# Cascade Attention Networks For Group Emotion Recognition with Face, Body and Image Cues

**Conference Paper** · October 2018

DOI: 10.1145/3242969.3264991

Kai Wang
National University of Singapore
**29** PUBLICATIONS **1,106** CITATIONS
SEE PROFILE

Jianfei Yang
Nanyang Technological University
**120** PUBLICATIONS **2,457** CITATIONS
SEE PROFILE

Debin Meng
Chinese Academy of Sciences
**7** PUBLICATIONS **546** CITATIONS
SEE PROFILE

Kaipeng Zhang
University of Windsor
**16** PUBLICATIONS **8,006** CITATIONS
SEE PROFILE

# Cascade Attention Networks For Group Emotion Recognition with Face, Body and Image Cues

Kai Wang
Shenzhen Key Laboratory of Virtual Reality and Human Interaction Technology
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
P.R. China

Xiaoxing Zeng
Shenzhen Key Laboratory of Virtual Reality and Human Interaction Technology
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
P.R. China

Jianfei Yang
School of Electrical and Electronic Engineering, Nanyang Technological University
Singapore

Debin Meng
Shenzhen Key Laboratory of Virtual Reality and Human Interaction Technology
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
P.R. China

Kaipeng Zhang
National Taiwan University
P.R. China

Xiaojiang Peng*
Shenzhen Key Laboratory of Virtual Reality and Human Interaction Technology
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
P.R. China

Yu Qiao*
Shenzhen Key Laboratory of Virtual Reality and Human Interaction Technology
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
P.R. China

## ABSTRACT

This paper presents our approach for group-level emotion recognition sub-challenge in the EmotiW 2018. The task is to classify an image into one of the group emotions such as positive, negative, and neutral. Our approach mainly exploits three types of visual cues for this task, namely face, body and global image with recent deep networks. Our main contribution is two-fold. First, we introduce body Convolutional Neural Networks (CNNs) into this task based on our previous winner method [18]. Specially, we crop all bodies in an image with the state-of-the-art human pose estimation method and train body CNNs with the image-level labels of group emotions. The body cue captures a full view of an individual. Second, we propose a cascade attention network for the face cue in images. This network exploits the importance of each face in an image to generates a global representation based on all faces. The cascade attention network is not only complementary with other models but also improves the naive average pooling method by about 2%. We finally achieve the second place in this sub-challenge with classification accuracies of 86.9% and 67.48% on the validation set and testing set, respectively.

## CCS CONCEPTS

• **Computing methodologies** → *Image representations*;

## KEYWORDS

Emotion Recognition, Group-level emotion recognition, deep learning, Convolutional Neural Networks, large-margin softmax

**ACM Reference Format:**
Kai Wang, Xiaoxing Zeng, Jianfei Yang, Debin Meng, Kaipeng Zhang, Xiaojiang Peng, and Yu Qiao*. 2018. Cascade Attention Networks For Group Emotion Recognition with Face, Body and Image Cues. In *2018 International Conference on Multimodal Interaction (ICMI '18), October 16–20, 2018, Boulder, CO, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3242969.3264991

*Corresponding author

*ICMI '18, October 16–20, 2018, Boulder, CO, USA*
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5692-3/18/10…$15.00
https://doi.org/10.1145/3242969.3264991

# 1 INTRODUCTION

Facial emotion recognition is a basic yet important problem with wide applications in human-computer interaction, virtual reality, entertainment, etc. The challenge of emotion recognition comes from the pose, illumination, occlusions and inter-person variations. Recently, group-level emotion in images has attracted increasing attention [6], which is even more challenging due to the clutter backgrounds, low-resolution faces and the inconsistency between faces and image label. Group-level emotion refers to use of certain visual cues to classify the image into a tone. Especially, the group emotion task of EmotiW 2018 is to classify the tone as positive, negative and neutral. It is especially useful for social image analysis, group people emotion prediction and risk behavior management.

**Related work**. Huang *et al.* [21] proposed Reisz-Based volume local binary pattern and a continuous conditional random fields model. Mou *et al.* [14] proposed group-level arousal and valence recognition from a view of the face, body and context. Unaiza *et al.* [2] used Hybrid-CNN to infer sentiment of social events images.

