02 Jul 2017

CNN Based 3D Facial Expression Recognition Using Masking And Landmark Features

Huiyuan Yang
Missouri University of Science and Technology, hyang@mst.edu

Lijun Yin

Recommended Citation

H. Yang and L. Yin, "CNN Based 3D Facial Expression Recognition Using Masking And Landmark Features," *2017 7th International Conference on Affective Computing and Intelligent Interaction, ACII 2017*, pp. 556 - 560, Institute of Electrical and Electronics Engineers, Jul 2017.
The definitive version is available at https://doi.org/10.1109/ACII.2017.8273654

CNN based 3D Facial Expression Recognition Using Masking and Landmark Features

Huiyuan Yang and Lijun Yin
Department of Computer Science
State University of New York at Binghamton

***Abstract*—Automatically recognizing facial expression is an important part for human-machine interaction. In this paper, we first review the previous studies on both 2D and 3D facial expression recognition, and then summarize the key research questions to solve in the future. Finally, we propose a 3D facial expression recognition (FER) algorithm using convolutional neural networks (CNNs) and landmark features/masks, which is invariant to pose and illumination variations due to the solely use of 3D geometric facial models without any texture information. The proposed method has been tested on two public 3D facial expression databases: BU-4DFE and BU-3DFE. The results show that the CNN model benefits from the masking, and the combination of landmark and CNN features can further improve the 3D FER accuracy.**

1. Introduction

Facial expression, as a non-verbal communication cue, is an important part for human interaction. A wide range of applications can benefit from the ability of automatic recognition of facial expressions, *e.g., human computer interaction, social behavior analysis, health-care application, and many others*. The goal of FER is to automatically distinguish a variety of facial expressions associated with various human emotions, including six basic expressions as shown in Figure 1. Although a lot of work have been done on 2D facial expression recognition, there are still some challenges, mainly caused by the variations of head-pose, illumination, registration, occlusion and identity [22]. Compared to 2D texture image, 3D face model is expected to contain more information about facial expressions, *e.g., invariant to pose and lighting conditions*. Therefore, analyzing facial expressions in a 3D space has a great potential in order to address those issues for FER [20] [24] [2].

Convolutional Neural Networks (CNNs) have been widely used in recent years, which show superior performance in a wide range of tasks, such as image classification [12], object detection [6] [21], face recognition [23] and more. One of the main factors for the successful applications of CNN is that it can automatically learn the complete representation from a large scale dataset, such as, ImageNet [12], VGG face [25] and FaceNet [23]. Inspired by this, different researchers start applying CNN to facial expression recognition. Tang [26] replaced the softmax layer with SVMs in the CNN model, and showed an improved performance for facial expression recognition. Mollahosseini et al. [17] proposed a novel CNN structure to improve the FER performance cross databases. Lopes et al. [15] trained a CNN model for facial expression recognition on a small face database by utilizing some specific image pre-processing steps. Pramerdorfer and Kampel [19] analyzed and compared several state-of-the-art CNN based methods for facial expression recognition. Although using CNN for 2D facial expression recognition is well studied in the literature, as far as we know, few works have been done for 3D facial expression recognition, due to the lack of large scale of 3D face expression databases as well as the inconsistency and uncertainty of 3D objects for input to the CNN model.

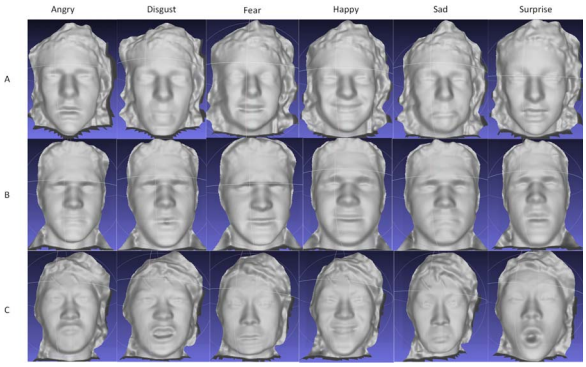

Figure 1. Example of 3D face models. A, B and C are different subjects, each of them shows six basic expression: anger, disgust, happiness, fear, sadness and surprise.

