electronics

MDPI

Article

Two-Stage Recognition and beyond for Compound Facial Emotion Recognition

Dorota Kamińska [1,*,†], Kadir Aktas [2,3,†], Davit Rizhinashvili [2,†], Danila Kuklyanov [2,†], Abdallah Hussein Sham [4,†], Sergio Escalera [5,6,7,†], Kamal Nasrollahi [6,†], Thomas B. Moeslund [6,†] and Gholamreza Anbarjafari [2,3,8,9,†]

1 Institute of Mechatronics and Information Systems, Lodz University of Technology, 90924 Lodz, Poland
2 iCV Lab, University of Tartu, 50090 Tartu, Estonia; kadir.aktas@ut.ee (K.A.); DAVIT.RIZHINASHVILI@ut.ee (D.R.); DANILA.KUKLYANOV@ut.ee (D.K.); shb@ut.ee (G.A.)
3 iVCV OÜ, 51011 Tartu, Estonia
4 Enactive Virtuality Lab, University of Tallinn, 19086 Tallinn, Estonia; Abdallah.Hussein@ee.ut
5 Computer Vision Center, 08193 Barcelona, Spain; sescalera@ub.edu
6 Visual Analysis and Perception Lab, University of Aalborg, 9220 Aalborg, Denmark; kn@create.aau.dk (K.N.); tbm@create.aau.dk (T.B.M.)
7 Department of Mathematics and Computer Science, Universitat de Barcelona, 08011 Barcelona, Spain
8 PwC Advisory, 00180 Helsinki, Finland
9 Institute of Higher Education, Yildiz Technical University, Istanbul 34349, Turkey
* Correspondence: dorota.kaminska@p.lodz.pl; Tel.: +48-631-25-78
† All authors contributed equally to this work.

Abstract: Facial emotion recognition is an inherently complex problem due to individual diversity in facial features and racial and cultural differences. Moreover, facial expressions typically reflect the mixture of people's emotional statuses, which can be expressed using compound emotions. Compound facial emotion recognition makes the problem even more difficult because the discrimination between dominant and complementary emotions is usually weak. We have created a database that includes 31,250 facial images with different emotions of 115 subjects whose gender distribution is almost uniform to address compound emotion recognition. In addition, we have organized a competition based on the proposed dataset, held at FG workshop 2020. This paper analyzes the winner's approach—a two-stage recognition method (1st stage, coarse recognition; 2nd stage, fine recognition), which enhances the classification of symmetrical emotion labels.

Keywords: compound emotion recognition; facial expression recognition; dominant and complementary emotion recognition; deep learning

Citation: Kamińska, D.; Aktas, K; Rizhinashvili, D.; Kuklyanov, D.; Sham, A.H.; Escalera, S.; Nasrollahi, K.; Moeslund, T.B.; Anbarjafari, G. Two-Stage Recognition and beyond for Compound Facial Emotion Recognition. *Electronics* **2021**, *10*, 2847. https://doi.org/10.3390/electronics10222847

Academic Editor: George A. Papakostas

Received: 27 October 2021
Accepted: 15 November 2021
Published: 19 November 2021

1. Introduction

Personal emotional state and intentions change according to the flow of surrounding social communication. There are various ways of detecting emotional state of a humans using an AI agent [1–6]. One of the most common and effective means of predicting emotional states for a human is using facial expressions, which are caused by changes in facial features. People use different types of facial expressions to demonstrate different types of social intentions. In [7], it is mentioned that facial expressions can contribute up to 55% of the information exchanged during an interaction. Due to the high ratio, facial expressions holds an important place in personal relationships and emotion-based communication [8]. Taking such arguments as a basis, researchers have shown that automatic recognition of facial expressions can improve human–machine interaction to make it more natural and effective [9].

For this reason, there have been many studies on automatic emotional state analysis using facial expressions. Recently, the most successful results were obtained using deep-learning-based methods. Deep Convolutional Neural Networks (DCNN) have produced

particularly promising results. In [10], the authors achieved 97.8% accuracy on some datasets using DCNN. Other studies increased the DCNN performance by introducing complementary features. For example, Wang et al. (2018) discussed how depth awareness can be obtained in DCNN [11]. Furthermore, Mohan et al. [12] extracted local features from face images using a local gravitational force descriptor and fed a DCNN with those features, resulting in improved performance. Other works benefited from object detector networks fine-tuned for facial recognition tasks [13–15].

Although there are some promising results for facial emotion recognition tasks, there is still room for improvement. Low inter-class variance makes feature extraction in this task a challenge that could improve the model's efficiency. Furthermore, until recently, research on this topic has been conducted by mainly considering six basic emotions. These emotions (happiness, surprise, sadness, disgust, anger, and fear) were described by Ekman and Friesen [16] in 1971. However, human emotions can be more complex than basic emotions, considering the factors, such as mental state, cultural background, and personal relationships, which affect the emotional state [2,17–21]. In [22], the concepts of compound emotions are discussed to detail the basic emotions. It was shown that a human emotional state consists of many emotions, which are called dominant and complementary emotions. Considering this, compound emotions create another big challenge for the facial emotion recognition tasks as it is needed to recognize both the dominant and complementary emotions. Furthermore, distinguishing these two expressions can be extremely difficult in some cases. To address the above-mentioned issues, this work presents the following contributions:

- We organized a competition on recognizing compound emotions that require, in addition to performing an effective visual analysis, dealing with recognition of micro emotions. The database includes 31,250 faces with different emotions of 115 subjects whose gender distribution is almost uniform. The challenge was held at the FG2020.
- We introduce the winning method: a two-stage recognition algorithm to recognize compound emotions. In the first stage, coarse recognition, a DCNN was used to extract appearance features, and they are afterward combined with facial-point features. Then, in the second stage, fine recognition is done using a binary classifier. This two-stage recognition method is our first contribution to the proposed method. The second one is multi-modality, using appearance information with facial-point information. Moreover, to improve the performance, a model-ensembling based on the label distribution was used.

The rest of this paper is organized as follows: Section 2 reviews the related works. Section 3 presents the compound-emotion-recognition challenge. Section 4 introduces the winning method. Section 5 presents the experimental results. Section 6 includes the discussion for the proposed method. Finally, the paper is concluded in the last section.

2. Related Work

In 1872, Darwin conducted the first study of facial expressions by comparing a human facial expression with an animal facial expression. In their book, he discussed the differences and how they are connected [23]. Much later, Ekman and Friesen (1971) [16] launched the modern facial expression recognition effort by defining the six basic expressions of a human. Their work provided a base for future works in emotions classification. A few years later, Suwa et al. (1978) [24] proposed a method to automatically analyze facial expressions from a face video animation or an image sequence. Later, K. Mase and A. Pentland [25] analyzed facial expressions using optical flow to extract muscle movements of the face during these expressions.

Years after these initial works, a significant development occurred due to methods based on deep neural networks. In 2012, AlexNet achieved a major success in the ImageNet competition [26], and deep learning related methods have been developing rapidly since then. Deep neural networks are applied to many computer vision tasks, including facial emotion recognition. Researchers have tried to extract features from images and videos

using methods such as convolutional neural networks (CNN) [27] or long short-term memory (LSTM) [28]. However, the low variance between classes has always been a challenge. To improve accuracy, researchers have implemented different approaches. In [29], Hu et al. used DCNN and improved it by adding a supervised scoring ensemble block. In 2017, Guo et al. [30] proposed a multi-modal method. They utilized CNN-based networks using both appearance and geometrical information. Their work showed that multi-modality could be helpful to extract features for micro expressions.

