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

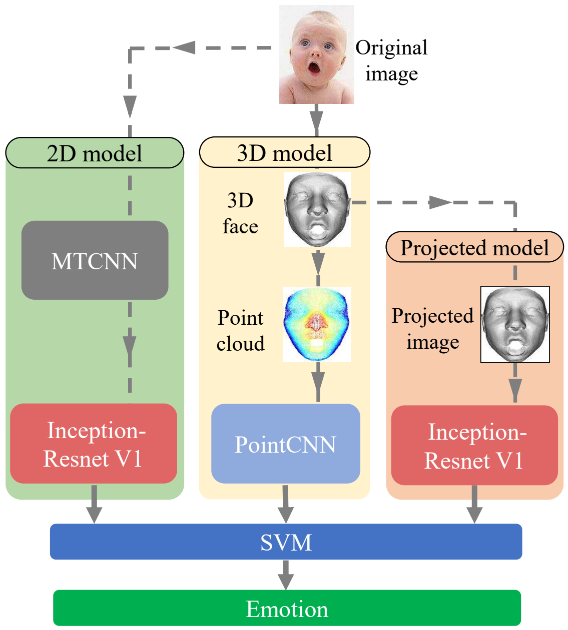

Figure 1. Pipeline of our approach. Each model was trained separately and the extracted features are then concatenated and learned by SVMs.

treated as point cloud on which the 3D geometry information is obtained using the $\mathcal{X}$-*Convolution* ($\mathcal{X}$-*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 LTNet 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:

$$\boldsymbol{s} = \bar{\boldsymbol{s}} + \boldsymbol{S}_{id}\boldsymbol{\alpha}_{id} + \boldsymbol{E}_{exp}\boldsymbol{\eta}_{exp}, \quad (1)$$

where $\bar{\boldsymbol{s}}$ is the mean 3D face shape, $\boldsymbol{S}_{id} \in \mathbb{R}^{3n \times s_p}$ is the orthonormal identity basis of s_p principal face shape components, $s_p = 99$, and the $\boldsymbol{\alpha}_{id} \in \mathbb{R}^{s_p}$ is the subject-specific shape weight. Similarly, the $\boldsymbol{E}_{exp} \in \mathbb{R}^{3n \times e_p}$ is the orthonormal expression basis of e_p principal expression components, $e_p = 29$, and the $\boldsymbol{\eta}_{exp} \in \mathbb{R}^{e_p}$ is the expression coefficient which estimated from input face image $\boldsymbol{I}$.

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

$$\boldsymbol{\Delta}(\boldsymbol{p}) = \theta(z'(\boldsymbol{p}) - z(\boldsymbol{p})), \quad (2)$$

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

$$z'(\boldsymbol{p}) = z(\boldsymbol{p}) + \theta^{-1}(\boldsymbol{\Delta}(\boldsymbol{p})), \quad (3)$$

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.

Figure 2. RAF examples of face reconstruction using [31], except for last sample which generated by [6].

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 χ-*Conv* in PointCNN [23], this study was capable of learning the 3D facial features. The χ-*Conv* operation could be mathematically described as:

$$\begin{aligned} &\chi - Conv(\boldsymbol{K}, p, \boldsymbol{P}, \boldsymbol{F}) = \\ &Conv(\boldsymbol{K}, MLP(\boldsymbol{P} - p) \times [MLP_\delta(\boldsymbol{P} - p), \boldsymbol{F}]), \end{aligned} \quad (4)$$

where $\boldsymbol{K}$ and $\boldsymbol{F}$ define the convolution kernels and feature map while $\boldsymbol{P}$ 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 $K \times K$ χ-transform matrix to subsequently transformed by the conventional convolution operation. These χ-*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 χ-*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 $\mathcal{X}$-*Conv* layers is formed as $\mathcal{X}$-*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 problems thus we performed up-sampling for under-represented classes by replicating its samples so that every class has the same number of sample. Furthermore, in $\mathcal{X}$-*Conv* operation, there are two ways to transform the local points to representative points, namely 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 study, the random down-sampling was used only in the comparison between sampling methods which described in next section. In addition, for all 3D model experiments, we conducted average model feature ensembling with last 10 checkpoints for experiments.

