Fusion of Hand-crafted and Deep Features for Empathy Prediction

Saurabh Hinduja, Md Taufeeq Uddin, Sk Rahatul Jannat, Astha Sharma, and Shaun Canavan
Department of Computer Science and Engineering, University of South Florida, Tampa, FL, USA

Abstract— **We propose an approach to the OMG-Empathy Challenge for predicting self-annotated, continuous values of valence within the range [-1,1]. We propose the fusion of hand-crafted and deep features, extracted from both actor and listener data, to predict these valence levels. The hand-crafted features include image level fusion, facial landmarks, and spectrogram features. Our proposed fusion approach can utilized in multiple parts (i.e. sub-modules), specifically utilized for the generalized track, leading to unique submissions to address the challenge problem. First, both actor and listener images are fused at the image-level. Secondly, facial landmarks from both the actor and listener are fused into one feature vector which is then used to train a random forest for prediction of continuous valence levels. Finally, we use a weighted fusion of the predicted values from both hand-crafted and deep features. We show competitive results on the OMG-Empathy challenge validation set.**

I. Introduction

When analyzing human intelligence, emotion is a necessary and important aspect to study [14]. Considering this, the field of Affective Computing [17], especially emotion recognition has been gaining in popularity in the past two decades. Many of these works have focused on posed expression [20], [19], however, more recently large-scale, multimodal datasets with spontaneous change in emotion have become available [25], [23], [24]. These new datasets have allowed for comparisons between emotion recognition with posed and spontaneous data. Fabiano et al. [8], found that 3D facial landmarks are a suitable modality for recognizing emotion in both, however, the statistically important landmarks were found in different areas of the face. In posed data, they are mostly near the mouth, while there is a more balanced set of important landmarks found in the mouth and eye regions with spontaneous 3D data.

Recent works have been successful in using both images and audio to recognize emotion. Liu et al. [12] investigated facial expressions using a deep belief network to perform feature learning, selection, and classifier construction. They did this iteratively using a unified loopy framework. Dhall et al. [7] conducted the Video and Image-based Emotion Recognition Challenge in the Wild (EmotiW2015). This challenge made use of the Acted Facial Expression in the Wild 5.0 [6], and the Static Facial Expression in the Wild 2.0 [5] datasets that contain audio and image data, to mimic real-world condition for emotion recognition. Subramaniam et al. [16], recognized human personality traits such as extraversion, agreeableness, conscientiousness, neuroticism, and openness. They trained a deep neural network on temporally ordered audio, as well as stochastic visual features. Zhang et al. [22] learned affective features by producing audio-visual features from a CNN and 3D-CNN. They then fused these features to train a deep-belief network. This network is trained to learn a discriminative audio-visual representation. Recently, an application, that can be run on a variety of devices (PC, tablet, phone), was developed showing that the fusion of audio and video can be effectively used for ubiquitous emotion recognition [10]. This work fused cropped face images, with raw audio signals that were plotted on the 2D image plane. The fused data was then used to train CNN for predicting emotions such as sad, and happy.

While the above works have focused on recognizing personality traits and emotions such as sad, happy, surprise, anger, disgust, and pain, a complementary problem is the continuous prediction of spontaneous affect in the arousal-valence space. Nicolaou et al. [13], detailed one of the first works to fuse face, shoulder gestures, and audio to predict continuous emotion. They showed that, on average, a Long Short-term Memory network outperformed Support Vector Machines for Regression for this problem. Using a feature-based approach, Wagner et al. [18] used physiological signals collected from a musical-induction to classify positive/negative arousal and valence levels. Greco et al. [9], used clustering algorithms, on electrodermal signals, to discern arousal and valence levels induced by sound stimuli. The sound stimuli were taken from the International Affective Digitized Sound dataset [2].