In the Group based Emotion Recognition in the Wild (EmotiW) 2016 challenge [6], happiness is measured in level 0 to 5. The winner proposed a scene feature extractor and a series of face feature extractors based LSTM for regression [10] . The second place proposed a bottom-up approach using geometric features and Partial Least Squares regression [19] . The third-place proposed LSTM for Dynamic Emotion and Group Emotion Recognition in the Wild [17].

The organizers refer to global and local image features as top-down and bottom-up components as the baseline. In their work [7], global features contain scene features related to factors external to group members' characteristics while local features contain face expressions, face attributes which related to intrinsic characteristics of the individuals in the group.

However, their proposed method relying on LBQ and PHOG features and CENTRIST, whose capture face representation and scene representation is limited. In the last year, our team (winner) [18] proposed an overall deep group emotion recognition framework that combines deep face level representation and deep scene level representation. Based on our previous work, we introduce two new methods, namely a body based CNNs and a face based cascade attention network.

# 2 OUR APPROACH

## 2.1 System Pipeline

Our system pipeline is shown in Figure 1. It mainly includes three types of CNNs, namely face based CNNs, images based CNNs, and body based CNNs. In particular, we train two types of face based emotion networks, namely the norm individual face CNNs and cascade attention networks. We average the scores of three types of CNNs to predict the final group emotion category.

## 2.2 Face based Emotion CNNs

We find that the facial emotion for each face is one useful cue for prediction of group emotion. For this reason, we utilize two kinds of networks to exploit this part, namely the aligned facial emotion CNN and cascade attention networks. Firstly, we briefly introduce the used face detector,large-margin softmax loss and our cascade attention networks as follows.

**Face detection and alignment**. We use MTCNN [24] to detect faces in the images. MTCNN is a CNN-based face detection method. It contains cascaded CNNs for accelerating and accurate detection and joints face alignment (five facial landmarks detection, i.e. two eyes, two mouth corners, and nose). It builds an image pyramid according to the input images and then feeds them to the following three-stage cascaded framework. The candidate regions are produced in the first stage and refine in the latter two stages, facial landmark location is produced in the third stage. We use five detected facial landmarks to make a similarity transform, then, get the aligned faces for all face based emotion CNNs.

**Large-margin softmax loss** (L-Softmax). L-Softmax [13] is introduced for discriminative learning and can alleviate the overfitting problem. L-Softmax can encourage intra-class compactness and inter-class separability between learned features by angular margin constraint. In the fine-tuning stage, for a face feature $x_i$ , the loss is computed by:

$$L_i = -\texttt{log}\frac{e^{||w_{y_i}||||x_i||\varphi(\theta_{y_i})}}{e^{||w_{y_i}||||x_i||\varphi(\theta_{y_i})} + \sum_{j\neq y} e^{||w_{y_i}||||x_i||\cos\theta_j}} \quad (1)$$

where $y_i$ is the label of $x_i$, $w_{y_i}$ is the weight of j class in a fully-connected layer, and

$$\cos(\theta_j) = \frac{w_j^T x_i}{||w_j||||x_i||}, \quad (2)$$

$$\varphi(\theta) = (-1)^k \cos m\theta - 2k, \theta \in [\frac{k\pi}{m}, \frac{(k+1)\pi}{m}], \quad (3)$$

where m is the pre-set angular margin constraint,$k$ is an integer and $k \in [0, m-1]$.

### 2.2.1 *Individual Facial Emotion CNNs.*

For individual facial emotion CNNs, we first transfer our previous facial emotion model, i.e. ResNet64 [18], into the task. To reduce overfit and enhance the generalization, we pre-train it using face recognition dataset and then training in EmotiW 2018 training dataset using L-softmax loss.

Besides ResNet64, we also use**VGG-FACE** [16], **ResNet34** [8], and **SE-net154** [9]. We pre-train those CNNs on the FERPlus expression dataset [3] and then fine-tune them in EmotiW 2018 training dataset using softmax loss.

