Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2018.DOI

Dominant and Complementary Emotion Recognition from Still Images of Faces

JIANZHU GUO[1], ZHEN LEI[1], (Senior Member, IEEE), JUN WAN[1,*], (Member, IEEE), EGILS AVOTS[2], (Student Member, IEEE), NOUSHIN HAJAROLASVADI[3], (Student Member, IEEE), BORIS KNYAZEV[4], ARTEM KUHARENKO[4], JULIO C. S. JACQUES JUNIOR[5,6], XAVIER BARÓ[5,6], HASAN DEMIREL[3], SERGIO ESCALERA[6,7,*], JURI ALLIK[8], and GHOLAMREZA ANBARJAFARI[2,9,*], (Senior Member, IEEE)

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, and University of Chinese Academy of Sciences, China
E-mail: {jianzhu.guo, zlei, jun.wan}@nlpr.ia.ac.cn
[2]iCV Research Lab, Institute of Technology, University of Tartu, Estonia
E-mail: {ea, shb}@icv.tuit.ut.ee
[3]Dept. of Electrical and Electronic Eng., Eastern Mediterranean University, Gazimagusa, KKTC, Turkey
E-mail: {noushin.hajarolasvadi,hasan.demirel}@emu.edu.tr
[4]NTechLab, Moscow, Russia
E-mail: {b.knyazev, a.kuharenko}@ntechlab.com
[5]Universitat Oberta de Catalunya, Spain
E-mail: {jsilveira,xbaro}@uoc.edu
[6]Computer Vision Center, Spain
[7]University of Barcelona, Spain
E-mail: sergio@maia.ub.es
[8]Institute of Psychology, University of Tartu, Estonia.
E-mail: juri.allik@ut.ee
[9]Department of Electrical and Electronic Engineering, Hasan Kalyoncu University, Turkey.

*Corresponding authors

This work has been partially supported by the Estonian Research Council Grant (PUT638 and IUT2-13), by the Estonian Center of Excellence in IT (EXCITE) funded by the European Regional Development Fund, by the Spanish projects TIN2015-66951-C2-2-R and TIN2016-74946-P (MINECO/FEDER, UE), by the European Commission Horizon 2020 granted project SEE.4C under call H2020-ICT-2015, by the CERCA Programme/Generalitat de Catalunya, by the National Key Research and Development Plan (Grant No. 2016YFC0801002), by the Chinese National Natural Science Foundation Projects ♯61502491, ♯61473291, ♯61572501, ♯61572536, ♯61673052, ♯61773392, ♯61403405. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPUs used for this research.

ABSTRACT Emotion recognition has a key role in affective computing. Recently, fine-grained emotion analysis, such as compound facial expression of emotions, has attracted high interest of researchers working on affective computing. A compound facial emotion includes dominant and complementary emotions (e.g. happily-disgusted, sadly-fearful), which is more detailed than the seven classical facial emotions (e.g. happy, disgust, etc.). Current studies on compound emotions are limited to use datasets with limited number of categories and unbalanced data distributions, with labels obtained automatically by machine learning-based algorithms which could lead to inaccuracies. To address these problems, we released the iCV-MEFED dataset, which includes 50 classes of compound emotions and labels assessed by psychologists. The task is challenging due to high similarities of compound facial emotions from different categories. In addition, we have organized a challenge based on the proposed iCV-MEFED dataset, held at FG workshop 2017. In this paper, we analyze the top three winner methods and perform further detailed experiments on the proposed dataset. Experiments indicate that pairs of compound emotion (e.g. surprisingly-happy *vs* happily-surprised) are more difficult to be recognized if compared to the seven basic emotions. However, we hope the proposed dataset can help to pave the way for further research on compound facial emotion recognition.

INDEX TERMS Dominant and complementary emotion recognition, compound emotions, fine-grained face emotion dataset.

I. INTRODUCTION

ARTIFICIAL intelligence agents such as robots and computers have become a prominent aspect of our lives and their presence will give rise to unique technologies. Therefore, Human-Computer Interaction (HCI) or Human-Robot Interaction (HRI) experiences become more realistic if computers/robots are capable of recognizing more detailed human expressions during the interaction. Hence, introducing techniques that enable automatic recognition of more detailed emotions than the classical ones is of significant interest. Emotion and expression recognition are natural and intuitive for humans, yet extremely complicated tasks during HCI, with applications ranging from mobile computing and gaming to health monitoring and robotics [1]–[6]. Automatic facial expression recognition can also be applied in vision-based automatic interactive machines [7]–[11], human emotion analysis [12]–[16], assistive robotics [17]–[20], and human-machine interfaces [21]–[24]. In general, facial expression recognition has become an important research topic within HCI/HRI communities and related areas, such as machine learning, computer vision, human cognition and pattern recognition.

Automatic recognition of facial expressions is a complex task because of significant variations in the physiognomy of faces with respect to person's identity, environment illumination conditions and head pose [25], [26]. When compound emotion recognition is considered, the task can be even harder. Currently, one of the main limitations to advance the research on automatic recognition of compound emotions is the lack of large and public labeled datasets in the field.

State-of-the art works for facial expression recognition usually focus on seven basic emotions, namely happy, surprised, fearful, sad, angry, disgust, and contempt [27]. However, there are some attempts to find out more precise and detailed facial emotion expressions [28]–[31] due to recent advances in the field of compound emotions [32]. Psychologists have come to the conclusion that different regions of the face convey different types of affective information [33]–[35]. This means that some parts of the face convey some emotions better than others. For instance, there is some evidence [33] that the upper part of the face, mainly the eyes and eyebrows, is more informative for human subjects in recognizing anger and fear. On the other hand, disgust and happiness appear to be mainly expressed with the mouth [31], [33], [36], whereas surprise can be conveyed equally with both parts of the face.

Compound emotion categories [32], [37], [38] have been introduced in order to investigate the emotional state of a person in a more detailed way through facial emotion expression analysis. Such pioneering works played an important role in the context of facial emotion expression analysis, as they propose to understand and recognize facial expressions from a different (fine-grained) point of view.

However, there are some limitations of fine-grained facial emotion recognition in existing works. First, the number of public databases in this field is limited [32], [39]. Second, current available public datasets have a small number of categories, i.e., 23 (EmotionNet [38]) and 22 (CFEE [32]), which may cover just a small portion of all possible compound emotions. Third, the labels provided with EmotionNet dataset are related to automatically detected Action Units (AU), which are used for compound emotion analysis. Although the AUs can be converted to compound emotion category, the results might not be accurate [38], [39] due to errors introduced by the AU recognition module.

To address the above mentioned limitations, we propose the following contributions:

- We released the iCV-MEFED dataset[1], which contains 50 compound emotion categories and has more than 30,000 images labeled with the support of psychologists, which should be able to provide labels with high accuracy. Although EmotionNet has about 1 million images, it contains noise labels (i.e., automatically obtained), as well as it is extremely unbalanced (as detailed in Sec. V).
- To push the research on the field, we organized a challenge [23] based on the iCV-MEFED dataset, held at the FG 2017. In this paper, we provide a substantial extension of our previous work [23]. In this sense, additional details are presented, and a more comprehensive and up-to-date literature review is provided. Furthermore, we introduce the top three winner methods in details and conduct additional experiments to analyze their performances.

The rest of this paper is organized as follows. Section II provides an overview of the related work on compound emotion recognition of facial expression. The dominant and complementary emotion recognition challenge is introduced in Section III, where the overall description and motivation of the iCV-MEFED dataset is presented. Section IV describes in short the top three winners' method from the organized competition, and Section V shows the performances of different methods on the iCV-MEFED dataset. Final discussions, suggestions for future work and conclusions are presented in Section VI.

II. RELATED WORKS

Past research on facial emotion expression recognition mainly focused on seven basic categories: happy, surprised, fearful, sad, angry, disgust, and contempt [27], [40]. However, there are many complex and more elaborated facial expressions humans do, built from the combination of different basic one, that started to attract more attention from the past few years within the computer vision and machine learning communities, i.e., the so called compound emotions.

Shichuan et al. [32] introduced compound facial emotion recognition and defined 22 emotion categories (e.g. happily-disgusted, sadly-fearful, sadly-angry, etc). They used Facial Action Coding System (FACS) [41] analysis to show the

[1] http://icv.tuit.ut.ee/icv-mefed-database.html

production of these distinct categories and released the Compound Facial Expressions of Emotion (CFEE) database. The CFEE dataset contains 5,060 facial images labeled with 7 basic emotions and 15 compound emotions for 230 subjects. Geometric and appearance information (extracted from landmark points captured from frontal face images) are combined with a nearest-mean classifier to recognize compound facial expressions. Authors reported accuracy performance on the CFEE database of 73.61% when using geometric features only, 70.03% when using appearance features, and 76.91% when both features are combined in a single feature space.

