive, neutral or negative. Experimental results show that the proposed hybrid network achieves 78.98% and 68.08% classification accuracy on the validation and testing sets, respectively. These results outperform the baseline of 64% and 61%, and achieved the first place in the challenge.

CCS CONCEPTS

• **Computing methodologies → Activity recognition and understanding**; *Scene understanding*; *Transfer learning*;

KEYWORDS

EmotiW 2018; Group-level Emotion Recognition; Multi-model; Scene Understanding; Visual Attention;

ACM Reference Format:
Xin Guo, Bin Zhu, Luisa F. Polanía, Charles Boncelet, and Kenneth E. Barner. 2018. Group-Level Emotion Recognition using Hybrid Deep Models based on Faces, Scenes, Skeletons and Visual Attentions. In *2018 International Conference on Multimodal Interaction (ICMI '18), October 16–20, 2018, Boulder, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3242969.3264990

ICMI '18, October 16–20, 2018, Boulder, CO, USA

ACM ISBN 978-1-4503-5692-3/18/10…$15.00
https://doi.org/10.1145/3242969.3264990

1 INTRODUCTION

Group-level emotion recognition has been a topic of interest in the social psychology community for decades [24]. The interest in this topic is motivated by the fact that group level emotions are essential in understanding the sense of social identity and how individuals relate to their social environment. Social psychology is not the only area interested in group emotions. This topic also has applications in shot selection [7], image retrieval [6], surveillance [2], event detection [32], and event summarization [7], which motivates the design of automatic systems capable of understanding human manifestations of emotional attributes at the group level. The EmotiW Group-level Emotion Recognition Sub-challenge was created with the aim of advancing group-level emotion recognition. In this annual sub-challenge, the collective emotion is classified as positive, neutral, or negative using the Group Affect Database 2.0 [8].

Prior work in the area includes the work of Dhall *et al.* [8], which introduces the Group Affect Database and a framework for group emotion recognition which includes the extraction of facial features using Facial Action Units, extraction of low-level features on the aligned faces, extraction of scene features using the GIST and CENTRIST descriptors and fusion using Multiple Kernel Learning. Another work that also uses hand-crafted features is described in [17]. Specifically, the authors used the Riesz transform and the local binary pattern descriptor to exploit not only neighbouring changes in the spatial domain of a face but also along the different Riesz faces.

Different approaches based on deep neural networks have been proposed to solve the task of the Group-level Emotion Recognition Sub-challenge. In 2016, the winner of the sub-challenge proposed the extraction of features from both the whole image and facial regions using the combination of a convolutional neural network (CNN) and a long short-term memory network (LSTM) [19]. Similarly, the winner of the 2017 version of the sub-challenge proposed to fuse the results of CNNs trained individually on faces and whole images [30]. The authors used a large-margin softmax loss for discriminative learning. In [11], we presented a hybrid network that