In this paper, the focus is on the valence space, specifically the prediction of continuous valence levels from video and audio in the OMG-Empathy challenge dataset [1]. This challenge consists of two tracks: (1) generalized, and (2) personalized - we submit to both tracks. For this challenge, we propose the fusion of hand-crafted features(image-level, facial landmarks, and spectrogram features), along with deep features from the images of the actors and listeners (Fig. 1). Our proposed approach fuses data at 3-levels: image-level; facial landmarks; and hand-crafted and deep features. Using this approach, we show competitive results, on the validation set. Our contribution can be summarized as follows:

1) We propose the fusion of hand-crafted and deep features for predicting empathy. This approach includes sub-modules that can be used as a single approach to predicting empathy.
2) We detail the results for 3 submissions for the generalized track of the challenge, and 1 submission for the personalized track of the challenge. Each submission details a sub-module of our fusion approach.
3) We obtain competitive results on the validation set for

978-1-7281-0089-0/19/$31.00 ©2019 IEEE

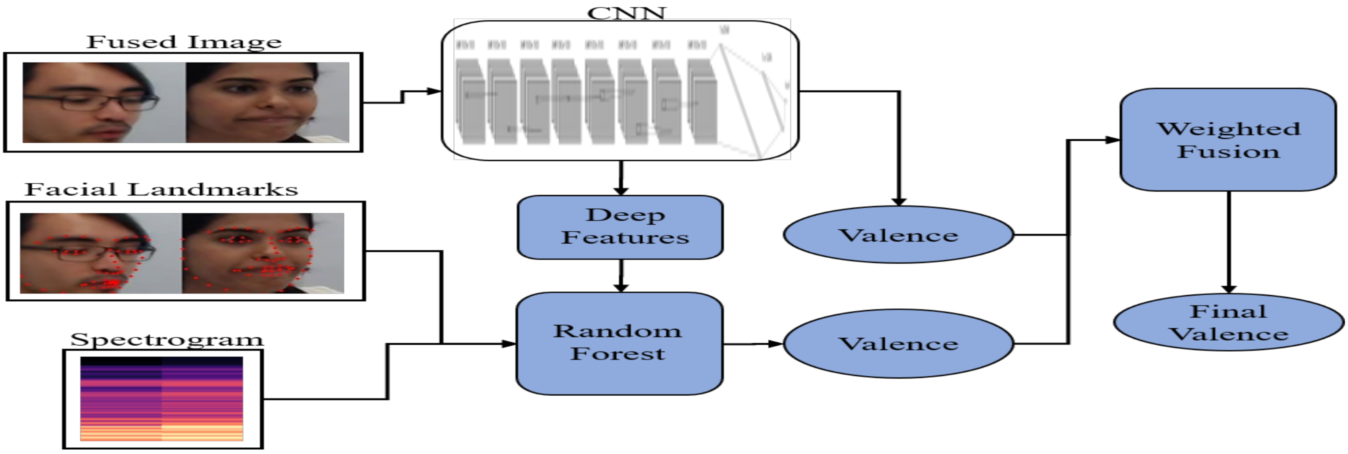

Fig. 1. A pictorial representation of submission 3 of generalized track (Section III). Here, we feed fused image of both actor and listener to CNN to predict valence score, and to extract deep features. We also feed landmarks (both actor and listener), spectrogram, and deep features to random forest to predict valence score. Finally, we perform weighted fusion of valence scores obtained from CNN and random forest to produce final valence score.

the OMG-Empathy dataset, for both the generalized and personalized tracks.

II. Method

We propose the fusion of hand-crafted and deep features extracted from both actor and listener data. Each of the feature types along with the dataset used to extract the features is detailed below.

A. Dataset

The One-Minute Graduate Empathy Prediction Challenge (OMG-Empathy) is meant to be a guide for the creation of empathy prediction models. To facilitate this, the dataset contains 80 videos that involve a semi-scripted talk that includes both an actor, as well as a listener. In total, there are 8 stories: (1) Childhood friend; (2) starting a band; (3) relationship with dog; (4) bad flight; (5) traveling adventure; (6) cheated on exam; (7) won martial arts contest; and (8) ate bad food. The actor was free to improvise these stories as they saw fit. Due to this, the length of each of the videos is not consistent. On average each story/video is 5 minutes and 12 seconds long. After the conversation between the actor and listener was finished, the listener watched the video and rated how they felt giving the ground-truth valence levels in the range of negative one to positive one. For the challenge, the dataset was pre-separated into training, validation, and testing sets.