Aleix and Du [37] defined a continuous model consistent with compound facial expression analysis. The continuous model explains how expressions of emotion can be seen at different intensities. In their work, multiple (compound) emotion categories can be recognized by linearly combining distinct continuous face spaces. Authors showed how the resulting model can be employed for the recognition of facial emotion expressions, and proposed new research directions from which the machine learning and computer vision communities could keep pushing the state-of-the-art on the field.

Fabian et al. [38] proposed an approach to quickly annotate Action Units (AUs) and their intensities, as well as their respective emotion categories for facial expression recognition. Thus, the EmotioNet dataset was released. In their work, geometric and shading features are extracted. Geometric features are defined as second-order statistics of facial landmarks (i.e., distances and angles between facial landmarks). Shading features, extracted using Gabor filters, model the shading changes due to local deformations of skin regions. This way, each AU is represented with shape and geometric features. Afterwards, Kernel Subclass Discriminant Analysis (KSDA) [42] is used to determine whether or not a specific AU is active. Authors [39] reported obtained AU annotation accuracy about 81%. Finally, according to different AU combinations, 23 emotion categories were defined.

The recently proposed EmotionNet Challenge [39] included two tracks. The first track was related to automatic detection of 11 AUs, whereas the second one addressed compound emotion. As the focus of our work is on the recognition of compound emotions, only the second track is reviewed. Briefly describing, the EmotionNet challenge employed the dataset defined in [38]. The training, validation and test sets were carefully defined to include 950K, 2K and 40K facial images, respectively. The validation and test sets were manually annotated. However, the training set were automatically annotated using the algorithm proposed in [38]. Finally, 16 basic and compound emotion categories have been defined.

Li and Deng [43] presented the RAF-DB (in the wild) database, containing 29,672 images. Each image was independently labeled by about 40 annotators based on the crowd-sourcing annotation. The dataset consisted of 7 basic emotions and 12 compound emotions. Authors also proposed a deep locality-preserving learning method for emotion recognition. Experiments showed that the average accuracy of compound emotion recognition was about 44.55%, which demonstrated that the compound emotion recognition (in the wild) was a very challenging task.

The main limitation of [32], [38], [39], [43] is that they provided very distinct compound facial emotion with limited categories (ranging from 16 to 23, as it can be seen in Table 1). In addition, the annotated labels provided in [39] were automatically obtained (in terms of recognized AU), which could undesirably add noise to the problem.

TABLE 1: Available public datasets on compound facial expression. Note that "Contr. env." means Controlled Environment.

Database	#images ($\times 10^3$)	#classes	#identities	Contr. env
CFEE [32]	5.06	22	230	Yes
EmotioNet [38]	1,000	23	-	No
EmotioNet Challenge [39]	992	16	-	No
RAF-DB [43]	29.672	19	-	No
iCV-MEFED	31.25	50	125	Yes

III. DOMINANT AND COMPLEMENTARY EMOTION RECOGNITION CHALLENGE

A. OVERALL DESCRIPTION

The iCV-MEFED dataset is designed to investigate compound emotion recognition. All emotion categories covered by the iCV-MEFED dataset are shown in Table 2. The motivation in creating such dataset, beyond to help pushing the research on the topic, is to explore how well emotion expression-based models can perform on this relatively novel and challenging task. The dataset includes 31250 frontal face images with different emotions captured from 125 subjects, whose gender distribution is relatively uniform. The subjects' age range from 18 to 37, as well as different ethnicity and appearance (e.g., hair styles, clothes, accessories, etc) are presented. Images were obtained in a controlled environment in order to focus on compound emotions and reduce problems introduced by, for example, background noise, strong head pose variations, illumination changes, etc, which could bias the results/analysis. The room where the images have been obtained was illuminated with uniform light, hence the variation of light changes can be ignored. Each subject acted 50 different emotions (Table 2) and for each emotion 5 samples have been taken. Note that face ID's are recorded within the dataset structure, so that one can analyze different performed emotions from a given individual. The images were taken and labeled under the supervision of psychologists, and the subjects were trained/instructed to express such wide range emotions.

B. ACQUISITION DETAILS

For each subject in the iCV-MEFED dataset, five sample images were captured (for each compound emotion) by a Canon 60D high resolution camera. In total, 50 distinct compound emotions have been considered. All images were

TABLE 2: 49 Dominant - complementary emotion combinations (*the* 50^{th} *emotion is neutral*).

	Angry	Contempt	Disgust	Fear	Happy	Sadness	Surprise
Angry	**angry**	contemptly angry	disgustingly angry	fearfully angry	happily angry	sadly angry	surprisingly angry
Contempt	angrily contempt	**contempt**	disgustingly contempt	fearfully contempt	happily contempt	sadly contempt	surprisingly contempt
Disgust	angrily disgusted	contemptly disgusted	**disgust**	fearfully disgusted	happily disgusted	sadly disgusted	surprisingly disgusted
Fear	angrily fearful	contemptly fearful	disgustingly fearful	**fearful**	happily fearful	sadly fearful	surprisingly fearful
Happy	angrily happy	contemptly happy	disgustingly happy	fearfully happy	**happy**	sadly happy	surprisingly happy
Sadness	angrily sad	contemptly sad	disgustingly sad	fearfully sad	happily sad	**sad**	surprisingly sad
Surprise	angrily surprised	contemptly surprised	disgustingly surprised	fearfully surprised	happily surprised	sadly surprised	**surprised**

captured under the same environment. The lightening condition was uniform, with a fixed background. Image resolution was set to 5184×3456. The motivation of using such controlled environment is to reduce pre-processing steps (such as face alignment, denoising, etc), which could introduce noise/errors to the problem, and focus on the compound emotions recognition task. Moreover, high resolution images can provide detailed information for compound emotion recognition approaches. Finally, the dataset is divided into training, validation and test sets, with $17,500$, $7,500$ and $6,250$ images, respectively. An illustration of the capturing setup can be seen in Fig. 1. Few samples of the iCV-MEFED dataset are shown in Fig. 2.

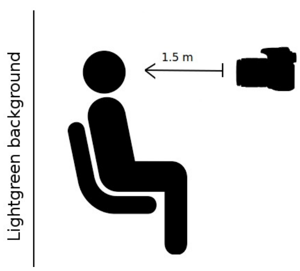


FIGURE 1: Illustrative sketch of the iCV-MEFED recordings setup.

During recording, the subject were also instructed to avoid excessive head movement and occlude face regions (e.g., with hair and/or upper body movements). When recording a specific emotion, a similar emotion example is simultaneously displayed as stimulus. If a person has any trouble in expressing a specific emotion, the corresponding common traits of this emotion are given so that he/she can train and improve his/her action. For example, tightening the lips is usually related to the contempt emotion.

Finished the capturing process, all sample images are given to psychologists for assessment of the truthfulness of the expressions. During this process, subject samples that do not managed to sufficiently convey their emotions are discarded. Even though participants are ordinary people (i.e., they are not professional actors), the captured images have

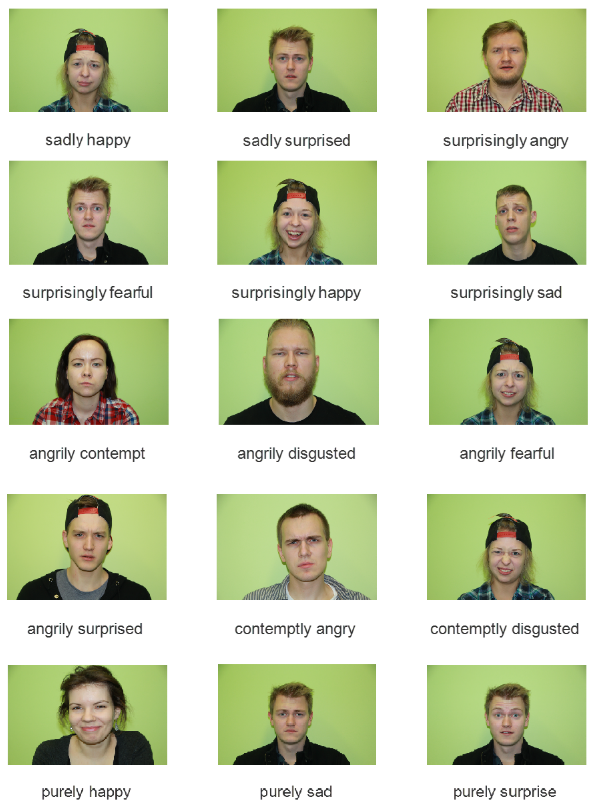


FIGURE 2: Few samples of the proposed dataset.

natural looking and can benefit and help to push the research in the field of compound emotion recognition and analysis.

