Exploring Regularizations with Face, Body and Image Cues for Group Cohesion Prediction

Da Guo[*,1,2], **Kai Wang**[*,1,2], **Jianfei Yang**[3], **Kaipeng Zhang**[4], **Xiaojiang Peng**[†,1], **Yu Qiao**[1] *
[1] ShenZhen Key Lab of Computer Vision and Pattern Recognition, SIAT-SenseTime Joint Lab,Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China.
[2] University of Chinese Academy of Sciences, China.
[3] School of Electrical and Electronic Engineering Nanyang Technological University, Singapore.
[4] The University of Tokyo, Japan.

ABSTRACT

This paper presents our approach for the group cohesion prediction sub-challenge in the EmotiW 2019. The task is to predict group cohesiveness in images. We mainly explore several regularizations with three types of visual cues, namely face, body ,and global image. Our main contribution is twofold. First, we jointly train the group cohesion prediction task and group emotion recognition task using multi-task learning strategy with all visual cues. Second, we elaborately design two regularizations, namely a rank loss and a hourglass loss, where the former aims to give a margin between the distance of distant categories and near categories and the later to avoid centralization predictions with only MSE loss. With careful evaluations, we finally achieve the second place in this sub-challenge with MSE of 0.43821 on the testing set. https://github.com/DaleAG/Group_Cohesion_Prediction

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Computer vision tasks**; **Biometrics**.

KEYWORDS

Group Cohesion Prediction, Deep Learning, Convolutional Neural Networks

ACM Reference Format:
Da Guo[*,1,2], Kai Wang[*,1,2], Jianfei Yang[3], Kaipeng Zhang[4], Xiaojiang Peng[†,1], Yu Qiao[1] . 2019. Exploring Regularizations with Face, Body and Image Cues for Group Cohesion Prediction. In *2019 International Conference on Multimodal Interaction (ICMI '19), October 14–18, 2019, Suzhou, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3340555.3355712

[*]Da Guo and Kai Wang contributed equally to this research.
† Corresponding author.

ICMI '19, October 14–18, 2019, Suzhou, China

ACM ISBN 978-1-4503-6860-5/19/10…$15.00
https://doi.org/10.1145/3340555.3355712

1 INTRODUCTION

Group cohesiveness is one of the essential indicators for evaluating the success of a team. A team with high group cohesiveness is often easier to achieve higher efficiency and productivity, which has been shown in the research of social psychology and management. Beal et al. [1] believe that the most important factor leads to success in a group is group cohesion. Myers [10] finds that the people in high group cohesiveness with more positive attitude. According to the relevant research of psychology, group cohesion also has a close relationship with group members' similarity[14], group size[3] and group success[18]. Due to the great influence of group cohesion on group level performance, group level success is highly correlated with high group cohesion[1].

Therefore, it is very meaningful to build an automatically predict group cohesiveness system. Shreya Ghosh et al. [5] indicate that holistic scene information contributes more to the perception of cohesion than face-level information, and the perceived group emotion can offer a prior knowledge for group cohesion prediction. To make full use of the cues from the image, Wang et al.[16] propose a cascaded attention network with three types of visual cues, namely image, body and face cues, for group emotion recognition.

Inspired by the aforementioned studies, we use three types of information, including face, body, and global image, and fuse them to predict group cohesiveness. Specifically, we jointly train the group cohesion prediction task and group emotion recognition task using multi-task learning strategy with all visual cues. Additionally, we elaborately design two regularizations, namely rank loss and hourglass loss, where the former aims to give a margin between the distance of distant categories and near categories and the later to avoid centralization predictions with only MSE loss. With careful evaluations, we finally achieve second place in this sub-challenge with MSE of 0.43821 on the testing set.

