Group-level emotion recognition based on faces, scenes, skeletons features

Conference Paper · January 2020

DOI: 10.1117/12.2557175

Luo Ruiming
Huawei Technologies

Group-level emotion recognition based on faces, scenes, skeletons features

Li, Dejian, Luo, Ruiming, Sun, Shouqian

Dejian Li, Ruiming Luo, Shouqian Sun, "Group-level emotion recognition based on faces, scenes, skeletons features," Proc. SPIE 11373, Eleventh International Conference on Graphics and Image Processing (ICGIP 2019), 1137307 (3 January 2020); doi: 10.1117/12.2557175

Event: Eleventh International Conference on Graphics and Image Processing, 2019, Hangzhou, China

Group-Level Emotion Recognition Based on Faces, Scenes, Skeletons Features

Dejian Li, Ruiming Luo, Shouqian Sun
College of Computer Science and Technology, Zhejiang University
dejianli@zju.edu.cn, joeluo@zju.edu.cn, ssq@zju.edu.cn

ABSTRACT

In this paper, we propose a deep neural network based approach for the group-level emotion recognition in 6th Emotion Recognition in the Wild Challenge (EmotiW 2018). The task of this challenge is to classify a group's perceived emotion as Positive, Neutral or Negative. Like the most of current researchers on visual emotion recognition, we mainly focus on facial, scene and body clues in images. We treat each clue as mono-model feature and apply early fusion method to combine them together. Experimental results show that our proposed method has outperformed the baseline techniques with the overall test accuracy of 62.90%.

Keywords: Image processing, group emotion recognition, deep convolutional network.

1. INTRODUCTION

Human emotion recognition can provide crucial information in many applications such as human-centered computing, behavior analysis and cognitive science. Along with the techniques development and merging of computer, the Internet and telegraphy, a huge amount of images are being produced and uploaded to the Internet everyday. Thus far, the vision community has made an increasing research efforts for understanding high level visual concepts such as objects and scenes in images [1]. However, the problem of affective comprehension for an image has been less widely studied and remains as an open research problem [2]. Group-level emotion recognition (GER) is challenging due to the inter-class similarities among different facial expressions and large intra-class variabilities such as changes in illumination, pose, scene, and expression.

In this paper, we present our method in the EmotiW 2018 GER sub-challenge [3, 4]. This challenge has been successfully held for six years since 2013 and has made great influences in the emotion recognition area. The images in this subchallenge are from the Group Affect Database 2.0 [3]. This year's GER sub-challenge is to classify a group's perceived emotion as Positive, Neutral or Negative. Literature in social psychology suggests that emotions observed from a group of people arise from its "bottom-up" components or its "downtop" components [5]. This implies that group emotions as the sum of its parts. Inspired by this, we propose a multi feature deep learning based approach, which combines top-down or up-bottom components (see Fig. 1). Besides, experiments in [6] show that when using both context and body information, performance of emotion recognition outperforms that of using only context image or only body image. Follow this idea, we split whole group image into three clues: faces, scene and body pose. Most of the recent vision recognition methods are based on deep convolutional neural network (DCNN). Therefore, we build three deep convolutional neural networks for each clue separately. Our proposed method outperforms the baseline approach and achieves an overall accuracy of 62.90 % on the test set.

The rest of the paper is organized as follows. We first provide some backgrounds on image features with an emphasis on recent popular deep neural networks in Section2. And then Section 3 gives the details of the proposed method. Section 4 describes the experimental settings and discusses results. And the final conclusion is given in Section 5.

2. RELATED WORK

Analysis of group emotions in the wild has been an important concern in a wide variety of fields nowadays and is studied by many researchers. In general, the existing methods can be roughly divided into the handcraft feature based approaches and the DCNN feature based approaches. There is also extensive work on modeling affective body expressions, here we only review methods presented during the EmotiW challenges.