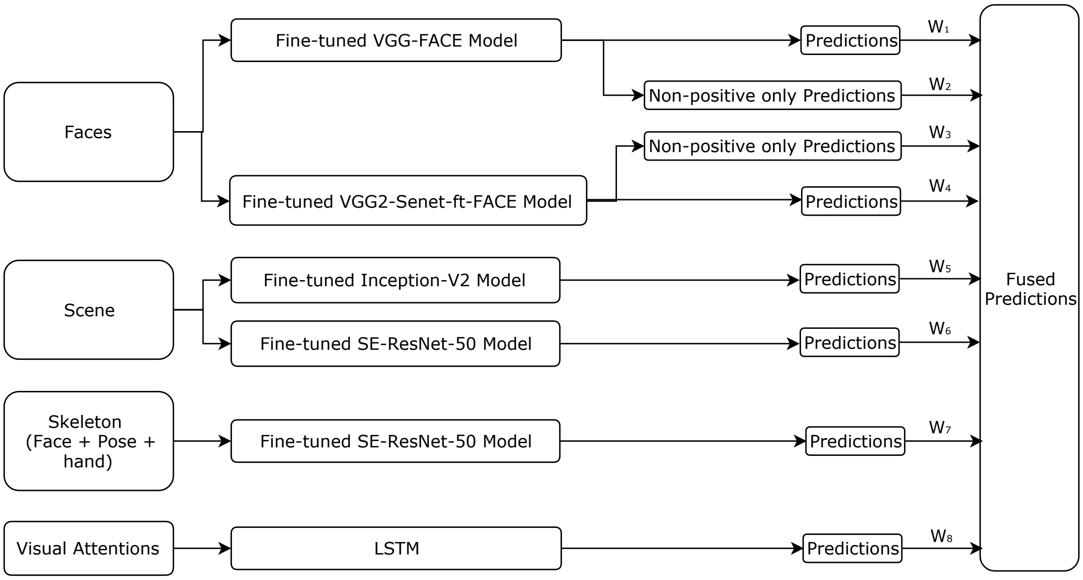
Faces
Fine-tuned VGG-FACE Model
Fine-tuned VGG2-Senet-ft-FACE Model
Predictions
Non-positive only Predictions
Non-positive only Predictions
Predictions
Scene
Fine-tuned Inception-V2 Model
Fine-tuned SE-ResNet-50 Model
Predictions
Predictions
Skeleton (Face + Pose + hand)
Fine-tuned SE-ResNet-50 Model
Predictions
Visual Attentions
LSTM
Predictions
W₁
W₂
W₃
W₄
W₅
W₆
W₇
W₈
Fused Predictions

Figure 1: The overall structure of the proposed hybrid network. The network contains 8 deep models trained on scenes, faces, skeletons and visual attentions separately. $\sum_i W_i = 1$, and $0 \le i \le 8$. Details are described in Section 2 and 3.

exploited information from whole images, faces and the skeleton representation of the subjects in the image using CNNs.

The problem of group emotion recognition is challenging due to face occlusions, illumination variations, head pose variations, varied indoor and outdoor settings, and faces at different distance from the camera which may lead to low-resolution face images [23]. In this paper, we propose a method to address this challenging problem by using a hybrid network that fuses the predictions of models trained individually on faces, scenes, skeletons and regions extracted with visual attention mechanisms. One of the models is trained on faces because facial expressions convey powerful discriminating information for emotion recognition. The model trained on scenes captures context information while the model trained on skeleton representations captures body gestures, which convey important affective information according to research results on experimental psychology and affective computing [26]. Visual attention is exploited by one of the models to enable deeper image understanding through fine-grained analysis by focusing on regions of the image that are salient and relevant for the emotion recognition task. The proposed hybrid network achieved the first place in the Group-level Emotion Recognition Sub-challenge, reporting a classification accuracy of 78.98% and 68.08% on the validation and testing sets, respectively. Code is available at *https://github.com/gxstudy/EmotiW2018_Group-level_Emotion_Recognition.*

2 THE PROPOSED METHOD

The general idea of the proposed method is to train multiple deep models to learn high-level abstractions of the input from different perspectives, so that the learned models can complement each other and be fused into a high-performance hybrid network (Figure 1).

2.1 Face Prediction

The VGG-FACE model is a 16-layer VGG architecture trained on a large-scale dataset (VGGFace dataset [25]), containing 2.6M images of 2.6K celebrities, for face recognition. It has been shown that fine-tuning the VGG-FACE model for the task of facial emotion recognition achieves better results than using hand-engineered features or deep models trained from scratch [11, 12].

A new face-related model, VGG2-Senet-ft-FACE, was recently proposed by the VGG group in [3]. VGG2-Senet-ft-FACE is a ResNet-50 [14] network with Squeeze-and-Excitation (SE) blocks [16], which was initially pretrained on the Ms-Celeb-1M dataset [13] and fine-tuned on the VGGFace2 [3] training set. The VGGFace2 dataset contains 3.31 million images of 9131 subjects.

The VGG2-Senet-ft-FACE model and VGG-FACE architectures are modified in this paper to address the problem of group-level emotion recognition. The modification consists of changing the number of neurons in the last fully-connected layer to 3. The modified architectures are initialized with the weights of the original VGG-FACE and VGG2-Senet-ft-FACE model individually, with the exception of the last fully-connected layer, which is initialized with weights sampled from a Gaussian distribution of zero mean and variance 1×10^{-4}. The learning parameters of each architecture, such as the overall learning rate, the weight decay, and the learning policy are set the same as in the original models, except that the last fully connect layer is trained with a learning rate for the weight and bias terms that is set to be 10 times larger than the overall learning rate.