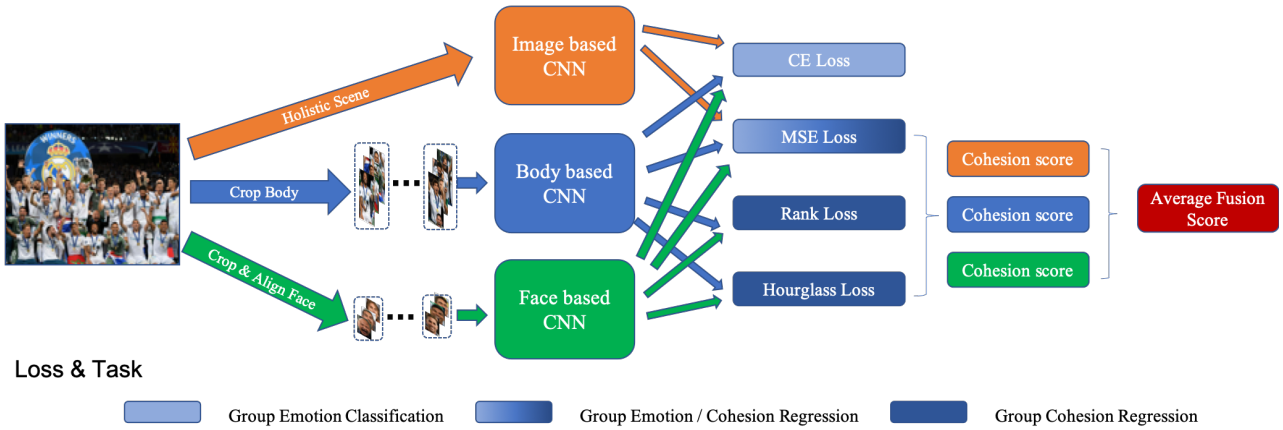

Figure 1: The system pipeline of our approach. It contains three kinds of CNN, namely image-based CNN, body-based CNN and face-based CNN. Particularly, we use cascade attention network structure to train our body-based CNN and face-based CNN. The final prediction is made by averaging all the scores of CNNs from three visual cues.

2 RELATED WORK

Group cohesion prediction and group emotion prediction. Hung et al. [8] use an SVM based classifier to predict cohesion score by the audio and video features of the audiovisual-based group meeting data. [5] contribute the GAF-Cohesion dataset, which is an image-level group cohesion dataset based on group affect dataset. On this dataset, they use image-level information by Inception V3[13] and use face-level information by CapsNet[11] to predict group cohesion. [16] used three types cues including total image, body and face to predict group emotion by attention CNNs, and won the second place in the EmotiW2018. Different visual cues are complementary predict group emotion and cohesion [9, 15], so we add these three cues to our architecture.

Multi-task learning. In the field of face detection, Zhang et al.[21] enhanced face alignment performance by using facial attribute recognition as an auxiliary task. In the field of group cohesion prediction. [5] indicate that jointly trained group-level emotion classification and group cohesion regression helps in increasing the performance for the group cohesion prediction task.

3 APPROACH

System Pipeline

Our system pipeline is shown in Figure 1. Following [16], we also use three types of visual cues including face, body and global images, to predict by fusing multimodal features. For different visual cues, we jointly train multi-task and add regularizations in loss function to improve cohesion prediction performance. Finally, we average the prediction scores of different visual cues as the final prediction score for the input image.

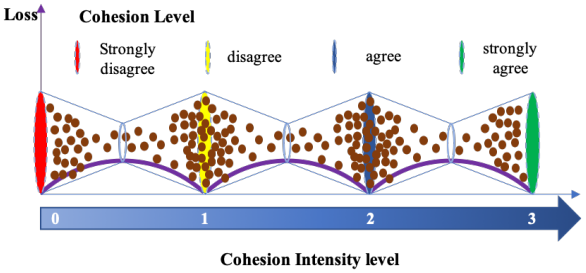

Figure 2: Illustration of our hourglass loss. The loss value becomes larger when the predicted GCS is near the middle of two adjacent levels.

CNNs for Three Visual Cues

Considering the characteristics and differences of different visual cues, we use different CNNs to extract features for different visual cues. Due to the interaction between group emotion and group cohesion, we joint train group emotion classification and group cohesion regression in all visual cues.

Face-based CNN. Facial emotion of each face in image is an essential cue for group cohesion prediction. We use S^3FD [20] and MTCNN [19] to detect face and 5 facial landmarks from GAF-Cohesion dataset, and align faces by the 5 facial landmarks. Then we use these aligned faces to train our face-based CNN model. In order to learn the correlation between the emotion of different members of the group and group cohesion, We use cascade attention network[16] with ResNet18 [6] backbone. We use the model pretrained on FERPlus dataset and finetune it on the GAF-Cohesion dataset.