### 2.2.2 *Aggregated face representation with Cascade Attention Network (CAN).*

For individual facial emotion CNNs, we usually feed every face in an image and average the scores for image-level prediction. We realize that some faces can be irrelated to the image label in practice. To this end, we propose the cascade attention network to aggregate all faces which can capture the importance of each face in an image. Our CAN is shown in Figure 2. We randomly choose several faces from all detected faces in an image and then input them into a feature extraction network, i.e. ResNet18. Suppose we get a feature $P_i$ after the final global average pooling layer for each face in an image, then we feed it into a fully-connected (FC) layer with one-dimension output $\mu_i$ which we expect to capture the importance of this face. We call the $\mu_i$ as self-attention. With $\mu_i$, we then calculate a weighted mean $P_m$ which indicates the coarse face based global representation for an image. Afterward,

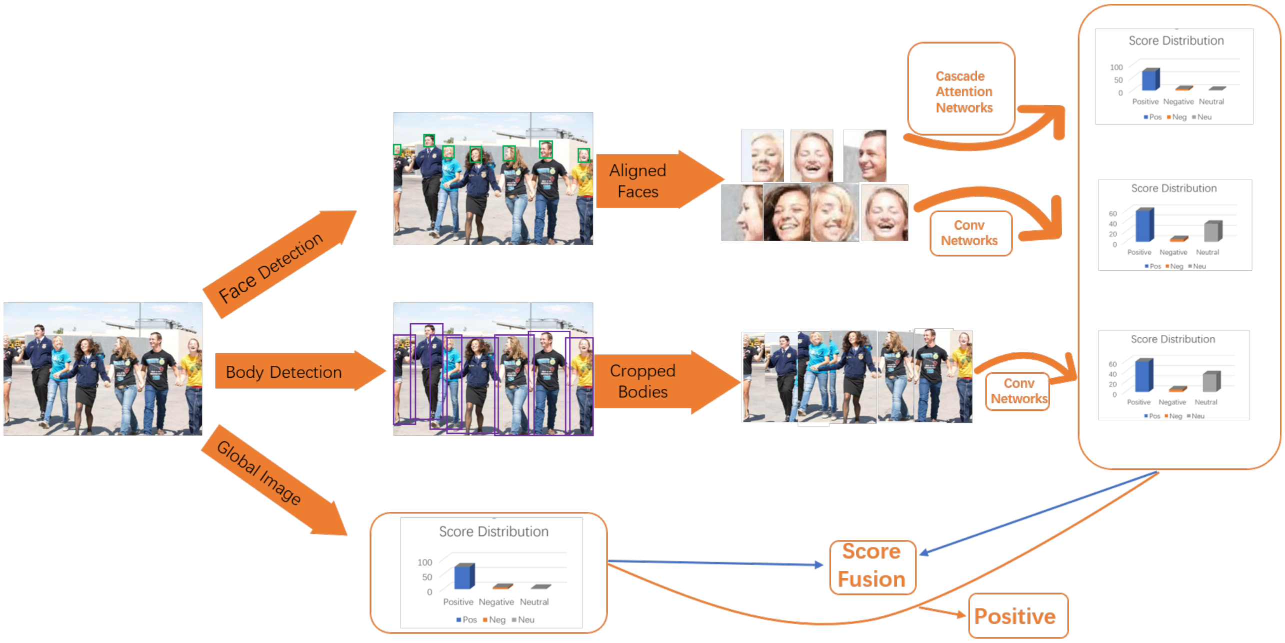

Figure 1: The system pipeline of our approach. It contains three kinds of CNNs, namely the face based emotion CNNs, the global image based CNNs and the body based CNNs. Particularly, we train two types of face-based emotion networks, namely the norm individual face CNNs and cascade attention networks. The final prediction is made by averaging all the scores of CNNs from all faces, the global image and all bodies.