Eleventh International Conference on Graphics and Image Processing (ICGIP 2019), edited by
Zhigeng Pan, Xun Wang, Proc. of SPIE Vol. 11373, 1137307 · © 2020 SPIE
CCC code: 0277-786X/20/$21 · doi: 10.1117/12.2557175

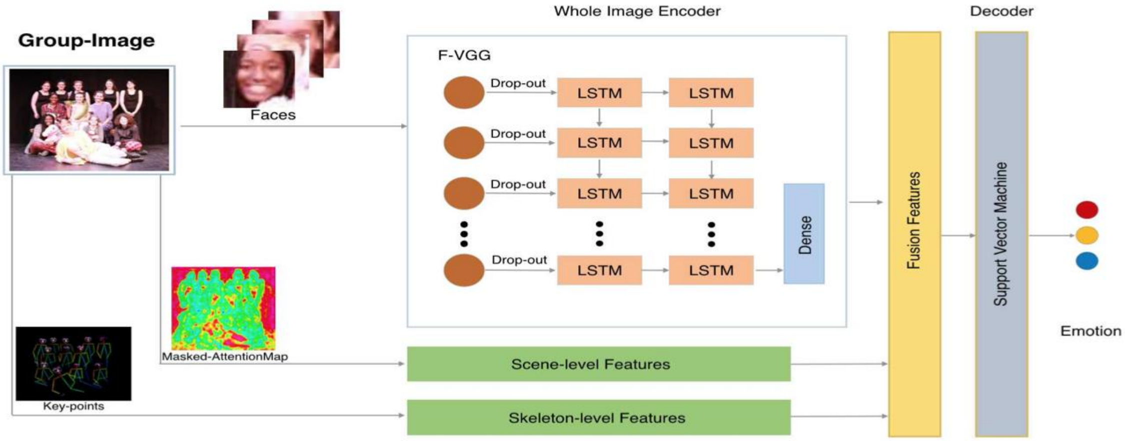

Figure.1. The proposed deep features based framework.

The handcraft feature based approaches extract emotion images based on various kinds of image features. Dhall et al. [3] used a multiple kernel learning based hybrid affect inference method. Vonikakis et al. [7] used geometric features derived from 49 facial points to train a regression model.

Since 2012, AlexNet [8] achieves substantially performance improvements on the ImageNet, deep neural networks became popular for many vision related tasks. Li et al. [9] proposed a recurrent neural network to incorporates both global scene components and local facial components. Cerekovic [10] proposed a novel pipeline which combines finetuned CNN feature extractor and LSTM emotion classification model together. Tan et al. [11] proposed two types of CNNs with large-margin softmax loss, one for individual facial emotion and the other for global image context. Guo et al. [12] used 7 models including global and local features to gird search proper late fusion results. Rassadin et al. [13] used an ensemble of Random Forest classifiers based on VGG-19, VGGFace, Resnet-50 and Xception features. Balaji et al. [14] proposed early fusion method which fused low-level components like SURF and mid-level components like LSD together.

3. PROPOSED APPROACH

3.1 Attention-Based Scene-Level Classification

Attention mechanism has been proposed to improve the performance of language translation in [15], and the idea of this is derive from the human brain. When the human brain receives external information, it often does not process and understand all the information, but only focuses on the part of the interesting areas. This helps to filter out unimportant information and improve the efficiency of information process. We add the attention module for the reason that even though a CNN has great capability in finding different patterns across images, it is not flexible enough to know which regions of the images need more attentions.

The structure of attention module that we explore is derived from the hourglass design [16]. It has been shown worked well in human pose estimation through downsampling in input patch then upsampling back to original to generate a heatmap added on input. Specifically, it consists of two branches: a “trunk” branch T(x) composed of two consecutive residual modules, and an hourglass mask branch M(x). Then final output can be represented as:

$$F(x) = ((1 + M(x)) * T(x) \ . \tag{1}$$

Where hourglass mask M(x) contains two max pooling and two up-sampling operation and the last sigmoid activation to get the final mask attention map. Even though this attention module could be incorporated in any CNN architecture, in this paper we used ResNet [17] as base network architecture to implement. Residual networks have shown great robustness and better performances than other networks when networks gradually go deeper.

