Efficient Group-Based Cohesion Prediction in Images Using Facial Descriptors

Ilya Gavrikov and Andrey V. Savchenko
HSE University, Laboratory of Algorithms and Technologies for Network Analysis, Nizhny Novgorod, Russia
ilsgavrikov@gmail.com, avsavchenko@hse.ru

Abstract. In this paper we study the problem of predicting the cohesiveness and emotion of a group of people in photo. We proposed a fast approach, consisting of face detection by using MTCNN, aggregation of facial features (age, gender and embeddings) extracted by multi-task MobileNet, prediction of group cohesion and classification of emotional background using multi-output convolution neural network. Experimental study on the Group Affect Dataset from EmotiW 2019 challenge demonstrated that our approach allows to achieve an improvement of quality and even to reduce the running time of an algorithm’s work when compared to known solutions. As a result, we obtained mean squared error 0.63 for cohesion prediction, which is 0.21 lower when compared to baseline CapsNet.

Keywords: Group cohesiveness prediction · Group emotion recognition · Facial analysis · Deep Convolutional Neural Network (CNN)

1 Introduction

Prediction of group-level emotion [1] and cohesiveness is very useful for various companies in order to analyze employee’s emotional state throughout the day and build a relationship between their emotional state and group cohesion [2]. As the cohesiveness of a group is a crucial indicator of success of a group of people, the problem of predicting the perceived cohesiveness of a group of people in image becomes one of the main tasks in the EmotiW (Emotion Recognition in the Wild) 2019 challenge [3]. The usage of the CapsNet (Capsule Network) fitted for emotion recognition with seven labels, made it possible to obtain the baseline with MSE (mean squared error) equal to 0.84 [3] by feeding aggregated emotions into regression CNN (Convolutional Neural Network). The fusion of three models with face detection and feature extraction for support vector regression and aggregation of predictions for all faces lead to MSE 0.66 on validation set [4]. An ensemble of three branches that process global image, poses based on skeleton
images and faces reached MSE 0.6493 on the validation set [5]. The first place in this challenge (MSE 0.52 on validation set) was obtained by a hybrid CNN with analysis of scene, faces, skeletons and UV coordinates [6]. Analysis of faces, bodies and the whole image lead to MSE 0.56 [7].

Unfortunately, all these techniques are very slow and cannot be used in embedded solutions for video analytics. Hence, in this paper we propose a lightweight solution based on facial feature extraction with MobileNet v1 [8] that simultaneously extract facial embeddings [9] and predict its gender and age [10]. We propose to extend this network by computing the average of its outputs for all detected faces and feed the overall descriptor of a group of people into a simple multi-task neural network for group-level emotion recognition and cohesiveness prediction. The main contribution of this paper is to experimentally demonstrate that information about facial features from very large external dataset of celebrities [11] may be used to train emotion classifier using rather small Group Affect Dataset [3]. As a result, in contrast to existing studies [5], we obtain rather accurate group-based cohesion prediction technique based on processing of facial features only.

The rest of the paper is organized as follows. In Sect. 2 we introduce the proposed algorithm. Experimental results for the dataset from EmotiW 2019 are presented in Sect. 3. Concluding comments are given in Sect. 4.

2 Proposed Approach

The task of this paper may be formulated as follows. Given an input image of a group of people, it is necessary to predict their cohesiveness, i.e. a measure
of bonding between group members [3]. It is an ordinal regression task with 4 variants of bonding (very weak, weak, strong and very strong, see Fig. 1). In addition, it is necessary to classify emotion of a group of people. We will use three classes from Group Affect Dataset: positive, neutral and negative (Fig. 2) [2].

The complete pipeline of the proposed Algorithm 1 is shown in Fig. 3. In this paper we test the hypothesis that all necessary information to solve this problem is reflected in the faces of persons for a group. Hence, at first, it is necessary to detect $R$ faces in a given photo. We will use MTCNN (multi-task CNN) facial detector [12] which is fast and accurate for rather large faces, though it does not obtain the state-of-the-art results for more complex photos [13]. After that, the features are extracted from every $r$-th facial region ($r = 1, 2, ..., R$). As the faces are observed in unconstrained conditions, modern transfer learning and domain adaptation techniques can be used. According to these methods, the large external dataset of celebrities, e.g., VGGFace2 [14], is used to train a deep CNN in order to let the neural network to learn reliable facial features. The outputs of one of the last layers of this CNN form $D$-dimensional features (embeddings/descriptor) $x_r$ of the $r$-th face. It is typical to use embeddings from pre-trained CNNs, e.g., VGGFace-16 [15], ResNet-50 [11], ArcFace [16], etc.