Recently, there have been studies that add to the basic emotions, namelycompound facial emotion recognition. As mentioned by [31], by merging several components of emotional categories to create new ones, compound emotions can be constructed. In 2015, Du and Martinez [31] have identified and studied 17 compound emotions, including some distinctly biased towards one of the component emotions.

Many different researchers are currently sutdying facial emotion recognition. The same interest exists in the field of compound emotion recognition. In general, the task of compound facial emotion recognition is crucial, as such emotions assist in solving specific behavioral or social problems [31]. Compound emotions consist in dominant and complementary emotions and are more detailed than the basic facial emotions [22,32,33]. Based on the research of Du and Martinez (2015) [31], we can see that facial expressions of these emotions are consistent across people and differential between emotional categories; moreover, the associated facial expressions can be universally visually recognized by observers. However, according to Guo et al. (2018) [22], the compound emotions are more challenging recognize in comparison with the basic ones.

Emotion perception and classification can be classified into two leading models: continuous and categorical [34]. A categorical model is based on C classifiers that are tuned to specific emotion categories [34]. According to Du and Martinez (2012), compound emotion recognition can be done by linearly combining C face spaces whose dimensions are mostly configural [34]. However, in this case, the classification of facial expressions is focused on the precise detection of facial landmarks rather than on recognition [34].

Facial emotion recognition can be done using different methodologies [35–38]. For example, in the research of Liliana et al. (2017), the authors used twelve types of compound facial emotion recognition in a sequence of images using two-stage learning, which was a combination of Support Vector Machines (SVM) and Conditional Random Fields (CRF) as sequence classifiers [39]. In this approach, SVM classifies each image frame and produces an emotion label output, which becomes the input for CRF [39]. CRF, in turn, yields the mixed emotion label of the corresponding observation sequence [39].

Zhang et al. (2020) propose a two-stage compound-emotion-recognition method that includes coarse and fine recognition stages [40]. This method allows the enhancement of the classification for symmetrical emotion labels [40]. Li et al. (2017) in their research propose using a Deep Locality-Preserving CNN (DLP-CNN) method for expression recognition [41]. This method enhances the discriminative power of deep features by preserving the locality closeness while maximizing the inter-class scattering [41]. Based on the experiments conducted, the authors show that the DLP-CNN method produces state-of-the-art results for emotion recognition in the wild.

3. Compound Emotion Recognition Challenge

3.1. Compound Emotions Database

Human–computer interaction performs more realistically if the computer can recognize more complex expressions of the human interacting with it. However, most publicly available datasets are mainly limited to six basic emotions, where each face image has only one from a set of six labels assigned. For this reason, we have developed a large facial expression database called iCV Multi-Emotion Facial Expression Dataset (iCV-MEFED) designed for multi-emotion recognition. This database is unique as it tries to investigate dominant and complementary emotions. The main aim for making this database was to understand which automated process can recognize combinations of seven primary emo-

tions. The database includes 31,250 images of facial expressions with different emotions from 115 subjects whose gender distribution is almost uniform. Subjects' age ranged from 18 to 37 years old with different ethnicity and hairstyles. The room has uniform lighting conditions; hence, the variation of light changes can be ignored. Each subject acts with 50 different emotions (Table 1), and for each of these emotions, five samples have been taken. The images are taken and labeled under the supervision of psychologists, and the subjects have been trained about acting emotions that they posed.

Table 1. 49 Dominant–complementary emotion combinations (the 50th 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**

As mentioned above, for each subject in the database, five samples for the set of 50 emotions have been captured by a Canon 60D camera under uniform lighting conditions with a relatively unchanged background and a resolution of 5184×3456. This was done so that the challenge solutions can focus on the main problem and avoid additional preprocessing requirements. For each subject, the purpose of the database and all the seven main emotions were thoroughly explained. Before capturing, subjects were instructed to remove any hair from their face and not to move their heads too much. For each emotion, examples were simultaneously displayed. If a person had trouble expressing emotion, common traits about the emotion were given, for example, thin lips for contempt. However, the combinations of emotions were left up to the subjects themselves. A few examples from the database can be seen in Figure 1.

After the data, were captured, they were given to psychologists to assess the realness or truthfulness of the expressions. During this process, few subjects that did not manage to convey their emotions well enough were eliminated. Even though the subject was mostly everyday people, not actors, the data captured are relatively natural-looking and could provide a huge benefit for researchers who are interested in combined emotions and how they could be recognized.

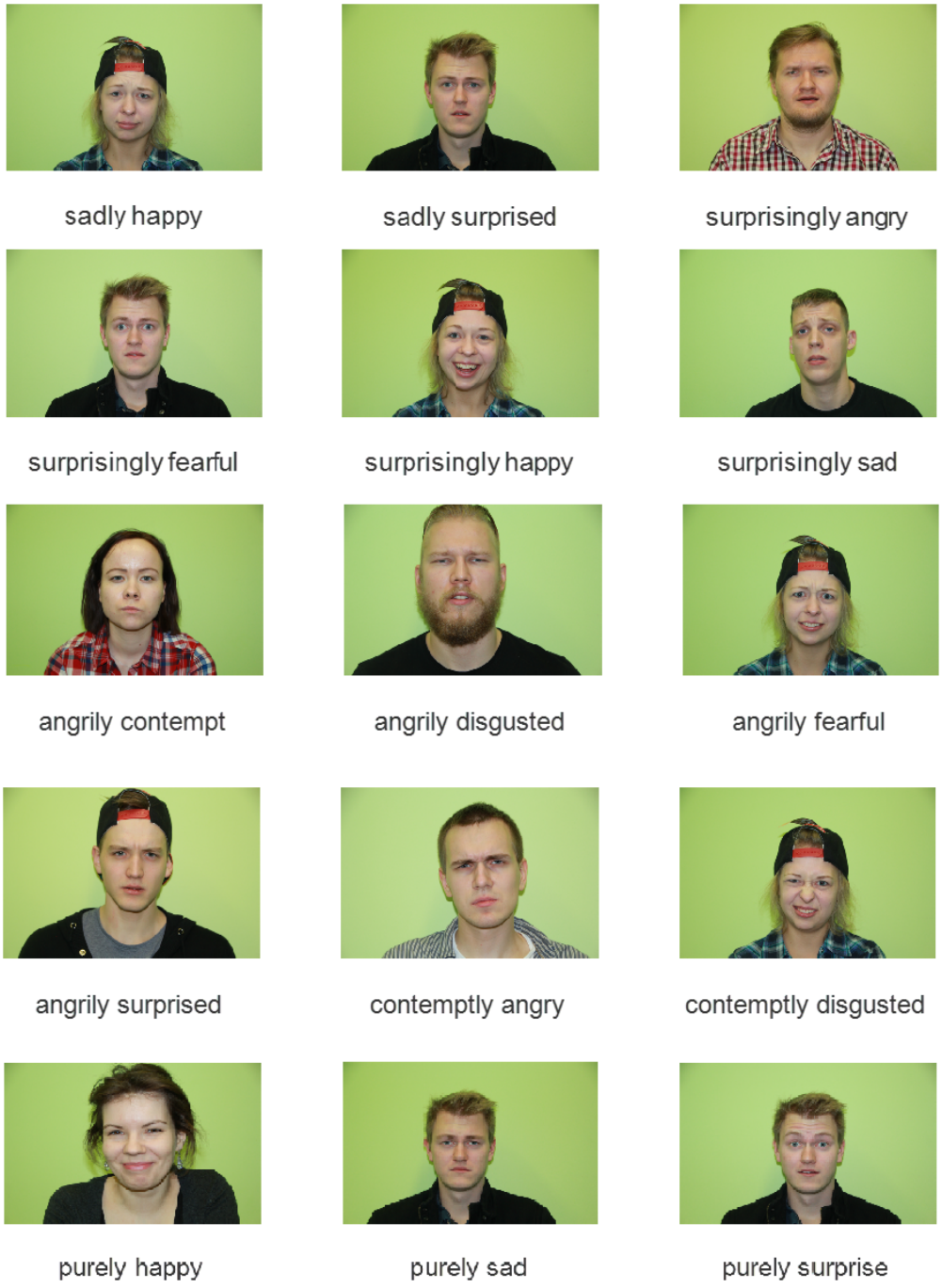

Figure 1. Some example emotions from the database.

