Group-Level Emotion Recognition using Deep Models on Image Scene, Faces, and Skeletons

Xin Guo
Department of Electrical and Computer Engineering, University of Delaware
Newark, DE, USA
guoxin@udel.edu

Luisa F. Polanía
American Family Mutual Insurance Company
Madison, WI, USA
polania@amfam.com

Kenneth E. Barner
Department of Electrical and Computer Engineering, University of Delaware
Newark, DE, USA
barner@udel.edu

ABSTRACT

This paper presents the work submitted to the Group-level Emotion Recognition sub-challenge, which is part of the 5th Emotion Recognition in the Wild (EmotiW 2017) Challenge. The task of this sub-challenge is to classify the emotion of a group of people in each image as positive, neutral or negative. To address this task, a hybrid network that incorporates global scene features, skeleton features of the group, and local facial features is developed. Specifically, deep convolutional neural networks (CNNs) are first trained on the faces of the group, the whole images and the skeletons of the group, and then fused to perform the group-level emotion prediction. Experimental results show that the proposed network achieves 80.05% and 80.61% on the validation and testing sets, respectively, outperforming the baseline of 52.97% and 53.62%.

CCS CONCEPTS

• **Computing methodologies → Activity recognition and understanding**; *Scene understanding*; *Transfer learning*; Neural networks; Feature selection;

KEYWORDS

EmotiW 2017; Emotion Recognition; Group level happiness prediction; Deep learning; Multi-model; Decision fusion

ACM Reference Format:
Xin Guo, Luisa F. Polanía, and Kenneth E. Barner. 2017. Group-Level Emotion Recognition using Deep Models on Image Scene, Faces, and Skeletons. In *Proceedings of 19th ACM International Conference on Multimodal Interaction (ICMI'17).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3136755.3143017

1 INTRODUCTION

The problem of emotion recognition for individuals has been widely studied due to its importance in affective computing, security and human-computer interaction [4, 19, 24]. Although the problem of emotion recognition for a group of people has been less extensively studied and remains as an open research problem, it is gaining popularity due to the huge amount of data available on social network sites, which contain images of groups of people participating in events and social gatherings. In addition, group-level emotion recognition (GER) has interesting applications in image retrieval [6], shot selection [7], surveillance [1], event summarization [7], and event detection [31], among others. Analysis of the emotion expressed by a group of people is also challenging due to head and body pose variations, face occlusions, illumination variations, varied indoor and outdoor settings, and interactions taking place between various number of people [21].

A pioneering work in the area of GER was proposed by Dhall *et al.* [9]. They collected the AFEW database, which consists of videos of multiple subjects with each frame labeled at both group level and individual level with respect to seven emotion classes. Similarly, the database HAPPEI [7] was built for overall happiness at the group level by collecting images from social networks, such as Facebook and Flickr. In a more recent work [10], Dhall *et al.* collected the Group Affect Database (GAD), which encompasses Google and Flickr images related to key words, which describe groups and events. The database contains images exhibiting heterogeneous expressions across subjects of a group.

Works in the area of GER can be classified into three categories: bottom-up methods, top-down methods and the combination of both. The attributes of subjects are used to infer emotion at the group level in the bottom-up methods while the group information is used as a prior for inference of subject level attributes in the top-down methods. An example of bottom-up methods is the work by Hernandez *et al.* [14], which refers to a system to detect happiness of the passerby at different locations of the Massachusetts Institute of Technology campus. The authors used the Shore Framework [18] to detect the faces in a crowd and extracted geometric facial features. The estimated level of happiness of the scene was calculated as the average over the happiness level of the group members. However, an averaging model ignores both global information, *e.g.*, the relative position of people in the image, and local information, *e.g.*, the level of occlusion of a face, and therefore it is not optimal for group emotion. An example of a top-down method is the work of Mou *et al.* [21], which extracts context features and uses k-nearest neighbor to predict emotion. In another top-down method [7], a minimum spanning tree is used to represent a group by having faces as the vertices and distances between two faces as the weights of the edges. The combination of top-down and bottom-up methods have shown better performance than using one method alone. The winner of the group-based emotion recognition sub-challenge in EmotiW

