Single-Image Facial Expression Recognition Using Deep 3D Re-Centralization

Zhipeng Bao  
Tsinghua University, China  
bzp15@mails.tsinghua.edu.cn

Shaodi You  
Data61-CSIRO, Australia  
shaodi.you@data61.csiro.au

Lin Gu  
National Institute of Infomatics, Japan  
ling@nii.ac.jp

Zhenglu Yang  
Nankai University, China  
yangzl@nankai.edu.cn

Abstract

Facial expression recognition (FER) aims to encode expression information from faces. Previous studies often hold the assumption that human subjects should properly face the camera. Such a laboratory-controlled condition, however, is too rigid for in-wide applications. To tackle this issue, we propose a single image facial expression recognition method that is robust to face orientation and light conditions. We achieved this by proposing a novel face re-centralization method by reconstructing a 3D face model from a single image. We then propose a novel end-to-end deep neural network that utilizes both re-centralized 3D model and landmarks for FER task. A comprehensive evaluation on three real-world datasets illustrates that the proposed model outperforms the state-of-the-art techniques in both large-scale and small-scale datasets. The superiority of our model on effectiveness and robustness is also demonstrated in both laboratory conditions and wild images.

1. Introduction

Facial expression recognition (FER) is an important computer vision task. According to the subjects of the problem, the tasks can be divided into two categories: image sequence-based and single image-based problems [8, 20]. The single image-based approaches provide the basic model for facial expression recognition problems, and they can be extended to sequence-based approaches with a few modifications.

Most of the existing works are based on the assumption that the given faces should be in the right orientation and lighting conditions (i.e., right front of the camera and in normal lighting conditions). These methods work well for standard datasets in laboratory conditions like CK+ [15], JAFFE [16] and OULU-CASIA [29].

However, the facial images collected in real practice are usually in various orientations and lighting conditions. Different orientations can affect the extraction of key facial features; likewise, the shadow caused by different lighting conditions can also cause some confusions. Owing to these deviations, the previous models, especially the neural network-based ones, have resulted in inaccurate findings. Figure 1 shows two sample images under challenging conditions and the classification results of some popular methods.

Figure 1. Facial images in challenging conditions and the classification results: (a) an image with side face and (b) an image with dark shadow.
struct the 3D face from a single image under various orientation and lighting condition. After that, we re-align the face such that it is rightly facing the camera and then generate the shading. In such a case, the facial expression would be much easier to infer as the shading could efficiently depict the geometric features.

As shown in Figure 6, we have proposed a full pipeline for single-image FER tasks which integrates our 3D re-centralization sub-network and 2D facial landmarks sub-network into a multi-modal deep neural network. The facial landmarks characterize the key points in the face, and the 3D sub-network takes advantage of the best perspective and the geometry of the image for FER. Our experiments on three widely used datasets show the superior performance of the proposed approach among state-of-the-art in terms of both accuracy and robustness.

In this paper, our contributions are three-fold:

- We propose a novel 3D facial reconstruction method to centralize the single still face image. This process can significantly reduce the influence of orientations and shadows for a wide range of FER tasks.

- We propose a novel end-to-end neural networks for single image-based FER. This method incorporates the re-aligned 3D facial geometry and landmark features to achieve robust expression detection.

- The experimental results on three widely used datasets demonstrate that our model outperforms the existing state-of-the-art approaches in terms of accuracy and robustness.

2. Related Work

2.1. DNN for FER Problems

In recent years, well-known deep neural network (NN) algorithms such as convolution neural networks (CNNs), recurrent neural networks (RNNs) and the long short-term memory (LSTM) models have gained popularity in FER tasks [25, 10, 21]. With the development of deep learning, complicated deep neural networks (DNNs) have further improved the performance, especially for large datasets. [8] first applies DNNs in combination with landmark features. Mollahosseini et al. use three inception structures in convolution for FER to extract the deeper features of the images [17]. Zhao et al. propose a Peak-Piloted network, which is proven to be robust and effective for sequence-based FER problems [30]. The work of [11] considers the context information and their method works well for processing wild facial images. Zhu et al. improve the traditional DNN model by combining RNN and CNN to carry out the classification [32]. [24] supposes the facial expression can be seen as the expressive component and the neutral component, and they designed a de-expression residue learning method to separate these two components. Zhang et al. consider the influence of poses and perspectives and propose an adversarial network for FER [27], [19] provides a novel method to compare the similarity among different facial expressions. Jia et al. [7] apply local low-rank label correlations in their networks.