In order to speed-up the decision process, we propose to use lightweight multi-task MobileNet [10]. In addition to facial embeddings, it can simultaneously predict age and gender of a person, so that we decided to concatenate two values (estimate of posterior probability for a male gender and predicted age) and extracted facial features. Next, it is necessary to aggregate $R$ facial features into a single descriptor of a photo. We will compute the mean of $L_2$-normalized features of individual faces [17]. We propose to fed the resulting mean descriptor into the multi-task neural network for predicting group cohesion and emotional background. Our neural net consists of the fully connected (dense) layer with
Algorithm 1. Proposed procedure of video-based cohesion prediction.

Require: Video frames or images \( \{X(t)\}, t = 1, 2, ..., T \)
Ensure: Cohesion and Emotion label of the given video

1: \hspace{1em} for each frame \( t = 1, 2, ..., T \) do
2: \hspace{2em} Obtain \( R \geq 0 \) facial regions using, e.g., MTCNN face detector
3: \hspace{2em} for each facial area \( r = 1, 2, ..., R \) do
4: \hspace{3em} Extract embeddings and simultaneously predict age and gender using the multi-output MobileNet \[10\]
5: \hspace{3em} Concatenate embeddings and predicted age and estimate of male gender posterior probability into a single descriptor \( x_r(t) \)
6: \hspace{2em} end for
7: \hspace{2em} Compute the frame feature vector \( x(t) \) as an average of embeddings \( \{x_r(t)\}, r = 1, 2, ..., R \) for all facial regions and normalize it
8: \hspace{2em} Feed the features into the multi-output neural network
9: \hspace{2em} Assign the vector of scores \( s_{c;\text{cohesion}}(t) \) and \( s_{c;\text{emotion}}(t) \) from the output of regression and classification layers for cohesion and emotion prediction, respectively
10: \hspace{1em} end for
11: \hspace{1em} Compute the cohesion scores \( s_{c;\text{cohesion}} \) as an average of scores \( \{s_{c;\text{cohesion}}(t)\}, t = 1, 2, ..., T \) for all frames
12: \hspace{1em} Compute the group-level emotion scores \( s_{c;\text{emotion}} \) as an average of scores \( \{s_{c;\text{emotion}}(t)\}, t = 1, 2, ..., T \) for all frames
13: \hspace{1em} Return the cohesion and group emotion categories with the maximal scores \( \text{argmax} s_{c;\text{cohesion}} \) and \( \text{argmax} s_{c;\text{emotion}} \).

600 neurons and ReLU activation, conv1d, MaxPooling1d, flattening, dense with 200 neurons and two output layers, namely, a layer with softmax activation and 3 outputs for emotion recognition and one linear output for group cohesiveness prediction.

Our approach (Fig. 3) has been implemented in a simple demo application\(^1\), which can predict group cohesiveness and emotions given an input photo and video from web camera. Sample screen shots of our application for high and low predicted cohesiveness are shown in Fig. 4.

3 Experiments

In the experimental study we used the Group Affect Dataset from EmotiW 2019 [3] with 9,815 images. The training set contains 3100 images for each class of emotions and 1141 pictures for strongly disagree (ground-truth label ‘0’), 1561 for disagree (label ‘1’), 4601 for agree (‘2’) and 1997 for strongly agree (‘3’) levels of group cohesion. The averaged facial features extracted by a CNN are fed into two fully connected layers to predict group emotions and cohesiveness (Fig. 3). The output of the latter linear layer is in the range between 0 and 3.

\(^1\) https://github.com/gbcodes/GroupCohesiveness.
Table 1. Results for group cohesion prediction, VGGFace2 facial features

| Supervised learning methods | MSE   |
|-----------------------------|-------|
| Ordinal Ridge [16]          | 1.01  |
| Ordinal Ridge Regularized [16] | 0.92  |
| Logistic All Threshold [16]  | 1.07  |
| Logistic All Threshold Regularized [16] | 0.87  |
| Logistic Immediate Threshold [16] | 1.20  |
| Logistic Immediate Threshold Regularized [16] | 0.97  |
| Logistic SE (Squared Error) [16] | 1.06  |
| Logistic Squared Error Regularized [16] | 0.85  |
| Multi-class Logistic Regression | 1.03  |
| Least Absolute Deviation    | 1.05  |
| Catboost [17]               | 0.96  |
| Catboost-based [17] Ordinal Classifier | 0.87  |

The predicted cohesiveness is rounded to the nearest number (0, 1, 2, 3) and the MSE with the ground-truth cohesiveness level is calculated.