3.2. Baseline–Compound Emotion Recognition Using Multi-Modality Network with Visual and Geometrical Information

This method combines the visual and geometrical information in an end-to-end convolutional neural network (CNN) [22,30]. First, each face ID's mean landmarks were estimated. Then, the geometrical representation of landmarks displacement is extracted. The visual features are extracted by modified AlexNet [42], and the geometrical features are represented as the face landmark displacement. In this way, both the geometrical and the visual information are fully utilized for emotion recognition.

Geometrical Representation. Dlib (http://dlib.net/, accessed on 14 November 2021) [43] was used to extract facial landmarks, and the algorithm in [44] was used to perform face alignment by using center point from two eyes and center point from the upper lip.

Network Structure. The network structure of the method is shown in Figure 2. The visual feature is a vector $p_1 \in \mathbb{R}^{256}$, and the geometrical feature is $p_2 \in \mathbb{R}^{136}$. Both p_1 and p_2 are concatenated into $p \in \mathbb{R}^{392}$ (see Figure 2). Then, the concatenated feature p is fed into a fully connected layer, and the hinge loss is adopted to optimize the whole network. In addition, the last fully connected layer with hinge loss is similar to a Supported Vector Machine (SVM) classifier, which is often used to classify geometric representations.

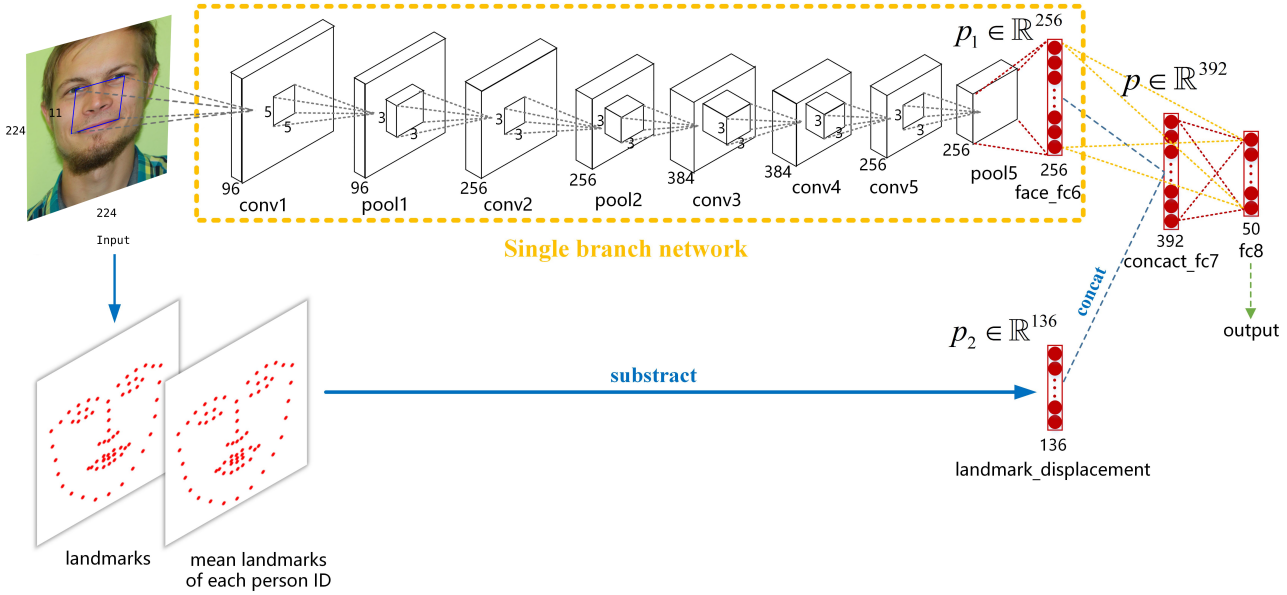

Figure 2. Overview of the 1st-place winner method of the competition. The upper part is a single-branch CNN network. The whole architecture constructs the multimodality network. It is similar to the single CNN, but one more input, named landmark displacement, is added into classifier in the last two layers of the network.

The CNN's extracted feature p_1 spans a vector space V_1, where decision boundary can correctly divide some samples, while its ability may reach the ceiling. Once the landmark displacement vector p_2 is embedded into the lower vector space, V_1 is mapped from a lower dimension into a higher dimension space V. Then V becomes more divisible because of the effectiveness of p_2, being similar to the kernel function in SVM but not nonlinear.

In the training phase, the size of the input image is 224×224, and the landmark displacement 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 is divided by 5 every 20,000 iterations. A weight decay of 5×10^{-4} and a momentum of 0.9 are adopted. In testing, the feature p_1 and each ID's landmark displacements p_2 are first calculated, and then they are concatenated as the input of the classifier, following the same pipeline as Figure 2.

Apart from multimodality network with visual and geometrical information, other baselines such as unsupervised Learning of CNN [45] and face alignment images trained with inception-V3 model and center loss [46] can be used. For the 50 labels as shown in Table 2, the baselines for the aforementioned three methods' accuracy of each category of emotion on the testing set of the iCV-MEFED dataset is shown in Figure 3.

Table 2. The 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		

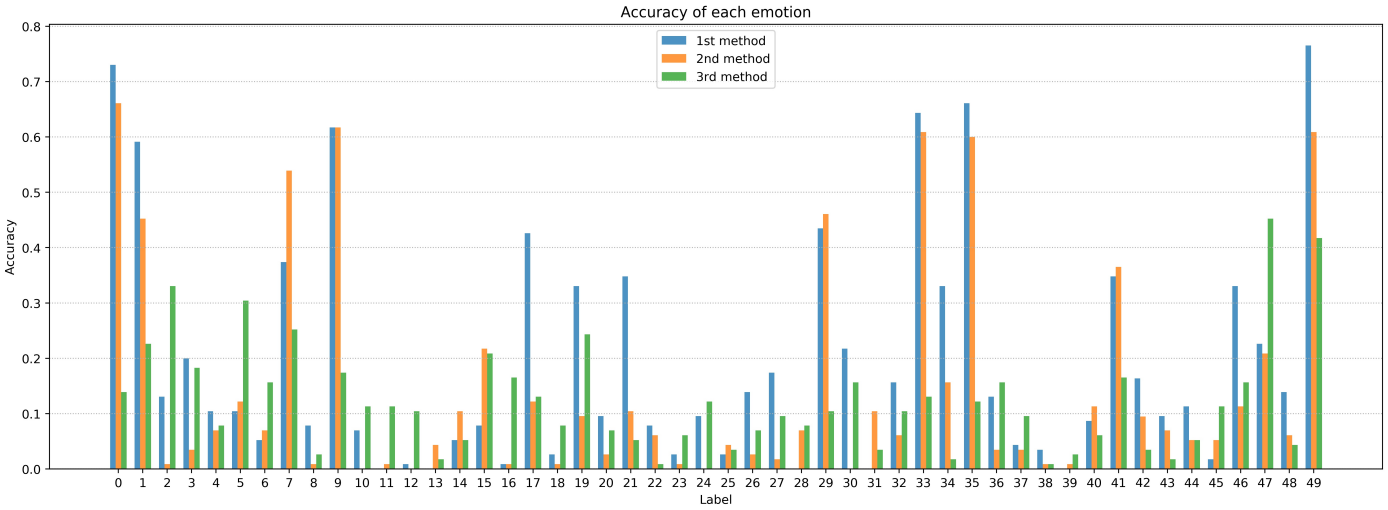

Figure 3. The accuracy of the three baseline methods (multimodality network with visual and geometrical information, unsupervised learning of CNN and face alignment images trained with inception-V3 model and centre loss, respectively) of each category of emotion on the testing set of the iCV-MEFED dataset.

