Multimodal 2D and 3D for In-the-wild Facial Expression Recognition

Son Thai Ly, Nhu-Tai Do, Guee-Sang Lee, Soo-Hyung Kim, Hyung-Jeong Yang
Department of Electronics and Computer Engineering
Chonnam National University, Korea
{thaisonly123, donhutai}@gmail.com, {gslee, shkim, hjyang}@jnu.ac.kr

Abstract

In this paper, unlike other in-the-wild facial expression recognition (FER) studies which only focused on 2D information, we present a fusion approach for 2D and 3D facial data in FER. In particular, the 3D facial data are first reconstructed from image datasets. The 3D information are then extracted by deep learning technique that could exploit the meaningful facial geometry details for expression. We further demonstrate the potential of using 3D facial data by taking the 2D projected images of 3D face as an additional input for FER. These features are fused with that of 2D features from a typical network. Following the experiment procedure in recent studies, the concatenated features are classified by linear support vector machines (SVMs). Comprehensive experiments are further conducted on integrating facial features for expression prediction. The results show that the proposed method achieves state-of-the-art recognition performances on both RAF database and SFEW 2.0 database. This is the first time such a deep learning combination of 3D and 2D facial modalities is presented in the context of in-the-wild FER.

1. Introduction

Among communication channels, facial expression is one of the most effective modality to convey human emotions. Many studies have been conducted to address the challenges in in-the-wild facial expression recognition (FER) such as occlusion, large head poses or illumination variations. Apart from what computing approach could be chosen, most of the existing and analysis research primarily rely on static or/and dynamic sequences data from various facial expression databases. In the recent years, many static and videos sequence databases have been sourced from the Internet to form in-the-wild facial expression databases [11, 14, 22, 26]. Being associated with released databases, there are public challenges or competitions [10, 11, 14, 32] that draw attention and make use of the community resource to tackle the problems. However, these problems persist as addressed in [21, 32].

Unlike its 2D counterpart, most of available 3D facial expression databases [7, 28, 37, 38] were established by capturing human actors’ faces using special devices such as camera system and Kinect RGB-D in lab-controlled environment. Although these essential setups usually offer high quality 3D face surface geometry and surface texture, these databases are restricted in terms of the diversity of the participants in regards to gender and ethnic-racial ancestries and the consistencies between in-the-wild and constrained environment. In other words, such a lab-constructed dataset will not only tend to have the similar 3D facial surface features of the same actors throughout the dataset but also the common pre-designed expressions behaviors (e.g., smile, laughter, or cries) are often overlapped by the same actors.

In approaching aspect, 2D FER studies are well established with both hand-crafted and deep learned features, whose comprehensive survey on 2D FER is reported in [21]. 3D FER using deep learning algorithm is still an untouched field. Conventional methodologies have been widely employed for 3D FER such as depth-SIFT [4], normal-LBP [18], and curvature-HOG [17]. Deep learning approaches are left behind and only a few attempts to learn the 3D facial expression representation as referred in recent studies [8, 20]. One of the major reasons for this downside could be because available the 3D facial expression databases contain only a small number of data samples. For example, the well-known BU-3DFE database [38] and Bosphorus 3D Face database [28] contain only 2,500 and 4,666 data samples, respectively. These numbers are far from enough for deep-learning-based approach. Consequently, these aforementioned constraints hinder achieving higher performance on in-the-wild FER with 3D facial data.

This study presents a potential method which aimed to tackle the above-described by combining 2D and 3D information for in-the-wild FER. In particular, to exploit the 3D facial information from in-the-wild image datasets, facilitated by recent advances in 3D face reconstruction [31], the 3D facial expression was constructed from available 2D facial expression datasets. The 3D facial data were then
treated as point cloud on which the 3D geometry information is obtained using the $\chi$-Convolution ($\chi$-Conv) operation [23]. Moreover, in order to demonstrate the benefits of 3D facial data, the 3D face frontalized and its 2D frontal projected image used as additional input to recognize the expression. These features were then fused with that of a 2D FER typical network. The overview of proposed approach is shown in Fig. 1. Our contributions are in three-fold as follow:

- For the first time, a novel and competent deep learning approach for fusion 2D and 3D modalities in in-the-wild FER is proposed in Sec. 3.

- Comprehensive ablation experiments are further conducted to show how to optimally learn the 2D and 3D facial representation in a deep learning manner on RAF [22] and SFEW 2.0 [11] dataset in Sec. 4.

- Despite of using standard approach, the proposed approach achieves state-of-the-art recognition accuracy on RAF [22] and SFEW 2.0 [11] dataset compared with recent studies in Sec. 4.2.