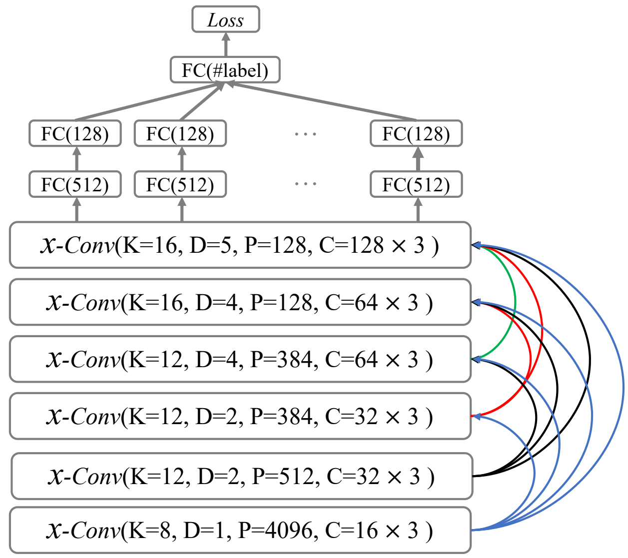

Figure 3. The set of hyper-parameters for each layer in 3D model with DenseNet-like links.

Note that, although the 3D reconstructed data also reserve the facial expression via principal expression components $\boldsymbol{E}_{exp}\boldsymbol{\eta}_{exp}$ as stated in Eqs. 1, the inconsistencies between expression labels from original image datasets and 3D facial data still exist. For instance, the reconstructed 3D face does not express the same expression as its corresponding original image. This would produce noises in training the 3D facial data. However, the objective of this study was neither presenting 3D data reconstruction algorithm specifically for facial expression nor constructing 3D facial expression database. Therefore, the experiments were carried out with the expression labels for training 3D facial data are automatically taken from the original 2D dataset without carefully inspecting the consistency between expression labels and 3D data.

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 $\mathcal{X}$-*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

Models	RAF	SFEW 2.0
2D model	85.1	55.3
3D model	65.2	39.7
Projected model	57.8	33.5
Fusion 2D and 3D model	86.7	56.4
Fusion three models	**87.5**	**56.9**

Table 1. Result of proposed method on RAF and SFEW 2.0 datasets.

Sampling methods	RAF	SFEW 2.0
Random sampling	62.5	38.2
Farthest point sampling	**65.2**	**39.7**

Table 2. Comparison between sampling methods.

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 $\mathcal{X}$*-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.

Classifiers	RAF	SFEW 2.0
Naive bayes	86.6	56.3
Random forest	86.4	56.7
K-nearest neighbor	85.9	56.5
Softmax	86.8	56
Linear SVMs	**87.5**	**56.9**

Table 3. Comparison between different classifiers.

Fusion strategies	RAF	SFEW 2.0
Score-level fusion	86.7	56.2
Feature-level fusion	**87.5**	**56.9**

Table 4. Comparison between fusion strategies.

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

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

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

[1] D. Acharya, Z. Huang, D. P. Paudel, and L. V. Gool. Covariance pooling for facial expression recognition. *IEEE Proc. Computer Vision and Pattern Recognition Workshops*, pages 480–4807, 2018.

[2] S. Albanie and A. Vedaldi. Learning grimaces by watching tv. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2016.

[3] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. Openface 2.0: Facial behavior analysis toolkit. *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 59–66, 2018.

[4] S. Berretti, A. D. Bimbo, P. Pala, B. B. Amor, and M. Daoudi. A set of selected sift features for 3d facial expression recognition. *Proc. Int. Conf. on Pattern Recognition*, pages 4125–4128, 2010.

[5] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[6] F.-J. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, and G. G. Medioni. Expnet: Landmark-free, deep, 3d facial expressions. *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 122–129, 2018.

[7] S. Cheng, I. Kotsia, M. Pantic, and S. P. Zafeiriou. 4dfab: A large scale 4d database for facial expression analysis and biometric applications. *IEEE Proc. Computer Vision and Pattern Recognition*, pages 5117–5126, 2018.

[8] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 38(8):1548–1568, 2016.

[9] M. Dahmane and J. Meunier. Emotion recognition using dynamic grid-based hog features. In *Face and Gesture 2011*, pages 884–888. IEEE, 2011.

[10] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. D. Gedeon. Emotion recognition in the wild challenge 2014: Baseline, data and protocol. In *ICMI*, 2014.

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

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

recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 118–126. IEEE, 2017.