3.3. Evaluation Metric

The CodaLab platform was adopted for the challenge for participants to submit their predictions and track their progress. The participants were provided with training (with labels) and validation data (without labels) during the development stage. Finally, during the test stage, test data (without labels) and validation labels were released. This way, training and validation data are used for internal validation, while test data are used for external validation. The data division was of 70%, 20%, and 10% of the total data for training, validation, and testing, respectively. The participants submitted the code and all dependencies via CodaLab, and the organizers ran the codes. The evaluation was based on the average correct emotion recognition. The evaluation metric used in the Challenge was defined as the percentage of misclassified instances. The final rank was provided based on the misclassification rate on the test set.

4. The Winning Approach

In [30], the authors show sthat the visual and the geometrical information of the face can be used to differentiate between facial expressions that have a slight variance between them. The winning method used appearance features and the facial-point features together for the compound emotion recognition task. In the proposed method, there is data preprocessing that is followed by two stages. In the first stage, the winners use the aforementioned features together to make a coarse recognition. Then, they follow with a fine recognition to enhance the recognition results in the second stage. Moreover, they use ensembling on the labels to improve recognition accuracy. In this section, we present the proposed method in detail.

4.1. The Two-Stage Recognition Model

The proposed two-stage recognition method (shown in Figure 4) includes a coarse recognition and a fine recognition, respectively, in the first-stage and in the second-stage. A preprocessing is also included to crop and align the face regions from the images to reduce the computational overhead.

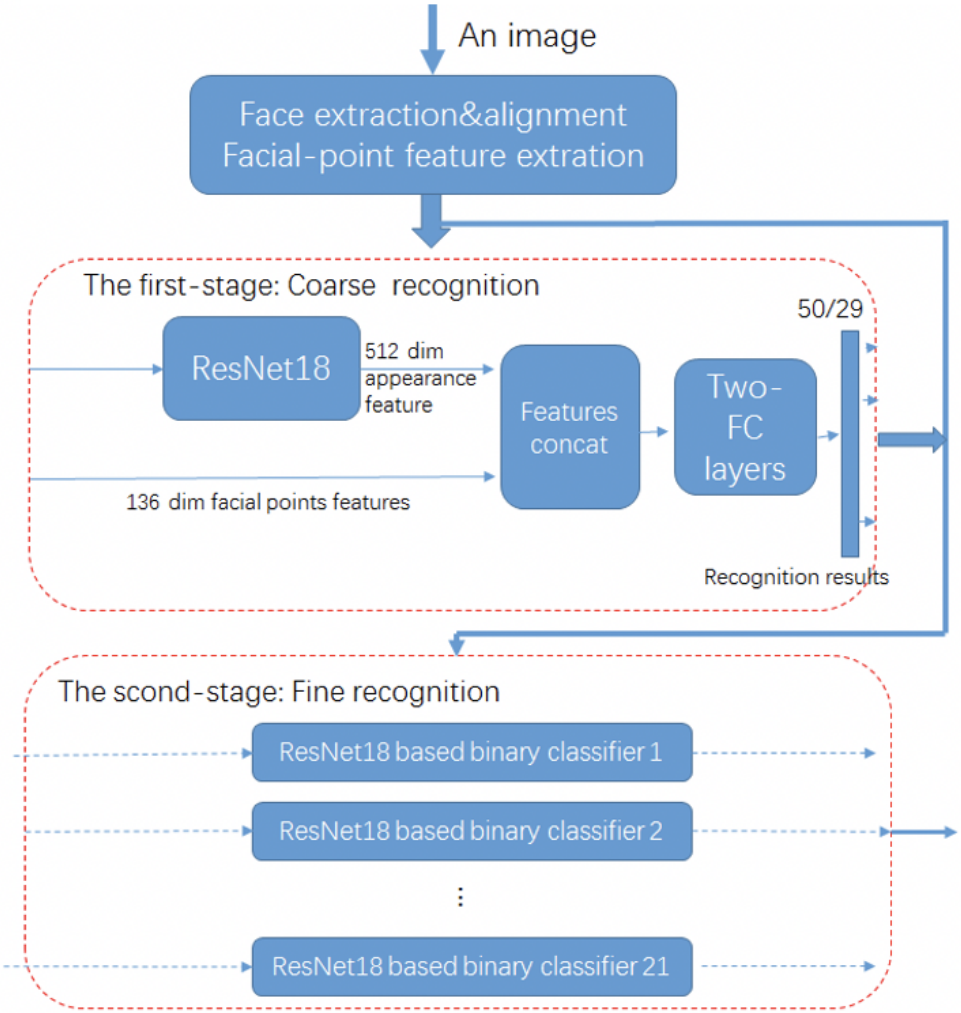

Figure 4. Framework of the proposed method.

In the first stage, classification is done using the appearance features concatenated with the facial-point features. Appearance features are extracted using a DCNN. For facial point features, the winners followed the same approach taken in [30], which aims to reduce the variance of the same facial expression in different faces. It was observed that the first-stage classifier performs poorly in classifying the symmetrical emotion types such as

angry–surprised and *surprised–angry*. To reduce such errors, the second-stage recognition was introduced. In this stage, the winners cluster the symmetrical labels and train a binary classifier for each cluster. The difference with the first stage is that it uses the symmetrical emotions data for each classifier and uses a sigmoid function in the last layer instead of the softmax function. Furthermore, the winners use a different feature extractor to extract the appearance feature between the symmetrical emotions.

4.1.1. Preprocessing

For this competition, the compound emotional recognition task is defined as a single-label-recognition task and thus has a multi-class classification problem with a large number of classes. Therefore, the original complementary label and dominant label were transferred into a single label. Following the proposal of [30], the following formula for the label transformation was used:

$$n = 7 \times (c - 1) + d \tag{1}$$

where c is the complementary emotion and d is the dominant emotion. By doing that, the original problem is converted into a classification problem with 50 classes.

In addition to label preparation, faces from the images are aligned and cropped using the method proposed in [43]. Firstly, facial landmarks are extracted using an ensemble of regression trees. Then, the faces are aligned according to their landmarks and cropped to 224×224. This image size is selected after trying different sizes as larger image resolution does not provide a significant gain in the performance. This preprocessing reduces the size of the image, which saves computational resources. At the same time, face alignment prevents irrelevant factors such as background or head pose to influence the results.

4.1.2. The First-Stage: Coarse Recognition

Facial appearance features are one of the standard features used in facial expression tasks as they provide excellent information due to the nature of the task. However, since the iCV-MEFED dataset is a small dataset for such a complex task, more information to train the classifier besides the facial appearance features was needed. Hence, facial-point information is exploited as another set of features.

In previous studies, it has been shown that facial-point information is essential for emotion recognition. Therefore, in the proposed method, the facial-point information extraction method from [30] was adopted. Firstly, the mean facial point value for each ID was calculated, and then the difference between the facial point value and its mean value was calculated as the facial-point features. Details are explained in [30]. Finally, concatenated appearance features and facial-point features to complement each other were used for better recognition. Since the training and the test samples are created from completely different people, we did not decorrelate the features.

