Group Emotion Recognition with Individual Facial Emotion CNNs and Global Image Based CNNs

Lianzhi Tan
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
P.R. China

Kaipeng Zhang
National Taiwan University
Taiwan, P.R. China

Kai Wang
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
P.R. China

Xiaoxing Zeng
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
P.R. China

Xiaojiang Peng*
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
P.R. China

Yu Qiao
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
P.R. China

ABSTRACT

This paper presents our approach for group-level emotion recognition in the Emotion Recognition in the Wild Challenge 2017. The task is to classify an image into one of the group emotion such as positive, neutral or negative. Our approach is based on two types of Convolutional Neural Networks (CNNs), namely individual facial emotion CNNs and global image based CNNs. For the individual facial emotion CNNs, we first extract all the faces in an image, and assign the image label to all faces for training. In particular, we utilize a large-margin softmax loss for discriminative learning and we train two CNNs on both aligned and non-aligned faces. For the global image based CNNs, we compare several recent state-of-the-art network structures and data augmentation strategies to boost performance. For a test image, we average the scores from all faces and the image to predict the final group emotion category. We win the challenge with accuracies 83.9% and 80.9% on the validation set and testing set respectively, which improve the baseline results by about 30%.

CCS CONCEPTS

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

KEYWORDS

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

ACM Reference Format:
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 19th ACM International Conference on Multimodal Interaction (ICMI'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3136755.3143008

*Corresponding author

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

ACM ISBN 978-1-4503-5543-8/17/11…$15.00
https://doi.org/10.1145/3136755.3143008

1 INTRODUCTION

Facial emotion recognition has wide applications in human-computer interaction, virtual reality, and entertainment. It is a challenging problem due to the variance of actors, large head poses, illumination, occlusion, etc. Recently, group-level emotion recognition in images has attracted increasing attention [5], which is even more challenging because of the clutter backgrounds and low-resolution faces. Group-level emotion refers to the moods, emotions and dispositional affects of a number of people. The task in the paper is to classify a group's perceived emotion as positive, neutral or negative. It is especially useful for social image analysis and user emotion prediction.

Related work. Huang *et al.* [16] proposed Reisz-Based volume local binary pattern and a continuous conditional random fields model. Mou *et al.* [12] proposed group-level arousal and valence recognition from view of face, body and context. Unaiza *et al.* [2] used Hybrid-CNN to infer image sentiment of social events. In the Group based Emotion Recognition in the Wild (EmotiW) 2016 challenge [5], 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 [9] . The second proposed a bottom-up approach using geometric features and Partial Least Squares regression [15] . The third proposed LSTM for Dynamic Emotion and Group Emotion Recognition in the Wild [14]. The organizers refer to global and local image features as top-down and bottom-up components as baseline. In their work [6], 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. Therefore, we propose an overall deep group emotion recognition framework that combines deep face level representation and deep scene level representation to infer group emotion.

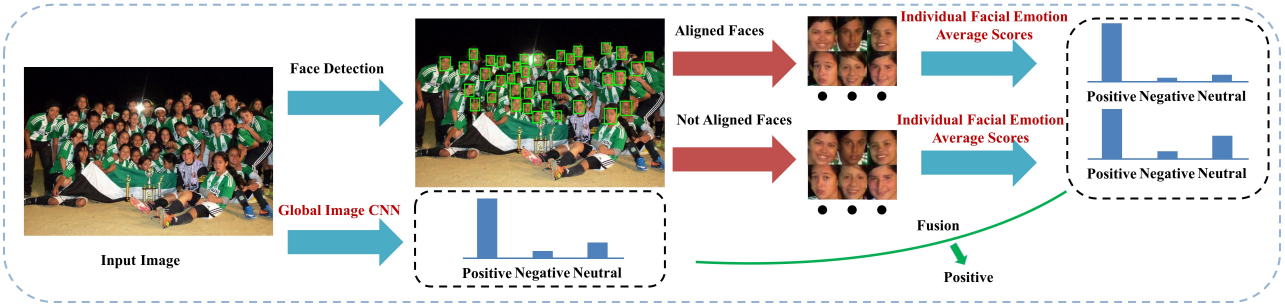

Figure 1: The system pipeline of our approach. It contains two kinds of CNNs, namely the individual facial emotion CNNs and the global image based CNNs. The final prediction is made by averaging all the scores of CNNs from all faces and the image.