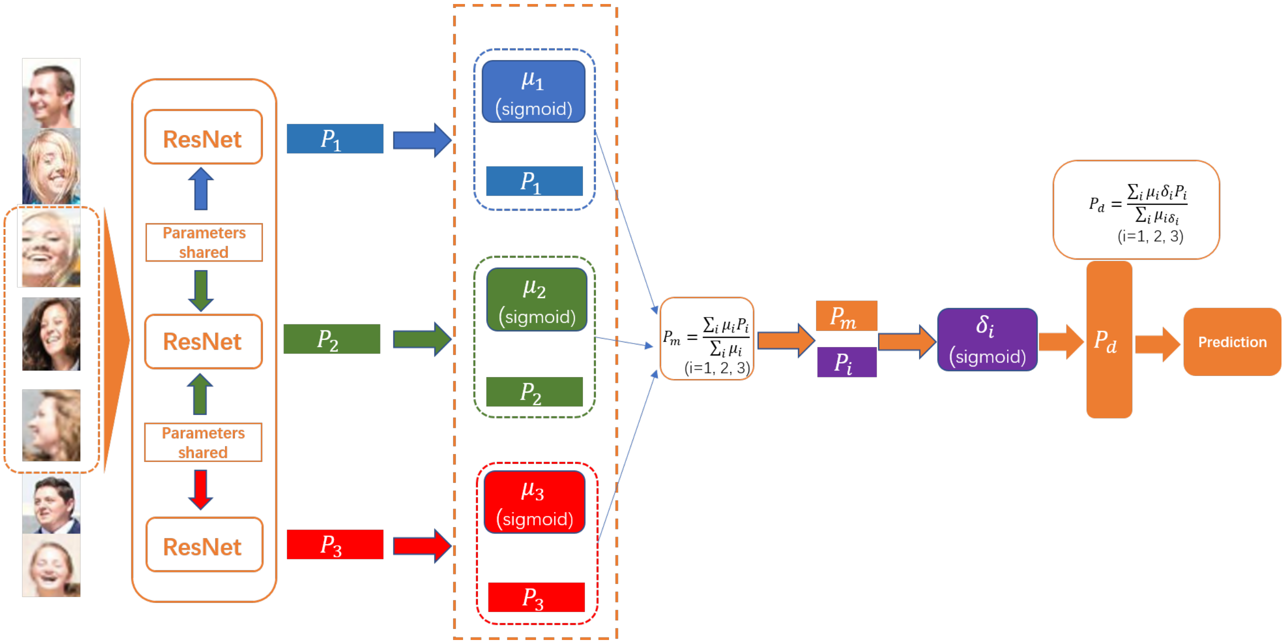

Figure 2: The structure of our cascade attention networks. It contains two stages,one is calculate each face contribution for the whole image, the other is use these contributions to predict the result of the whole image.

we concatenate $P_m$ and $P_i$ and feed it into another FC layer with one-dimension output $\delta_i$ which indicates the relationship between $P_i$ and $P_m$. We call $\delta_i$ as contextual attention since this weight depends on the face descriptor and face based global descriptor. With this cascade attention scheme, we obtain the final aggregated face based global representation $P$ as

$$P = \frac{\sum_i \delta_i \mu_i P_i}{\sum_i \delta_i \mu_i}. \tag{4}$$

## 2.3 Global Image Based CNNs

A global image can also provide an important clue for group-level emotion prediction. It is easy to know an image taken from a wedding party is more likely to be positive and from a meeting room is more likely neutral. Inspire of some advanced applications of image [11]and video [22].Therefore, we train global image based CNNs with three recent state-of-the-art network architectures, namely **VGG19** [16], **ResNet101** [8], and **SE-net154** [9].

## 2.4 Body Based CNNs

Body information can be helpful from previous work [14]. We argue that the body region which also includes the human face can be view as a large contextual region of a face. To extract all the human bodies, we use **OPENPOSE** [4, 15, 20] to get the human poses. Each pose is composed of 18 key points. We crop the human body as the maximum outer rectangle region.

Finally, we train **ResNet101** [8] and **SE-net154** [9] for human bodies. Both networks are pretrained on ImageNet and then fine-tuned on EmotiW 2018 training dataset.

# 3 EXPERIMENTS

In this part, we first present the EmotiW 2018 dataset and the implementation details, and then show the evaluation of different kinds of CNNs, finally, we present our submissions.

## 3.1 Dataset

This task is to classify a group's perceived emotion into Positive, Neutral or Negative. The dataset used for group-level emotion recognition of EmotiW 2018 [1] is collected from the internet. The data is distributed into three sets: Train, Validation, and Test. The three sets contain 9136, 4346, and 3011 images, respectively.