2. Related works

2.1. Facial expression recognition from 2D image data

In the recent years, most of the advances in FER has been studied based on static/dynamic 2D databases that contain images or sequences of images and the term in-the-wild facial expression recognition has been mainly used to refer to FER on 2D image data. The reason is that real-life large-scale facial expression databases are constructed by sourcing 2D images from the Internet [10, 11, 14, 26] and thus more suitable for deep learning approaches which become well known in recent years. Therefore, numerous deep feature learning approaches have been employed to effectively improve the performance on image FER task. Using various standard as well as modified network architectures is the most popular approach [2, 12, 35]. These works often included pre-training on similar and larger datasets to capture the useful deep facial features. Model ensembling is another straightforward and proved effective approach that has been widely used in many challenges. For example, in the recent EmotiW2018 challenge where the original dataset was relatively small, high ranking teams [13, 24, 33] carried out fine-tuning of their networks on FER-2013 [14], RAF [22], and AffectNet [26] and fused the predicted score of multiple networks to attain the final score. Aside from above common strategies of learning the deep feature via training, there are research specifically focused on investigating the deep feature systematically. While in [1], the authors utilized the manifold networks along with covariance pooling to capture the second-order statistics for feature extraction in a deep learning fashion, in [39], the authors proposed LT-Net to learn the truth label from noisy datasets, thus, could employ multiple inconsistently labeled and large scale unlabeled datasets for training procedure.

2.2. Facial expression recognition from 3D data

Since the most popular BU-3DFE databases [38] was presented, many studies on 3D FER were proposed to leverage the usefulness of 3D information. Conventional methodologies were widely employed for 3D FER such as depth-SIFT [4], normal-LBP [18], curvature-HOG [17]. Still, there are limited studies implementing the 3D data with deep learning methods. Although few proposals claimed to be implementing the 3D data with deep learning methods, they were conducted based on the projected image [20] or the 2D depth map of the 3D face data [15], or using the 3D features extracted from traditional methods. Reason being that the available 3D facial expression databases contain only a small number of data samples but lack sufficient deep learning algorithms for learning 3D data information. Therefore, compared to its counterpart, 3D FER’s achievements are relatively insignificant, which is evident by the number of related studies, open-source-code repositories, and the attention of the community. While 2D FER could be well-established by both hand-crafted [9, 41, 42] and deep learned features [25, 27, 34, 36], 3D FER using deep learning algorithm is still an untouched field.

These studies seek to enhance the performance by analyzing the deep feature of in-the-wild facial images or learning 3D facial information by tradition techniques. On the
other hand, our proposed approach takes the advantage of 3D reconstruction into account via deep learning method in conjunction with existing 2D image FER method for in-the-wild FER.

3. Proposed method

3.1. Constructing 3D facial expression data

This study benefits from Tran et al. [31] study for reconstructing the 3D face from the original image dataset. The reason is that their study could reconstruct the mid-level features that are meaningful for expression recognition. They first modelled the foundation face shape in PCA form:

\[ s = \bar{s} + S_{id}\alpha_{id} + E_{exp}\eta_{exp}, \]

where \( \bar{s} \) is the mean 3D face shape, \( S_{id} \in \mathbb{R}^{3n \times s_p} \) is the orthonormal identity basis of \( s_p \) principal face shape components, \( s_p = 99 \), and the \( \alpha_{id} \in \mathbb{R}^{s_p} \) is the subject-specific shape weight. Similarly, the \( E_{exp} \in \mathbb{R}^{3n \times e_p} \) is the orthonormal expression basis of \( e_p \) principal expression components, \( e_p = 29 \), and the \( \eta_{exp} \in \mathbb{R}^{e_p} \) is the expression coefficient which estimated from input face image \( I \).

In the other hand, the bump map \( \Delta(p) \) of mid-level detail corresponds to pixel coordinate \( p \) in image \( I \) is computed as follow:

\[ \Delta(p) = \theta(z'(p) - z(p)), \]

The linear function \( \theta \) encodes the different in depth of the estimated depth \( z' \) and the depth of foundation shape \( z \) at pixel \( p \) to intensity range \([0, 255]\). Thus, given a bump map \( \Delta \) and foundation shape \( s \), the estimated depth \( z' \) could be simply calculated as:

\[ z'(p) = z(p) + \theta^{-1}(\Delta(p)), \]

The training and combination of foundation shape and mid-level features which heavily rely on 2D image facial landmarks detection are further discussed in [31].