2.2. Facial Landmarks

Facial landmarks are the key points in a human face. These points include the contour of the face and the positions of the eyes, mouth, nose, and eyebrows. Numerous studies have been conducted on extracting facial landmark points, and several robust methods have been proposed [22, 9]. Many approaches for FER tasks based on landmark features have also gained great achievements.

For example, [3] builds a system only based on landmark points. They use both geometric and shading features generated from landmark points to conduct the classification. This approach has achieved high accuracy for specific datasets. However, it is not robust to classify the expressions merely based on landmark points because of the limited information they can provide. These points are generally taken as a secondary channel to provide auxiliary information for FER tasks in other situations [8, 23].

2.3. 3D Facial Features for FER Systems

Although 3D features have been applied to FER systems for many years, most of the previous works use either 3D landmarks or 3D coordinates to carry out the classification [2]. [27] also takes advantage of the 3D model, but they use the 3D model to generate multi-poses and multi-perspectives facial projection to improve the FER.

Different from the previous approaches, the proposed method uses the 3D model to re-align the facial images, which can reduce the influence of orientation and lighting, thus improve the performance of expression classifiers.

3. Single Image 3D Facial Re-centralization

As mentioned above, in-wild facial images may be of various positions and perspective. They may also contain different degrees of facial shadows. Figure 3 lists several examples of real-world facial images.

Noted that the 3D geometry, which is highly related to the expressions, is invariant from shadows and orientations. We propose a pipeline to reconstruct and align the 3D face model from a single image through the following three steps: 3D model generation, face re-centralization and shading generation. The whole process is demonstrated in Figure 2.

3.1. 3D Model Generation

The foundation of the whole process is to build a reliable 3D model from one single image. We adopt the method
initially proposed by Feng et al. [4] in our study. This model applies a Position map Regression Network (PRN) with an encoder-decoder architecture to reconstruct the 3D face. This network incorporates the context information by tracking the facial landmarks. As shown in Figure 2, this step delivers a position-color point cloud from a single image. We adopt a pre-trained model on 300-WLP dataset [31] in this paper.

3.2. Face Re-centralization

We re-centralize the 3D face by tracking the 3D landmarks. Assume \( \mathbf{P} \) records the positions of landmark points in the generated face and \( \mathbf{P}^* \) denotes the locations of referred landmarks of centralized face. Then the problem is to find a Matrix \( \mathbf{R} \) so that:

\[
\mathbf{R} = \arg \min_{\mathbf{R}} \| \Omega \mathbf{P} - \mathbf{P}^* \|.
\]

Since the problem is also essentially a 3D rotation problem and \( \mathbf{R} \) is the rotation matrix, \( \mathbf{R} \) would be an orthogonal matrix. Then this problem becomes an orthogonal Procrustes problem [5] and could be solved by singular value decomposition (SVD) of \( \mathbf{P} \cdot \mathbf{P}^T \).

The solution of this equation requires at least six pairs of corresponding points. In our method, sixty-eight pairs of points extracted by [22, 9] are used.

A potential concern for the re-centralization process is the extravagant distortions when images are of extreme orientations. However, in practice, these images are also challenging for state-of-the-art FER methods. Compared with the original 2D image, the re-aligned face model can provide more reliable information for expression recognition even though it is not completely precise. We show some examples in Figure 4.

Above model-based 3D reconstruction and re-centralization provides us the occluded face part. However, we do not use it in the FER task, because it still lacks reliable information through hallucination. Specifically, before the 3D rotation, we check the 3D vertices first. For the points in the point cloud, if several points share the same \( x, y \) coordinates, we only keep the outer-most two points and mask the others. Then the generated parts, which are invisible from the original image, will be discarded.

3.3. Shading Generation

The re-centralized 3D model is represented by a 3D point cloud. A straightforward approach is to build a 3D cube based on the point cloud. However, introducing the 3D cube requires considerable memory and calculation time, which poses a heavy burden for real-time expression recognition.