B. Hand-crafted Features

For our hand-crafted features, we extracted facial landmarks from the actor and listener images, and spectrogram features from the audio signals. Details on both are given below.

Facial Landmarks. We used a pretrained facial landmarks detector [11], [15] to identify 136 landmark coordinates (x, y) of the face representing eyes, eyebrows, nose, mouth and jaw of subject (Fig. 1). We extracted 68 (x, y) pairs of points, given its success on solving many other affective computing problems [15], from the images of both actor and listener, and used the raw x and y coordinates to train a random forest regressor [3] to predict valence levels.

Spectrogram Features. Generally, spectrograms are used to visualize the strength of frequency over time, however, we extracted spectrogram features [26] to determine how the changes of frequency over time, through actor and listener conversation, impacts the listener's empathy. In order to get features for 25 frames per second of we segmented the audio time series with a specific length. Each specific segment produces 366 features. As the audio has two channels, the 366 features represent the combination frequencies of two channels per frame. These features are used to facilitate the prediction of empathy (Section II-D).

C. Deep Network and Features

We used a convolutional neural network (CNN) in our submissions for both tracks. The CNN (Fig. 1) has 8 convolutional layers and 3 fully connected layers, with the Adadelta optimizer [21] used along with RMSE for the error. For the personalized track, the output of the network was a single neuron. For the generalized track, two submissions were made using our CNN. First, similar to the personalized track the network has an output of a single neuron. Second, 256 deep features were extracted that were combined with the hand-crafted features to train a random forest.

D. Fusion Techniques

Our proposed fusion approach allows for unique submissions to the challenge. Each of the fusion approaches (i.e. sub-modules) are detailed below.

Actor and listener image-level fusion. Our CNN is trained with fused images of both the actor, on the left and the subject on the right. Both the actor and subject images are cropped to 100×100 and image-level fusion [4] is performed on both to create a new image of size 200×100 (Fig. 1). Subsequently, the new image of size 200×100 is used to train our CNN. We have chosen this type of fusion, compared

TABLE I

PERSONALIZED TRACK RESULTS ON VALIDATION SET. HERE, SUB INDICATES SUBJECT.

Model	**Sub 1**	**Sub 2**	**Sub 3**	**Sub 4**	**Sub 5**	**Sub 6**	**Sub 7**	**Sub 8**	**Sub 9**	**Sub 10**
CNN trained on fused face images	**0.07**	0.01	-0.05	-0.04	-0.05	-0.03	**0.01**	-0.02	0.01	**0.12**
Baseline (Listener)	0.01	**0.11**	**0.04**	**0.1**	**0.11**	**0.35**	-0.01	**0.05**	**0.05**	0.10
Baseline (Actor)	0.00	-0.12	-0.06	0.00	-0.04	0.01	-0.05	-0.03	-0.07	-0.00

TABLE II

PERSONALIZED TRACK RESULTS ON TESTING SET USING CNN MODEL.

Subject	**Sub 1**	**Sub 2**	**Sub 3**	**Sub 4**	**Sub 5**	**Sub 6**	**Sub 7**	**Sub 8**	**Sub 9**	**Sub 10**	**Mean**
CCC	-0.11	0.10	0.15	0.10	0.04	-0.03	-0.06	-0.01	-0.02	0.00	0.02

to fusing the 2 images at the fully connected layers, as it has been shown to be effective when changes in pose are found in the data [4].

Actor and listener facial landmark fusion. We extracted 136 facial landmark features from actor and listener (68 landmarks each), and then we fused the landmarks as:

$$FL = \big(xa_i, ..., xa_N, ya_i, ..., ya_N,\\ xl_i, ..., xl_N,\ yl_i, ..., yl_N\big), \tag{1}$$

where xa_i and ya_i are the i^{th} facial landmarks (x, y) from the actor, xl_i and yl_i are the i^{th} facial landmarks (x, y) from the listener, and $N = 68$ (total number of landmarks).

Weighted fusion of valence levels from deep and hand-crafted features. We trained a random forest on fused spectrogram features, facial landmarks, and deep features extracted from our CNN, to predict valence levels. These features were fused as:

$$multimodal_{fusion} = \big(SF_i, ..., SF_{sn}, FL,\\ DF_j, ..., DF_{dn}\big), \tag{2}$$

where SF_i is the i^{th} spectrogram feature (Section II-B), FL is the fused landmark feature vector (Equation 1), DF_j is the j^{th} deep feature (Section II-C), and $sn = 366$ (total number of spectrogram features) and $dn = 256$ (total number of deep features).