Since the training dataset size is not very large in this competition, the ResNet18-like structure used in the face recognition domain as our appearance features extraction backbone was selected. The resNet18-like framework includes 18 deep layers. The input size is 224×224. After two linear layers, the final appearance feature vector is 512. This vector is concatenated with 136 dim facial-point feature vector to make the images' final feature representation vector. Table 3 lists the ResNet18-like structure that was used.

Table 3. Network structure of ResNet 18-like.

Layer Name	Output Size	Layer
Input layer	224×224	$3 \times 3, 64, \text{stride} = 1$
Layer1	112×112	$[3 \times 3, 64] \times 2$ $[3 \times 3, 64] \times 2$
Layer2	56×56	$[3 \times 3, 128] \times 2$ $[3 \times 3, 128] \times 2$
Layer3	28×28	$[3 \times 3, 256] \times 2$ $[3 \times 3, 256] \times 2$
Layer4	14×14	$[3 \times 3, 512] \times 2$ $[3 \times 3, 512] \times 2$

4.1.3. The Second-Stage: Fine Recognition

After training the first-stage recognition, the trained model was tested on the training dataset. The confusion matrix is illustrated in Figure 5.

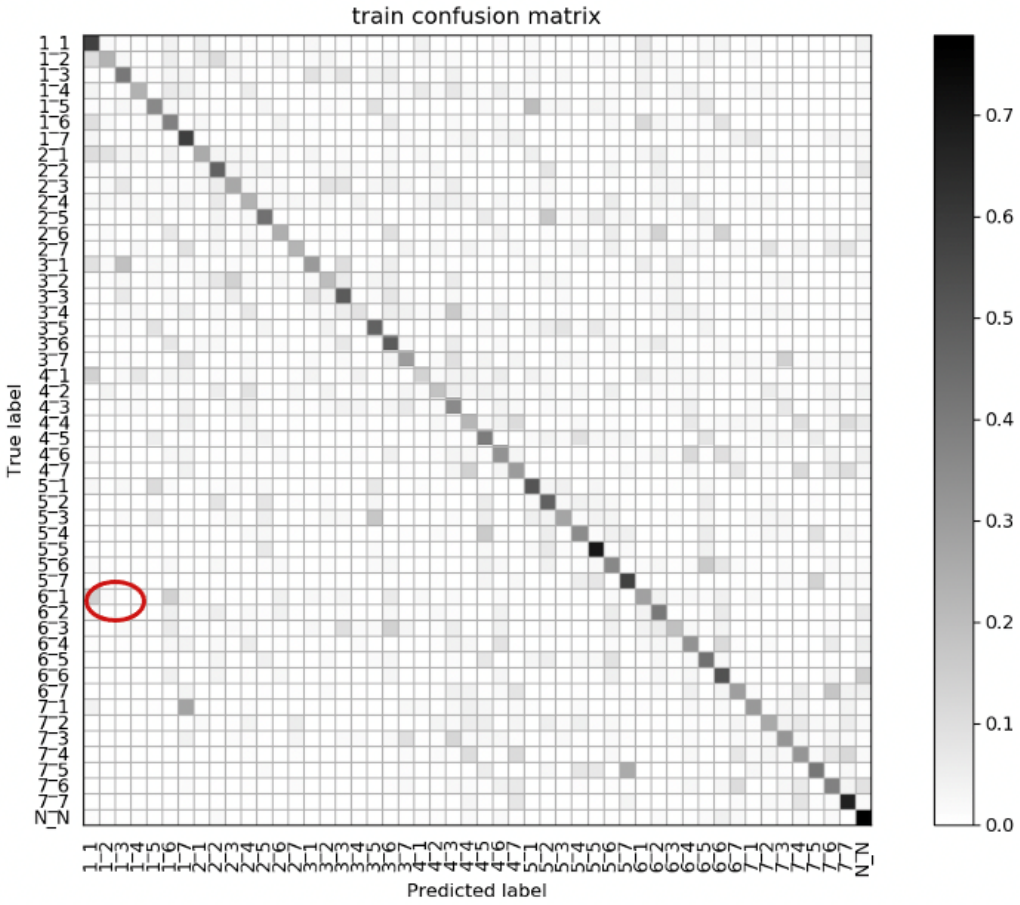

Figure 5. Confusion matrix of the training dataset.

From this confusion matrix, it was found that a large proportion of false-detections occur between the symmetrical emotions such as *angry–surprised* and *surprised–angry* (red circle (28%) in Figure 5). This motivated the winners to use a second stacked recognition stage to enhance the recognition of such symmetrical emotions. In this stage, firstly, they cluster the symmetrical labels such as (1_7, 7_1) together; after that, 50 labels are clustered to 29 labels. In these 29 labels, there are 21 symmetrical labels, since there are 8 labels in which the complementary emotion label and the dominant emotion label are the same. Then, features were extracted and a binary classifier was trained for each symmetrical label,

which helped mitigate the error caused by the symmetrical labels. In this stage, symmetrical emotion samples were used to train the classifier. The feature extractor and the binary classifier used in this stage are similar to the methods used in the first stage. A ResNet18-like backbone is used to extract the appearance features between the symmetrical emotions. Then, the same fully connected structure is used as the binary classifier. However, in this stage, a sigmoid function is used in the last layer instead of a softmax function. Thanks to the fine recognition in this stage, recognition performance improved while the false cases caused by symmetrical labels are better handled. Figure 6 summarizes this procedure.

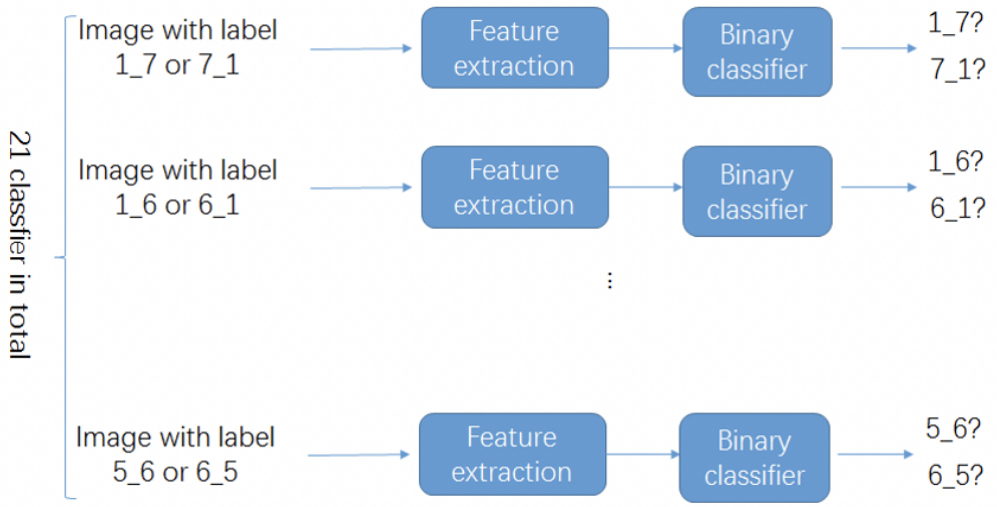

Figure 6. Binary classifications for classifying the symmetrical labels.

4.1.4. Ensembling and Voting

The dataset includes 31,250 facial images with different emotions of 125 subjects, each subject acting out 50 different emotions. Note that five images were taken with a camera for each emotion of a subject. These five images of each emotion can be captured in a continuous way as shown in Figure 7.