Therefore, we generate 2D shading image to represent 3D geometry to support our FER network. Based on Lambertian reflection [1], the shading image of each point on the given lighting can be modeled as:

\[
I = \mathbf{N} \cdot \mathbf{L},
\]

where \( I \) is the shading image density, \( \mathbf{N} \) is the surface normal, and \( \mathbf{L} \) represents incoming light direction. To reduce the influence of lighting, we relight the face with a standard \( \mathbf{L} \).

Now we estimate the normals \( \mathbf{N} \) of the given 3D face. For each point in the centralized point, we select \( n \) nearest
Figure 4. 3D face re-centralization better aligns the face and preserves the facial expression: images in the first two rows are selected from existing database, those in the last two rows are collected by us. The occluded part are also visualized for illustration but will not be used in the estimation.

points of it. Then we fit a plane with the minimum error based on them, the normal of that plane is considered as the normal of the selected point. In our experiments, we set \( n \) as 10.

Once we obtained the normals, we relight the face with \( L = [1, 0, 0] \), which is right facing the front face. Then we can generate a shading \((N \cdot L)\) map [18] that effectively reflects geometric facial features as shown in Fig 5.

4. Facial Expression Recognition

Based on the proposed face re-centralization method, we propose a novel end-to-end deep model for single image facial expression recognition. As shown in Figure 6, our model integrates the re-aligned 3D facial geometry and facial landmarks to recognize facial expression from a single image.

4.1. 3D Re-centralization Sub-network

Given the generated shading image of the size \( 100 \times 100 \), we extract the features with 3 convolution layers of the filter numbers as 256, 128 and 64. The kernel size of each convolution layer is 3, and the stride is 2. Then two fully-connected layers are applied to encode the 3D geometric feature vector into \( 1 \times 128 \).

4.2. 2D Landmark Sub-network

Since facial landmarks convey much information on expression [3], we incorporate 2D Landmark information for expression recognition.

We determine the face contour by [22] and locate the facial key points by [9]. Following [3], we then combine them to generate Sixty-eight landmark points recorded in total as follows:

\[
LP = [x_1, y_1; x_2, y_2; \cdots; x_{68}, y_{68}]. \tag{3}
\]

With the extracted points, we normalize their coordinates by max-min normalization with the position of the nose as the origin point. We then measure the distances between each pair of points as

\[
D = [d_{1,2}, d_{1,3}, \cdots, d_{67,68}]. \tag{4}
\]

In the formula, \( d_{i,j} \) means the normalized distance between the \( i^{th} \) and the \( j^{th} \) landmark points.

After that, we encode the landmark features via four fully connected layers whose hidden dimensions are 1024, 512, 256 and 128. The final dimension of this path is 128.

4.3. Network Structure

Given an input image, we at first detect the human face with Haar cascades [13] to crop it and resize into \( 100 \times 100 \). Then we encode the details with ResNet [6] structure. We
apply a separate convolution layer followed by one identity block and one convolution block to extract the global features of the detected 2D face. There are 7 convolution layers in total, and the number of filters for each layer are 128, 64, 64, 128, 64, 64, and 256. This is followed by a max-pooling layer with the pooling size 4. We further encode the 2D global feature by adding three fully connected layers whose hidden layers are 1024, 1024 and 128. The shape of the bottleneck output feature is also 128.

Finally, we fuse the above ResNet bottleneck feature with 3D geometry and landmark features by concatenating them together into a [384, 1] vector. The concatenated features are passed into two fully connected layers. The hidden layers of these two layers are 512 and 128. Finally, we classify the expression with a softmax layer.

5. Experiments

5.1. Implementation Details

In our experiments, the batch size is 128. The optimizer is Adam with learning rate initialized to 0.01. The dropout rate is 0.4 and we apply an early stopping method with training patience as 10. To attain a more robust model, we introduce Gaussian noise in the training process. The variance of Gaussian noise is set to 0.5. We use categorical cross entropy as loss function in our experiments if not specially declared. Besides, all the subnets have not been pre-trained before fusion.

5.2. Experiment Setting

Datasets We evaluate our experiments on RAF [12], OULU-CASIA [29] and CK+ dataset [15].