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

ACM ISBN 978-1-4503-5543-8/17/11…$15.00
https://doi.org/10.1145/3136755.3143017

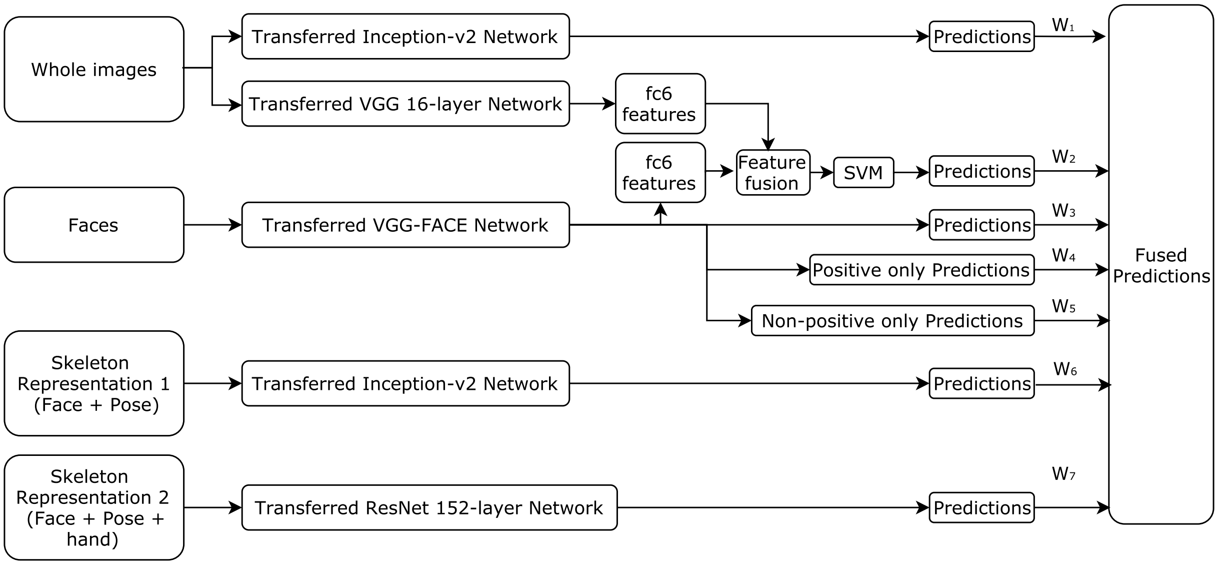
Whole images
Transferred Inception-v2 Network
Transferred VGG 16-layer Network
fc6 features
fc6 features
Feature fusion
SVM
Faces
Transferred VGG-FACE Network
Predictions
Positive only Predictions
Non-positive only Predictions
Skeleton Representation 1 (Face + Pose)
Transferred Inception-v2 Network
Skeleton Representation 2 (Face + Pose + hand)
Transferred ResNet 152-layer Network
W₁
W₂
W₃
W₄
W₅
W₆
W₇
Fused Predictions

Figure 1: The overall structure of the proposed hybrid network. Note that multiple models have been trained on the scene, faces and skeletons separately, then a decision fusion is performed on all the models to learn an optimal combination. The final network configuration consists of 7 models since the redundant models are removed after decision fusion. Details are described in Section 2 and 3.

2016 [20] proposed a long short-term memory (LSTM) to selectively learn the important features from both the whole image and the facial regions.

In this paper, we also propose to incorporate both top-down and bottom-up methods. Instead of the CENTRIST [36] scene descriptor used by the winner of the 2016 sub-challenge [20], we propose to fine-tune state-of-the-art deep neural architectures using the Group Affect Database 2.0 for the GER task. The proposed scene classifiers are computed from the whole images and efficiently exploit both global and local attributes.