In general, it is possible that some captured emotions may appear weird/rare. Nevertheless, we believe they can also help researchers to analyze any existing relationship (such as the frequency) in comparison with other generated emotions, and any other relationship that may exist in real life.

C. EVALUATION METRIC

The evaluation metric used in the Challenge [23] was defined as the percentage of misclassified instances. Note that during the challenge, the final rank is given according to the misclassification rate on the test set. However, since two emotions (both complementary and dominant) needed to be correctly recognized in order to be considered a precise prediction, in general, participants did not achieve high scores. For instance, sometimes they were able to recognize the dominant emotion but failed to recognize the complementary one (or vice-versa). Nevertheless, even though other evaluation metric can be considered, we believe this was the most direct way to rank participants.

IV. WINNER METHODS FROM PARTICIPANTS

In this section, we introduce the top three winners' methods submitted to the challenge. All three methods adopted Convolutional Neural Network (CNN) approaches to extract features. Their main and general ideas are summarized as follows: 1) The first ranked method exploited landmark displacement as geometric representation of emotions, thus leading to better results compared with texture-only information; 2) The second ranked method adopted unsupervised learning combined with multiple SVM classifiers; 3) The third ranked method combined CNN inception-v3 with a discriminative loss function (center loss). Next, further details of the top three winner's methods are given.

A. MULTI-MODALITY NETWORK WITH VISUAL AND GEOMETRICAL INFORMATION (1ST PLACE)

The method proposed in [44][2] combined texture and geometrical information in an end-to-end trained CNN. Texture features are extracted using the AlexNet [45], and geometrical features are represented by facial landmarks displacements. Such fusion strategy achieved better result when compared to texture-only or geometric-only based approaches.

1) Proposed Solution

Geometrical Representation. Winners' method used Dlib [46][3] library for facial landmark extraction, and face alignment following [47]. Then, facial landmarks are refined after face alignment. In their approach, each face i (i.e., face ID) is first represented by an average $lm^{(i)}$ landmark face:

$$lm^{(i)} = \frac{1}{N}\sum_{j=1}^{N} l_j^{(i)}, \tag{1}$$

where N is each face ID's number of samples, which is about 250 in iCV-MEFED dataset, and l represents the flattened vector of landmark. Finally, the geometrical representation is extracted as the landmarks displacement:

$$lr^{(i)} = l^{(i)} - lm^{(i)}. \tag{2}$$

where lr is landmark residual (or displacement).

Network Structure. The network structure of this method is shown in Fig. 3. Texture features are represented by the vector $p_1 \in \mathbb{R}^{256}$ and geometrical feature by $p_2 \in \mathbb{R}^{136}$. Both p_1 and p_2 are concatenated into $p \in \mathbb{R}^{392}$, as illustrated in Fig. 3. The concatenated feature p is fed into a fully connected layer before hinge loss optimization.

In a nutshell, p_1 can span a vector space V_1 and its decision boundary provided by classifier can correctly divide some samples, but the discriminative ability is limited. Once the landmarks displacement vector p_2 is embedded, V_1 can be mapped from a lower dimension into a higher dimension space V. Then V becomes more divisible because of the effectiveness of p_2. This map from low dimension to high dimension is similar to kernel function in SVM.

2) Implementation Details

In the training phase, the input image size is set to 224×224, and the size of landmark displacement vector is 136×1. The method uses stochastic gradient descent (SGD) with a mini-batch size of 32 and the max iteration is 1×10^5. The learning rate starts from 5×10^{-4}, and it is divided by 5 every 20,000 iterations. A weight decay of 5×10^{-4} and a momentum of 0.9 are adopted. At test stage, p_1 and p_2 are computed, then concatenated as given as the input to the classifier.

B. UNSUPERVISED LEARNING OF CONVOLUTIONAL NEURAL NETWORKS (2ND PLACE)

Similarly to the winner approach, the second top-ranked method also extracts and aligns all faces using the Dlib [46] library. Then, face images are resized to $96 \times 96 \times 3$. Next, an unsupervised learning model described in [48] is applied. It is a CNN model with filters trained layer-wise using k-means clustering. While being a simple model, it turned out to be very effective to address the problem proposed in this challenge. Obtained results also indicate that wider shallow networks can achieve better accuracy performances than deeper ones. Fig. 4 illustrates the pipeline of this method.

1) Proposed Solution

The CNN structure consists of a batch-norm layer, convolutional layer with 1024 filters ($15 \times 15 \times 3$), max-pooling (12×12), a Rectified Linear Units (ReLU) rectifier and a rootsift normalization. Principal component analysis (PCA) is applied to extracted features. The number of principal components was set to 500. Participants then take 10 subsets from these 500 dimensional features. In the first subset, features are projected on the first 50 principal components. In the second subset, features are projected onto the first 100 principal components, and so on. Thus, instead of training just one classifier on 500 dimensional feature vectors, 10 classifiers for different subsets of features are trained. A linear SVM is chosen as a classifier and all 50 emotions are treated as independent.

Note that in [48], one of the core steps during filter learning is recursive autoconvolution applied to images patches. However, participants did not find it useful on the task of compound emotion recognition and choose to learn filters without recursive autoconvolution.

[2] https://github.com/cleardusk/EmotionChallenge
[3] http://dlib.net/

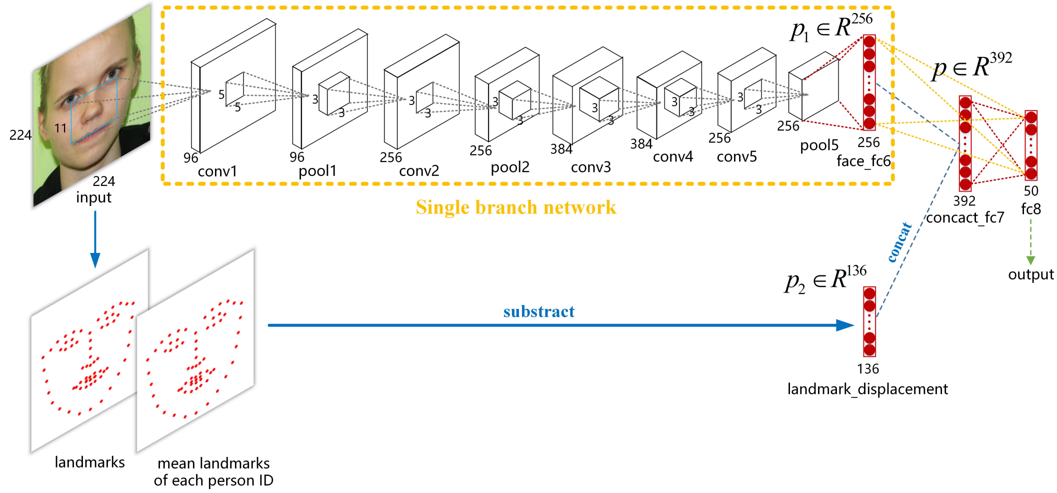


FIGURE 3: Overview of the winner method (1st place) of the competition. The upper branch is a single CNN network. The whole architecture constructs the multi-modality (fusion) network.

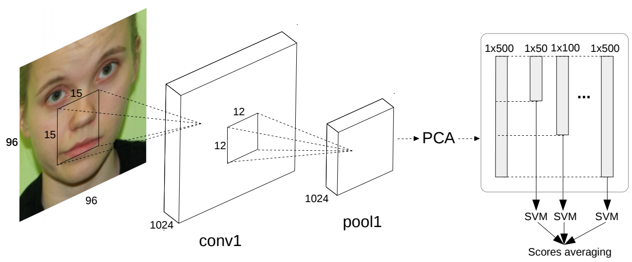


FIGURE 4: Overview of the 2nd ranked method. Filters of the convolutional layer are trained using *k*-means. Then, different SVMs are trained and combined to improve the results.

2) Implementation Details

Filters are trained using *k*-means with ZCA-whitening following [48]. Filter size, pooling type and size, as well as the SVM regularization constant are selected by 5 fold cross-validation. At feature extraction stage, a mini-batch size of 25 and augmentation of horizontal flipping are adopted. As different SVMs are employed, based on the distinct sets of extracted features, final prediction is obtained by averaging individual SVM scores.