RAF dataset [12] contains many challenging cases, particularly including large face orientation and shadow. Notice that we did not clean up for RAF dataset that we remove invalid label (images containing more than 2 major faces) and tiny faces (the facial region is smaller than 50 pixels). For clarity, we name the clean dataset as RAF-subset.

OULU-CASIA dataset [29] contains image sequences in laboratory conditions, the lighting of which is standard but not perfect. For each image sequence, we pick the first, third and fifth images in reverse order in our experiments. The reason is that these images contain the peaks of the expressions and also differ from one another. CK+
<table>
<thead>
<tr>
<th>Model</th>
<th>SU</th>
<th>FE</th>
<th>DI</th>
<th>HA</th>
<th>SA</th>
<th>AN</th>
<th>NE</th>
<th>Acc</th>
<th>C-Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Li-2017</td>
<td>60.51</td>
<td>44.12</td>
<td>42.27</td>
<td>91.02</td>
<td>67.22</td>
<td>63.64</td>
<td>87.56</td>
<td>78.85</td>
<td>65.19</td>
</tr>
<tr>
<td>Jung-2015</td>
<td>73.85</td>
<td>41.76</td>
<td>37.11</td>
<td>92.65</td>
<td>72.24</td>
<td>69.70</td>
<td>80.73</td>
<td>80.14</td>
<td>66.78</td>
</tr>
<tr>
<td>Zhang-2017</td>
<td>69.74</td>
<td>38.24</td>
<td>31.96</td>
<td>91.78</td>
<td>72.91</td>
<td>57.58</td>
<td>72.93</td>
<td>76.92</td>
<td>64.01</td>
</tr>
<tr>
<td>Fabian-2016</td>
<td>56.41</td>
<td>29.41</td>
<td>24.74</td>
<td>90.52</td>
<td>47.83</td>
<td>49.5</td>
<td>80.49</td>
<td>71.83</td>
<td>54.13</td>
</tr>
<tr>
<td>INC-2016</td>
<td>62.05</td>
<td>26.47</td>
<td>37.11</td>
<td>86.85</td>
<td>66.22</td>
<td>55.56</td>
<td>55.37</td>
<td>69.28</td>
<td>55.66</td>
</tr>
<tr>
<td>Zeng-2018</td>
<td>67.69</td>
<td>47.06</td>
<td>42.27</td>
<td>92.41</td>
<td>68.90</td>
<td>61.62</td>
<td>83.41</td>
<td>78.85</td>
<td>65.19</td>
</tr>
<tr>
<td>Our Model</td>
<td>73.33</td>
<td>64.71</td>
<td>57.73</td>
<td>92.92</td>
<td>72.58</td>
<td>75.76</td>
<td>81.95</td>
<td>71.83</td>
<td>54.13</td>
</tr>
<tr>
<td>Our Model(FL)</td>
<td>75.38</td>
<td>41.18</td>
<td>38.14</td>
<td>94.31</td>
<td>71.91</td>
<td>70.71</td>
<td>83.17</td>
<td>81.95</td>
<td>74.14</td>
</tr>
</tbody>
</table>

Table 2. Results on RAF-subset dataset: “SU”, “FE”, “DI”, “HA”, “SA”, “AN”, “NE” stand for the accuracy for “Surprise”, “Fear”, “Disgust”, “Happiness”, “Sadness”, “Anger” and “Neutral” expressions in the dataset. “Acc” is the accuracy for this dataset and “C-Acc” is the class accuracy. “Our Model(FL)” means our model with focal loss. All the values in the table are percentages (%).

Dataset [15] is relatively small, and it is used to test the robustness of our model in the small dataset. Following previous works [8, 25], we apply the 10-fold cross validation method on the last two datasets.

Table 1 lists the details of all the three datasets.


During our experiments, the landmark points used the compared approaches are extracted by the same method as ours. We retrain all the compared models based on the original articles or the released codes. A single image is treated as a sequence with only one frame for the sequence-based models (Jung-2015 [8] and Zhang-2017 [28]). Besides, we do not apply data augmentation or cross-database learning for all the models.