According to Eqs. 2 and 3, the accurate facial landmarks are crucial for extracting the mid-level features that are meaningful for expression recognition. For a better result, the state-of-the-art landmarks detection OpenFace 2.0 Toolkit [3] was used for detecting facial landmarks on 2D image datasets. Note that, the reconstruction error sometimes occurs due to the faulty landmark detection or reconstruction. In that case, ExpNet [6], which could reconstruct 3D face without the need of landmarks, was applied for that specific sample data to obtain 3D face. However, the ExpNet does not deliver a detailed geometry 3D face as [31]. Examples of 3D face reconstruction are illustrated in Fig. 2.

![RAF examples of face reconstruction using [31], except for last sample which generated by [6].](image)

3.2. Facial expression models

**Image models.** We describe the learning procedure for 2D images which include the original benchmark datasets and the projected image of frontalized 3D face, denoted as 2D model and projected model, respectively, in Fig. 1. Following the procedure in [1], the Inception-ResnetV1 [30] was used to train the benchmark datasets from scratch as well as fine-tuned on a model pre-trained on VGGFace2 [5] and AffectNet dataset [26]. The output of trained embedding layer is treated as input for fitting Support Vector Machines (SVMs). Note that, the Inception-ResnetV1 and SVMs were trained separately.

**3D model.** In various studies, the 3D facial information was only exploited using hand-crafted methods, and yet never a deep learning one, the possible reasons is due to the lacking of available learning approaches. Taking advantage from state-of-the-art point cloud learning algorithm \( \chi\)-Conv in PointCNN [23], this study was capable of learning the 3D facial features. The \( \chi\)-Conv operation could be mathematically described as:

\[
\chi - Conv(K, p, P, F) = Conv(K, MLP((P - p)) \times [MLP_p((P - p), F)], \]

where \( K \) and \( F \) define the convolution kernels and feature map while \( P \) and \( p \) correspond to the point in local coordinate system and representative point in feature map. The local points are “lifted” to be representative points by the multilayer perceptron \( MLP \), it is then weighted and permuted by the \( K \times K \) \( \chi \)-transform matrix to subsequently transformed by the conventional convolution operation. These \( \chi\)-Conv layers are then stacked to create a deep network making PointCNN capable of learning the spatial-local correlation between points better without being affected by ordering. The 3D facial expression model is denoted as 3D model in Fig. 1.

**Feature extraction and fusion.** After training, the features were extracted from each model and concatenated as input for training SVMs. While features from 2D models are extracted from the last embedding layer, those of 3D models are the output of last \( \chi\)-Conv layer.
4. Experiments

4.1. Datasets and training

**Image data.** To compare the proposed approach for in-the-wild facial expression against previous studies, we evaluate the models on the RAF [22] dataset contains 12271 and 3068 images for training and validation, respectively. SFEW 2.0 [11], a static subset of videos of AFEW dataset [10], contains 958 images for training and 438 for validation. Both of them were labeled with seven discrete expressions (anger, disgust, fear, happiness, sadness, surprise, and neutral).

Before training, face detection and alignment were performed using Multi-task Cascade Convolutional Neural Networks (MTCNN) [40] on original image datasets. The 2D model was trained with Adam optimizer [16], batch size of 128 for 100 epochs. The training process also included standard image augmented techniques as random flipping, cropping and rotating.

**3D data.** After reconstructing the 3D face, the number of vertex of reconstructed 3D face spans from 145k to 170k and the result of reconstructing are frontalized and occlusion free. However, the 3D data need to be down-sampled and normalized for learning. Therefore, the preprocessing for 3D facial data was performed as follow: 1) First, the 3D reconstructed facial data were trimmed to remove the inessential parts, such as ears, remaining 20,000 points, empirically. 2) They were then uniformly down-sampled, for reserving the point distribution, to 4,096 points. 3) Finally, the 3D facial data were normalized so that the coordinates are all in the interval [-1, 1].

The 3D model was trained with learning rate 3e-2, decay every 8000 steps, batch size of 32 for 100 epochs with early stopping if, the validation loss has not decreased in last 5 epochs. The hyper-parameters were set as shown in Fig. 3. Each $\chi$-Conv layers is formed as $\chi$-Conv($K$, $D$, $P$, $C$), where $K$ is the neighborhood size, $D$ is the dilation rate, $P$ is the representative point number in the output, and $C$ is the output channel number. The DenseNet-like links between layers were also used to fight vanishing gradient problem along with drop out and rotation augmentation. The 3D model also suffers from the imbalanced data problem along with drop out and rotation augmentation.