## 3.2 Implementation Details

For the individual facial emotion CNN model, we take the corresponding model (i.e., ResNet64) from [12] which is trained on the Webface dataset [23] and used an Angular Softmax loss, and then finetune it on the ExpW facial expression dataset [25] with L-softmax, finally we finetune it on the EmotiW 2017 dataset. We use the batch size of 60 and set $m$ (i.e., the angular margin constraint, see Eq. (3)) to 4. In fine-tune step, the start learning rate is 0.0015, and times 0.1 at 6k, 8k and 10k iterations. A complete training is stopped at 12k iterations. We also use the same initial learning rate to train**VGG-FACE** [16], **ResNet34** [8], and **SE-net154**.

For the cascade attention networks, we first pre-train the ResNet18 model on FERPlus expression dataset [3]. The learning rate is initialized by 0.01, and and divided by 10 for every 3 epochs. We stop training at the eighth epoch. Then, we fine-tune it on EmotiW 2017 dataset with a learning rate of 0.001 and the same strategy as training. On both steps, we set a dropout of 0.9 after the average pooling of ResNet18. Finally, we fix all the convolutional layer and train the cascade attention networks. To construct the input of CAN, we randomly select 3 faces for each image in the training step. Inspire by [18], we keep all the faces for emotion recognition in the test step.

For the global image based CNN models, the VGG19 is pre-trained on the Places dataset [26], the ResNet101 is pre-trained on the ImageNet dataset [5], and the SE-net154 is pre-trained on the ImageNet dataset. In the fine-tune step, we fix all the batch normalization layers and set dropout of 0.9 after the average pooling layer of ResNet101, drop out of 0.5 for both the FC6 and the FC7 layer of VGG19, and drop out of 0.2 after the average pooling layer of SE-net154. We resize all the images to have a minimum size of 256, and randomly crop 224 × 224 regions for fine-tuning. In test step, we average the scores of the center crop images.

For the body based CNN models, the ResNet101 and SE-net154 are pre-trained on the ImageNet dataset. We use the same training strategy as a global image based CNN models. Since our last year article [18], L-SOFTMAX has been discussed with SoftMax. This year, we will no longer compare the differences between them, we also did not use L-softmax in some networks.

## 3.3 Experimental Results

In this section, we evaluate our approaches on both the validation set and test set.

**Evaluation of individual facial emotion CNNs and cascade attention networks** Table 1 shows the results of four types of individual facial emotion CNN models and one cascade attention networks on the EmotiW 2018 validation set. All the models are achieved about the accuracy of 70%. For the aligned faces, we use two kinds of loss, L-Softmax for ResNet64, Softmax for VGG-FACE, ResNet34, ResNet18, and SE-net154. Particularly, we use cascade attention networks improve the performance about 2% than the baseline of resnet18. In another word, our cascade attention networks is effective when we train with the facial data. We also find the cascade attention networks has some complementariness with facial emotion CNNs. therefore, we can get a better accuracy when we fuse the cascade attention networks and the facial emotion CNNs.

Although the performance of those models are not very well, we find those models are complementary to the global image-based model and bodies image-based model, see Table 4.

**Evaluation of global image based CNNs**. Table 2 presents three types of global image based CNN models on the EmotiW 2018 validation set. SE-net154 is a state-or-the-art networks in recognition, it gives each channel a predict weight,then get the feature recalibration. It can be flexibly combined with a variety of neural networks, such as SE-ResNet and SE-Inception. SE-net is also the champion of ILSVRC 2017.

**Evaluation of body-based CNNs**.

Table 3 presents the results of two types of bodies based models on EmotiW 2018 validation set. We try to use ResNet101 and SE-net154, first, cropped images are feed into ResNet101 and SE-net154, then, get each body image score, finally, average all the scores as

**Table 1: Results of individual facial emotion CNN models and Cascade Attention Networks on the EmotiW validation set.**

| | Aligned Faces | | Cascade Attention Networks |
|---|---|---|---|
| | Softmax | L-Softmax | Softmax |
| ResNet64 | - | 70.5 | - |
| VGG-FACE | 72.5 | - | - |
| ResNet34 | 72.3 | - | - |
| ResNet18 | 69.5 | - | 71.9 |
| SE-net154 | 70.6 | - | - |

**Table 2: Results of global image based CNN models on the EmotiW validation set.**

| | VGG19 | | ResNet101 | | SE-net154 | |
|---|---|---|---|---|---|---|
| | Softmax | L-Softmax | Softmax | L-Softmax | Softmax | L-Softmax |
| Accuracy | 67.2 | **73.0** | 72.8 | | **74.9** | |

**Table 3: Results of body based CNN models on the EmotiW validation set.**

| | ResNet101 | SE-net154 |
|---|---|---|
| | Softmax | Softmax |
| Accuracy | 69.05 | **70.5** |

the whole image prediction. The performance in SE-net154 better than ResNet101.