A significant amount of emotion recognition research is based on the extraction of geometric and appearance facial features [33]. However, face information alone is not enough to succeed in the problem of group-level emotion recognition in the wild [21]. Although body features are not as widely employed as facial features, they have proven useful to predict emotions [16, 30]. Motivated by the success of face and body-level features in emotion recognition, the proposed approach uses face and skeleton features, in addition to the scene classifiers, for the GER problem. Multiple models are trained separately based on scene, face and skeleton features. Late decision fusion is optimized to hybridize those models. The results exhibit an overall classification accuracy of 80.61% on the testing set, compared to 53.62% provided by the baseline[1].

The structure of this paper is as follows. Section 2 presents the details of the proposed approach. Section 3 validates the performance of the proposed approach on the Group Affect Database 2.0, and compares it with the baseline. The paper concludes in Section 4 with final remarks.

2 THE PROPOSED METHOD

The proposed method is a hybrid network (Figure 1) that combines predictions from deep neural network models learned on scene, faces and skeletons.

2.1 Scene Classification

Li *et al.* [20] has demonstrated that an average of predicted happiness intensities of all the faces in the image alone is inadequate to predict the happiness level of the image. Therefore, holistic scene descriptors plays an important role in the emotion prediction of a group. The CENTRIST descriptor used in the sub-challenge baseline method only achieves 52.97% and 53.62% classification accuracy on the validation and test sets of the Group Affect Database 2.0, respectively. In this paper, we demonstrate that a superior classification accuracy can be achieved by exploiting state-of-the-art deep models.

Deep neural network architectures, such as AlexNet [17], VGG [25], GoogLeNet [26], Inception-v2 [27] and ResNet [13], have achieved state-of-the-art performance on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [5] in recent years. In this paper, we propose to use these deep architectures as scene classifiers to predict the overall happiness level of the group. We argue that since these state-of-the-art neural networks are able to identify objects in large scale (1000 classes in ImageNet), the information they learned, such as the class label, the location and the shape of the object, is useful to infer high level knowledge, such as people layout, activity and the background environment, which in turn, is useful to predict group emotion.

[1] The source code is available at https://github.com/gxstudy/EmotiW2017_Group

Specifically, whole images are used as input to state-of-the-art deep networks. Each architecture is modified by changing the number of neurons in the last layer to 3, indicating a ternary classification, having as targets Negative, Neutral and Positive emotions for the group. With the exception of the last layers, the modified architectures are initialized with the models trained on ImageNet, which are expected to have learned essential parameters to identify objects. The last layer in each architecture is initialized in the same way as in the original setup of each architecture, and trained with a learning rate for the weight and bias terms which is set to be 10 times larger than the overall learning rate. The learning parameters of each architecture, such as the overall learning rate, the weight decay, and the learning policy are set the same as in the original submissions to the ImageNet challenge.

2.2 Face Prediction

Faces and facial landmarks are first detected using the method described in [15], then a 2D affine transformation where the left and right eye corners of all the images are aligned to the same positions is performed (the code for the the face detection and alignment algorithms is developed based on [28]). The Viola-Jones [32] face detector, which is better at detecting small faces, is also used to detect the faces ignored by the first algorithm. False positives are filtered out by a face model trained on CNN networks provided by [11].

The VGG-FACE model was presented as the result of training the 16-layer VGG architecture on a large-scale dataset containing 2.6M images of 2.6K celebrities and public figures for face recognition in [23]. As the state-of-the-art for face recognition, a modified version of the VGG-FACE model is employed to perform emotion recognition in this paper. The modification consists of changing the number of neurons in the last fully-connected layer to 3. The modified architecture is initialized with the weights of the original VGG-FACE model, with the exception of the last fully-connected layer, which is initialized with weights sampled from a Gaussian distribution of zero mean and variance 1×10^{-4}. The features learned by the first CNN layers typically correspond to generic features, such as contours and edges. Therefore, the weights of all the convolutional layers are kept the same as in the original VGG-FACE model, while the weights of the first two fully-connected layers are fine-tuned and the last fully-connected layer is trained from scratch.