Faces are extracted and aligned using MTCNN [39]. The modified VGG-FACE and VGG2-Senet-ft-FACE models are first fine-tuned on a combined facial emotion dataset (30205 samples in total), which includes images from the facial expression recognition 2013 (FER-2013) dataset [10] and the GENKI-4K dataset [35]. The negative collection is formed with the angry and sad classes from the FER-2013 database, the neutral collection is formed by combining neutral images from both datasets, and the positive is formed by combining the happy images from both datasets.

The models are further fine-tuned on the detected faces of the Group Affect Database 2.0. During training, all the extracted faces are re-scaled to 256 × 256 pixels and have the same weight when fine-tuning the parameters of the network. For scoring, the mean of the prediction of individual faces is used as the final prediction.

2.2 Scene Prediction

Deep models based on whole images (scene) have demonstrated superior performance compared to hand-engineered features, such as CENTRIST [36], in [11, 30]. A new state-of-the-art model, SENet [16], which is the winner of the ImageNet 2017 Large Scale Visual Recognition Challenge (ILSVRC 2017) [5], is based on the idea of stacking Squeeze-and-Excitation (SE) blocks together to improve the representational power of a network by explicitly modeling the interdependencies between the channels of its convolutional features. Squeeze-and-Excitation blocks can be applied to state-of-the-art deep convolutional networks such as AlexNet [18], VGGNet [28] and ResNet [14]. Inception-V2 [29] and SE-ResNet-50, which is a 50-layer ResNet with SE blocks, trained on ImageNet-1K database, are used as scene models in this paper.

Specifically, whole images are used as input to SE-ResNet-50 and Inception-V2 deep networks. Each architecture is modified by changing the number of neurons in the last layer to 3, indicating a ternary classification, having as targets negative, neutral and positive emotions for the group. With the exception of the last layer, the modified architectures are initialized with the models trained on ImageNet. The last layer in each architecture is initialized in the same way as in the original setup of each architecture, and trained with a learning rate for the weight and bias terms that is set to be 10 times larger than the overall learning rate. The learning parameters of each architecture, such as the overall learning rate, the weight decay, and the learning policy are set the same as in the original submissions to the ImageNet challenge.

2.3 Skeleton Classification

Skeletons corresponding to the collection of keypoints of human face, body and hands were first used for the problem of group-emotion recognition in [11]. As shown in Figure 2, skeleton features demonstrate salient patterns of different categories through facial expression, people layout, pose and gestures. In this paper, we propose to fine-tune an SE-ResNet-50 model on skeletons. The skeleton of each image is extracted using OpenPose [4, 31, 34], which can jointly detect human body, hand and facial keypoints (130 keypoints in total for each person) on single images, invariant to the number of people in the image. Model modification, training parameters and training procedure are the same as in the SE-ResNet-50 model trained on scene images.

2.4 Visual Attention Classification

Visual attention mechanisms have been widely explored in image captioning [1, 21, 27], visual question answering [37, 38] and image classification [22, 33]. In this paper, we propose to use areas of images that contain salient objects or features to classify group-level emotions. Specifically, 16 salient regions are extracted using the combined bottom-up and top-down attention mechanism proposed in [1], then a SENet-154 model[1] trained on the ImageNet-1k database is used to extract a 2048-dimensional vector from each region at layer *pool5/7x7_s1*[2]. As shown in Figure 3, the attention mechanism is able to detect and crop salient features such as people, white signs, red signs, pink mats and bouquets.