C. INCEPTION-V3 STRUCTURE WITH AUXILIARY CENTER LOSS (3RD PLACE)

This method directly predicts emotion categories using a Inception-V3 network structure. In order to increase the discrimination of features for similar emotion classes, they also adopted the center loss [49] as an auxiliary optimization function. Proposed pipeline is shown in Fig. 5.

1) Proposed Solution

Base Pipeline. First, Multi-task CNN (MTCNN) [50] is adopted to parse face bounding boxes and landmarks. Then, face images are aligned by affine transformation and resized to $224 \times 224 \times 3$. Features are then extracted using the Inception-V3 CNN. Finally, cross-entropy loss is applied to for optimization.

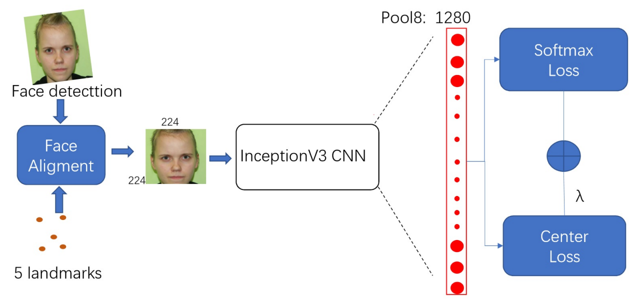


FIGURE 5: Overview of the 3rd ranked method.

Discriminative Training. The cross-entropy loss works fine when the predicted labels are mutually exclusive. However, the labels of the iCV-MEFED dataset are interrelated (e.g., happily-angry and surprisingly-angry). To address this problem, participants adopted the center loss function as an auxiliary loss to reduce the effect of similar label. The center loss can simultaneously learns each class center of deep features and penalizes the distances between the deep features and their corresponding class center. This loss enhances the ability of model to distinguish similar samples and improves the overall performance.

2) Implementation Details

The network is optimized by SGD and maximum number of iteration is set to 1×10^5. For the first 3×10^4 iterations, the learning rate is fixed to be 10^{-3}. For the rest 7×10^4 iterations, the learning rate stays at 10^{-4}. Weight decay is 4×10^{-4}, momentum is 0.9 and all layers are initialized following [51].

V. EXPERIMENT ANALYSIS

In this section, we perform a thorough comparison of the three top-ranked methods on the iCV-MEFED dataset from the organized challenge [23]. A detailed analysis including misclassification rates, execution time, accuracy in relation to each category and confusion matrix are provided and discussed.

As previously mentioned, there are mainly four public datasets for compound facial emotion recognition: CFEE [32], EmotionNet [39], RAF-DB [43] and the proposed iCV-MEFED dataset. As the size of CFEE dataset is small for CNN based methods, we opted to not use it in the experiments. Although RAF-DB dataset contains about 30k face images with 19 kinds of emotion categories, the distribution is not well balanced. There are only 8 images of "fearfully disgusted" and 86 images of "sadly surprised", but the emotion "happy" includes 5,957 images and "sad" includes 2,460 images. Thus RAF-DB is not considered because of its unbalanced emotion distribution.

According to the transformation rule from AUs to emotion category presented in [39], with respect to EmotioNet dataset, we could obtain 105,203 images with emotion labels. The distribution of each emotion category is shown in Table 3. As it can be seen, the distribution of emotion categories is extremely unbalanced. Most images have been assigned to *happy* category, and the number of images of other categories are very small and sometimes close to zero. The results might be caused by inaccurate AUs [39] provided with the dataset. Therefore, we also consider EmotionNet is not a proper dataset to be used in our experiments.

TABLE 3: Label distribution (Emotion *vs* number of images) of EmotioNet dataset [39] after transformation from AU to emotion category.

Happy	Angrily Surprised	Surprised	Sad	Owed	Others
104511	388	300	3	1	0

A. OVERALL RECOGNITION ACCURACY

The emotion recognition for the three top-ranked methods described not so far is treated as a classification task of the 50 classes shown in Table 2. Complementary and dominant labels are indexed according to Table 4 to facilitate the evaluation process.

Obtained results of the top-3 ranked methods on the iCV-MEFED dataset are shown in Table 5, using the evaluation metric described in Sec. III-C. It can be observed that fine-grained emotion recognition is very challenging, and the accuracy still has a big room for improvement. The winner method (1st place), which are based on multi-modality network with texture and geometrical information, outperformed other two methods by a large margin.

Fig. 6 shows the top-k obtained accuracies of three methods on the iCV-MEFED dataset. As it can be observed, the performance gap between the winner (1st method) and other two methods is greater with the growth of k, demonstrating its effectiveness in recognizing compound emotions.

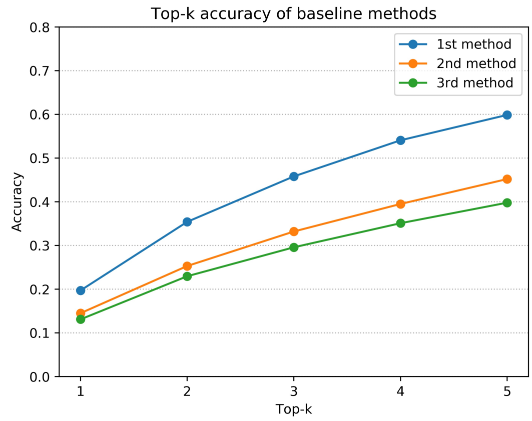


FIGURE 6: Top-5 accuracy of three competition methods on testing set of the iCV-MEFED dataset.

B. ACCURACY OF EACH EMOTION CATEGORY

In this section we analyze the accuracy of each compound emotion category. Fig. 7 shows the performance of there methods on different emotion categories. For the top ranked approach, the following emotions demonstrated to be better recognized: 0 (neutral), 1 (angry), 9 (contempt), 33 (happy), 35 (happily surprised), 46 (surprisingly fearful), 49 (surprised). In relation to the second ranked approach, the classification accuracy of the following emotions achieved higher accuracy when compared to other methods: 7 (angrily surprised), 15 (disgustingly angry), 29 (happily angry), 41 (sad) is better. For the third one, 2 (angrily contempt), 5 (angrily happy), 47 (surprisingly happy) were better recognized. From Fig. 7 it can also be observed that some emotion categories are easy to be recognized (i.e., with high accuracy associated values) whereas others are very hard to be recognized. In general, the classification results of three methods demonstrated to complement each other. Future work combining the best of the three methods would be an interesting way to improve the recognition rates and advance the research on the field of compound emotions.

C. CONFUSION MATRIX ANALYSIS

In order to analyze the statistics of emotion misclassification, we generated the confusion matrix of different emotion recognition methods among different categories (Fig. 9, Fig. 10 and Fig. 11). We first analyzed each confusion matrix individually and found that all methods easily misrecognize dominant and complementary emotions. It means that these algorithms may correctly find that both emotions are present (e.g., angry and sad), but they fail to recognize which one is the dominant (e.g., sadly-angry instead of angrily-sad) with high probability. It demonstrates that dominant and complementary emotion recognition is a very

TABLE 4: Label conversion table.

Label	Emotion	Label	Emotion	Label	Emotion	Label	Emotion
0	neutral	14	contemptly surprised	28	fearfully surprised	42	sadly surprised
1	angry	15	disgustingly angry	29	happily angry	43	surprisingly angry
2	angrily contempt	16	disgustingly contempt	30	happily contempt	44	surprisingly contempt
3	angrily disgusted	17	disgust	31	happily disgust	45	surprisingly disgust
4	angrily fearful	18	disgustingly fearful	32	happily fearful	46	surprisingly fearful
5	angrily happy	19	disgustingly happy	33	happy	47	surprisingly happy
6	angrily sad	20	disgustingly sad	34	happily sad	48	surprisingly sad
7	angrily surprised	21	disgustingly surprised	35	happily surprised	49	surprised
8	contemptly angry	22	fearfully angry	36	sadly angry		
9	contempt	23	fearfully contempt	37	sadly contempt		
10	contemptly disgusted	24	fearfully disgust	38	sadly disgust		
11	contemptly fearful	25	fearful	39	sadly fearful		
12	contemptly happy	26	fearfully happy	40	sadly happy		
13	contemptly sad	27	fearfully sad	41	sad		