The modified VGG-FACE model is first fine-tuned on a combined facial emotion dataset. The facial emotion dataset (30205 samples in total) is a combination of the facial expression recognition 2013 (FER-2013) dataset [12] and the GENKI-4K dataset [35], with the negative collection being the angry and sad classes from the FER-2013 database, the neutral being the combination of neutral classes from the FER-2013 and the GENKI-4K databases, and the positive being the combination of the happiness class from both datasets.

The FACE model is further fine-tuned on the detected faces of the Group Affect Database 2.0. During training, all the faces are re-scaled to 256 × 256 pixels and have the same weight when fine-tuning the parameters of the network. During testing, however, the overall happiness level is a weighted sum of the prediction of individual faces, where the weight of each face is proportional to the face size (width × height). The motivation behind the weighted summation is that faces closer to the camera contain more details, and therefore, should be given more weigh than the ones away from the camera.

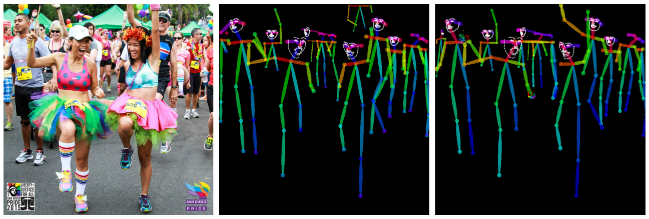

Figure 2: Samples of skeleton representations. Left: original image; Middle: skeleton representation 1 (includes faces and bodys); Right: skeleton representation 2 (includes faces, bodys and hands)

2.3 Skeleton Classification

Facial landmarks have been widely used to recognize emotions either directly as location vectors or by computing geometric distances among them [33]. Meanwhile, body features are encoded mostly as hand-crafted features of the body regions in the image [22]. To explore both the facial landmarks and body features without losing the relative position of keypoints, we propose to use a feature representation that is referred to as skeleton features, and corresponds to the collection of keypoints of human face, body and hands (Figure 2). We argue that skeleton is an effective way to learn the overall emotions of the group by emphasizing facial expressions, layout, pose, and the gesture of the group.

The skeleton of each image is extracted using OpenPose [2, 29, 34], which can jointly detect human body, hand and facial keypoints (130 keypoints in total for each person) on single images, invariant to the number of detected people in the image. The detection is not perfect, as shown in Figure 2, as keypoints can be neglected or misdetected due to either occlusion or false positive texture similarities. However, the results show a clear mouth shape, pose, gesture, and layout of humans in the image. Two type of skeleton features, facial and body landmarks without and with hand keypoints (which are referred to as skeleton representations 1 and 2, respectively), are used in this paper. The extracted images are of the same size as the original images and only contain skeleton features of the group. Inception-v2 and ResNet are fine-tuned on the skeleton images to predict happiness level of the group. We followed the training procedure described in Section 2.1. Experimental results show that both skeleton representations 1 and 2 play an important role in classifying overall emotions.

2.4 Feature Fusion and Decision Fusion

A single classifier is generally unable to handle the scalability and variability of modern pattern recognition tasks, as fusing the decisions of different classifiers have demonstrated superior performance. In this paper, we explore both decision fusion and feature fusion on the GER task.

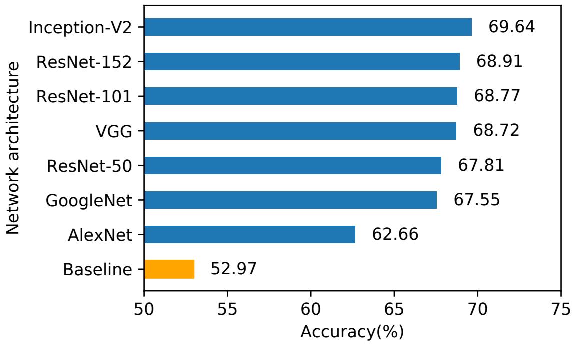

Figure 3: Comparison between the classification accuracy of the baseline and the proposed scene classifiers on the validation set of the Group Affect Database 2.0.

Since the fine-tuned VGG-FACE model and the fine-tuned VGG scene model have exactly the same network architecture, the fc6 layers (4096 features in total) of both models are extracted and concatenated. Support vector machines (SVMs) [3] based on radial basis function (RBF) kernels are trained on the fused feature representation to classify the overall happiness level of the group.