In the first experiment we used conventional ResNet-50 [11] trained on VGGFace2 dataset [14] in order to extract $D = 2048$ facial features. We used ordinal regression methods from mord package [18] and Catboost [19]. Their hyper-parameters (regularization coefficient, learning rate, etc.) have been tuned on the 20% of training set using cross validation (GridSearchCV). The results on validation set are shown in Table 1.

Here the lowest MSE is obtained for LogisticSE ordinal regression [18] with regularization, which will be used in the next experiment to choose the best facial descriptor. Logistic SE obtained not the best result, but regularization greatly improves the quality of prediction. The text on page 4 is about Logistic
Table 2. Results for facial descriptors for group cohesion prediction, LogisticSE

| Facial Descriptor                    | Feature extraction time, ms |
|-------------------------------------|-----------------------------|
|                                     | MSE | CPU       | GPU       |
| VGGFace (VGG-16) [15]               | 1.12| 109.74    | 8.61      |
| VGGFace2 (ResNet-50) [14]           | 0.85| 57.14     | 11.54     |
| Multi-output MobileNet [10]         | 0.80| 19.94     | 4.76      |

SE with regularization. It has already been mentioned. We compared three pre-trained CNNs for feature extraction, namely, 1) VGGFace-16 [15] that extracts $D = 4096$ features; 2) VGGFace2 (Res-Net-50) [11] ($D = 2048$ features); and 3) multi-output MobileNet [10] trained on the same VGGFace-2 dataset [14]. The dimensionality of the feature vector extracted by the latter network ($D = 1024$) is the lowest one. Keras framework with TensorFlow 1.15 backend were used in all experiments. The results are presented in Table 2.

Here we additionally measure an average inference time to extract features of one face by using CPU (12 core AMD Ryzen Threadripper) or GPU (Nvidia GTX1080TI) on a special server. It is surprisingly that the lightweight descriptor (MobileNet that simultaneously extracts facial features and predicts age and gender) outperforms the deeper CNNs including VGGFace2 that achieves the state-of-the-art results in face recognition [14].

In order to speed up decision-making, we examined the usage of principal component analysis (PCA) and processed only small number of first principal components. The dependence of MSE on the number of components for facial features from MobileNet [10] is shown in Fig. 5. Here we compared our simple multi-task neural network with one convolutional layer with a similarly engineered networks with 2 and 3 sequentially connected conv1d layers. These models were fitted with batch size 32 during 100 epochs with early stop.

Here the lowest MSEs are achieved by a network with one convolutional layer. Unfortunately, regression for small number of principal components causes significant (up to 0.3) increase in MSE. However, usage of all principal components in the proposed approach leads to the lowest MSE for cohesion prediction (0.63). It is 0.03 and 0.02 lower when compared to ensembles of several complex models [4,5]. Moreover, our approach much better than the baseline multi-task CNN of the challenge’s organizers [3].
4 Conclusion

In this paper we proposed an efficient approach (Fig. 3) to process the images of a group of people, for which MSE of cohesiveness prediction on a validation set is 0.21 lower when compared to MSE (0.84) of the baseline from the EmotiW 2019 [3]. The proposed approach is implemented in a publicly-available demo application (Fig. 4). In this demo, we predict age and gender of each person and predict the cohesiveness and emotion of the whole group. It is rather fast (+10 FPS for at most 16 persons in a group using Nvidia GTX1080 Ti GPU) due to the usage of MobileNet. Our preliminary results demonstrated that our model can be used even at Android mobile device with 5 FPS for a small group of 3 persons.

Our approach is not obviously the best one, as our MSE on validation set is 0.11 and 0.07 greater than MSEs of the first [6] and second [7] places in the EmotiW 2019 challenge. However, their running time is much worth: slower than 0.35 and 0.2 FPS for ensembles from [6] and [7], respectively. Another advantage of our model is the possibility to simultaneously recognize gender and age of each person in a photo/video, which can be important for potential usage of cohesiveness prediction in industrial applications. Finally, our results prove that the claim from [5] that “faces features may not reflect the group cohesiveness as well as other features” is incorrect. Indeed, it is necessary to use the pre-trained facial descriptors in order to obtain the high-quality group-based cohesion prediction. However, it is important to emphasize that in this task our MobileNet [10] obtains lower MSE than much more powerful ResNet-50 that is
characterized by the state-of-the-art results in face identification [14]. It seems that the lightweight CNN with low number of parameters is more appropriate for such tasks with rather small number of training examples.