3.2 CNN-LSTM Face-level Classification

Simple average of predicted happiness intensities of all the faces in the image alone is inadequate to predict the happiness level of the image. Inspired from encoder-decoder framework, which was proposed in [18] to perform machine translation as a sequence-to-sequence model. We propose a CNN-RNN pipeline which makes prediction relied on all input image features. As the name defines, CNN-RNN pipeline can be divided into two parts. One is the CNN feature extraction part, the other is the RNN feature selection and final prediction part.

3.2.1 FVGG: VGG Fine-tuning Model

Pre-trained ImageNet models and transfer learning have found significant applications in many areas [19]. This is because that the low level convolutional layers extract similar features across different recognition tasks, which means the learned filters can be transferred. VGG-16 net is employed as it is one of the very effective CNNs, we follow original network architecture with only changing neural units of last fully connected layer into task specific number. The whole fine-tune has two process, first trained on AffectNet database, which contains more than 1,000,000 facial images from the Internet divided into eight emotion categories, and to reflect exact human feeling of emotion, we only use manually labeled part facial expression data. Then the net is further fine-tuned in the Group Affect Database 3.0 (GAF-3) cropped faces. We keep the parameters of the first three groups of convolutional layers unchanged and update the rest of the layers during training. Note that we make the cropped faces label to keep consistent with the whole scene label.

3.2.2 Train LSTM Network

Recurrent neural network (RNN) is able to recognize patterns in sequences of data, and as a variation of RNN, the Long Short-Term Memory (LSTM) is widely used these years due to its improvements in dealing with gradient vanishing problem. It can selectively memorize the previous state and feed that to the next neural node. In this group-facial expression LSTM model, the model takes a sequence of faces fc6's features extracted by the fine-tuned VGG-16 model, experiments showed that stacked of two LSTM layers with 50, 100 nodes in first layer and second layer achieves best validation performance. Dropout is used between LSTM and dense layers during training process. In the end, a softmax layer predicts the label of the whole group-level image.

3.3 RtPose Skeleton-level Classification

There is an extensive literature showing that body pose is as informative as facial expression as an affective channel [20]. Therefore, body feature could be an important clue in GER. We use the open-sourced toolkit RtPose [21], which could jointly detect multi human body, hand and facial keypoints on a single image (Fig. 2). We make that the extracted images are of the same size as the original images, therefore, there is zero information losing of the relative position of key points. And for classify model, different architectures of DenseNet (121,169,201) are fine-tuned on those extracted skeleton images. Experimental results show that skeleton feature plays a vital role in classifying overall emotions.

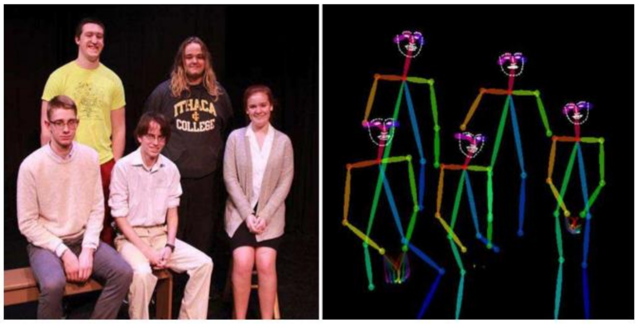

Figure.2. Sample of RtPose skeleton representation. The left is original group image and the right is skeleton results.

4. EXPERIMENTS

4.1 Data Pre-processing and Parameter Setting

Following the challenge database split, we select the whole validation data for local validation, which means that we use total 9,815 images for training and the remained 4,346 images for validation. We first use the open-sourced toolkit

Frontalize algorithm with facial landmarks to crop out the faces. Each cropped face is then resized to 256x256 pixels and is passed as an input to the network. Note that due to some faces in images are not front or quite small, the Frontalize can not detect them at well. Therefore, there is actually 8,425 images for training and 3,635 images for validations.