Evaluation metric: We evaluate the performance by the classification accuracy on individual dataset that describes the overall performance of the compared models. We denote it as Accuracy and reported in percentages (%). We also report Class Accuracy on average accuracy for all the classes as RAF-subset and CK+ suffer class imbalance.

5.3. Results

5.3.1 RAF-subset dataset

RAF-subset contains many cases with large face orientation or strong shadow. The representative results are shown in Figure 7. As shown in Table 2, our model reaches the best performance in terms of both Accuracy and Class Accuracy.

We also provide the confusion matrix of our model in the small dataset. Following previous works, we apply the 10-fold cross validation method on the last two datasets.
most likely to be misclassified as "surprise".

**Focal loss** Since the number of each category in the dataset is imbalanced (Table 1), there is a gap between Accuracy and Class Accuracy. To relieve the imbalanced class, we introduce focal loss [14]. The normal categorical cross entropy loss has the following form:

$$CE(p_t) = -\log(p_t),$$

where $p_t$ is a function of $y$ and $p$, $y \in \{\pm 1\}$ indicates the ground-truth class, and $p \in [0, 1]$ is the model's estimated probability for certain class $y = 1$:

$$p_t = \begin{cases} 
p & \text{if } y = 1, \\
1 - p & \text{otherwise}.
\end{cases}$$

Then the focal loss for categorical cross entropy is:

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t).$$

In the formula, $\alpha_t$ is the balanced coefficient, which has a similar definition as $p_t$:

$$\alpha_t = \begin{cases} 
\alpha & \text{if } y = 1, \\
1 - \alpha & \text{otherwise}.
\end{cases}$$

and $\gamma$ controls the rate of weight conduction. In our experiment, $\alpha$ is set as 0.25 and $\gamma$ is set as 2.

When trained with the above focal loss, the performance, particularly the class accuracy has been significantly improved as reported in Table 2.

### Table 3. Results on CK+ dataset. “CO” stands for “Contempt”, “Acc” is the accuracy and “C-Acc” is the class accuracy. All the values in the table are percentages (%).

<table>
<thead>
<tr>
<th>Model</th>
<th>SU</th>
<th>FE</th>
<th>DI</th>
<th>HA</th>
<th>SA</th>
<th>AN</th>
<th>CO</th>
<th>Acc</th>
<th>C-Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>LI-2017</td>
<td>100.00</td>
<td>56.00</td>
<td>93.22</td>
<td>97.10</td>
<td>57.14</td>
<td>82.22</td>
<td>55.56</td>
<td>86.24</td>
<td>77.32</td>
</tr>
<tr>
<td>Jung-2015</td>
<td>100.00</td>
<td>20.00</td>
<td>89.83</td>
<td>98.55</td>
<td>75.00</td>
<td>80.00</td>
<td>50.00</td>
<td>89.60</td>
<td>83.63</td>
</tr>
<tr>
<td>Zhang-2017</td>
<td>97.59</td>
<td>20.00</td>
<td>96.61</td>
<td>100.00</td>
<td>60.71</td>
<td>91.11</td>
<td>44.44</td>
<td>90.52</td>
<td>73.53</td>
</tr>
<tr>
<td>Fabian-2016</td>
<td>100.00</td>
<td>84.00</td>
<td>96.61</td>
<td>98.55</td>
<td>75.00</td>
<td>80.00</td>
<td>50.00</td>
<td>89.60</td>
<td>83.63</td>
</tr>
<tr>
<td>INC-2016</td>
<td>95.00</td>
<td>48.00</td>
<td>88.14</td>
<td>97.10</td>
<td>60.71</td>
<td>75.56</td>
<td>50.00</td>
<td>82.57</td>
<td>73.53</td>
</tr>
<tr>
<td>Zeng-2018</td>
<td>95.18</td>
<td>48.00</td>
<td>89.83</td>
<td>94.20</td>
<td>60.71</td>
<td>88.89</td>
<td>50.00</td>
<td>84.10</td>
<td>75.26</td>
</tr>
<tr>
<td>Our Model</td>
<td>100.00</td>
<td>84.00</td>
<td>94.92</td>
<td>100.00</td>
<td>85.71</td>
<td>97.78</td>
<td>83.33</td>
<td>95.41</td>
<td>92.25</td>
</tr>
</tbody>
</table>