Body-based CNN. Group cohesion is highly correlated with the members of the group, so it is necessary to use

body information to train a model. The body region contains more semantic information about a member, which helps CNN learn the feature of group members to predict more accurate group cohesion score. The strategy of body region cropping follows the design of [16]: we use OPENPOSE [2, 12, 17] to detect 18 human body keypoints, and crop human body with the maximum outer rectangular area as the body region. Since body information is more complicated than face, we chose SE-Net154 [7] to train the body-based CNN model. Also, we use cascade attention network structure in SE-Net154 to learn the correlation between the body state of different members of the group and group cohesion.

Image-based CNN. Global image provides the holistic scene and contextual information about the group which is more complicated and contribute more to the perception of cohesion, than face and body information. Therefore, we use the SE-Net154 [7] to train our image-based CNN model, and we also find that the more lightweight CNN such as ResNet101 [6], cannot handle the holistic scene information very well to predict group cohesion score. In addition, both body-based CNN and image-based CNN use the model which is pre-trained on ImageNet datasets in advance.

Regularizations

Regularizations are techniques used to reduce the error by fitting a function appropriately on the given training set and avoid overfitting. In our paper, we propose two type regularizations, namely rank loss and hourglass loss. The intuition of rank loss is that the margin of adjacent cohesion intensity levels should be smaller than distant cohesion intensity levels. Although rank loss improves the performance of prediction, most of the predict values are concentrated between 1 and 2. we propose hourglass loss to make the prediction results more dispersed and closer to the corresponding ground truth. In general, we jointly use MSE, rank and hourglass losses to optimize the CNNs.

RMSE loss. In statistics, the mean squared error (MSE) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors — that is, the average squared difference between the estimated values and what is estimated. We calculate the MSE between prediction and the ground truth of cohesion. The standard MSE loss is defined as follows:

$$L_m = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y_i})^2 \tag{1}$$

where y_i and $\tilde{y_i}$ are the prediction of our CNNs and the ground truth of the image.

Rank loss. Intuitively, images with high and low cohesion should have a obvious differences in the feature space. Based on such intuition, we design rank loss which aims to give a margin between the distance of distant cohesion levels and near cohesion levels.

Given the distance of different intensity as follows:

$$d_i^1 = ||C_i - C_{i+1}||, i = 0, 1, 2 \tag{2}$$

$$d_i^2 = ||C_i - C_{i+2}||, i = 0, 1 \tag{3}$$

$$d_i^3 = ||C_i - C_{i+3}||, i = 0 \tag{4}$$

where d_i^1, d_i^2, d_i^3 are the L_2 distances between the centers (C_i) of each cohesion intensity level in feature space. The intuition is that d_i^1 should be smaller than d_i^2 and d_i^2 should be smaller than d_i^3. Formulaically, the rank loss is as follows:

$$L_{rank1} = \sum_{i=0}^{1}\sum_{j=0}^{2} max(0, \delta - (d_i^2 - d_j^1)) \tag{5}$$

$$L_{rank2} = \sum_{i=0}^{0}\sum_{j=0}^{2} max(0, 2\delta - (d_i^3 - d_j^1)) \tag{6}$$

$$L = L_m + L_{rank1} + L_{rank2} \tag{7}$$

We set a margin δ between different intensity level samples in feature space. Note that, the margin is depended on the intensity level difference. For example, the distance between *strongly disagree* and *disagree* should be smaller than the distance between *strongly disagree* and *strongly agree*. Here, we set margin δ is 0.75.

Trick. Additionally, we use a trick to manually adjust the predicted score to achieve higher performance on the validation. We appropriately reduce the predicted group cohesion score(GCS) in the interval [1.35, 1.95] and appropriately increase the predicted GCS in the interval [2.1,2.7]. However, this behavior may lead to over-fitting the validation set. Inspired by this method, we appose a new and effective regularization approach namely hourglass loss.

Table 1: Evaluation of regularizations with face-based CNN on the GAF-Cohesion validation set.