The model is implemented by using Caffe and is trained with the SGD optimizer, which is used with a learning rate of 0.001 and momentum 0.9. All new layers' weights are initialized from a zero-centered normal distribution with a standard deviation of 0.01. As for RNN part, SGD is used with learning rate 0.001 with mini-batch size 64. The weights of feature extractor CNNs are fixed while training. There is a fully connected layer ensure feature size to size 4096 as the input of RNN, the face RNN is a two layers LSTM with 50, 100 hidden neural units.

4.2 Evaluation of the Attention Mechanism

To evaluate the effectiveness of the hourglass like attention module, the following experiments have been carried out. Three networks have been investigated, including the ResNet-50 as the baseline, attention module added after res2c branch and added after res3c branch. The experiment results are given in Table 1.

Table 1. ResNet-50 V.S. Attention-based ResNet-50.

Subject	Module Setting	Validation Accuracy
1	origin Res-50	0.685
2	attention in res2c	0.705
3	attention in res3c	0.699

From the Table 1) we can see a steady increase of weighted average accuracy after the attention module is added. This shows the effectiveness of the added hourglass like attention module. ii) accuracy gains slightly concussion at different level of CNN feature, the accuracy was 0.705 in the res2c while it was only 0.699 in the res3c. This may be due to with the network goes deeper, the more abstract information is gained along with more low-level information lost.

4.3 Evaluation of the CNN-LSTM

Two fundamental group expression models are estimated as baseline model after CNN features are extracted. One is mean group expression model and the other is weighted group expression model. The main difference between those two models are whether considering the face sizes into feature average calculation. And since the LSTM is often used for temporal task as a dynamic model. In GER, we assume a similar architecture LSTM model is capable of remembering important faces in a group-level image.

Table 2. Different face combination settings.

Subject	Combine type	Validation Accuracy
1	mean VGGFace	0.726
2	weighted VGGFace	0.731
3	two LSTM layers	0.746

The results of average classification accuracy on the validation dataset associated to the different settings is shown in Table 2. It is apparent from the table that the LSTM net reaches a much better accuracy. There are chiefly two reasons for the two accuracy increases. One is that big faces usually are closer to the camera which contains more details. The other is that not all faces in images contains useful information.

4.4 Comparison with Other Methods

To comprehensively evaluate the performance of the proposed method, two methods, i.e. the CENsus TRansform hISTogram (CENTRIST) descriptor and the Inception V3 network are adopted for investigation. CENTRIST is based on the Census transform and can capture both the top-down and bottom-up attributes. For fusion setting, we use early fusion strategy to concatenate more than one feature together. This means for each image, we have at least one of three types features or both. Unweighted classification accuracy is used as the evaluation metric. A closer look at the confusion matrix in Fig. 3, our model performed the best on Positive (Pos) among the three emotions. This may attribute to positive class occupy the largest number and model favor of predict positive.

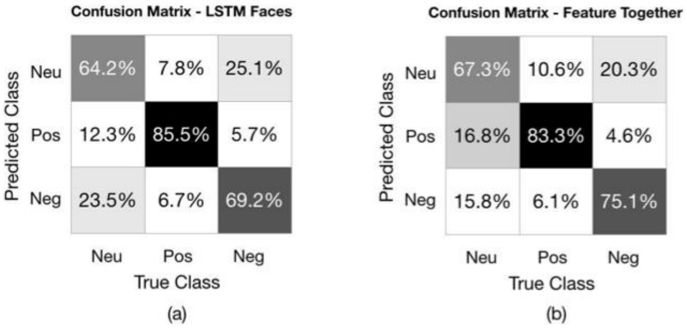

Figure.3. Confusion matrixes for GAF-3 database using different features.

From the Table 3, the following issues can be concluded: i) the DCNN based methods have better performance than the handcraft based approach. ii) the models fusing both scene features, face features and skeleton features outperform the models using only fusing two of them. This tells that learning multiple different features can be more useful. iii) Among all the approaches investigated, the proposed method achieves the best performance.

Table 3. Accuracies of model early fusions on the Validation Set.