For the experiments that are presented later on in this study, unless stated otherwise, the experiment results are the result of fusion of three models (2D, 3D and projected model) in which all images models were fine-tuned, the fps method was used in $\chi$-Conv operation, and joint features were classified by SVMs.

4.2. Results and discussion

For the experiments that are presented later on in this study, unless stated otherwise, the experiment results are the result of fusion of three models (2D, 3D and projected model) in which all images models were fine-tuned, the fps method was used in $\chi$-Conv operation, and joint features were classified by SVMs.

**Result analysis.** Table 1 shows the total accuracy of each models in the proposed method on RAF and SFEW 2.0 database. Compared to the projected model and 3D model, the 2D model has better performances. One possible reason lies in the different input of these model. While the 2D model is taken the original image as input, the projected model and 3D model are learned from projected images and
### Table 1. Result of proposed method on RAF and SFEW 2.0 datasets.

<table>
<thead>
<tr>
<th>Models</th>
<th>RAF</th>
<th>SFEW 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D model</td>
<td>85.1</td>
<td>55.3</td>
</tr>
<tr>
<td>3D model</td>
<td>65.2</td>
<td>39.7</td>
</tr>
<tr>
<td>Projected model</td>
<td>57.8</td>
<td>33.5</td>
</tr>
<tr>
<td>Fusion 2D and 3D model</td>
<td>86.7</td>
<td>56.4</td>
</tr>
<tr>
<td>Fusion three models</td>
<td>87.5</td>
<td>56.9</td>
</tr>
</tbody>
</table>

### Table 2. Comparison between sampling methods.

<table>
<thead>
<tr>
<th>Sampling methods</th>
<th>RAF</th>
<th>SFEW 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random sampling</td>
<td>62.5</td>
<td>38.2</td>
</tr>
<tr>
<td>Farthest point sampling</td>
<td>65.2</td>
<td>39.7</td>
</tr>
</tbody>
</table>

### Table 3. Comparison between different classifiers.

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>RAF</th>
<th>SFEW 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive bayes</td>
<td>86.6</td>
<td>56.3</td>
</tr>
<tr>
<td>Random forest</td>
<td>86.4</td>
<td>56.7</td>
</tr>
<tr>
<td>K-nearest neighbor</td>
<td>85.9</td>
<td>56.5</td>
</tr>
<tr>
<td>Softmax</td>
<td>86.8</td>
<td>56</td>
</tr>
<tr>
<td>Linear SVMs</td>
<td>87.5</td>
<td>56.9</td>
</tr>
</tbody>
</table>

### Table 4. Comparison between fusion strategies.

<table>
<thead>
<tr>
<th>Fusion strategies</th>
<th>RAF</th>
<th>SFEW 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score-level fusion</td>
<td>86.7</td>
<td>56.2</td>
</tr>
<tr>
<td>Feature-level fusion</td>
<td>87.5</td>
<td>56.9</td>
</tr>
</tbody>
</table>

### Table 5. Comparison between state-of-the-art studies on RAF and SFEW 2.0 datasets.

<table>
<thead>
<tr>
<th>Models</th>
<th>RAF</th>
<th>SFEW 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>LTNet [39]</td>
<td>86.7</td>
<td>58.2</td>
</tr>
<tr>
<td>Cov. Pooling [1]</td>
<td>87</td>
<td>58.1</td>
</tr>
<tr>
<td>Transfer learning [33]</td>
<td>80</td>
<td>55.8</td>
</tr>
<tr>
<td>DLP-CNN [22]</td>
<td>74.2</td>
<td>51</td>
</tr>
<tr>
<td>DSN [13]</td>
<td>84</td>
<td>-</td>
</tr>
<tr>
<td>Multimodal fusion [24]</td>
<td>83.8</td>
<td>-</td>
</tr>
<tr>
<td>[1]’s baseline</td>
<td>84.6</td>
<td>52.5</td>
</tr>
<tr>
<td>Fusion [1]’s baseline and our 3D model</td>
<td>85.8</td>
<td>53.7</td>
</tr>
<tr>
<td>Proposed method</td>
<td>87.5</td>
<td>56.9</td>
</tr>
</tbody>
</table>

3D geometry data, respectively, in which contain none facial texture (skin) information as illustrated in Fig. 2. However, the 3D model accuracies are still 7% and 5% higher than projected model on both RAF and SFEW 2.0 dataset, respectively. These results validate the benefit of using 3D geometry information on FER. Furthermore, the combination of all three models contributes to improve the total accuracies over those of 2D model on both RAF and SFEW 2.0. This proves the advantage of employing 3D facial data on increasing the facial expression recognition result.