[13] Y. Fan, J. C. Lam, and V. O. Li. Video-based emotion recognition using deeply-supervised neural networks. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pages 584–588. ACM, 2018.

[14] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. C. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. T. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64:59–63, 2015.

[15] D. Kim, M. Hernandez, J. Choi, and G. Medioni. Deep 3d face identification. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 133–142. IEEE, 2017.

[16] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

[17] P. Lemaire, M. Ardabilian, L. Chen, and M. Daoudi. Fully automatic 3d facial expression recognition using differential mean curvature maps and histograms of oriented gradients. *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 1–7, 2013.

[18] H. Li, L. Chen, D. Huang, Y. Wang, and J.-M. Morvan. 3d facial expression recognition via multiple kernel learning of multi-scale local normal patterns. *Proc. Int. Conf. on Pattern Recognition*, pages 2577–2580, 2012.

[19] H. Li, H. Ding, D. Huang, Y. Wang, X. Zhao, J.-M. Morvan, and L. Chen. An efficient multimodal 2d+ 3d feature-based approach to automatic facial expression recognition. *Computer Vision and Image Understanding*, 140:83–92, 2015.

[20] H. Li, J. Sun, Z. Xu, and L. Chen. Multimodal 2d + 3d facial expression recognition with deep fusion convolutional neural network. *IEEE Trans. on Multimedia*, 19(12):2816–2831, 2017.

[21] S. Li and W. Deng. Deep facial expression recognition: A survey. *CoRR*, abs/1804.08348, 2018.

[22] S. Li, W. Deng, and J. Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *IEEE Proc. Computer Vision and Pattern Recognition*, pages 2852–2861, 2017.

[23] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen. Pointcnn: Convolution on x-transformed points. In *NeurIPS*, 2018.

[24] C. Liu, T. Tang, K. Lv, and M. Wang. Multi-feature based emotion recognition for video clips. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pages 630–634. ACM, 2018.

[25] P. Liu, S. Han, Z. Meng, and Y. Tong. Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1805–1812, 2014.

[26] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *arXiv preprint arXiv:1708.03985*, 2017.

[27] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza. Disentangling factors of variation for facial expression recognition. In *European Conference on Computer Vision*, pages 808–822. Springer, 2012.

[28] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun. Bosphorus database for 3d face analysis. pages 47–56, 2008.

[29] A. Savran, B. Sankur, and M. T. Bilge. Facial action unit detection: 3d versus 2d modality. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 71–78. IEEE, 2010.

[30] C. Szegedy, S. Ioffe, and V. Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, 2016.

[31] A. T. Tran, T. Hassner, I. Masi, E. Paz, Y. Nirkin, and G. G. Medioni. Extreme 3d face reconstruction: Seeing through occlusions. *IEEE Proc. Computer Vision and Pattern Recognition*, pages 3935–3944, 2018.

[32] M. F. Valstar, E. Sánchez-Lozano, J. F. Cohn, L. A. Jeni, J. M. Girard, Z. Zhang, L. Yin, and M. Pantic. Fera addressing head pose in the third facial expression recognition and analysis challenge. pages 839–847, 2017.

[33] V. Vielzeuf, C. Kervadec, S. Pateux, A. Lechervy, and F. Jurie. An occam's razor view on learning audiovisual emotion recognition with small training sets. In *ICMI*, 2018.

[34] V. Vielzeuf, S. Pateux, and F. Jurie. Temporal multimodal fusion for video emotion classification in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 569–576. ACM, 2017.

[35] F. Wang, X. Xiang, C. Liu, T. D. Tran, A. Reiter, G. D. Hager, H. Quon, J. Cheng, and A. L. Yuille. Regularizing face verification nets for pain intensity regression. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1087–1091. IEEE, 2017.

[36] H. Yang, U. Ciftci, and L. Yin. Facial expression recognition by de-expression residue learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2168–2177, 2018.

[37] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale. A high-resolution 3d dynamic facial expression database. *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 1–6, 2008.

[38] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. In *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 211–216. IEEE, 2006.

[39] J. Zeng, S. Shan, and X. Chen. Facial expression recognition with inconsistently annotated datasets. In *ECCV*, 2018.

[40] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[41] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu. Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. In *Proceedings Third IEEE International Conference on Automatic face and gesture recognition*, pages 454–459. IEEE, 1998.

[42] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):915–928, 2007.