A grid search is performed across the predictions[2] of all the models to learn the weight of each model. Even though it is simply an exhaustive search through a manually specified subset of the hyper-parameter space and it's not guaranteed to be optimal, it is an effective and widely used way to fuse decisions. Furthermore, redundant models can be identified by analyzing the resulting weight. Models whose weight is 0 are redundant, and therefore, are removed from the hybrid network. Figure 1 shows the overall structure of the hybrid network after the removal of all the redundant models.

3 EXPERIMENTS

3.1 Group-Level Emotion Recognition Sub-challenge

Group-level emotion recognition is one of the sub-challenges in the fifth Emotion Recognition in the Wild (EmotiW 2017) Grand Challenge [8]. The images in this sub-challenge are from the Group Affect Database 2.0 [10], which contains 3630, 2065, 772 images in the training, validation and testing sets, respectively. These images are collected from social events, such as convocations, marriages, parties, meetings, funerals, protests, etc. Participants compete on the accuracy of classifying the group perceived emotion as Positive, Neutral or Negative on the testing data[3].

Table 1: Confusion matrix on the validation set of the fine-tuned Inception-v2 scene classifier, with overall accuracy being 69.64% and unweighted sum of accuracies per class being 70.14%.

	Neg	Neu	Pos
Neg	76.06	14.36	9.57
Neu	22.80	65.80	11.40
Pos	6.60	24.84	68.56

Table 2: Confusion matrix of the combined scene classifiers on the validation set, with overall accuracy being 72.35% and unweighted sum of accuracies per class being 72.72%.

	Neg	Neu	Pos
Neg	77.48	14.54	7.98
Neu	20.47	67.71	11.81
Pos	5.05	21.99	72.96

3.2 Scene Classification Results

AlexNet, VGG 16-layer network, GoogLeNet, Inception-v2 and ResNet are explored as the deep learning architectures to classify group-level emotion on the whole images. Experimental results (Figure 3) show that deep neural networks perform significantly better than CENTRIST for scene classification on the validation set. The confusion matrix corresponding to the fine-tuned Inception-v2 model is shown in Table 1.

It is worth noting that the performance of these fine-tuned networks is consistent with the performance of their original versions on the 1000-label classification problem of ImageNet, which suggests that transfer learning from ImageNet classification is worthy and higher accuracy on ImageNet corresponds to better initialized parameters and network structure for the GER problem. However, decision fusion of these models only increases the performance by 2.71% on the overall accuracy (Table 2).

There is redundancy between the different deep networks employed for scene classification. However, as discussed in Section 2.4, such redundancy is removed using the weights generated by the decision fusion stage.

3.3 Face Prediction Results

The confusion matrix of the classification on the validation dataset associated to the fine-tuned VGG-FACE model is shown in Table 3. This model exhibits better performance on the positive and negative classes than on the neutral class. The reason is that, unlike neutral-emotion and sad faces, smiley and angry faces usually come with clear muscular indication. That is why additional information related to the scene and to the activities performed in the image are necessary to improve emotion recognition. For example, a group of people with neutral faces may be either neutral as in a meeting or sad as in a silent protest, which are difficult to distinguish without information other than faces.

Given that the fine-tuned VGG-FACE model has high accuracy and low false-positive rate on the positive class, two new predictors were developed, positive-only predictor and non-positive predictor, to use in combination with the fine-tuned VGG-FACE model. The idea behind it is that the prediction is expected to be more reliable if the VGG-FACE model predicts a sample as positive instead of non-positive, since the fine-tuned VGG-FACE model is better at predicting the positive class.

[2] By predictions we mean the probabilities of an image belonging to each class.

[3] Note that since the class distribution is unbalanced (the test data contains 311, 165, and 296 images in the negative, neutral and positive classes, respectively), the accuracy we compete on is the weighted sum of accuracies per class, where the weight for each class is the number of samples in that class, we refer to as the overall accuracy. However, the unweighted sum of accuracies per class is also provided in this paper to clarify the confusion that readers may have when they compute the unweighted sum of accuracies per class directly from the confusion matrices and find it different from the overall accuracy.