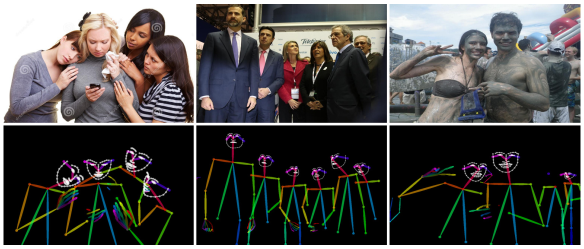

Figure 2: Samples of skeleton representations. Left: original image. Middle: skeleton representation 1 (includes faces and bodies). Right: skeleton representation 2 (includes faces, bodies and hands)

After feature extraction, each image is represented by 16 feature vectors (each vector of dimension 2048 × 1). Keeping only 16 regions can be viewed as a hard attention mechanism, as only a small number of image regions are selected from a large number of region proposals. A single-layer LSTM [15] with 128 neurons is trained on the extracted attention features. The network takes a sequence of 16 feature vectors and learns to combine features from 16 salient areas and predict the overall emotion of the image. The learning rate is set to 0.001, the bias terms are initialized to 0, and the weights of the LSTM are initialized with values drawn from a truncated normal distribution with mean 0 and standard deviation 0.1. The training process runs for 10 epochs using Adam optimizer.

3 EXPERIMENTAL RESULTS

3.1 Group-Level Emotion Recognition Sub-challenge

Group-level emotion recognition is one of the sub-challenges in the 6th Emotion Recognition in the Wild (EmotiW 2018) Grand Challenge [9]. The images in this sub-challenge are from the Group Affect Database 2.0 [8], which contains 9815, 4346 and 3011 images in the training, validation and testing sets, respectively[3]. These images are collected from social events, such as convocations, marriages, parties, meetings, funerals, protests, etc. Participants compete on the accuracy of classifying the group perceived emotion as positive, neutral or negative on the testing data.

3.2 Experimental Details

The prediction accuracy of the models tested on the validation dataset is shown in Table 1. Large-margin softmax loss [20] is used in each model, with the exception of the attention-LSTM model, to

[1] Model downloaded from https://github.com/hujie-frank/SENet
[2] *pool5/7x7_s1* is a layer name
[3] Note that in 2017 the numbers of images were 311, 165, and 296, respectively

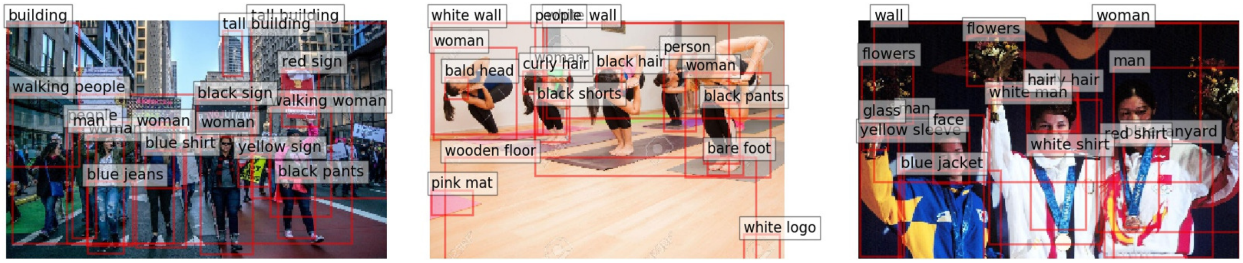

Figure 3: Examples of visual attentions. Left: negative. Middle: neutral. Right: Positive.

Table 1: Performance of each model on the validation set.

Models	Acc	+LargeMargin
VGG-FACE Model	68.28	70.07
VGG2-Senet-ft-Face Model	68.94	70.55
Inception-V2 Scene Model	67.21	68.09
SE-ResNet-50 Scene Model	68.16	68.38
SE-ResNet-50 Skeleton Model	64.42	65.87
Attention LSTM Model	67.46	–

Table 2: Accuracies of model fusions on the validation set.

Fused Models	Acc
VGG-FACE Model + VGG-FACE Non-positive	71.79
+ VGG2-Senet-ft-Face Model	73.49
+VGG2-Senet-ft-Face Model Non-positive	74.23
+Inception-V2 Scene Model	76.39
+SE-ResNet-50 Scene Model	76.85
+SE-ResNet-50 Skeleton Model	78.36
+Attention LSTM Model	78.98

further improve accuracy. The hybrid network is built by fusing the predictions[4] from individual models. Exhaustive grid search is used to calculate the weights of the predictions of each model. The weights range from 0 to 1, with increments of 0.05 and their sum is constrained to be 1. The resulting weight of a redundant model is 0.