ResNet-18 (Face-based CNN)		MSE
rank loss	hourglass loss	face model
-	-	0.6615
✓	-	0.6227
✓	✓	**0.6058**

Hourglass Loss. By observing the prediction results of the validation set, we find that the GCS of most samples are predicted to be within the interval of 1 to 2, which leads to a high MSE on the validation set and makes it difficult to distinguish different group cohesion levels. To this end, we propose hourglass loss to optimize this problem, so that

Table 2: Evaluation of different visual cues and their fusion on GAF-Cohesion validation set.

Visual Cues	MSE
Image-based CNN	0.6108
Body-based CNN	0.6327
Face-based CNN	0.6058
fusion CNN	**0.5588**

samples of different group cohesion levels are more dispersed and closer to their corresponding ground truth.

Assume P_k is the predicted GCS, g_i and g_j are the actual GCS level of the current sample and another GCS level in the interval where the sample is located, respectively. Then we define the hourglass loss as follows:

$$L_h = |P_k - g_i||g_j - P_k| \tag{8}$$

$$L = L_m + L_{rank1} + L_{rank2} + L_h \tag{9}$$

As is shown in Figure 2, hourglass can make the data distribution more dispersed.

4 EXPERIMENTS

In this section, we first present the GAF-Cohesion dataset in EmotiW 2019 and the implementation details. Then we show the evaluations of our regularizations and our submission results.

Dataset

The GAF-Cohesion dataset used for group cohesion prediction of EmotiW 2019 is extended the images from the GAF 3.0 database[4], which is collected from Internet. Each image of GAF-Cohesion dataset is labelled in the range [0-3] as its cohesiveness. The dataset contains 14,175 images and is divided into three parts: train set, validation set and test set. The number of images in these three sets is 9,815, 4,349 and 3,011 respectively.

Implementation Details

For face-based CNN, we firstly use ResNet-18 [6] which is pretrained on the ImageNet dataset to train a emotion classification model on the FERPlus dataset, then finetune it on the aligned faces of GAF-Cohesion dataset with batch size 64. In the finetuning step, we use 0.1 learning rate for the first 8 epochs, then continue training for 4 epochs with 0.01 and 0.001.

For body-based CNN, we finetune SENet-154 which is pretrained on the ImageNet dataset on the cropped body from GAF-Cohesion dataset with batch size 16. During the training, the start learning rate is 0.01, and times 0.1 at 6, 10 and 16 epochs.

For image-based CNN, we also finetune SENet-154 which is pretrained on the ImageNet dataset on the GAF-Cohesion dataset with batch size 64. And the learning rate setting is same to the design of body-based CNN. Our method is all implemented in PyTorch./

Experimental Results

In this subsection, we evaluate our regularizations on the validation set, and present our final submission results on the valodation set and test set.

Evaluation of regularizations on the face-based CNN. Table 1 shows the results of different regularizations on the GAF-Cohesion 3.0 validation set. Rank loss reduces the MSE by 0.0388 on the validation set. Hourglass loss and rank loss together reduce the MSE by 0.0557 on the validation set. And the results indicate that our regularizations can effectively improve the performance of group cohesion prediction.

Evaluation of different visual cues. Table 2 presents the MSE of different visual cues and their fusion on the GAF-Cohesion validation set. The Face-based CNN gets the best performance followed by image-based CNN and body-based CNN. Since the differences and complementarities between different visual cues, we achieve better performance by combining multimodal features.

Results of final submission. Table 3 shows the details of our final 5 submissions. In order to make use of all data, we randomly select 90% data from validation set, and add it to the training set for training. The final submitted results are shown in Table 4, and we win the second place with MSE of 0.43821 on the test set. From the results of different cohesion intensity levels, we found that there is an imbalance in the data of different cohesion intensity levels. Besides, the results proves our previous thought that trick method is definitely over-fitting the validation set.

5 CONCLUSIONS

We present our approach for the group cohesion prediction in the EmotiW 2018. We mainly explore several regularizations with three types of visual cues, namely face, body, and global image. Particularly, we elaborately design two regularizations, namely a rank loss and a hourglass loss, where the former aims to give a margin between the distance of distant categories and near categories and the latter to avoid centralization of predictions with only MSE. With careful evaluations, we finally achieve the second place in this subchallenge with MSE of 0.43821 on the test set.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (U1813218, U1613211), Shenzhen Research Program(JCYJ20170818164704758,CXB201104220032A, JSGG20180507182100698), and Joint Lab of CAS-HK.