Table 3: Confusion matrix on the validation set of the fine-tuned VGG-FACE model, with overall accuracy being 72.78% and unweighted sum of accuracies per class being 72.85%.

	Neg	Neu	Pos
Neg	76.24	16.31	7.45
Neu	33.52	60.16	6.32
Pos	4.79	13.07	82.15

Table 4: Confusion matrix of the SVM predictions on fused features, with overall accuracy being 74.00% and unweighted sum of accuracies per class being 74.51%.

	Neg	Neu	Pos
Neg	81.38	13.83	4.79
Neu	28.43	66.48	5.08
Pos	6.47	17.85	75.68

Table 5: Confusion matrix of the Inception-v2 Skeleton 1 classifier on the validation set, with overall accuracy being 63.20% and unweighted sum of accuracies per class being 62.98%.

	Neg	Neu	Pos
Neg	59.22	26.24	14.54
Neu	18.68	70.74	10.58
Pos	10.87	30.14	58.99

Table 6: Confusion matrix of the ResNet-152 Skeleton 2 classifier on the validation set, with overall accuracy being 64.99% and unweighted sum of accuracies per class being 64.45%.

	Neg	Neu	Pos
Neg	59.22	22.70	18.09
Neu	21.43	63.74	14.84
Pos	7.89	21.73	70.38

The confidence level is indicated by the weight assigned to the predictions. In order to assign different weights, the positive and non-positive predictions have to be split apart. The splitting rule is as follows: (1) if the fine-tuned VGG-FACE model favors the positive class, then the probability for positive class of the positive-only predictor is 1, and the probabilities for the negative and neutral classes are 0; while in the non-positive predictor, all the probabilities are 0; *e.g.*, if the prediction of the fine-tuned VGG-FACE model is $[0.15, 0.05, 0.8]$, then the positive-only prediction is $[0, 0, 1]$ and the non-positive prediction is $[0, 0, 0]$; (2) if the fine-tuned VGG-FACE model favors either the negative or the neutral class, then all the output probabilities of the positive-only predictor are set to 0; while the output probabilities of the non-positive predictor for both negative and neutral are set to 0.5; *e.g.*, if the prediction of the fine-tuned VGG-FACE model is $[0.9, 0.05, 0.05]$, then the positive-only prediction is $[0, 0, 0]$ and the non-positive prediction is $[0.5, 0.5, 0]$. Note that the unweighted sum of the positive-only predictor and the non-positive predictor should always sum up to 1, indicating they are from the same prediction.

Optimal weights can be learned to combine the original fine-tuned VGG-FACE model, the positive-only predictor and the non-positive predictor. Note that the negative and neutral predictions are not being separated even though the VGG-FACE model exhibits higher accuracy on the negative class than on the neutral class. The reason is that scene classifiers and the skeleton classifiers are better at distinguishing the negative from the neutral class than the VGG-FACE model, so the non-positive predictor serve as a way to favor the non-positive classes but let the scene classifiers and skeleton classifiers decide the exact class.

3.4 Skeleton Prediction Results

Inception-v2 and ResNet-152 network architectures are fine-tuned with the skeleton representations. Table 5 and 6 show the confusion matrices of the Inception-v2 trained on the Skeleton Representation 1 and of the ResNet-152 trained on the Skeleton Representation 2, respectively. Even though the overall accuracy of the skeleton classifiers is lower than the one of the face and scene classifiers, the skeleton classifiers show superior performance on the neutral class. It verifies the conjunction that the face landmarks, pose, gesture and layout of the group convey information about the overall emotion.

Table 7: Accuracies of model fusions on the validation set.