Since face models have high accuracy and low false-positive rate on the positive class, the non-positive predictor, described in [11], is used in combination with the fine-tuned VGG-FACE and VGG2-Senet-ft-Face models[5] so that weights can be separately assigned to positive and non-positive classes during grid search. Table 2 shows the performance gain by adding one model at a time to previously fused models[6].

3.3 Submission Results

The challenge allows 7 submissions in total. For the first submission, we trained models on the training data only and learned the weights of the decision fusion by favoring the highest accuracy on the validation data. Table 3 shows it overfits the testing set. The second submission is a single Inception-V2 model on scene. The 3-4 submissions and 5-7 submissions are trained on the combination of training and validation datasets with and without large-margin softmax loss, respectively. Confusion matrices of submission 1 and 7 are shown in Table 4 and Table 5, respectively.

Table 3: Submission Results

Sub	Training Data	Val	Test
1	Training Set Only	78.98	64.96
2	Training Set Only	67.21	59.81
3	Training + Val	-	64.73
4	Training + Val	-	65.86
5	Training + Val	-	67.59
6	Training + Val	-	67.32
7	Training + Val	-	**68.08**

Table 4: Confusion matrix of submission 1, with overall accuracy being 64.96% on the testing data.

	Neg	Neu	Pos
Neg	472	167	190
Neu	147	510	259
Pos	164	128	974

Table 5: Confusion matrix of submission 7, with overall accuracy being 68.08% on the testing data.

	Neg	Neu	Pos
Neg	559	128	142
Neu	156	529	231
Pos	188	116	962

4 CONCLUSIONS

In this paper, a hybrid network that combines 8 models for group-level emotion recognition in the wild is proposed. To the best of our knowledge, visual attention mechanism is presented and explored for the group emotion recognition problem for the first time in this paper. The overall accuracy of the proposed method, which scored the first place in the EmotiW Group-level Emotion Recognition Sub-challenge, is 68.08% on the test data.

ACKNOWLEDGMENTS

The work is supported by the National Science Foundation under Grant No. 1319598.

[4] By predictions we mean the probabilities of an image belonging to each class.
[5] Positive-only predictor was also applied, but the weight of it ended up being 0 after grid search.
[6] Note that due to the length limitations of the paper, redundant models have been removed and are not mentioned in this paper.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-Up and Top-Down Attention for Image Captioning and VQA. *CoRR* abs/1707.07998 (2017). arXiv:1707.07998 http://arxiv.org/abs/1707.07998

[2] J. Bullington. 2005. Affective computing and emotion recognition systems: the future of biometric surveillance?. In *Proceedings of the 2nd annual conference on Information security curriculum development*. ACM, 95–99.

[3] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. 2017. VGGFace2: A dataset for recognising faces across pose and age. *CoRR* abs/1710.08092 (2017). arXiv:1710.08092 http://arxiv.org/abs/1710.08092

[4] Z. Cao, T. Simon, S. Wei, and Y. Sheikh. 2016. Realtime multi-person 2D pose estimation using part affinity fields. *arXiv preprint arXiv:1611.08050* (2016).

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*.

[6] A. Dhall, A. Asthana, and R. Goecke. 2010. Facial expression based automatic album creation. In *International Conference on Neural Information Processing*. Springer, 485–492.

[7] A. Dhall, R. Goecke, and T. Gedeon. 2015. Automatic group happiness intensity analysis. *IEEE Transactions on Affective Computing* 6, 1 (2015), 13–26.

[8] A. Dhall, J. Joshi, K. Sikka, R. Goecke, and N. Sebe. 2015. The more the merrier: Analysing the affect of a group of people in images. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, Vol. 1. IEEE, 1–8.

[9] Abhinav Dhall, Amanjot Kaur, Roland Goecke, and Tom Gedeon. 2018. EmotiW 2018: Audio-Video, Student Engagement and Group-Level Affect Prediction *(ACM International Conference on Multimodal Interaction 2018 (in press))*. ACM.