### Table 4. Confusion matrix of our model on RAF-subset dataset. All the values are reported in percentage (%).

<table>
<thead>
<tr>
<th></th>
<th>SU</th>
<th>FE</th>
<th>DI</th>
<th>HA</th>
<th>SA</th>
<th>AN</th>
<th>NE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SU</td>
<td>75.4</td>
<td>0.0</td>
<td>2.1</td>
<td>4.6</td>
<td>2.6</td>
<td>3.6</td>
<td>11.8</td>
</tr>
<tr>
<td>FE</td>
<td>23.5</td>
<td>41.2</td>
<td>0.0</td>
<td>0.0</td>
<td>14.7</td>
<td>11.8</td>
<td>8.8</td>
</tr>
<tr>
<td>DI</td>
<td>1.0</td>
<td>1.0</td>
<td>37.1</td>
<td>15.5</td>
<td>9.3</td>
<td>13.4</td>
<td>22.7</td>
</tr>
<tr>
<td>HA</td>
<td>0.5</td>
<td>0.0</td>
<td>0.6</td>
<td>94.3</td>
<td>1.0</td>
<td>0.3</td>
<td>3.3</td>
</tr>
<tr>
<td>SA</td>
<td>1.0</td>
<td>0.3</td>
<td>5.7</td>
<td>6.4</td>
<td>71.9</td>
<td>2.0</td>
<td>12.7</td>
</tr>
<tr>
<td>AN</td>
<td>5.1</td>
<td>0.0</td>
<td>10.1</td>
<td>7.1</td>
<td>1.0</td>
<td>70.7</td>
<td>6.1</td>
</tr>
<tr>
<td>NE</td>
<td>1.7</td>
<td>0.2</td>
<td>3.9</td>
<td>4.9</td>
<td>5.7</td>
<td>12.7</td>
<td>83.2</td>
</tr>
</tbody>
</table>

### Table 5. Accuracy on OULU-CASIA dataset. All the values in the table are percentages (%).

<table>
<thead>
<tr>
<th>Model</th>
<th>SU</th>
<th>FE</th>
<th>DI</th>
<th>HA</th>
<th>SA</th>
<th>AN</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>LI-2017</td>
<td>49.58</td>
<td>95.00</td>
<td>72.92</td>
<td>85.83</td>
<td>72.08</td>
<td>66.25</td>
<td>73.61</td>
</tr>
<tr>
<td>Jung-2015</td>
<td>45.83</td>
<td>76.67</td>
<td>61.25</td>
<td>97.50</td>
<td>62.92</td>
<td>56.67</td>
<td>66.81</td>
</tr>
<tr>
<td>Zhang-2017</td>
<td>87.08</td>
<td>93.33</td>
<td>61.67</td>
<td>92.08</td>
<td>68.33</td>
<td>90.83</td>
<td>82.22</td>
</tr>
<tr>
<td>Fabian-2016</td>
<td>68.75</td>
<td>92.92</td>
<td>65.00</td>
<td>91.67</td>
<td>72.50</td>
<td>79.17</td>
<td>78.33</td>
</tr>
<tr>
<td>INC-2016</td>
<td>50.00</td>
<td>80.83</td>
<td>61.25</td>
<td>52.50</td>
<td>62.08</td>
<td>70.00</td>
<td>62.78</td>
</tr>
<tr>
<td>Zeng-2018</td>
<td>66.25</td>
<td>61.67</td>
<td>82.92</td>
<td>55.83</td>
<td>62.92</td>
<td>52.08</td>
<td>61.02</td>
</tr>
<tr>
<td>Our Model</td>
<td>87.50</td>
<td>98.33</td>
<td>72.08</td>
<td>89.17</td>
<td>83.33</td>
<td>78.75</td>
<td>84.86</td>
</tr>
</tbody>
</table>

### 5.3.2 OULU-CASIA Dataset

OULU-CASIA dataset is collected in laboratory conditions with face properly facing the camera, but the lighting condition is not satisfactory which results in facial shadows.

The experimental result on OULU-CASIA dataset is shown in Table 5. As OULU-CASIA dataset is a balanced dataset, C-Acc has a same value as Acc and we only report Acc in Table 5.