Given the valence levels from this fusion, along with the valence levels obtained from our CNN, to obtain our final valence levels, we took a weighted fusion of these 2 sets of valence levels as:

$$V_i = \frac{(w \times fv_i) + dv_i}{2}, \tag{3}$$

where w is the weight, fv_i is the i^{th} predicted valence level of the fused features (Equation 2), and dv_i is the i^{th} predicted valence level from our CNN. We have empirically found that a weight of 2.5 is sufficient for a strong weighted fusion, which we used in our submission.

III. SUBMISSIONS AND RESULTS

A. Personalized Track Submission

For the personalized track, we adopt a non-fusion-based technique. First, each listener image (training data) is cropped to 100×100 around the face of the subject. We then train our CNN (Section II-C) on each listener separately (e.g. create a deep model for each listener). The valence levels are calculated directly from the CNN (i.e. single neuron output).

TABLE III

GENERALIZED TRACK RESULTS ON VALIDATION SET.

Submission Method	**CCC Score**
CNN trained on fused face images	0.10
RF trained on actor and listener facial landmarks	0.067
Weighted fusion of deep and hand-crafted features	0.061
Baseline (Listener)	**0.1111**
Baseline (Actor)	-0.04

TABLE IV

GENERALIZED TRACK RESULTS ON TESTING SET.

Submission Method	**CCC Score**
CNN trained on fused face images	-0.03
RF trained on actor and listener facial landmarks	**0.04**
Weighted fusion of deep and hand-crafted features	0.03

B. Generalized Track Submissions

The generalized track includes the following submissions:

1) Valence levels are calculated by our CNN using the image-level fusion of cropped actor and listener images.
2) Valence levels are calculated by a random forest (RF) that is trained on fused facial landmarks of the actor and listener.
3) Valence levels are calculated from the weighted fusion of valence levels calculated from the image-level fusion of the actor and listener, as well as the valence levels calculated from the fusion of deep-features and hand-crafted features.

C. Results

As can be seen in Table III, the proposed approaches to our 3 submissions perform comparable to the listener baseline, while outperforming the actor baseline results on the validation set. Calculating the valence levels directly from the CNN on the fused actor and listener images resulting in the highest CCC score of 0.10. While the proposed weighted fusion method achieved a lower CCC score compared to the

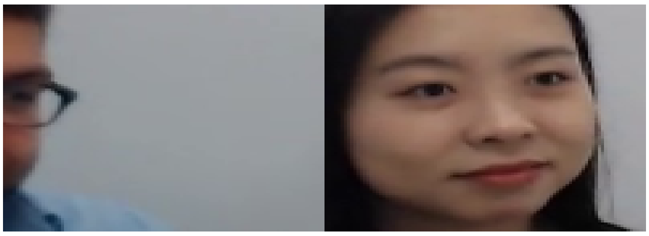

Fig. 2. Incorrect cropping of actor from testing data.

CNN, it shows comparable results to the baseline, showing significant improvement over the actor only baseline (weighted fusion contains features from both listener and actor). These results are encouraging, showing the power of fusing both the listener and actor.

As can be seen in Table IV, the fusion of the facial landmarks from the actor and listener achieved the highest CCC score at 0.04, however, the proposed weighted fusion of valence levels calculated from hand-crafted and deep features, performed comparatively well with a CCC score of 0.03. It is interesting to note, that while the image-level fusion performed the best on the validation set, it showed the worst performance of the three submissions on the testing set achieving a CCC score of -0.03. This can partially be explained due to the inaccurate cropping of faces, especially the actor images attributed to 1. Motion blur, 2. Actor looking in different directions and 3. Placement of the subject camera close to the actor. More images were incorrectly cropped in the testing set (e.g. the face was not found in the image) compared to the training and validation sets. An example of incorrect cropping, of the actor, from the testing set, can be seen in Fig. 2.