Fused Models	Acc (%)
Incetion-v2 Scene classifier + VGG-FACE classifier	77.29
+ Positive-only predictor	78.06
+ Non-positive predictor	78.26
+ SVM feature fusion classifier	79.52
+ Inception-v2 Skeleton 1 classifier	79.90
+ Resnet152 Skeleton 2 classifier	80.05

3.5 Results of the Feature Fusion, Decision Fusion and Final Submission

The confusion matrix of the SVM prediction using the fused fc6 features from the fine-tuned VGG scene classifier and the fine-tuned VGG-FACE model is shown in Table 4. The overall accuracy is higher than the accuracy obtained when using the individual features.

After removal of the redundant models which have weights equal to 0 after decision fusion, the hybrid network ends up with 7 models, as shown in Figure 1. To demonstrate the contribution of each model, decision fusion of these 7 models is performed gradually, with one model adding up every time, as shown in Table 7.

The challenge allows 7 submissions in total. For the first submission, we trained models on the training data only and learn the weights of the decision fusion by favoring the highest accuracy on the validation data, which resulted in 7 models. For the second submission, we trained these 7 models on the combination of training and validation data (the learning parameters of the models are kept the same as in the first submission with only an increase in the step size and maximum number of iterations since the size of the training set becomes larger), and kept the same weights as in the first submission. Since the models and the weights for submission 2 are not learned from decision fusion but are inherited from submission 1, they can hardly be optimal. However, the larger size of the training data always leads to better performance, so the overall accuracy of submission 2 increases. The rest of the submissions are subtle adjustments of the weights only, with the aim of finding a better local minimal each time. The accuracies of all the submissions are shown in Table 8, with the best submission being the sixth one. The confusion matrices for submissions 1 and 6 are shown in Table 9 and Table 10, respectively.

Table 8: Submission accuracies

Sub	Training Data	Val	Test
1	Training Set Only	80.05	78.25
2	Training + Val	-	78.67
3	Training + Val	-	80.06
4	Training + Val	-	80.47
5	Training + Val	-	79.91
6	Training + Val	-	80.60
7	Training + Val	-	78.81

Table 9: Confusion matrix of Submission 1, with overall accuracy being 78.25% **on the testing data.**

	Neg	Neu	Pos
Neg	88.75	6.75	4.5
Neu	13.33	54.55	32.12
Pos	8.11	24.66	67.23

Table 10: Confusion matrix of Submission 6, with overall accuracy being 80.61% **on the testing data.**

	Neg	Neu	Pos
Neg	87.14	6.75	6.11
Neu	8.48	57.58	33.94
Pos	2.37	24.66	72.97

4 CONCLUSIONS

In this paper, we propose a hybrid network that combines 7 models for group-level emotion recognition in the wild. To the best of our knowledge, skeleton representations, positive-only and non-positive predictors are presented and explored for the GER problem for the first time in this paper. The overall accuracy of the proposed method achieves 80.61% on the test data, which is significantly larger than the baseline of 53.62%.

ACKNOWLEDGMENTS

The work is supported by the National Science Foundation under Grant No. 1319598.

REFERENCES

[1] J. Bullington. 2005. Affective computing and emotion recognition systems: the future of biometric surveillance?. In *Proceedings of the 2nd annual conference on Information security curriculum development*. ACM, 95–99.

[2] Z. Cao, T. Simon, S. Wei, and Y. Sheikh. 2016. Realtime multi-person 2D pose estimation using part affinity fields. *arXiv preprint arXiv:1611.08050* (2016).

[3] C. Cortes and V. Vapnik. 1995. Support-Vector Networks. *Mach. Learn.* 20, 3 (Sept. 1995), 273–297.

[4] R. Cowie et al. 2001. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine* 18, 1 (2001), 32–80.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*.

[6] A. Dhall, A. Asthana, and R. Goecke. 2010. Facial expression based automatic album creation. In *International Conference on Neural Information Processing*. Springer, 485–492.

[7] A. Dhall, R. Goecke, and T. Gedeon. 2015. Automatic group happiness intensity analysis. *IEEE Transactions on Affective Computing* 6, 1 (2015), 13–26.

[8] A. Dhall, R. Goecke, S. Ghosh, J.i Joshi, J. Hoey, and T. Gedeon. 2017. From Individual to Group-level Emotion Recognition: EmotiW 5.0 *(ICMI 2017)*. ACM.