### 5.3.3 CK+ dataset

CK+ dataset is a small dataset in proper laboratory situations. We also report the performance in Table 3 and demonstrates the superior performance of the proposed model.

### 5.4. Ablation Study

To analyze the effect of the individual component in our proposed method, we conduct an ablation study. Since the features from 3D re-centralization sub-net and landmark sub-net are fused into the main network, we could choose to connect the main network to specific sub-net or not. In this section, we name **3D** for the 3D re-centralization sub-net and **Landmark** for landmark sub-net. When neither **3D** or **Landmark** is fused, we call the rest as **Base**. Thus, we also evaluate these four types of networks and reported their performance in Table 6.

From the result, we can see that both re-centralized 3D geometry features and landmark features could help to improve the FER performances. Combing two features together could further boost the performance significantly.
### Table 6. Ablation study of individual component on three datasets.

```
<table>
<thead>
<tr>
<th>Model</th>
<th>RAF-s</th>
<th>OULU</th>
<th>CK+</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc</td>
<td>Acc</td>
<td>C-Acc</td>
</tr>
<tr>
<td>Base</td>
<td>78.69</td>
<td>78.69</td>
<td>78.71</td>
</tr>
<tr>
<td>Base+Landmark</td>
<td>79.62</td>
<td>79.62</td>
<td>79.62</td>
</tr>
<tr>
<td>Base+3D</td>
<td>79.73</td>
<td>79.73</td>
<td>79.73</td>
</tr>
<tr>
<td>Whole Model</td>
<td>81.60</td>
<td>81.60</td>
<td>81.60</td>
</tr>
<tr>
<td></td>
<td>64.96</td>
<td>68.37</td>
<td>67.80</td>
</tr>
<tr>
<td></td>
<td>77.36</td>
<td>79.62</td>
<td>84.86</td>
</tr>
<tr>
<td></td>
<td>85.93</td>
<td>93.88</td>
<td>95.41</td>
</tr>
</tbody>
</table>
```

Table 6: Ablation study of individual component on three datasets. “Acc” stands for accuracy and “C-Acc” stands for class accuracy. All the values in the table are percentages (%).

The only exception is for RAF-subset dataset where the class accuracy falls when including 3D geometric features. We believe it is due to the imbalanced class distribution as discussed in RAF-subset dataset. The ablation analysis validates the effect of our proposed 3D re-centralization and landmark sub-networks.

### 5.5. Time Consumption

We test and compare the whole executive time (including pre-processing time) with 2 Titan X GPU among all the models. The per-frame process time is reported in Table 7. As can be seen, the proposed method only takes slightly more time than existing methods but achieves much more accurate result.

<table>
<thead>
<tr>
<th>Model</th>
<th>Time(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Model</td>
<td>0.123</td>
</tr>
<tr>
<td>Li-2017</td>
<td>0.084</td>
</tr>
<tr>
<td>Jung-2015</td>
<td>0.086</td>
</tr>
<tr>
<td>Zhang-2019</td>
<td>0.090</td>
</tr>
<tr>
<td>Fabian-2016</td>
<td>0.089</td>
</tr>
<tr>
<td>INC-2016</td>
<td>0.097</td>
</tr>
</tbody>
</table>

Table 7: Time consumption analysis.

### 5.6. Failure Cases

Figure 8 lists three representative failure cases. These images contain a huge part of occlusions in the front face, which will cause serious distortion of 3D reconstruction and further influence the shading generation. However, as shown in Figure 8, these images would also confuse alternative existing methods.

### 6. Conclusions and Future Work

This study proposes a novel system for single image Facial Expression Recognition (FER). To reduce the influence of orientations and shadows, we propose a novel approach to reconstruct and re-centralize a 3D facial model from a single image. Re-aligned 3D facial geometry and landmarks are then integrated into the proposed network for the robust FER. The experiments on three datasets demonstrate that the proposed model obtains state-of-the-art performance.

For future work, we consider optimizing our model in two directions. We will extend our model in image sequence-based emotion classification, which is more practical in the real world. We will also commit to further enhancing the robustness of 3D face alignment for extreme orientation cases.

### References