**Evaluation of score combinations**. Table 4 shows the combinations of different models with varied weights. In Table 4, the number 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 represent VGG-FACE, ResNet34, ResNet101(body), ResNet64, ResNet101(global), SE-net154(body), VGG19, SE-net154(face), SE-NET154(global), and ResNet18, respectively. '*' denotes average score fusion while '**' weighted score fusion.

**Table 4: Results of different combinations on Validation set**

| | Acc. on validation set (%) |
|---|---|
| 1, 2, 5, 6, 7, 8, 9, 10 (*) | 86.7 |
| 1, 2, 4, 5, 6, 7, 8, 9, 10 (*) | 86.53 |
| 1, 2, 4, 5, 6, 7, 9 (**) | 86.26 |
| 1, 2, 4, 5, 6, 7, 9, 10 (**) | 86.53 |
| 1, 2, 3, 4, 6, 7, 9, 10 (*) | **86.9** |
| 1, 2, 4, 5, 7, 8, 9, 10 (*) | 86.28 |

We sumarize our 7 submissions as follows. Table 5 shows the results on the validation set and test set. The submissions mostly include all types of information from face, body, and global image. The best result on test set comes from run (1) which is a simple score average fusion of three kinds of CNNs. From the accuracy comparison between validation set and test set, we find there is a large perfermance gap which indicates the data distribution is very different between them.

(1) face: VGG-FACE, ResNet34, SE-net154, ResNet18(CAN); global: ResNet101, VGG19, SE-NET154; body: SE-net154 ***(average)***
(2) face: VGG-FACE, ResNet34, ResNet64, SE-net154, ResNet18(CAN); global: ResNet101, SE-net154, VGG19; body: SE-net154 ***(average)***
(3) face: VGG-FACE(1.0), ResNet34(1.0), ResNet64(1.0); global: ResNet101(1.2), SE-net154(1.0), VGG19(1.2); body: SE-NET154(1.0)
(4) face: VGG-FACE(1.0), ResNet34(1.0), ResNet64(1.0), ResNet18(CAN,1.0); global: ResNet101(1.2), SE-net154(1.5), VGG19(1.2); body: SE-NET154(1.0)
(5) face: VGG-FACE, ResNet34, ResNet64, ResNet18(CAN); global: ResNet101, VGG19, SE-net; body:SE-net154; ***(average)***
(6) face: VGG-FACE, ResNet34, SE-net154, ResNet64, ResNet18(CAN); global: ResNet101, VGG19, SE-NET154; ***(average)***
(7) face: VGG-FACE(1.0), ResNet34(1.0), ResNet64(1.0), SE-net154(1.0), ResNet18(CAN,1.2); global: ResNet101(1.0), SE-net154(1.2), VGG19(1.2); body: SE-NET154(1.0) ***(only train on the training set)***

**Table 5: Results of our final submissions.**

| Runs | Validation | Test | | | |
|---|---|---|---|---|---|
| | Overall | Positive | Neutral | Negative | Overall |
| 1 | 86.7 | 77.09 | 55.89 | 65.62 | **67.48** |
| 2 | 86.53 | 77.01 | 56.33 | 64.41 | 67.25 |
| 3 | 86.26 | 75.27 | 56.00 | 63.93 | 66.29 |
| 4 | 86.53 | 76.46 | 56.76 | 63.20 | 66.82 |
| 5 | 86.9 | 76.93 | 57.20 | 63.44 | 67.22 |
| 6 | 86.28 | 77.25 | 55.89 | 65.37 | **67.48** |
| 7 | 86.9 | 75.90 | 53.71 | 64.17 | 65.92 |

The results of the seven submissions are hierarchical, taking into account the integration of multiple features, taking into account the different weights given to each feature, and also discarding some features. This can better prevent overfitting. These weights are determined by the performance of the model on the validation set and a large number of experiments.