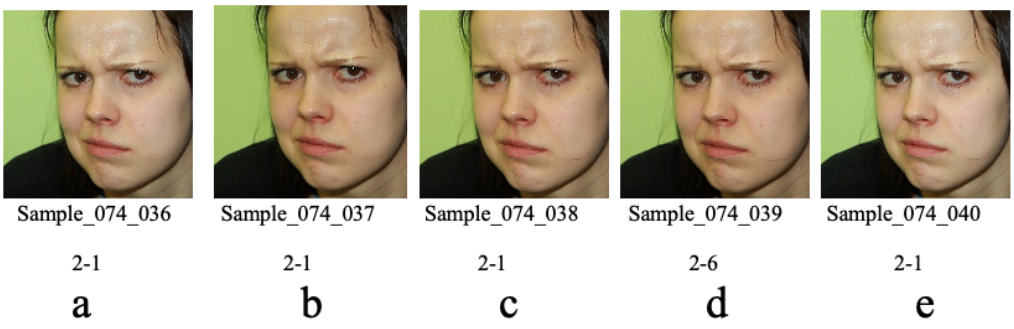

Figure 7. Continuous images for one label.

One can see that these five sample images of the same emotion look very similar. In the video object detection domain, multiple continuous frames were used to vote to decide the final object category. Such a method can improve the robustness of detection results for the task at hand. We observed that the recognition results for these burst images are sometimes disturbed. For example, image d in Figure 7 is recognized as 2_6, while images a, b, c and e in the same figure are recognized as 2_1 by the model at the same time.

To improve the robustness of the recognition task, the sequential information in the same way as in the video object detection domain was used.

In the method, these five burst images were taken as one sequence, and the majority voting recognition result as the final output of all the images in this sequence was used. For example, the prediction result for image d in Figure 7 will be changed to 2_1 after voting.

5. Experimental Results

We trained and tested the proposed method on the competition dataset, comprising both dominant and complementary labels for each sample/image, hence, increasing the complexity of the challenge.

In our experiments, we cropped and resized the input images into a resolution of 224 by 224 pixels. The facial-point representation uses a vector of 136 dimensions. We adopted the RADAM [47] optimization method and set the batch size to 512. The initial learning rate of RADAM was 0.0001 and the epochs number was 1000. Due to the relatively small size of the dataset, we used a dropout ratio of 0.8 in the linear layer of the network when training the above three models. We used PyTorch as both the training and testing platforms. To achieve a more accurate recognition rate, we trained three different models. We then used these models for model ensembling. For better understanding, we make the following two definitions:

- Dialog tag: we used the same complementary and dominant as a dialog tag to define the label. There are eight dialog tags in this competition such as N_N1_1
- Symmetrical tag: we defined the symmetrical labels pair as a symmetrical tag. There are 21 symmetrical tags in this competition such as $1\_7, 7\_1$

Next, we described these three models in detail:

- Model_1: One-stage model: In this model, we used only the first-stage recognition, and the output of this model is the prediction results of all 50 categories. The backbone of this Model is ResNet18-like, where we used both appearance and facial-point features for the classifier.
- Model_2: Two-stage recognition model-version1: The first-stage recognition is the same as Model 1, and the prediction results will go through the second-stage recognition. In this stage, we trained 21 binary classifiers corresponding to 21 Symmetrical tags to classify the final label for each symmetrical tag.
- Model_3: Two-stage recognition model-version2: We transferred the 50 categories into 29 tags for training data in this model and retrained the first stage recognition with 29 output categories. Furthermore, the prediction results would go through the corresponding binary classifier in the second stage of recognition for the final output.

We used the ensembling method explained in Section 4.1.4 to vote on the final output predict outcomes after we obtained the results from the three models. The ensembling process of our approach is depicted in Figure 8.

Finally, the three models were tested separately on the test dataset, and the ensembling results are reported in Table 4.

Table 4. The competition accuracy results of the proposed method.

Method	Test Set
Model_1	18.51
Model_2	19.71
Model_3	19.28
Ensembling	21.83

We can observe that the two-stage recognition approach performs better than the one-stage approach. Moreover, the ensembling method further improves the performance. Using ensembling method with the two-stage approach, we can achieve about 4% gain compared to model_1. Finally, we achieve 21.83% accuracy. Our method with ensembling performs also better than all the baseline methods in the competition and achieves over 2%

accuracy improvement compared to the best performing baseline, i.e., 1st baseline method. Figure 9 shows the training loss of model_1 and the first stage recognition in model_3.

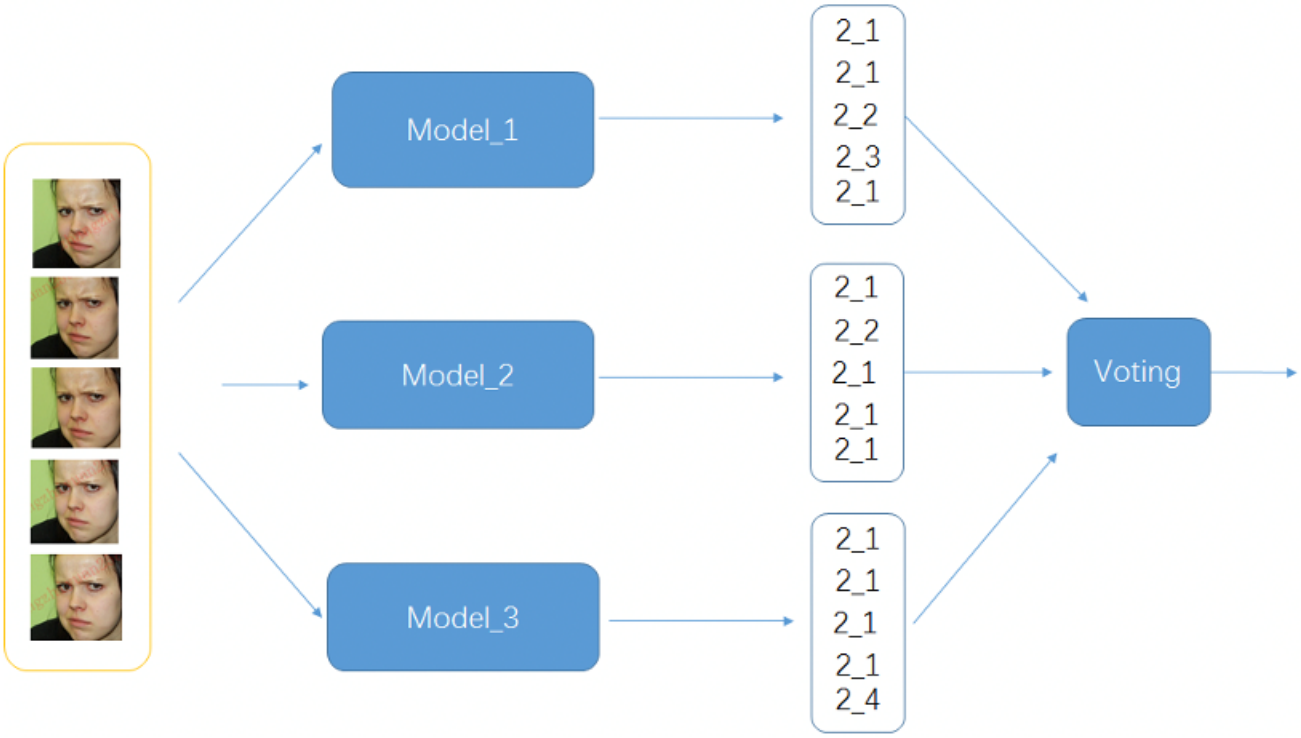

Figure 8. Assembling of three models.