[9] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. 2012. Collecting large, richly annotated facial-expression databases from movies. 19, 3 (July 2012), 34–41.

[10] A. Dhall, J. Joshi, K. Sikka, R. Goecke, and N. Sebe. 2015. The more the merrier: Analysing the affect of a group of people in images. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, Vol. 1. IEEE, 1–8.

[11] Y. Fan, X. Lu, D. Li, and Y. Liu. 2016. Video-based Emotion Recognition Using CNN-RNN and C3D Hybrid Networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI 2016)*. ACM, New York, NY, USA, 445–450.

[12] I.J. Goodfellow et al. 2013. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*. Springer, 117–124.

[13] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.

[14] J. Hernandez, M.E. Hoque, W. Drevo, and R.W. Picard. 2012. Mood meter: counting smiles in the wild. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 301–310.

[15] V. Kazemi and J. Sullivan. 2014. One Millisecond Face Alignment with an Ensemble of Regression Trees. In *CVPR*.

[16] A. Kleinsmith and N. Bianchi-Berthouze. 2013. Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing* 4, 1 (2013), 15–33.

[17] A. Krizhevsky, I. Sutskever, and G.E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. 1097–1105.

[18] C. Küblbeck and A. Ernst. 2006. Face detection and tracking in video sequences using the modified census transformation. *Image and Vision Computing* 24, 6 (2006), 564–572.

[19] O. Kwon, K. Chan, J. Hao, and T. Lee. 2003. Emotion recognition by speech signals. In *Eighth European Conference on Speech Communication and Technology*.

[20] J. Li, S. Roy, J. Feng, and T. Sim. 2016. Happiness Level Prediction with Sequential Inputs via Multiple Regressions. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI 2016)*. ACM, New York, NY, USA, 487–493.

[21] W. Mou, O. Celiktutan, and H. Gunes. 2015. Group-level arousal and valence recognition in static images: Face, body and context. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 5. IEEE, 1–6.

[22] W. Mou, O. Celiktutan, and H. Gunes. 2015. Group-level arousal and valence recognition in static images: Face, body and context. In *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 5. IEEE, 1–6.

[23] O.M. Parkhi, A. Vedaldi, A. Zisserman, et al. 2015. Deep Face Recognition.. In *BMVC*, Vol. 1. 6.

[24] B. Schuller, G. Rigoll, and M. Lang. 2003. Hidden Markov model-based speech emotion recognition. In *International Conference on Multimedia and Expo*, Vol. 1. IEEE, I–401.

[25] K. Simonyan and A. Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[26] C. Szegedy et al. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.

[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2818–2826.

[28] H. Tal, H. Shai, P. Eran, and E. Roee. 2015. Effective Face Frontalization in Unconstrained Images. In *CVPR*.

[29] S. Tomas, J. Hanbyul, M. Iain, and S. Yaser. 2017. Hand Keypoint Detection in Single Images using Multiview Bootstrapping. In *CVPR*.

[30] J. Van den Stock, R. Righart, and B. De Gelder. 2007. Body expressions influence recognition of emotions in the face and voice. *Emotion* 7, 3 (2007), 487.

[31] T. Vandal, D. McDuff, and R. El Kaliouby. 2015. Event detection: Ultra large-scale clustering of facial expressions. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 1. IEEE, 1–8.

[32] P. Viola and M. Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *CVPR*, Vol. 1. I, 511 , I, 518 vol.1.

[33] V. Vonikakis, Y. Yazici, V.D. Nguyen, and S. Winkler. 2016. Group Happiness Assessment Using Geometric Features and Dataset Balancing. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI 2016)*. ACM, New York, NY, USA, 479–486.

[34] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. 2016. Convolutional pose machines. In *CVPR*.

[35] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan. 2009. Toward practical smile detection. *IEEE transactions on pattern analysis and machine intelligence* 31, 11 (2009), 2106–2111.

[36] J. Wu and J.M. Rehg. 2011. CENTRIST: A Visual Descriptor for Scene Categorization. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 8 (2011), 1489–1501.