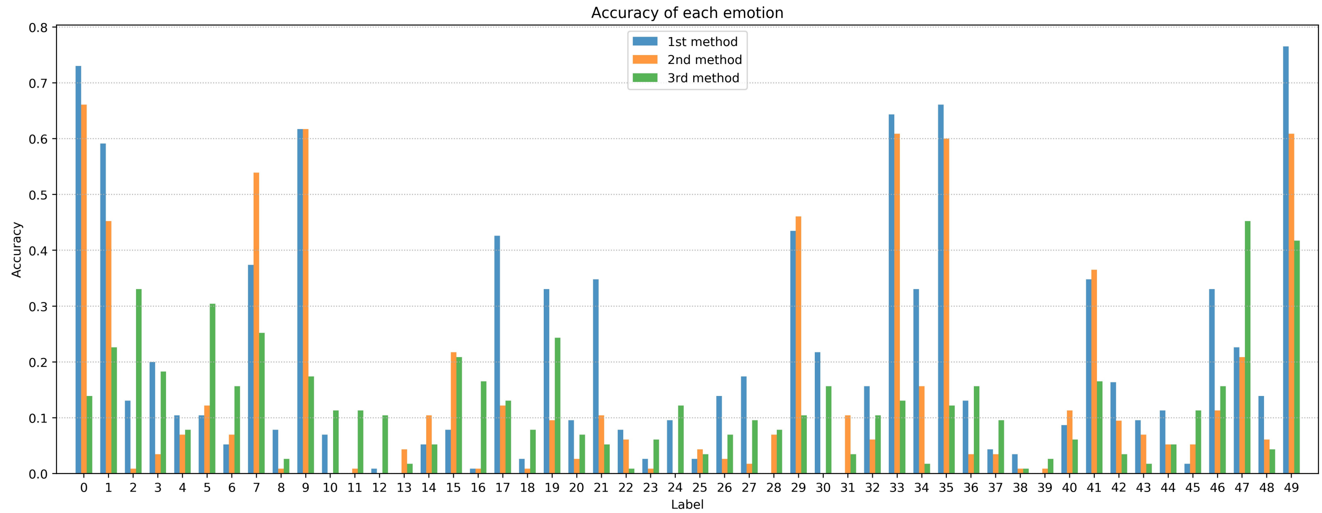


FIGURE 7: Accuracy performance obtained by each method on the test set of the iCV-MEFED dataset, in relation to each emotion category.

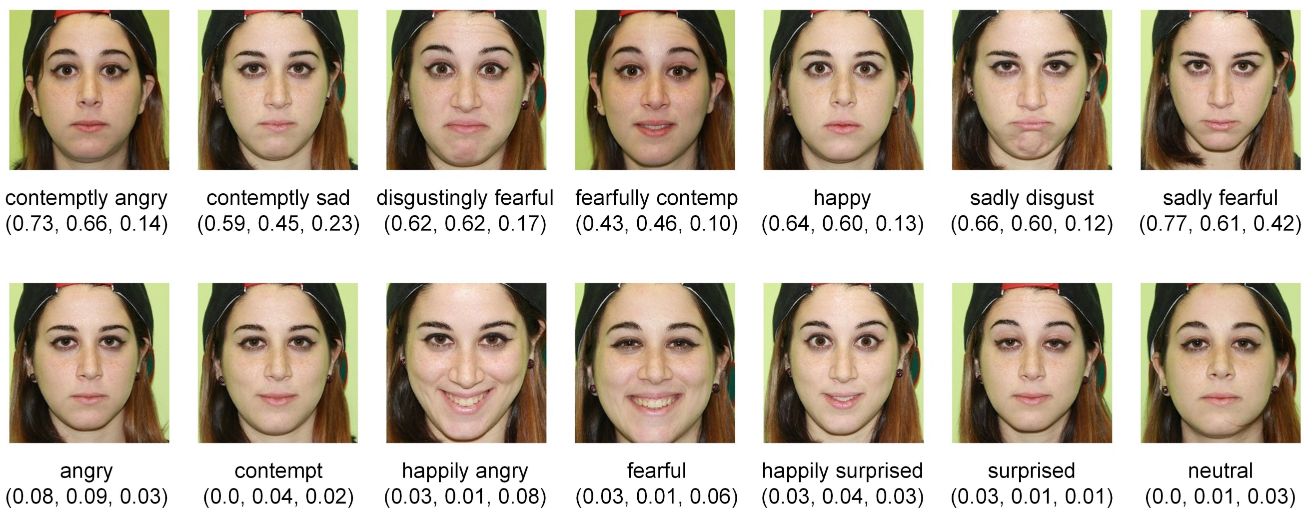


FIGURE 8: It shows some easy and difficult samples to recognize. Easy samples are shown in the first row and difficult samples are listed in the second row. The accuracy of three methods is given in brackets in order.

TABLE 5: The misclassification rates of three competition methods on the validation and test sets of the iCV-MEFED dataset.

Ranking	Misclassification (validation set)	Misclassification (test set)
1st	**0.793**	**0.802**
2nd	0.840	0.853
3rd	0.875	0.877

challenging task.

For instance, if we check in detail Fig. 7, we will see that the winner method (1st place) performed well on some specific emotion categories (e.g. neutral, angry, disgustingly happy, disgustingly sad, disgust, surprisingly fearful, surprised). More specifically, its average accuracy of seven basic emotions was 51.84%, while the average accuracy of compound emotion was 13.7%. This demonstrates that the basic emotions are more easier to recognize than dominant and complementary emotion (i.e., when combined). In addition, from the confusion matrix shown in Fig. 9, it can be observed that the winner method also obtained low accuracy performance in recognizing dominant and complementary emotions (listed in Table 6). From the Table 6, it can be seen that some compound emotions are easy to confuse with opposite compound emotions, such as surprisingly-happy and happily-surprised, as well as surprisingly-angry and angrily-surprised. This may happen due to the complexity of the task.

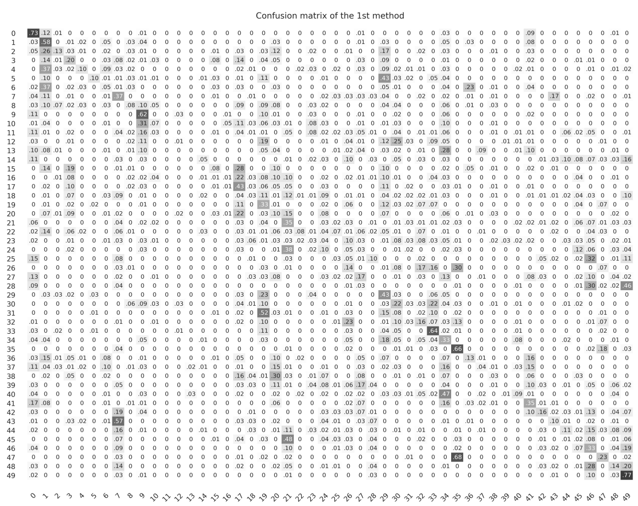


FIGURE 9: Confusion matrix of the first method. Each cell shows corresponding prediction's probability value, which is in range $[0, 1]$. The numbers of two axises are transformed labels following Table 4 (better view on electronic version).

Fig. 10 and 11 show the confusion matrices for the second and third ranked methods, respectively. Their corresponding top-10 hardest misclassified emotions are listed in Table 7 and Table 8, respectively. In general, the hardest misclassified emotions of the three proposed methods are similar. For instance, three kinds of emotion pairs were strongly misclassified for all three competition methods, which are surprisingly-angry *vs* angrily-surprised, surprisingly-happy *vs* happily-surprised, and angrily-happy *vs* happily-angry. Few samples of these compound emotions are shown in Fig. 12.

TABLE 6: The top-10 hardest misclassified emotion categories for the winner method (1st place).

Ground-truth	Misclassified category	Rate of misclassification
surprisingly happy	happily surprised	0.68
surprisingly angry	angrily surprised	0.57
happily disgust	disgustingly happy	0.52
surprisingly disgust	disgustingly surprised	0.48
sadly happy	happily sad	0.47
fearfully surprised	surprised	0.46
angrily happy	happily angry	0.43
fearfully disgust	disgustingly surprised	0.38
angrily fearful	angry	0.37
angrily sad	angry	0.37

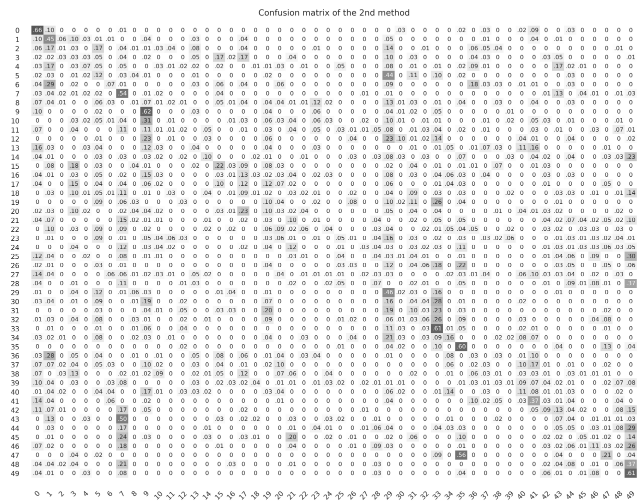


FIGURE 10: Confusion matrix of the second ranked method. Each cell shows corresponding prediction's probability value, which is in range $[0, 1]$. The numbers of two axises are transformed labels following Table 4 (better view on electronic version).