2. Research Questions

2.1. CNN based 3D Facial Expression Recognition

Recently, with deep models overcoming more and more challenges, researchers turn to embrace deep learning based methods, and find that combining the two stages of facial expression recognition (feature extraction and classification) could get a better result. Liu et al. [14] unified

This is a joint work with Prof. Lijun Yin. Huiyuan Yang is a Ph.D. student at the Graphics and Image Computing Lab advised by Dr. Lijun Yin. Contact: hyang51@binghamton.edu, lijun@cs.binghamton.edu

978-1-5386-0563-9/17/$31.00 ©2017 IEEE

the three training stages(feature learning, feature selection and classification) by using a boosted deep belief network, which was effective to learn the expression-related facial appearance/shape changes. Khorrami et al. [10] presented two schemes to learn multiple deep convolutional neural networks for static images classification, and generated state-of-the-art results on several datasets. Zhao et al. [34] proposed a peak-piloted deep network, which used peak expression to supervise the recognition of non-peak expression, but it only facilitated the recognition of same expression of the same subject.

Until we write this paper, 3D facial expression recognition is still not well studied. The lack of large labelled 3D database is the first reason; and how to pass a 3D face model to a CNN is yet not determined. Wu et al. [28] used a $30\times30\times30$ voxel grid representation of 3D model for object recognition, but that is too coarse for a high resolution 3D face model, and a higher voxel grid representation requires a massive labelled 3D training data, which is not available right now.

2.2. Identity-independent CNN for Facial Expression Recognition

We notice the performance gap between the training and testing in deep model for FER. In other words, the performance degrades on unseen subjects. Although the performance may be improved with extra subjects added to the training procedure, it isn't always easy to obtain more data for both the cost of the device and the availability of new subjects. So, current works focus on how to transfer the pre-trained model to the rather small dataset and alleviate the personal variations in the training dataset. Ding et al. [5] proposed a FaceNet2ExpNet to train an expression recognition network on a relatively small dataset, and was able to capture improved high-level expression semantics. Ng et al. [18] performed a supervised cascaded approach to fine-tune a pre-trained deep model, and showed comparable results on small datasets. Both of the beforementioned methods try to transfer the pre-trained model to a small dataset, but still suffer on unseen subjects due to high variations between subjects. Meng et al. [16] tried to alleviate inter-subject variations by using expression-sensitive contrastive loss, and achieved identity-invariant expression recognition on three public databases. But, this method just alleviates the inter-subject issues to some extent, and still suffers from large inter-subject variations.

2.3. Dynamic and Multi-modal Facial Expression Recognition

Some expressions such as sadness and anger are usually hardly recognized using only static images due to subtle face deformations involved in those cases [3], but we can generally obtain better understanding if we analyse the facial deformations over time. This is the core of dynamic approaches, which not only capture the appearance features but also the spatio-temporal features. Dapogny et al. [4] proposed a pairwise conditional random forests to learn spatio-temporal patterns, and showed significant improvements on several facial expression benchmarks. Le et al. [13] utilized the facial level curves based 3D shape representation for dynamic 3D facial expression recognition. Dapogny et al. [3] showed a promising result for dynamic FER by training a transition classifier that was fused with static estimation. Readers can find more details in [1].

Many works have also considered using multimodality (*e.g., audio, video and physiological data*) for facial expression recognition. The advantages of fusing multiple modalities are the increased robustness and complementary information. Zamzmi et al. [31] showed an improved performance to assess infants' pain by combining both behavioral and physiological pain indicators. He et al. [7] used a deep bidirectional long short-term memory recurrent neural network to fusion multimodal features, and showed promising results on AVEC 2015 challenge. Irani et al. [8] utilized the combination of RGB, depth and thermal facial images for pain level recognition. Because of the availability of multimodalities in some database, *e.g., BP4D [32], BP4D+ [33]*, we may also consider adding those information for 3D facial expression recognition.

3. Proposal and Experiments

We propose a CNN based 3D facial expression recognition method using masking and landmark features. Figure 2 is the framework of our proposed method. We first preprocess the 3D face model by cropping and face registration to remove pose variation; and then generate depth and curvature maps from 3D face model using orthogonal projection, which are combined with masks generated by landmarks to train a CNN model. After training, we output the fully connected layer (FC1) as features, which are further combined with the landmark features for facial expression recognition.

3.1. Data

BU-3DFE: The BU-3DFE [30] database contains 2500 3D face models of 100 subjects (56% female, 44%male), with a variety of ages and ethnics. Each subject has six basic expressions with four levels of intensity and a neutral expression.