2 OUR APPROACH

2.1 System Pipeline

Our system pipeline is shown in Figure 1. It contains two types of convolutional neural networks, one based on images and the other based on faces. In particular, we train two CNNs based on faces with alignment and without alignment. We fuse the scores of CNNs from all faces and the image to predict the final group emotion category.

2.2 Individual Facial Emotion CNNs

We realize that the facial emotion for each face is the most important cue for group emotion recognition. To this end, we utilize two complementary CNNs to exploit this part, namely the aligned facial emotion CNN and non-aligned facial emotion CNN. We first briefly introduce the used face detector and the large-margin softmax loss as following.

Face Detection. We use MTCNN [18] to detect faces in the images. MTCNN is a CNN-based face detection method. It uses three cascaded CNNs for fast and accurate detection and joints face alignment (five facial landmarks detection, i.e. two eyes, two mouth corners and nose) and face detection learning. It builds an image pyramid according to the input images and then feed them to the following three-stage cascaded framework. The candidate regions are produced in the first stage and refine in the latter two stages. The final detection results and according facial landmark location are produced in the third stage.

Large-margin softmax loss (L-Softmax). L-Softmax [11] 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, for a face feature x_i , the loss is computed by:

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

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

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

Table 1: The architecture of our aligned facial emotion CNN.

Type	Filter Size/Stride	Output Size
Conv1.x	3 x 3/2	48 x 56 x 64
	$\begin{bmatrix} 3\times3/1 \\ 3\times3/1 \end{bmatrix} \times 3$	48 x 56 x 64
Conv2.x	3 x 3/2	24 x 28 x 128
	$\begin{bmatrix} 3\times3/1 \\ 3\times3/1 \end{bmatrix} \times 8$	24 x 28 x 128
Conv3.x	3 x 3/2	12 x 14 x 256
	$\begin{bmatrix} 3\times3/1 \\ 3\times3/1 \end{bmatrix} \times 16$	12 x 14 x 256
Conv4.x	3 x 3/2	6 x 7 x 512
	$\begin{bmatrix} 3\times3/1 \\ 3\times3/1 \end{bmatrix} \times 3$	6 x 7 x 512
FC1_bn	512	512
FC2	3	3

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

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

2.2.1 Aligned Facial Emotion CNN. Talbe 1 shows the architecture of our aligned facial emotion CNN. Ths model takes 96×112 RGB aligned face images as input, feeds them into four stages of convolutional layers and two fully-connected layers. We use similarity transform based on five detected facial landmarks to align the faces. To alleviate the overfitting and enhance the generalization, we pre-train it using face recognition dataset and then fine-tune in EmotiW 2017 training dataset using L-softmax loss.

2.2.2 Non-aligned Facial Emotion CNN. Following [9], we train a non-aligned facial emotion CNN which is complementary to the aligned facial emotion CNN in our case. This model is based on the ResNet34 [7], and takes 48×48 gray face images as inputs.

Inspired by [9], we pre-train the model on the FERPlus expression dataset [3] where all the faces are cropped to 48×48. Some samples of this dataset are shown in Figure 2. We use MTCNN to detect faces and crop the faces to 48×48.

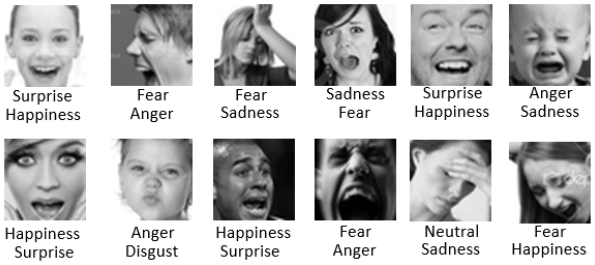

Figure 2: Some samples of the FERPlus dataset.