## 4 CONCLUSIONS

We present our approach for the group-level emotion recognition in the Emotion Recognition in the Wild Challenge 2018. We propose three types of Convolutional Neutral Networks(CNNs), namely face based emotion CNNs, global image based CNNs and body based CNNs. Particularly, we introduce an effective attention networks for face based emotion CNNs. We utilize a large-margin softmax loss for discriminative learning, and explore different fusion strategies. Experimental results show the effectiveness of our approach, and we win second place of the group-level emotion recognition task.

## ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (U1613211, 61502152), Shenzhen Basic Research

Program (JCYJ20150925163005055, JCYJ20170818164704758), and International Partnership Program of Chinese Academy of Sciences (172644KYSB20160033,172644KYSB20150019).

## REFERENCES

[1] Roland Goecke Abhinav Dhall, Amanjot Kaur and Tom Gedeon. 2018. EmotiW 2018: Audio-Video, Student Engagement and Group-Level Affect Prediction. In *ICMI*. ACM.

[2] Unaiza Ahsan, Munmun De Choudhury, and Irfan Essa. 2017. Towards using visual attributes to infer image sentiment of social events. In *International Joint Conference on Neural Networks (IJCNN)*. 1372–1379.

[3] Emad Barsoum, Cha Zhang, Cristian Canton-Ferrer, and Zhengyou Zhang. 2016. Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution. *CoRR* abs/1608.01041 (2016). http://arxiv.org/abs/1608.01041

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*. IEEE, 248–255.

[6] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. 2016. Emotiw 2016: Video and group-level emotion recognition challenges. In *ICMI*. 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 *Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 1. 1–8.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385 (2015). http://arxiv.org/abs/1512.03385

[9] 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

[10] Jianshu Li, Sujoy Roy, Jiashi Feng, and Terence Sim. 2016. Happiness level prediction with sequential inputs via multiple regressions. In *ICMI*. ACM, 487–493.

[11] Zheng Li, Jianfei Yang, Juan Zha, Chang-Dong Wang, and Weishi Zheng. 2016. Online visual tracking via correlation filter with convolutional networks. In *Visual Communications and Image Processing (VCIP), 2016*. IEEE, 1–4.

[12] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. 2017. SphereFace: Deep Hypersphere Embedding for Face Recognition. *CoRR* abs/1704.08063 (2017). http://arxiv.org/abs/1704.08063

[13] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. 2016. Large-Margin Softmax Loss for Convolutional Neural Networks.. In *ICML*. 507–516.

[14] Wenxuan Mou, Oya Celiktutan, and Hatice Gunes. 2015. Group-level arousal and valence recognition in static images: Face, body and context. In *Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 5. 1–6.

[15] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. 2017. Hand Keypoint Detection in Single Images using Multiview Bootstrapping. In *CVPR*.

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

[17] Bo Sun, Qinglan Wei, Liandong Li, Qihua Xu, Jun He, and Lejun Yu. 2016. LSTM for dynamic emotion and group emotion recognition in the wild. In *ICMI*. ACM, 451–457.

[18] Lianzhi Tan, Kaipeng Zhang, Kai Wang, Xiaoxing Zeng, Xiaojiang Peng, and Yu 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 (ICMI 2017)*. ACM, New York, NY, USA, 549–552. https://doi.org/10.1145/3136755.3143008

[19] Vassilios Vonikakis, Yasin Yazici, Viet Dung Nguyen, and Stefan Winkler. 2016. Group happiness assessment using geometric features and dataset balancing. In *ICMI*. ACM, 479–486.

[20] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional pose machines. In *CVPR*.

[21] Guoying Zhao Roland Goecke Xiaohua Huang, Abhinav Dhall and Matti Pietikäinen. 2015. Riesz-based Volume Local Binary Pattern and A Novel Group Expression Model for Group Happiness Intensity Analysis. In *BMVC*. 1–8.

[22] Jianfei Yang, , Kai Wang, Xiaojiang Peng, and Yu Qiao. 2018. Deep Recurrent Multi-instance Learning with Spatio-temporal Features for Engagement Intensity Prediction. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction (in press)*. ACM.

[23] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. 2014. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923* (2014).

[24] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.

[25] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2016. From Facial Expression Recognition to Interpersonal Relation Prediction. *CoRR* abs/1609.06426 (2016). http://arxiv.org/abs/1609.06426

[26] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *NIPS*. 487–495.