BU-4DFE: The BU-4DFE [29] is a high resolution 3D dynamic facial expression database, which contains 58 female and 43 male total 101 subjects, with a variety of ethnics. Six model sequences are captured for each subject, and each of them shows one of the six basic expressions(anger, disgust, happiness, fear, sadness and surprise) starting from neutral(zero intensity) to apex(highest intensity), and then ending with neutral again. Because no label information available to split the neutral frames from the whole sequence, we manually label and remove the neutral frames from both the beginning and ending of each sequence. Thus, an experimental database consisting of 45,000 frames

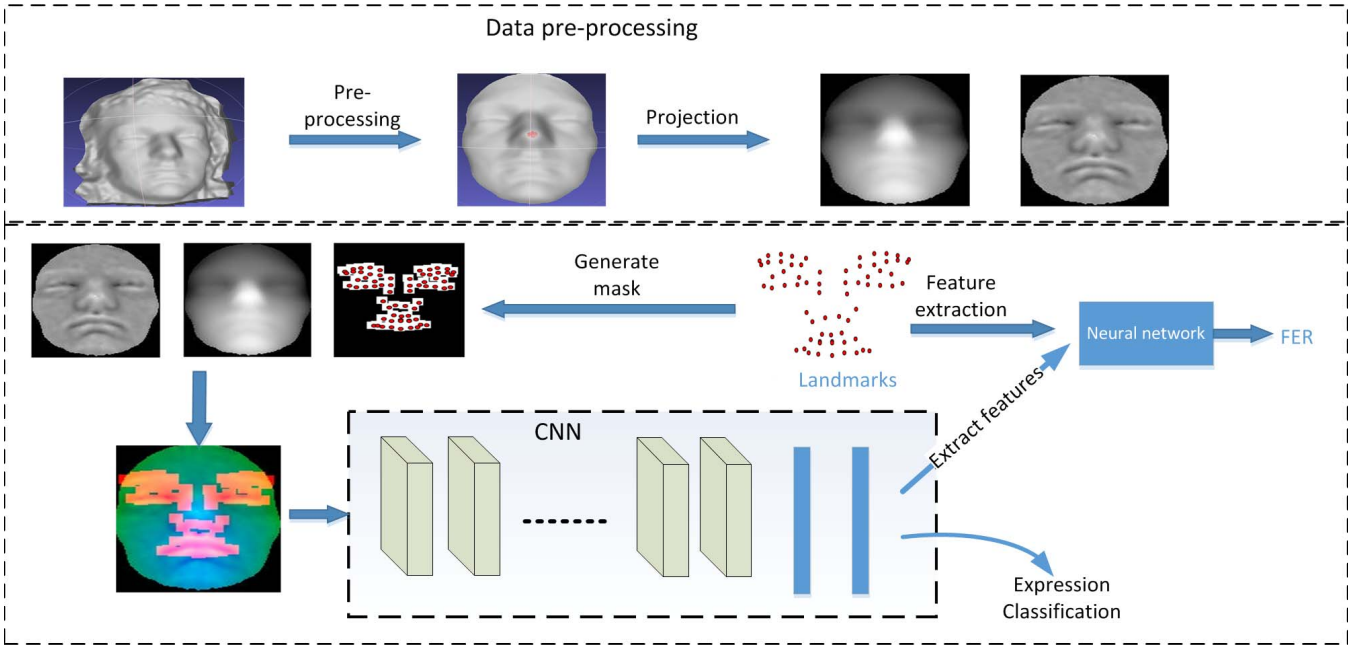

Figure 2. An overview of our method. (1) The upper part is data pre-processing steps. The nosetip is first detected on 3D face model, and then face crop and pose normalization are applied. Finally, depth and curvature feature maps are generated from the pre-processed face model. (2) The lower part is the combination of landmarks and CNN model. The landmark is first used to generate mask, which is combined with two other feature maps to generate a three channel [mean curvature map, depth map, mask] image. The image is then used to train a CNN model with six outputs. After that, we output the fully connected layer (FC1) as feature, which is further combined with the landmark features for facial expression recognition.

is built. The database is further divided into 10 subsets, where the subject in any two subsets are mutually exclusive. Then, following the general rule, a 10-fold cross-validation strategy is applied, where we use 8 subsets for training, and the remaining two subsets for validation and testing.

3.2. Data Pre-processing

Given a 3D face model, we first detect the tip of nose based on SHOT [27] feature with a specific search radius (25mm used in our experiments); then we crop the 3D face model using a sphere with a 90mm radius and centring in nosetip. Those steps keep only the facial region and also reduce the storage space. Finally, iterative closest point(ICP) [9] is used to frontalize the 3D face model, so the pose variations are eliminated. The processed 3D model is further projected orthogonally to generate 2D representations: depth and curvature maps. Here we use the mean curvature during our experiments. During the projection, we also use bilinear interpolation to fill the holes in 2D feature map.

Those steps firstly decrease the computation complexity by using less vertices; second, improve the performance by removing the pose variation.