Table 3: Model details of our final submissions.

Runs	Train Data	Trick	Model Details (R/H: Rank/Hourglass Loss)		
			Image-based CNN	Body-based CNN	Face-based CNN
1	Train + 90%Validation	✓	1∗SENet-154	2∗SENet-154(R+H)	1∗ResNet-18(R+H)
2	Train + 90%Validation	-	1∗SENet-154	2∗SENet-154(R+H)	1∗ResNet-18(R+H)
3	Train	-	1∗SENet-154	1∗SENet-154(R)	3∗ResNet-18(2∗R+1∗H)
4	Train	✓	1∗SENet-154	1∗SENet-154(R)	3∗ResNet-18(3∗R)
5	Train	✓	1∗SENet-154	1∗SENet-154(R)	4∗ResNet-18(3∗R+1∗H)

Table 4: Results of our final submissions.

Runs	Validation	Test				
	Overall	0	1	2	3	Overall
1	–	1.754	0.618	0.422	0.618	0.55340
2	–	2.150	0.548	0.162	0.676	**0.43821**
3	0.56888	2.659	0.646	0.101	0.763	0.46112
4	**0.49689**	2.102	0.734	0.310	0.643	0.52539
5	0.50026	2.056	0.681	0.335	0.701	0.55009

REFERENCES

[1] Daniel J Beal, Robin R Cohen, Michael J Burke, and Christy L McLendon. 2003. Cohesion and performance in groups: a meta-analytic clarification of construct relations. *Journal of applied psychology* 88, 6 (2003), 989.

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

[3] Albert V Carron and Kevin S Spink. 1995. The group size-cohesion relationship in minimal groups. *Small group research* 26, 1 (1995), 86–105.

[4] Abhinav Dhall, Roland Goecke, Shreya Ghosh, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. 2017. From individual to group-level emotion recognition: EmotiW 5.0. In *Proceedings of the 19th ACM international conference on multimodal interaction*. ACM, 524–528.

[5] Shreya Ghosh, Abhinav Dhall, and Nicu Sebe. 2018. Predicting Group Cohesiveness in Images. *arXiv preprint arXiv:1812.11771* (2018).

[6] 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

[7] 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

[8] Hayley Hung and Daniel Gatica-Perez. 2010. Estimating cohesion in small groups using audio-visual nonverbal behavior. *IEEE Transactions on Multimedia* 12, 6 (2010), 563–575.

[9] Sunan Li, Wenming Zheng, Yuan Zong, Cheng Lu, Chuangao Tang, Xingxun Jiang, Jiateng Liu, and Wanchuang Xia. 2019. Bi-modality Fusion for Emotion Recognition in the Wild. In *Proceedings of the 21th ACM International Conference on Multimodal Interaction (in press)*. ACM.

[10] Albert E Myers. 1961. *Team competition, success, and the adjustment of group members*. Technical Report. ILLINOIS UNIV URBANA GROUP EFFECTIVENESS RESEARCH LAB.

[11] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic routing between capsules. In *Advances in neural information processing systems*. 3856–3866.

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

[13] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.

[14] Henri Tajfel. 2010. *Social identity and intergroup relations*. Vol. 7. Cambridge University Press.

[15] Kai Wang, Jianfei Yang, Da Guo, Kaipeng Zhang, Xiaojiang Peng, and Yu Qiao. 2019. Bootstrap Model Ensemble and Rank Loss for Engagement Intensity Regression. In *Proceedings of the 21th ACM International Conference on Multimodal Interaction (in press)*. ACM.

[16] 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 *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 640–645.

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

[18] Stephen J Zaccaro and M Catherine McCoy. 1988. The effects of task and interpersonal cohesiveness on performance of a disjunctive group task 1. *Journal of applied social psychology* 18, 10 (1988), 837–851.

[19] 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.

[20] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. 2017. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE International Conference on Computer Vision*. 192–201.

[21] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2014. Facial landmark detection by deep multi-task learning. In *European conference on computer vision*. Springer, 94–108.