**Comparison between sampling methods.** As mentioned in Sec. 4.1, in $\chi$-Conv operation, there are two ways to transform the local points to representative points, farthest point sampling (fps) and random down-sampling. While fps could uniformly reserve the point distribution and thus retain the facial mid-level details, the random method could not. In this experiment, we compare the effectiveness of these two down-sampling methods on in-the-wild FER. As had been predicted, the performance that benefits from fps method is higher than that of random method as shown in Table 2.

**Comparison between different classifiers.** Experiments on popular classifiers, such as softmax, linear SVMs, naive bayes, random forest, k-nearest neighbor were conducted for classifying the joint features. As reported in Table 3, all the classifiers produce comparable results with linear SVMs performing the best. Therefore, linear SVMs is generally the best classifier for classifying fused deep features.

**Comparison between fusion strategies.** Table 4 reports the results of two fusion strategies: feature-level and score-level fusion. We can see that the feature-level fusion achieved better results than score-level.

**Comparison with recent state-of-the-art studies.** Table 5 presents the performances comparison between the best of proposed approach and state-of-the-art studies on RAF [22] and SFEW 2.0 [11] databases. To keep a fair comparison, our training and testing procedure were conducted by following the procedures in [1, 22] which use deep network to extract features and classify features into expression labels by SVMs. Despite of using common fusion approach, the proposed fusion model achieved best recognition accuracy on RAF dataset, compared with the state-of-the-art reports [1, 13, 24, 33, 39] which use complex algorithms. In the case of SFEW 2.0, the proposed approach obtained a competent result as well. It can be reasoned that the SFEW 2.0 data has less than 1,500 data samples in total, which is not enough for deep learning methods. Nevertheless, it might not be clear in case of RAF’s performances, the proposed method outperform the transfer learning result in [33] which transferred from VGG-face model fine-tuned on FER-2013 [14] on SFEW 2.0 dataset.

**Advantages of using 3D facial data.** As shown in Table 1, on both datasets, the feature of 3D model and projected...
models improve the fusion model performances. In addition, the accuracies of 3D models are better than those of projected models. This, again, clearly confirms the benefit of using 3D over 2D information for in-the-wild FER in deep learning manner. It is also worth mentioning that proposed 3D models were all trained from scratch. Given the fact that this study is one of the very first report which utilized 3D facial expression for in-the-wild FER dataset using deep learning, there is none available model for fine-tuning. Moreover, constructing an entire new 3D face expression dataset for fine-tuning also is not the scope of this study which exploits the 3D geometry information in in-the-wild FER context. Therefore, the results in Table 1 are reasonable. In addition, to demonstrate that proposed approach could be employed in other works, the features of the baseline model in [1] were fused with those of proposed 3D model, the results were shown in Table 5. Although the missing of texture information channel has depressed the capability of 3D data, it also suggests a great potential of achieving better recognition accuracy with 3D data coupled with texture information. On the other hand, despite of small contribution to the overall performance (less than 1%), the using of projected model demonstrates the promising of using 3D face on FER as it could be utilized in many other ways. For instance, since the projected image is frontalized, it would be easier for estimating the 2D and 3D facial landmarks and apply it for FER.

**Drawbacks of using 3D facial data.** One major downside of using 3D information in FER is that, currently, in-the-wild 3D facial expression database is not available. Existing 3D databases are either contain a small number of data samples or sampled in constrained laboratory environments. Therefore, the public databases are neither suitable for deep learning techniques nor appropriate for in-the-wild FER, which was the original purpose of this study. Alternatively, the data preparation step required more efforts from 3D face reconstruction to error checking and preprocessing. That is not to mention the inconsistency between reconstruction 3D faces and labels from the original benchmarks. In fact, the data preparation is a time-consuming task, which alone took two-third of total experiment time.

**5. Conclusion and further works**

This study explores the benefits of 3D facial modeling for in-the-wild FER for the first time. Despite of using conventional deep learning methods, the competent results justified the benefit of using 3D information for FER. It is also suggested that the 3D facial expression features could be harvested in many approaches and contributed to improve facial expression recognition performance.

As indicated above, the limitation of in-the-wild 3D facial expression databases makes the data preparation phase more complex than of that for 2D image datasets. Therefore, we plan to construct an in-the-wild 3D facial expression database for the sake of academic purpose.

**Acknowledgement**

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A3B05049058 & NRF-2017R1A4A1015559). The corresponding author is Guee-Sang Lee.

**References**


[12] H. Ding, S. K. Zhou, and R. Chellappa. Facenet2expnet: Regularizing a deep face recognition net for expression...