Subject	Model Setting	Validation Accuracy
1	Centrist descriptor	0.542
2	Inception V3	0.650
3	Scene feature + LSTM feature	0.744
4	Scene feature + skeleton feature	0.761
5	LSTM feature + skeleton feature	0.736
6	All features together	0.763

5. CONCLUSION

In this paper, we presented our proposed method for the group-level emotion recognition of the EmotiW 2018 challenge. Our method incorporates both bottom-up and top-down components in the Group-level emotion recognition. We carefully employ three level features named attention based scene level feature, CNN-LSTM face level feature and RtPose skeleton level feature respectively. Finally, the early fusion is explored to ensemble these three features. Our best submission achieves a test accuracy of 62.90%.

ACKNOWLEDGEMENTS

This research is funded by the National Natural Science Foundation of China (No.91748127).

REFERENCES

[1] Jia Deng, Alex Berg, Sanjeev Satheesh, H Su, Aditya Khosla, and L Fei-Fei, “Imagenet large scale visual recognition competition 2012 (ilsvrc2012),” Google Scholar, 2012.

[2] Scotty D Craig, Sidney D'Mello, Amy Witherspoon, and Art Graesser, "Emote aloud during learning with autotutor: Applying the facial action coding system to cognitive–affective states during learning," Cognition and Emotion, vol. 22, no. 5, pp. 777–788, 2008.

[3] Abhinav Dhall, Jyoti Joshi, Karan Sikka, Roland Goecke, and Nicu Sebe, "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. IEEE, 2015, vol. 1, pp. 1–8.

[4] Abhinav Dhall, Amanjot Kaur, Roland Goecke, and Tom Gedeon, "Emotiw 2018: Audio-video, student engagement and group-level affect prediction," in Proceedings of the 2018 on International Conference on Multimodal Interaction. ACM, 2018, pp. 653–656.

[5] Sigal G Barsade and Donald E Gibson, "Group emotion: A view from top and bottom.," 1998.

[6] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza, "Emotion recognition in context," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[7] Vassilios Vonikakis, Yasin Yazici, Viet Dung Nguyen, and Stefan Winkler, "Group happiness assessment using geometric features and dataset balancing," in Proceedings of the 18th ACM International Conference on Multimodal Interaction. ACM, 2016, pp. 479–486.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105.

[9] Jianshu Li, Sujoy Roy, Jiashi Feng, and Terence Sim, "Happiness level prediction with sequential inputs via multiple regressions," in Proceedings of the 18th ACM International Conference on Multimodal Interaction. ACM, 2016, pp. 487–493.

[10] Aleksandra Cerekovic, "A deep look into group happiness prediction from images," in Proceedings of the 18th ACM International Conference on Multimodal Interaction. ACM, 2016, pp. 437–444.

[11] Lianzhi Tan, Kaipeng Zhang, Kai Wang, Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao, "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, 2017, pp. 549–552.

[12] Xin Guo, Luisa F Polan´ıa, and Kenneth E Barner, "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, 2017, pp. 603–608.

[13] Alexandr Rassadin, Alexey Gruzdev, and Andrey Savchenko, "Group-level emotion recognition using transfer learning from face identification," in Proceedings of the 19th ACM International Conference on Multimodal Interaction. ACM, 2017, pp. 544–548.

[14] B Balaji and V Oruganti, "Multi-level feature fusion for group-level emotion recognition," in Proceedings of the 19th ACM International Conference on Multimodal Interaction. ACM, 2017, pp. 583–586.

[15] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.

[16] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang, "Residual attention network for image classification," arXiv preprint arXiv:1704.06904, 2017.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[18] Ilya Sutskever, Oriol Vinyals, and Quoc V Le, "Sequence to sequence learning with neural networks," in Advances in neural information processing systems, 2014, pp. 3104–3112.

[19] Wei Li, Farnaz Abtahi, and Zhigang Zhu, "A deep feature based multi-kernel learning approach for video emotion recognition," in Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. ACM, 2015, pp. 483–490.

[20] Andrea Kleinsmith and Nadia Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," IEEE Transactions on Affective Computing, vol. 4, no. 1, pp. 15–33, 2013.

[21] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh, "Convolutional pose machines," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4724–4732.