As noted in Section III-A, our submission for the personalized track consisted of cropped facial images of just the listener. The results shown in Table I are encouraging, showing the fusion of actor and listener images can be used to predict continuous level of valence. The results shown in Table II are also encouraging, showing similar prediction of continuous levels of valence with cropped facial images of just the listener. It also suggests that the listeners exhibited varying levels of expression in the collected videos, as there is a range of CCC scores across each subject. For example, Subject 3 achieved the highest personal CCC score with 0.15, and Subject 7 had the lowest with -0.06.

IV. Conclusion

In this paper, we have detailed our submission to the One-minute Gradual Empathy Prediction Challenge (OMG-Empathy). We proposed the fusion of hand-crafted and deep features that were extracted from actor and listener data from the supplied training, validation, and testing sets. The proposed fusion approach can be used in single submodules, which allowed us to make unique submissions to the challenge for the generalized and personalized tracks. We made 1 submission to the personalized track, where a CNN was trained on cropped facial images of only the listener, achieving competitive results. For the generalized track, we made 3 unique submissions: (1) image-level fusion of the cropped actor's and listener's faces; (2) fusion of facial landmarks from both actor and listener; and (3) weighted fusion of valence levels calculated from the spectrogram, image-level fusion of actor and listener, and deep features, as well as valence levels calculated from image-level fusion of the cropped actor and listener using CNN.

References

[1] P. Barros, E. Barakova, and S. Wermter. A deep neural model of emotion appraisal. *arXiv preprint, ArXiv:1808.00252*, 2018.

[2] M. Bradley and P. Lang. The international affective digitized sounds(iads): Affective ratings of sounds and instruction manual. *Technical Report, University of Florida, Gainseville*, 20007.

[3] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[4] S. Canavan, M. Kozak, Y. Zhang, and R. Sullins. Face recognition by multi-frame fusion of rotating heads in videos. *BTAS*, 2007.

[5] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *Intl. Conf. on Computer Vision Workshops*, 2011.

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

[7] A. Dhall and others. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *ICMI*, 2015.

[8] D. Fabiano and S. Canavan. Spontaneous and non-spontaneous 3d facial expression recognition using a statistical model with global and local constraints. *International Conference on Image Processing*, 2018.

[9] A. Greco, G. Valenza, L. Citi, and E. Scilingo. Arousal and valence recognition of affective sounds based on electrodermal activity. *IEEE Sensors Journal*, 17(3):716–725, 2017.

[10] R. Jannat et al. Ubiquitous emotion recognition using audio and video data. In *Intl. Joint Conf. and Intl. Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, 2018.

[11] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014.

[12] P. Liu, S. Han, Z. Meng, and Y. Tong. Facial expression recognition via a boosted deep belief network. *CVIU*, 2014.

[13] M. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Trans. on Aff. Com.*, 2(2):92–105, 2011.

[14] R. Picard, E. Vyzas, and J. Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on PAMI*, 23(10):1175–1191, 2001.

[15] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. *Image and vision computing*, 47:3–18, 2016.

[16] A. Subramaniam et al. Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features. In *ECCV*, 2016.

[17] J. Tao and T. Tieniu. Affective computing: A review. *Affective Computing and Intelligent Interaction*, pages 981–985, 2005.

[18] J. Wagner, J. Kim, and E. André. From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. *ICME*, pages 940–943, 2005.

[19] L. Yin et al. A 3d facial expression database for facial behavior research. *FG*, 2006.

[20] L. Yin et al. A high-res. 3d dynamic facial exp. database. *FG*, 2008.

[21] M. Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint, ArXiv:1212.5701*, 2012.

[22] S. Zhang et al. Learning affective features with a hybrid deep model for audio–visual emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):3030–3043, 2018.

[23] X. Zhang et al. A high-resolution spontaneous 3d dynamic facial expression database. *Face and Gestures*, pages 1–6.

[24] X. Zhang et al. Bp4d-spontaneous: A high resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

[25] Z. Zhang et al. Multimodal spontaneous emotion corpus for human behavior analysis. *CVPR*, 2016.

[26] L. Zheng, Q. Li, H. Ban, and S. Liu. Speech emotion recognition based on convolution neural network combined with random forest. In *CCDC*, 2018.