TABLE 7: The top-10 hardest misclassified emotion categories for the second ranked method.

Ground-truth	Misclassified category	Rate of misclassification
surprisingly happy	happily surprised	0.56
surprisingly angry	angrily surprised	0.5
angrily happy	happily angry	0.44
surprisingly sad	surprised	0.37
fearfully surprised	surprised	0.37
contemptly disgusted	contempt	0.31
fearful	surprised	0.3
angrily sad	angry	0.29
surprisingly contempt	surprised	0.29
happily contempt	happy	0.28

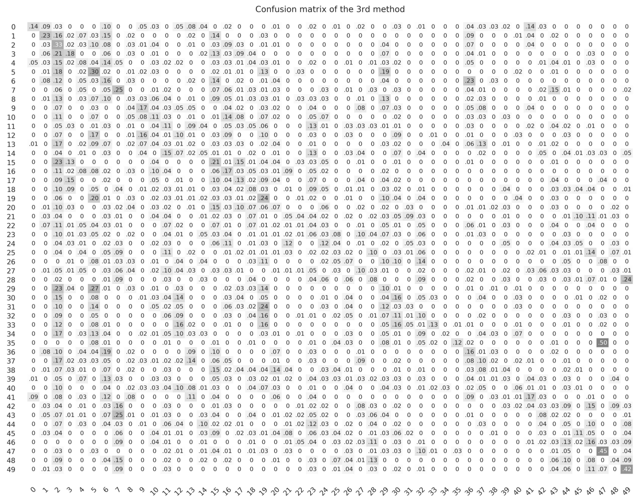


FIGURE 11: Confusion matrix of the third ranked method, and each cell shows corresponding prediction's probability value, which is in range $[0, 1]$. The numbers of two axises are transformed labels following Table 4 (better view on electronic version).

TABLE 8: The top-10 hardest misclassified emotion categories for the third ranked method.

Ground-truth	Misclassified category	Rate of misclassification
happily surprised	surprisingly happy	0.5
happily angry	angrily happy	0.27
surprisingly angry	angrily surprised	0.25
fearfully surprised	surprised	0.24
happily disgust	disgustingly happy	0.24
happily angry	angrily contempt	0.23
angrily sad	sadly angry	0.23
disgustingly angry	angrily contempt	0.23
angrily disgusted	angrily contempt	0.21
disgustingly happy	angrily happy	0.2

D. COMPUTATIONAL COST ANALYSIS

Table 9 shows computation time and the number of parameters among different methods. The proposed three methods were tested under the same environments (GPU: GTX TITAN X, CPU: Xeon E5-2660@2.20GHz).

It can be seen that the winner method achieved the fastest average inference time, requiring 1.57ms (using GPU) or 30 ms (using CPU). Winner approach also has relatively small number of parameters (compared with other approaches). Furthermore, the winner method adopted a modified version of AlexNet to extract facial features, while the third method employed the inception-V3 structure which is deeper and demonstrated to require more computational power. The second method used a shallow CNN to extract features, however, the 50 adopted classifiers increased computation time.

VI. CONCLUSION

In this work, we collected and released a new compound facial emotion dataset, named iCV-MEFED, which includes large number of labels, 50 categories to be specific, obtained with the support of psychologists. The recognition of compound emotions on the iCV-MEFED dataset demonstrated to be very challenging, leaving a large room for improvement. Top winners' methods from FG 2017 workshop have been analyzed and compared. As it could be observed, there are some compound emotions that are more difficult to be recognized. Reported methods treated all 50 classes of emotions independently, meaning that prior knowledge of dominant and complementary emotions were not considered. How to incorporate prior information of dominant and complementary categories into compound facial emotion recognition is one question we want to address in future work.

TABLE 9: Computation time per image, and the number of parameters of three competition methods. Note that M means megabytes.

Method	Input Size	Inference Time (GPU/CPU)	#params
1st	224×224	1.57ms/30ms	4.7M
2nd	96×96	42ms/570ms	34M
3rd	299×299	50ms/890ms	23M

REFERENCES

[1] M. Bianchi, G. Valenza, A. Greco, M. Nardelli, E. Battaglia, A. Bicchi, and E. P. Scilingo, "Towards a novel generation of haptic and robotic interfaces: integrating affective physiology in human-robot interaction," in 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), 2016. IEEE, 2016, pp. 125–131. 2

[2] C. M. Ranieri and R. A. Romero, "An emotion-based interaction strategy to improve human-robot interaction," in Robotics Symposium and IV Brazilian Robotics Symposium (LARS/SBR), 2016 XIII Latin American. IEEE, 2016, pp. 31–36. 2

[3] M. Bellantonio, M. A. Haque, P. Rodriguez, K. Nasrollahi, T. Telve, S. Escarela, J. Gonzalez, T. B. Moeslund, P. Rasti, and G. Anbarjafari, "Spatio-temporal pain recognition in cnn-based super-resolved facial images," in International Conference on Pattern Recognition (icpr). Springer, 2016. 2

[4] F. Noroozi, M. Marjanovic, A. Njegus, S. Escarela, and G. nbarjafari, "Fusion of classifier predictions for audio-visual emotion recognition," in International Conference on Pattern Recognition (icpr). Springer, 2016. 2

[5] N. B. Kar, K. S. Babu, A. K. Sangaiah, and S. Bakshi, "Face expression recognition system based on ripplet transform type ii and least square svm," Multimedia Tools and Applications, pp. 1–24, 2017. 2

[6] A. Bolotnikova, H. Demirel, and G. Anbarjafari, "Real-time ensemble based face recognition system for nao humanoids using local binary pattern," Analog Integrated Circuits and Signal Processing, pp. 1–9, 2017. 2

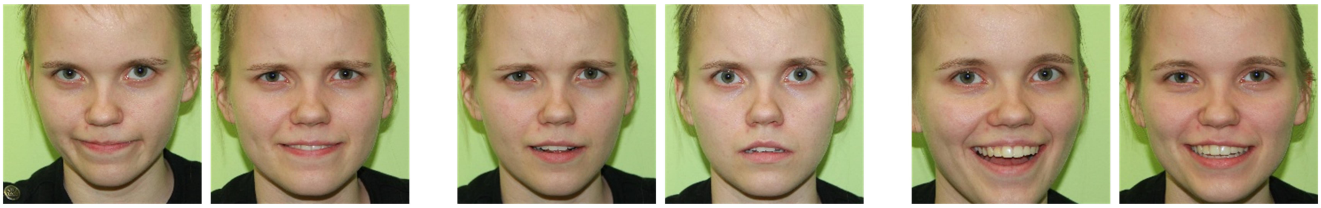
[7] R. Bowden, A. Zisserman, T. Kadir, and M. Brady, "Vision based interpretation of natural sign languages," in 3rd International Conference on Computer Vision Systems - ICVS 2003, April 2003. 2

[8] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," Artificial Intelligence Review, vol. 43, no. 1, pp. 1–54, 2015. 2

[9] I. Aydin, M. Karaköse, G. G. Hamsin, and E. Akin, "A vision based inspection system using gaussian mixture model based interactive segmentation," in Artificial Intelligence and Data Processing Symposium (IDAP), 2017 International. IEEE, 2017, pp. 1–4. 2

[10] R. E. Haamer, K. Kulkarni, N. Imanpour, M. A. Haque, E. Avots, M. Breisch, K. Nasrollahi, S. E. Guerrero, C. Ozcinar, X. Baro et al., "Changes in facial expression as biometric: A database and benchmarks of identification," in IEEE Conf. on Automatic Face and Gesture Recognition Workshops. IEEE, 2018. 2

[11] J. Gorbova, I. Lüsi, A. Litvin, and G. Anbarjafari, "Automated screening of job candidate based on multimodal video processing," in Proceedings



Left: surprisingly angry (43)
Right: angrily surprised(7)

Left: surprisingly happy(47)
Right: happily surprised(35)