Unfortunately, our group-level emotion recognition accuracy is rather low (0.69), so that it is necessary to further improve our model. Moreover, in future, it is important to extract the faces from a group photo, which significantly influence the overall cohesiveness score. In this case it will be possible to find persons who, e.g., are less concentrated on the task than the majority of other persons in a photo. Finally, it is necessary to examine the state-of-the-art face detectors, e.g., RetinaFace [13], instead of MTCNN in order to locate more faces accurately. However, it is still possible that better facial detectors won’t lead to the better quality of cohesiveness prediction, because very small faces do not have robust facial features. For example, experiments from paper [17] clearly demonstrate that usage of TinyFaces detector provide lower accuracy of group emotion recognition when compared to Viola-Jones detector, if VGGFace features are used [15]. In such case, it is possible that we should use other facial descriptors trained on facial images with small resolution [1].

Acknowledgements. The work of A.V. Savchenko is supported by RSF (Russian Science Foundation) grant 20-71-10010.

References

1. Tarasov, A.V., Savchenko, A.V.: Emotion recognition of a group of people in video analytics using deep off-the-shelf image embeddings. In: van der Aalst, W.M.P., et al. (eds.) AIST 2018. LNCS, vol. 11179, pp. 191–198. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-11027-7_19
2. Dhall, A., Goecke, R., Ghosh, S., Joshi, J., Hoey, J., Gedeon, T.: From individual to group-level emotion recognition: EmotiW 5.0. In: Proceedings of the 19th International Conference on Multimodal Interaction (ICMI), pp. 524–528. ACM (2017)
3. Ghosh, S., Dhall, A., Sebe, N., Gedeon, T.: Predicting group cohesiveness in images. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2019)
4. Zhu, B., Guo, X., Barner, K., Boncelet, C.: Automatic group cohesiveness detection with multi-modal features. In: Proceedings of the International Conference on Multimodal Interaction (ICMI), pp. 577–581. ACM (2019)
5. Zou, B., Lin, Z., Wang, H., Wang, Y., Lyu, X., Xie, H.: Joint prediction of group-level emotion and cohesiveness with multi-task loss. In: Proceedings of the 5th International Conference on Mathematics and Artificial Intelligence, pp. 24–28 (2020)
6. Xuan Dang, T., Kim, S.H., Yang, H.J., Lee, G.S., Vo, T.H.: Group-level cohesion prediction using deep learning models with a multi-stream hybrid network. In: Proceedings of the International Conference on Multimedia Interaction (ICMI), pp. 572–576. ACM (2019)
7. Guo, D., Wang, K., Yang, J., Zhang, K., Peng, X., Qiao, Y.: Exploring regularizations with face, body and image cues for group cohesion prediction. In: Proceedings of the International Conference on Multimodal Interaction (ICMI), pp. 557–561. ACM (2019)
8. Howard, A.G., et al.: MobileNets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
9. Savchenko, A.: Efficient statistical face recognition using trigonometric series and CNN features. In: Proceedings of the 24th International Conference on Pattern Recognition (ICPR), pp. 3262–3267. IEEE (2018)
10. Savchenko, A.V.: Efficient facial representations for age, gender and identity recognition in organizing photo albums using multi-output convNet. PeerJ Comput. Sci. 5, e197 (2019)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. IEEE (2016)
12. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Sig. Process. Lett. 23(10), 1499–1503 (2016)
13. Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: RetinaFace: single-shot multi-level face localisation in the wild. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5203–5212. IEEE (2020)
14. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: VGGFace2: a dataset for recognising faces across pose and age. In: Proceedings of 13th International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 67–74. IEEE (2018)
15. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition (2015)
16. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: ArcFace: additive angular margin loss for deep face recognition. In: Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4690–4699. IEEE (2019)
17. Rassadin, A., Gruzdev, A., Savchenko, A.: Group-level emotion recognition using transfer learning from face identification. In: Proceedings of the 19th International Conference on Multimodal Interaction (ICMI), pp. 544–548. ACM (2017)
18. Pedregosa-Izquierdo, F.: Feature extraction and supervised learning on fMRI: from practice to theory. Ph.D. thesis (2015)
19. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: CatBoost: unbiased boosting with categorical features. In: Advances in Neural Information Processing Systems (NIPS), pp. 6638–6648 (2018)