[10] I.J. Goodfellow 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] X. Guo, L.F. Polanía, and K.E. Barner. 2017. Group-level emotion recognition using deep models on image scene, faces, and skeletons. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 603–608.

[12] Xin Guo, Luisa F. Polania, and Kenneth E. Barner. 2018. Smile detection in the wild based on transfer learning. (2018).

[13] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. 2016. MS-Celeb-1M: A Dataset and Benchmark for Large Scale Face Recognition. In *European Conference on Computer Vision*. Springer.

[14] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.

[15] S. Hochreiter and J. Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (Nov. 1997), 1735–1780.

[16] Jie Hu, Li Shen, and Gang Sun. 2017. Squeeze-and-Excitation Networks. *CoRR* abs/1709.01507 (2017). arXiv:1709.01507 http://arxiv.org/abs/1709.01507

[17] Xiaohua Huang, Abhinav Dhall, Guoying Zhao, Roland Goecke, and Matti PietikÃďinen. 2015. Riesz-based Volume Local Binary Pattern and A Novel Group Expression Model for Group Happiness Intensity Analysis. In *BMVC*. 1–9.

[18] A. Krizhevsky, I. Sutskever, and G.E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. 1097–1105.

[19] J. Li, S. Roy, J. Feng, and T. 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] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. 2016. Large-Margin Softmax Loss for Convolutional Neural Networks. In *Proceedings of The 33rd International Conference on Machine Learning*. 507–516.

[21] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning.

[22] Volodymyr Mnih, Nicolas Heess, Alex Graves, and koray kavukcuoglu. 2014. Recurrent Models of Visual Attention. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 2204–2212. http://papers.nips.cc/paper/5542-recurrent-models-of-visual-attention.pdf

[23] W. Mou, O. Celiktutan, and H. Gunes. 2015. Group-level arousal and valence recognition in static images: Face, body and context. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 5. IEEE, 1–6.

[24] P.M. Niedenthal and M. Brauer. 2012. Social functionality of human emotion. *Annual review of psychology* 63 (2012), 259–285.

[25] O. M. Parkhi, A. Vedaldi, and A. Zisserman. 2015. Deep Face Recognition. In *British Machine Vision Conference*.

[26] F.E. Pollick, H.M. Paterson, A. Bruderlin, and A.J. Sanford. 2001. Perceiving affect from arm movement. *Cognition* 82, 2 (2001), B51–B61.

[27] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2016. Self-critical Sequence Training for Image Captioning. *CoRR* abs/1612.00563 (2016). arXiv:1612.00563 http://arxiv.org/abs/1612.00563

[28] K. Simonyan and A. Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2818–2826.

[30] L. Tan, K. Zhang, K. Wang, X. Zeng, X. Peng, and Y. Qiao. 2017. Group emotion recognition with individual facial emotion CNNs and global image based CNNs. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 549–552.

[31] S. Tomas, J. Hanbyul, M. Iain, and S. Yaser. 2017. Hand Keypoint Detection in Single Images using Multiview Bootstrapping. In *CVPR*.

[32] T. Vandal, D. McDuff, and R. El Kaliouby. 2015. Event detection: Ultra large-scale clustering of facial expressions. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 1. IEEE, 1–8.

[33] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. 2017. Residual Attention Network for Image Classification. *CoRR* abs/1704.06904 (2017). arXiv:1704.06904 http://arxiv.org/abs/1704.06904

[34] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. 2016. Convolutional pose machines. In *CVPR*.

[35] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan. 2009. Toward practical smile detection. *IEEE transactions on pattern analysis and machine intelligence* 31, 11 (2009), 2106–2111.

[36] J. Wu and J.M. Rehg. 2011. CENTRIST: A Visual Descriptor for Scene Categorization. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 8 (2011), 1489–1501.

[37] Huijuan Xu and Kate Saenko. 2016. Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII*. 451–466.

[38] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. 2015. Stacked Attention Networks for Image Question Answering. *CoRR* abs/1511.02274 (2015). http://arxiv.org/abs/1511.02274

[39] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. 2016. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters* 23, 10 (Oct 2016), 1499–1503. https://doi.org/10.1109/LSP.2016.2603342