Left: angrily happy (5)
Right: happily angry (29)

FIGURE 12: Illustration of few emotion pairs with high misclassification rates.

of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 29–35. 2

[12] R. Adolphs, “Cognitive neuroscience of human social behaviour,” Nature Reviews Neuroscience, vol. 4, no. 3, pp. 165–178, 2003. 2

[13] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, “Audio-visual emotion recognition in video clips,” IEEE Transactions on Affective Computing, 2017. 2

[14] F. Mormann, J. Niediek, O. Tudusciuc, C. M. Quesada, V. A. Coenen, C. E. Elger, and R. Adolphs, “Neurons in the human amygdala encode face identity, but not gaze direction,” Nature neuroscience, 2015. 2

[15] I. Lüsi, S. Escarela, and G. Anbarjafari, “Sase: Rgb-depth database for human head pose estimation,” in Computer Vision–ECCV 2016 Workshops. Springer, 2016, pp. 325–336. 2

[16] C. Nagaraju, D. Sharadamani, C. Maheswari, and D. Vishnu Vardhan, “Evaluation of lbp-based facial emotions recognition techniques to make consistent decisions,” International Journal of Pattern Recognition and Artificial Intelligence, vol. 29, no. 06, p. 1556008, 2015. 2

[17] D. Feil-Seifer and M. J. Mataric, “Defining socially assistive robotics,” in 9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005. IEEE, 2005, pp. 465–468. 2

[18] A. Guler, N. Kardaris, S. Chandra, V. Pitsikalis, C. Werner, K. Hauer, C. Tzafestas, P. Maragos, and I. Kokkinos, “Human joint angle estimation and gesture recognition for assistive robotic vision,” in European Conference on Computer Vision. Springer, 2016, pp. 415–431. 2

[19] E. S. John, S. J. Rigo, and J. Barbosa, “Assistive robotics: Adaptive multimodal interaction improving people with communication disorders,” IFAC-PapersOnLine, vol. 49, no. 30, pp. 175–180, 2016. 2

[20] A. Bolotnikova, K. Chappellet, A. Paolillo, A. Escande, G. Anbarjafari, A. Suarez-Roos, P. Rabaté, and A. Kheddar, “A circuit-breaker use-case operated by a humanoid in aircraft manufacturing,” in 13th IEEE Conference on Automation Science and Engineering, 2017. 2

[21] Y. Dai, Y. Shibata, T. Ishii, K. Hashimoto, K. Katamachi, K. Noguchi, N. Kakizaki, and D. Cai, “An associate memory model of facial expressions and its application in facial expression recognition of patients on bed,” in IEEE International Conference on Multimedia and Expo. IEEE, 2001, pp. 591–594. 2

[22] G. Anbarjafari and A. Aabloo, “Expression recognition by using facial and vocal expressions,” Proceedings of the Third Workshop on Vision and Language, pp. 103–105, 2014. 2

[23] I. Lüsi, J. C. J. Junior, J. Gorbova, X. Baró, S. Escalera, H. Demirel, J. Allik, C. Ozcinar, and G. Anbarjafari, “Joint challenge on dominant and complementary emotion recognition using micro emotion features and head-pose estimation: Databases,” in 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), 2017. IEEE, 2017, pp. 809–813. 2, 4, 7

[24] K. Nasrollahi, S. Escalera, P. Rasti, G. Anbarjafari, X. Baro, H. J. Escalante, and T. B. Moeslund, “Deep learning based super-resolution for improved action recognition,” in International Conference on Image Processing Theory, Tools and Applications (IPTA), 2015. IEEE, 2015, pp. 67–72. 2

[25] W. Gu, C. Xiang, Y. Venkatesh, D. Huang, and H. Lin, “Facial expression recognition using radial encoding of local gabor features and classifier synthesis,” Pattern Recognition, vol. 45, no. 1, pp. 80–91, 2012. 2

[26] Y. Rahulamathavan, R. C.-W. Phan, J. Chambers, D. J. Parish et al., “Facial expression recognition in the encrypted domain based on local fisher discriminant analysis,” IEEE Transactions on Affective Computing, vol. 4, no. 1, pp. 83–92, 2013. 2

[27] P. Ekman, “Are there basic emotions?” Psychological Review, vol. 99, no. 3, pp. 550–553, 1992. 2

[28] A. Bellocchi, “Methods for sociological inquiry on emotion in educational settings,” Emotion Review, vol. 7, no. 2, pp. 151–156, 2015. 2

[29] R. Berrios, P. Totterdell, and S. Kellett, “Eliciting mixed emotions: a meta-analysis comparing models, types, and measures,” Frontiers in psychology, vol. 6, 2015. 2

[30] S. Du and A. M. Martinez, “Wait, are you sad or angry? large exposure time differences required for the categorization of facial expressions of emotion,” Journal of vision, vol. 13, no. 4, pp. 13–13, 2013. 2

[31] C. Loob, P. Rasti, I. Lüsi, J. C. J. Junior, X. Baró, S. Escalera, T. Sapinski, D. Kaminska, and G. Anbarjafari, “Dominant and complementary multi-emotional facial expression recognition using c-support vector classification,” in 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), 2017. IEEE, 2017, pp. 833–838. 2

[32] S. Du, Y. Tao, and A. M. Martinez, “Compound facial expressions of emotion,” Proceedings of the National Academy of Sciences, vol. 111, no. 15, pp. E1454–E1462, 2014. 2, 3, 7

[33] M. G. Calvo, A. Fernández-Martín, and L. Nummenmaa, “Facial expression recognition in peripheral versus central vision: role of the eyes and the mouth,” Psychological research, vol. 78, no. 2, pp. 180–195, 2014. 2

[34] A. J. Calder, A. W. Young, J. Keane, and M. Dean, “Configural information in facial expression perception.” Journal of Experimental Psychology: Human perception and performance, vol. 26, no. 2, p. 527, 2000. 2

[35] C. Blais, C. Roy, D. Fiset, M. Arguin, and F. Gosselin, “The eyes are not the window to basic emotions,” Neuropsychologia, vol. 50, no. 12, pp. 2830–2838, 2012. 2

[36] J. N. Bassili, “Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face.” Journal of personality and social psychology, vol. 37, no. 11, p. 2049, 1979. 2

[37] A. Martinez and S. Du, “A model of the perception of facial expressions of emotion by humans: Research overview and perspectives,” Journal of Machine Learning Research, vol. 13, no. May, pp. 1589–1608, 2012. 2, 3

[38] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, “Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5562–5570. 2, 3

[39] C. F. Benitez-Quiroz, R. Srinivasan, Q. Feng, Y. Wang, and A. M. Martinez, “Emotionet challenge: Recognition of facial expressions of emotion in the wild,” arXiv preprint arXiv:1703.01210, 2017. 2, 3, 7

[40] C. A. Corneanu, M. Oliu, J. F. Cohn, and S. Escalera, “Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications,” 2016. 2

[41] P. Ekman and W. Friesen, “Facial action coding system: a technique for the measurement of facial movement,” Palo Alto: Consulting Psychologists, 1978. 2

[42] D. You, O. C. Hamsici, and A. M. Martinez, “Kernel optimization in discriminant analysis,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 3, pp. 631–638, 2011. 3

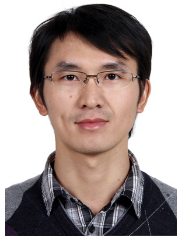
[43] S. Li and W. Deng, “Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. IEEE, 2017. 3, 7

[44] J. Guo, S. Zhou, J. Wu, J. Wan, X. Zhu, Z. Lei, and S. Z. Li, “Multi-modality network with visual and geometrical information for micro emotion recognition,” in International Conference on Automatic Face & Gesture Recognition (FG), 2017, pp. 814–819. 5

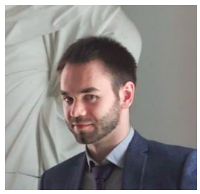
[45] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, 2012, pp. 1097–1105. 5
[46] V. Kazemi and J. Sullivan, “One millisecond face alignment with an ensemble of regression trees,” in IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2014, pp. 1867–1874. 5
[47] Z. Tan, S. Zhou, J. Wan, Z. Lei, and S. Z. Li, “Age estimation based on a single network with soft softmax of aging modeling,” in Asian Conference on Computer Vision. Springer, 2016, pp. 203–216. 5
[48] B. Knyazev, E. Barth, and T. Martinetz, “Recursive autoconvolution for unsupervised learning of convolutional neural networks,” in International Joint Conference on Neural Networks (IJCNN), 2017. IEEE, 2017, pp. 2486–2493. 5, 6
[49] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, “A discriminative feature learning approach for deep face recognition,” in European Conference on Computer Vision. Springer, 2016, pp. 499–515. 6
[50] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499–1503, Oct 2016. 6
[51] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in The IEEE International Conference on Computer Vision (ICCV), December 2015. 6

JIANZHU GUO received the B.E. degree from the school of Transportation, Southeast University (SEU), Nanjing, China in 2016. He is currently studying for a PHD degree at the Institute of Automation, Chinese Academy of Science (CASIA). His main research interests include deep learning, face attribute analysis, face recognition and 3D morphable model.