2.3 Global Image Based CNNs

In some sense, the global feature also refects the group-level emotion. For example, an image taken from a wedding party is likely to be positive and from a funeral negative. To this end, we train global image based CNNs with sevaral state-of-the-art network architectures, namely **VGG19** [13], **BN-Inception** [8], and **ResNet101** [7].

3 EXPERIMENTS

In this section, we first present the EmotiW 2017 dataset [1] and the implementation details, and then show the evaluations of the metioned CNNs, finnaly we evaluate socre combinations.

3.1 Dataset

The dataset used for group-level emotion recognition of EmotiW 2017 [1] is collected from web images, and consists of three sets, namely training, validation and testing set. The three sets contain 3630, 2068, and 772 images respectively.

3.2 Implementation Details

For the aligned facial emotion CNN model, we take the corresponding model (i.e., ResNet64) from [10] which is trained on the Webface dataset [17] using an Angular Softmax loss, and then finetune it on the ExpW facial expression dataset [19] 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 finetuning step, the learning rate is started from 0.01, and divided by 10 at the 1.2K, 1.8K iterations. A complete training is finished at 2.2K iterations.

For the non-aligned facial emotion CNN model, we first pre-train the ResNet34 model from scratch on FERPlus expression dataset [3]. The learning rate is initilized by 0.01, and divided by 10 every 4 epochs. We stop training at the tenth epoch. After we finetune it on EmotiW 2017 dataset with a learning rate 0.001 and the same strategy as training. On both steps, we set a dropout of 0.9 after the average pooling of ResNet34.

In test step of facial emotion CNNs, we remove small faces since they are confused for emotion recognition. In particular, we remove those faces whose minimum sides are less than 24. If there is no faces left, we keep the largest 3 faces.

For global image based CNN models, the VGG19 is pretrained on the Places dataset [20], and the others are pre-trained on the ImageNet dataset [4]. In the finetuning step, we fix all the batch normalization layers and set a dropout of 0.9 after the average pooling of both ResNet101 and BN-Inception models, and a dropout of 0.5 for both the FC6 layer and the FC7 layer of VGG19. We scale all the images to have a minimum side of 256, and randomly crop 224×224 regions with random horizontal flipping for finetuning. In test step, we average the scores of the center crop and its horizontal flipping.

Table 2: Results of individual facial emotion CNN models on the EmotiW validation set.

	Aligned(96×112)		Non-aligned (48×48)
	Softmax	L-Softmax	Softmax
rm small faces	73.8	79.7	70.73
keep all faces	74.1	80.19	69.97

Table 3: Results of global image based CNN models on the EmotiW validation set.

	VGG19		BN-Inception		ResNet101	
	Softmax	L-Softmax	Softmax	L-Softmax	Softmax	L-Softmax
+extra data	-	73.2	-	67.3	**74.7**	73.2
-	67.2	70.6	60.54	65.3	71.04	-

Data collection. In addition the above metioned data augmentation strategies, we also search similar images using the training and validation set from internet. We apply the "search by image" engin of Baidu and Google, and carefully select about 2K images for each class. *Those extra images are only used in the global image based CNN models.* We will make it available at mmlab.siat.ac.cn.

3.3 Experimental Results

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

Evaluation of individual facial emotion CNNs. Table 2 shows the results of individual facial emotion CNN models on the EmotiW validation set. L-Softmax boosts performance for aligned facial emotion CNN with a large margin. Removing small faces helps slightly for the non-aligned model. We fix removing small faces for the non-aligned model while keepingall faces for the aligned model later in this paper unless other statement. Comparing both facial emotion CNNs, we observe two good practice namely i) pre-training on large dataset with deeper networks, and ii) large resolution inputs. Although the low performance of the non-aligned model, we find it is complementary to the global image based model and the aligned model, see Table 4.

Evaluation of global image based CNNs. Table 3 presents the results of three global image based CNN models on the validation set. L-Softmax improves the VGG19 and BN-Inception models significantly while degrades the ResNet101 slightly. Adding more data improves all the models significantly. The best performance is observed from the ResNet101 with softmax loss and our extra data.