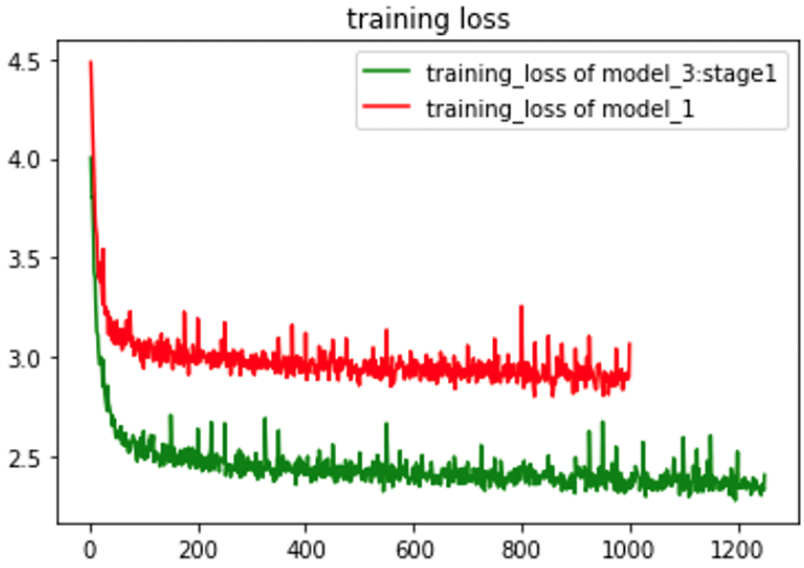

Figure 9. Training loss of model_1 and model_3: stage1.

According to Figure 9, the average loss of both models is steady in the range [2.2, 2.7], which is our model selection criterion.

6. Discussion

6.1. Backbone Selection

The backbone selection is vital for the model quality. Since the training dataset size is relatively small, we selected the following three candidate backbones, namely Alexnet, ResNet18-like (the input layer is slightly different; see Table 1), and ResNet34. As was

shown in the experimental results, ResNet18-like backbone is the best option for the compound emotion recognition. For a fair comparison, we used the same facial-point features in all of the above models. Since the GPU usage of the ResNet34 model is more extensive than ResNet18-like and Alexnet, we set the batch size of ResNet34 to 256 for the training completion. It was observed that the ResNet backbone over perform than Alexnet because ResNet structure usually can extract more discriminated features, which has been shown in other computer-vision-related tasks [48–50]. However, the predicted results of ResNet34 is not as good as ResNet18, possibly because the training process's batch size is small due to the small training dataset size.

6.2. Image Resolution

In order to investigate the effect of image resolution on the recognition performance, the following two image resolutions were investigated: 112×112 and 224×224. We utilized the backbone network as the ResNet18-like framework. The experimental results have shown that a higher resolution can provide slight improvement by 0.3% for this dataset.

7. Conclusions and Future Work

Compound emotion recognition is a new topic in automation that aims to overcome the shortcomings of basic emotion representation such as limited expressions represented. However, there are challenges that come with compound emotions, such as the high number of classes and complexity of the expression, which is yet to be tackled satisfactorily. We aimed to handle these challenges by presenting a two-stage recognition approach.

In this paper, we introduced the proposed method we used in the Compound Emotion Recognition Competition. The proposed method includes three major contributions:

- A two-stage strategy is used to mitigate the symmetrical label misclassification.
- We benefit from both appearance and facial-point information for compound emotion recognition.
- We ensemble one-stage and two-stage base models to further enhance the performance.

Our experimental results showed that by combining facial appearance features and facial-point features, emotion recognition performance improved compared to baseline methods in average correct predictions.

One of the main limitations of the proposed method is noise handling. During our training process, we found out that the dataset may include some noisy data, which affects the training process. We observed that the noisy data occur when the actor "acted" the emotion differently from how a person ordinarily would. Therefore, we think that different people may understand "emotion" differently, thus making the "cognition noise" task hard. When the multiple-emotion problems are involved, it will be even harder to say which emotion is dominant and complementary. As a future work, the noisy data can be eliminated from the dataset, or the consequences of the noisy data can be reduced by improving the method. Moreover, a dataset with the compound emotions in the wild can be collected to evaluate the performance on a more naturally expressed sample set.

Author Contributions: Investigation, D.K. (Dorota Kamińska), K.A., D.R. and D.K. (Danila Kuklyanov); Supervision, D.K. (Dorota Kamińska), A.H.S., S.E., K.N., T.B.M. and G.A.; Validation, D.K. (Dorota Kamińska) and A.H.S.; Writing—original draft, D.K. (Dorota Kamińska), K.A., A.H.S. and S.E.; Writing—review and editing, D.K. (Dorota Kamińska), D.R., D.K. (Danila Kuklyanov), A.H.S., S.E., K.N., T.B.M. and G.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: We want to express our special thanks to Zhiyuan Zhangs, Jianping Shen, Miao Yi, Juan Xu, and Rong Zhang from Pingan Life Insurance of China for participating in the competition held at the FG workshop 2020. This work has been partially supported by the Spanish project PID2019-105093GB-I00 (MINECO/FEDER, UE), CERCA Programme/Generalitat de Catalunya), ICREA under the ICREA Academia programme, and the Estonian Centre of Excellence in IT (EXCITE) funded by the European Regional Development Fund.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Alarcao, S.M.; Fonseca, M.J. Emotions recognition using EEG signals: A survey. *IEEE Trans. Affect. Comput.* **2017**, *10*, 374–393. [CrossRef]
2. Sapiński, T.; Kamińska, D.; Pelikant, A.; Ozcinar, C.; Avots, E.; Anbarjafari, G. Multimodal database of emotional speech, video and gestures. In *International Conference on Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 153–163.
3. Noroozi, F.; Kaminska, D.; Corneanu, C.; Sapinski, T.; Escalera, S.; Anbarjafari, G. Survey on emotional body gesture recognition. *IEEE Trans. Affect. Comput.* **2018**, *12*, 505–523. [CrossRef]
4. Tammvee, M.; Anbarjafari, G. Human activity recognition-based path planning for autonomous vehicles. *Signal Image Video Process.* **2021**, *15*, 809–816. [CrossRef]
5. Saxena, A.; Khanna, A.; Gupta, D. Emotion recognition and detection methods: A comprehensive survey. *J. Artif. Intell. Syst.* **2020**, *2*, 53–79. [CrossRef]
6. Deng, J.; Ren, F. A Survey of Textual Emotion Recognition and Its Challenges. *IEEE Trans. Affect. Comput.* **2021**. [CrossRef]
7. Zhou, J.; Zhang, S.; Mei, H.; Wang, D. A method of facial expression recognition based on Gabor and NMF. *Pattern Recognit. Image Anal.* **2016**, *26*, 119–124. [CrossRef]
8. Zeng, J.; Shan, S.; Chen, X. Facial expression recognition with inconsistently annotated datasets. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 222–237.
9. Cohen, I.; Sebe, N.; Garg, A.; Chen, L.S.; Huang, T.S. Facial expression recognition from video sequences: Temporal and static modeling. *Comput. Vis. Image Underst.* **2003**, *91*, 160–187. [CrossRef]
10. Karnati, M.; Seal, A.; Krejcar, O.; Yazidi, A. FER-net: Facial expression recognition using deep neural net. *Neural Comput. Appl.* **2021**, *33*, 9125–9136. [CrossRef]
11. Wang, W.; Neumann, U. Depth-aware cnn for rgb-d segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 135–150.
12. Mohan, K.; Seal, A.; Krejcar, O.; Yazidi, A. Facial Expression Recognition Using Local Gravitational Force Descriptor-Based Deep Convolution Neural Networks. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–12. [CrossRef]
13. Levi, G.; Hassner, T. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, DC, USA, 9–13 November 2015; pp. 503–510.
14. Zhao, X.; Liang, X.; Liu, L.; Li, T.; Han, Y.; Vasconcelos, N.; Yan, S. Peak-piloted deep network for facial expression recognition. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 425–442.
15. Grobova, J.; Colovic, M.; Marjanovic, M.; Njegus, A.; Demire, H.; Anbarjafari, G. Automatic hidden sadness detection using micro-expressions. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 828–832.
16. Ekman, P.; Friesen, W.V. Constants across cultures in the face and emotion. *J. Personal. Soc. Psychol.* **1971**, *17*, 124. [CrossRef] [PubMed]
17. Alameda-Pineda, X.; Ricci, E.; Sebe, N. Multimodal behavior analysis in the wild: An introduction. In *Multimodal Behavior Analysis in the Wild*; Elsevier: Amsterdam, The Netherlands, 2019; pp. 1–8.
18. Izdebski, K. *Emotions in the Human Voice, Volume 3: Culture and Perception*; Plural Publishing: San Diego, CA, USA, 2008; Volume 3.
19. Keltner, D.; Sauter, D.; Tracy, J.; Cowen, A. Emotional expression: Advances in basic emotion theory. *J. Nonverbal Behav.* **2019**, *43*, 133–160
20. Haamer, R.E.; Rusadze, E.; Lsi, I.; Ahmed, T.; Escalera, S.; Anbarjafari, G. Review on emotion recognition databases. *Hum. Robot Interact. Theor. Appl.* **2017**, *3*, 39–63.
21. Noroozi, F.; Marjanovic, M.; Njegus, A.; Escalera, S.; Anbarjafari, G. Audio-visual emotion recognition in video clips. *IEEE Trans. Affect. Comput.* **2017**, *10*, 60–75. [CrossRef]
22. Guo, J.; Lei, Z.; Wan, J.; Avots, E.; Hajarolasvadi, N.; Knyazev, B.; Kuharenko, A.; Junior, J.C.S.J.; Baró, X.; Demirel, H.; et al. Dominant and complementary emotion recognition from still images of faces. *IEEE Access* **2018**, *6*, 26391–26403. [CrossRef]
23. Darwin, C.; Prodger, P. *The Expression of the Emotions in Man and Animals*; Oxford University Press: Oxford, UK, 1998.
24. Sown, M. A preliminary note on pattern recognition of facial emotional expression. In Proceedings of the 4th International Joint Conferences on Pattern Recognition, Kyoto, Japan, 7–10 November 1978.
25. Mase, K. An Application of Optical Flow-Extraction of Facial Expression. In *MVA*; Tokyo, Japan, 1990; pp. 195–198. Available online: https://www.cvl.iis.u-tokyo.ac.jp/mva/proceedings/CommemorativeDVD/1990/papers/1990195.pdf (accessed on 1 November 2021).
26. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
27. Hasani, B.; Mahoor, M.H. Facial expression recognition using enhanced deep 3D convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 30–40.
28. Yu, Z.; Liu, G.; Liu, Q.; Deng, J. Spatio-temporal convolutional features with nested LSTM for facial expression recognition. *Neurocomputing* **2018**, *317*, 50–57. [CrossRef]