ZHEN LEI received the B.S. degree in automation from the University of Science and Technology of China, in 2005, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, in 2010, where he is currently an Associate Professor. He has published over 100 papers in international journals and conferences. His research interests are in computer vision, pattern recognition, image processing, and face recognition in particular. He served as an Area Chair of the International Joint Conference on Biometrics in 2014, the IAPR/IEEE International Conference on Biometric in 2015, 2016, 2018, and the IEEE International Conference on Automatic Face and Gesture Recognition in 2015.

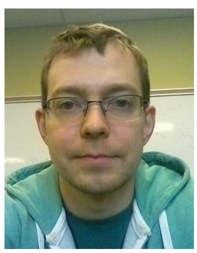


JUN WAN received his B.S. degree from the China University of Geosciences, Beijing, China, in 2008, and the PhD degree from the Institute of Information Science, Beijing Jiaotong University, Beijing, China, in 2015. Since January 2015, he has been an assistant professor at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Science (CASIA). He received the 2012 ChaLearn One-Shot-Learning Gesture Challenge Award, sponsored by Microsoft, ICPR 2012. He also received the 2013, 2014 Best Paper Award from the Institute of Information Science, Beijing Jiaotong University. His main research interests include computer vision, machine learning, especially for gesture and action recognition, facial attribution analysis (i.e. age estimation, facial expression, gender and race classification). He has published papers in top journals, such as JMLR, TPAMI, TIP, and TCYB. He has served as the reviewer on several top journals and conferences, such as JMLR, TPAMI, TIP, TMM, TSMC, PR, ICPR2016, CVPR2017, ICCV2017, FG2017.



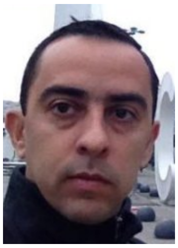
EGILS AVOTS received his B.S. and M.S. degree in Electronics in 2013 and 2015, from the Ventspils University College and M.S. degree in Robotics and Computer Engineering in 2017 from University of Tartu. At present moment he is a PhD student in iCV Research Lab, at University of Tartu. His research works involves application of machine learning in computer vision. He has been involved in project for development of novel retexturing technique for virtual fitting room solution. He has been a reviewer on several top journals and conferences including FG2016, ICCV2017, SIVP, PRL, IET CV and JIVP.

NOUSHIN HAJAROLASVADI is a PhD student at Department of Electrical and Electronic Engineering of Eastern Mediterranean University. Her research works involves human behaviour analysis with focus on facial expression analysis and deceptive emotion recognition. She has been involved in several scientific short term missions and have served as a reviewer at various conferences and journals.



BORIS KNYAZEV received a Diploma with Honours in Information Technology at Bauman Moscow State Technical University in 2011. From 2011 to 2016 he worked as a research engineer in the computer vision area with application to biometrics. He was a visiting researcher at the University of Luebeck in 2015-2016 under the DAAD Michail-Lomonosov scholarship. He worked as a machine learning engineer at VisionLabs from 2014 to 2015 and NTechLab from 2016 to 2017. Currently, he is a PhD student at the University of Guelph within the Machine Learning Research Group

ARTEM KUKHARENKO received his M.S. degree in Computer Science at the Moscow State University in 2012. Now he is the founder and head of research in NTechLab.



JULIO C. SILVEIRA JACQUES JUNIOR received his M.S. degree in Applied Computing in 2006, from Universidade do Vale do Rio dos Sinos (UNISINOS) and his Ph.D. degree in Computer Science in 2012, from Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), Brazil. His Ph.D. Thesis was awarded by the IBM Ph.D. Fellowship Awards Program in 2011. Currently, he is a postdoctoral researcher at the Computer Science, Multimedia and Telecommunications department at Universitat Oberta de Catalunya (UOC), within the Scene Understanding and Artificial Intelligence (SUnAI) group. He is a research collaborator within the Human Pose Recovery and Behavior Analysis (HuPBA) group, at Universitat de Barcelona and within the Computer Vision Center (CVC), at Universitat Autónoma de Barcelona (UAB). His research interests include, between others, image processing and computer vision-based applications, with special interest in human behavior analysis from multi-modal data.

XAVIER BARÓ received his B.S. degree in Computer Science at the UAB in 2003. In 2005 he obtained his M.S. degree in Computer Science at UAB, and in 2009 the PhD degree in Computer Engineering. At the present he is associate professor and researcher at the Computer Science, Multimedia and Telecommunications department at Universitat Oberta de Catalunya (UOC).

HASAN DEMIREL received his PhD and MSc degrees from Imperial College London in 1998 and King's College London in 1993 respectively after graduating from the Electrical and Electronic Engineering Department of Eastern Mediterranean University in 1992. He joined the Electrical and Electronic Engineering Department, Eastern Mediterranean University, as an Assistant Professor in 2000. He was appointed as Associate Professor in 2009 and Professor in 2014. He has served as the Vice chairman of the department from 2003 to 2005 and from 2007 to 2014. He has been elected Chairman of the department since May 2014. His main research is in the field of Image Processing and he is currently involved in research projects related to biomedical image processing, image/video resolution enhancement, face recognition/tracking, facial expression analysis, low-bit rate video coding. He had been the deputy director of the Advanced Technologies Research and Development Institute. Furthermore, he had served as the member of the Executive Council of the EMU Technopark.

SERGIO ESCALERA is an associate professor at the Department of Mathematics and Informatics, Universitat de Barcelona. He is an adjunct professor at Universitat Oberta de Catalunya, Aalborg University, and Dalhousie University. He obtained the PhD degree on Multi-class visual categorization systems at the Computer Vision Center, UAB. He obtained the 2008 best Thesis award on Computer Science at Universitat Autónoma de Barcelona. He leads the Human Pose Recovery and Behavior Analysis Group at UB and CVC. He is also a member of the Computer Vision Center at Campus UAB. He is vice-president of ChaLearn Challenges in Machine Learning, leading ChaLearn Looking at People events. He is Chair of IAPR TC-12: Multimedia and visual information systems. His research interests include affective computing, human behavior analysis, deep learning, visual object recognition, and HCI systems, including human human analysis from multi-modal data.

JÜRI ALLIK was Candidate of Science (PhD), University of Moscow (1976) and obtained PhD in psychology from the University of Tampere, Finland (1991). He has been Chairman of Estonian Science Foundation (2003-2009), Professor of Psychophysics (1992-2002) and Professor of Experimental Psychology (2002-) at the University of Tartu. He was also Dean of Faculty of Social Sciences (1996-2001), President (1988-1994) and Vice-President (1994-2001) of the Estonian Psychological Association. He served as a Foreign Member of the Finnish Academy of Science and Letters (1997). He has has received many awards including Estonian National Science Award in Social Sciences category (1998, 2005). He was a member of the Estonian Academy of Sciences. His research interests are psychology, perception, personality and neuroscience and his research works have received over 14,000 citations.

GHOLAMREZA ANBARJAFARI is heading the intelligent computer vision (iCV) research lab in the Institute of Technology at the University of Tartu. He is also Deputy Scientific Coordinator of he European Network on Integrating Vision and Language (iV&L Net) ICT COST Action IC1307. He is Associate Editor and Guest Lead Editor of several journals, Special Issues and Book projects. He is an IEEE Senior member and the Vice Chair of Signal Processing/Circuits and Systems/Solid-State Circuits Joint Societies Chapter of IEEE Estonian section. He has got Estonian Research Council Grant (PUT638) in January 2015 and has been involved in many international industrial and European projects. He is expert in computer vision, human-robot interaction, graphical models and artificial intelligence. He has been in the organizing committee and technical committee of IEEE Signal Processing and Communications Applications Conference in 2013, 2014 and 2016 and TCP of conferences such as ICOSST, ICGIP, SampTA, SIU and FG. He has been organizing challenges and workshops in FG17, CVPR17, and ICCV17. He is a leading guest editor of several special issues in top venues such as JIVP and MVA.