Evaluation of score combinations. Table 4 shows the combinations of different models with varied weights. Averaging the scores of the metioned three models degrades performance on validation set. By cross-validation, we find the best combination

weights for those models are 3.0, 0.5, and 0.5, see the 4th row in Table 4.

Table 4: Results of different combinations on Validation set

	Acc. on validation set (%)
Aligned	80.2
Aligned (3.0) + Global (0.5)	81.0
Aligned (3.0) + Global (0.5) + Non-aligned (0.5)	82.7
Aligned (1.0) + Global (1.0) + Non-aligned (1.0)	78.4

We sumarize our 7 submissions as follows. Table 5 shows the corresponding results on the validation set and test set. We find the best combination weights for the three models are 3.2, 0.8, and 0.6 on the validation set when adding our collected data. The align facial emotion model in the 3rd, 4th, and 7th runs is trained only on training set. The main reason for this model is that the training step with L-Softmax loss is not stable.

(1) Aligned facial emotion model (3.0) +Global image based ResNet101 (0.5) + Non-aligned facial emotion model (0.5)
(2) Aligned facial emotion model (3.0) + Global image based VGG19 (0.5) + Non-aligned facial emotion model (0.5)
(3) Aligned facial emotion model *trained only on trainset* (3.0) + Global image based ResNet101 (0.5) + Non-aligned facial emotion model (0.5)
(4) Aligned facial emotion model *trained only on trainset* (3.0) + Global image based VGG19 (0.5) + Non-aligned facial emotion model (0.5)
(5) Aligned facial emotion model (3.2) + Global image based ResNet101 (0.8,+extra data) + Non-aligned facial emotion model (0.6)
(6) Aligned facial emotion model (3.2) + Global image based VGG19 (0.8,+extra data) + Non-aligned facial emotion model (0.6)
(7) Aligned facial emotion model *trained only on trainset* (3.2) + Global image based ResNet101 (0.8,+extra data) + Non-aligned facial emotion model (0.6)

Table 5: Results of different combinations for submissions.

Runs	Validation	Test			
	Overall	Positive	Neutral	Negative	Overall
1	82.7	82.0	57.6	76.7	79.9
2	81.7	82.3	57.0	74.7	79.1
3	-	85.2	63.6	67.6	78.9
4	-	85.2	64.84	66.6	78.8
5	83.7	83.9	58. 8	76.4	**80.9**
6	83.8	83.9	58.2	74.0	79.8
7	-	85.5	59.4	70.0	79.1

4 CONCLUSIONS

We presented our approach in this paper for the group-level emotion recognition in the Emotion Recognition in the Wild Challenge 2017. We explored two types of Convolutional Neural Networks (CNNs), namely individual facial emotion CNNs and global image based CNNs. We expanded the training data by searching similar images on the web, utilized a large-margin softmax loss for discriminative learning, and explored different combinations. Experimental results indicate the effectiveness of our approach, and we win the group-level emotion recognition task.

ACKNOWLEDGMENTS

This work was partly supported by the National Natural Science Foundation of China (U1613211, 61502152), and External Cooperation Program of BIC Chinese Academy of Sciences (172644KYSB2016 0033).

REFERENCES

[1] Shreya Ghosh Jyoti Joshi Jesse Hoey Abhinav Dhall, Roland Goecke and Tom Gedeon. 2017. From Individual to Group-level Emotion Recognition: EmotiW 5.0. 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] 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.
[5] 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.
[6] 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.
[7] 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
[8] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *CoRR* abs/1502.03167 (2015). http://arxiv.org/abs/1502.03167
[9] Jianshu Li, Sujoy Roy, Jiashi Feng, and Terence Sim. 2016. Happiness level prediction with sequential inputs via multiple regressions. In *ICMI*. ACM, 487–493.
[10] 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
[11] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. 2016. Large-Margin Softmax Loss for Convolutional Neural Networks.. In *ICML*. 507–516.
[12] 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.
[13] 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
[14] 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.
[15] 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.
[16] 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.
[17] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. 2014. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923* (2014).
[18] 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.
[19] 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
[20] 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.