29. Hu, P.; Cai, D.; Wang, S.; Yao, A.; Chen, Y. Learning supervised scoring ensemble for emotion recognition in the wild. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, Glasgow, UK, 13–17 November 2017; pp. 553–560.
30. Guo, J.; Zhou, S.; Wu, J.; Wan, J.; Zhu, X.; Lei, Z.; Li, S.Z. Multi-modality network with visual and geometrical information for micro emotion recognition. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 814–819.
31. Du, S.; Martinez, A.M. Compound facial expressions of emotion: From basic research to clinical applications. *Dialogues Clin. Neurosci.* **2015**, *17*, 443.
32. Loob, C.; Rasti, P.; Lüsi, I.; Jacques, J.C.; Baró, X.; Escalera, S.; Sapinski, T.; Kaminska, D.; Anbarjafari, G. Dominant and complementary multi-emotional facial expression recognition using c-support vector classification. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 833–838.
33. Lüsi, I.; Junior, J.C.J.; Gorbova, J.; Baró, X.; Escalera, S.; Demirel, H.; Allik, J.; Ozcinar, C.; Anbarjafari, G. Joint challenge on dominant and complementary emotion recognition using micro emotion features and head-pose estimation: Databases. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 809–813.
34. Martinez, A.; Du, S. A model of the perception of facial expressions of emotion by humans: Research overview and perspectives. *J. Mach. Learn. Res.* **2012**, *13*, 1589–1608.
35. Wan, J.; Escalera, S.; Anbarjafari, G.; Jair Escalante, H.; Baró, X.; Guyon, I.; Madadi, M.; Allik, J.; Gorbova, J.; Lin, C.; et al. Results and analysis of chalearn lap multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenges. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 29 October 2017; pp. 3189–3197.
36. Kulkarni, K.; Corneanu, C.; Ofodile, I.; Escalera, S.; Baro, X.; Hyniewska, S.; Allik, J.; Anbarjafari, G. Automatic recognition of facial displays of unfelt emotions. *IEEE Trans. Affect. Comput.* **2018**, *12*, 377–390. [CrossRef]
37. Haamer, R.E.; Kulkarni, K.; Imanpour, N.; Haque, M.A.; Avots, E.; Breisch, M.; Nasrollahi, K.; Escalera, S.; Ozcinar, C.; Baro, X.; et al. Changes in facial expression as biometric: A database and benchmarks of identification. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 621–628.
38. Gorbova, J.; Lusi, I.; Litvin, A.; Anbarjafari, G. Automated screening of job candidate based on multimodal video processing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 26 July 2017; pp. 29–35.
39. Liliana, D.Y.; Basaruddin, C.; Widyanto, M.R. Mix emotion recognition from facial expression using SVM-CRF sequence classifier. In Proceedings of the International Conference on Algorithms, Computing and Systems, Jeju, Korea, 10–13 August 2017; pp. 27–31.
40. Zhang, Z.; Yi, M.; Xu, J.; Zhang, R.; Shen, J. Two-stage Recognition and Beyond for Compound Facial Emotion Recognition. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, 16–20 November 2020; pp. 900–904.
41. Li, S.; Deng, W.; Du, J. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 26 July 2017; pp. 2852–2861.
42. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
43. Kazemi, V.; Sullivan, J. One millisecond face alignment with an ensemble of regression trees. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23 June 2014; pp. 1867–1874.
44. Tan, Z.; Zhou, S.; Wan, J.; Lei, Z.; Li, S.Z. Age Estimation Based on a Single Network with Soft Softmax of Aging Modeling. In *Asian Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 203–216.
45. Knyazev, B.; Barth, E.; Martinetz, T. Recursive autoconvolution for unsupervised learning of convolutional neural networks. In Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 14–19 May 2017; pp. 2486–2493.
46. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. A Discriminative Feature Learning Approach for Deep Face Recognition. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 499–515.
47. Liu, L.; Jiang, H.; He, P.; Chen, W.; Liu, X.; Gao, J.; Han, J. On the variance of the adaptive learning rate and beyond. *arXiv* **2019**, arXiv:1908.03265.
48. Liu, S.; Tian, G.; Xu, Y. A novel scene classification model combining ResNet based transfer learning and data augmentation with a filter. *Neurocomputing* **2019**, *338*, 191–206. [CrossRef]
49. McNeely-White, D.; Beveridge, J.R.; Draper, B.A. Inception and ResNet features are (almost) equivalent. *Cogn. Syst. Res.* **2020**, *59*, 312–318. [CrossRef]
50. Sapiński, T.; Kamińska, D.; Pelikant, A.; Anbarjafari, G. Emotion recognition from skeletal movements. *Entropy* **2019**, *21*, 646. [CrossRef] [PubMed]