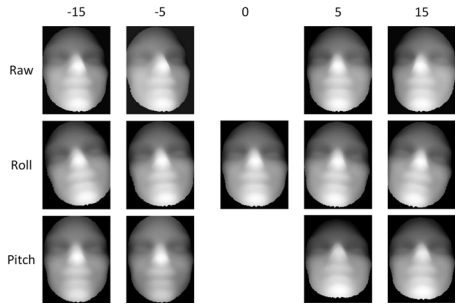

Figure 3. Examples of generated feature maps after Rotating the 3D face model along raw, roll and pitch directions.

3.3. Implementation

Our implementation is based on TensorFlow[1]. To avoid overfitting, data augmentation procedures are employed to train the CNN model, where a 3D face model is rotated $[-10^o, -5^o, 0^o, 5^o, 10^o]$ along each of Pitch, Yaw and Roll directions, as shown in Figure 3; then a 10x10 random black block and gaussian noise with a 0.01 standard deviation are also added to the feature maps, resulting in a 2.6M training data.

Our implementation for the deep model follows the practice in VGG face. The feature map is first resized to 64x64. The weight is initialized by a normal distribution

1. https://www.tensorflow.org

TABLE 1. ACCURACY USING DIFFERENT FEATURES ON BU-4DFE DATABASE.

Name	Model	Accuracy
Landmarks	residual error	69.3%
3D face model	depth map	62.1%
	curvature map	67.3%
	depth + curvature	69.0%
Combination	depth + curvature + mask	73.3%
	all together	**75.9%**
Joint [3]	2D texture	75.8%
PCRF [4]	2D texture	75.2%

with mean and standard deviation (0, 0.1), and all the biases are given a 0.1 initial value. We use the Adam algorithm [11] with a 0.0001 learning rate for optimization and a mini-batch size of 100. The model is trained for 10,000 iterations.

3.4. Performance on BU-4DFE Database

3.4.1. 3D FER Using CNN Model. With using depth map and mean curvature map only, the CNN model shows 62.1% and 67.3% accuracy respectively. The deep model using mean curvature map performs better than solely using depth map because the curvature is invariant to head-pose, while the depth maps still relies on an accurate face registration. And, depth map and mean curvature compensate each other for facial expression recognition, we get a 69.0% accuracy after combining those two feature maps.

3.4.2. Combination of Landmark Features and CNN Model. After adding the mask generated from landmarks, the deep model reaches an accuracy of 73.3%. The mask, which is generated based on the landmark positions, is used to force the CNN model to focus on the areas that closely related to expression changes. For example, the areas around the mouth, eyes, and eyebrows. Finally, the residual error based landmark features and the 3D face model features that extracted by the CNN model are further combined, and a simple neural networks is used for expression classification. Our proposed method achieves a comparable accuracy of 75.9% on BU-4DFE database. It worth to note that both Joint [4] and PCRF [3] report their results based on 2D texture image on BU-4DFE, while our proposed method using 3D face shade model. Figure 4 is the confusion matrix for recognizing six basic expressions. We can note that the happy expression has a higher accuracy rate than others. While the fear expression only has a 46% recognition rate; Looking at the confusion matrix, we may find that the fear expression was confused majority with happy and disgust.

3.5. Evaluation on BU-3DFE Database

After training on BU-4DFE, we also test the model on BU-3DFE. It worth to note that none of the BU-3DFE database is used for training the model. We report our

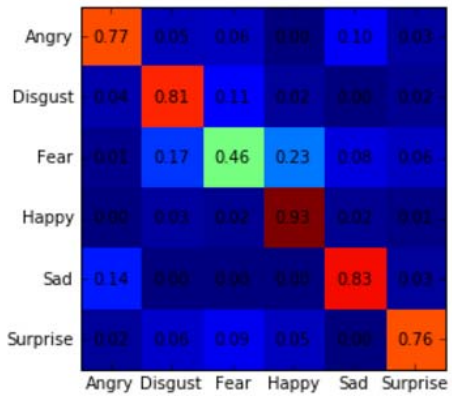

Figure 4. Confusion matrix for recognizing six basic expressions.

TABLE 2. ACCURACY ON RECOGNIZING SIX EXPRESSIONS WITH DIFFERENT LEVELS OF INTENSITY ON BU-3DFE DATABASE.

	Expression intensity level			
	I	II	III	IV
Accuracy	49.8%	56.9%	56.8%	66.7%

results based on the expression intensity level. As we can see in Table 2, the expression intensity IV has the highest accuracy than others; while the expression intensity I only has a 49.8% recognition rate. The average accuracy over four intensity levels is 57.6%.

4. Conclusion and Future Work

In this paper, we propose a 3D facial expression recognition (FER) algorithm using convolutional neural network (CNN) and landmark features/masks. The algorithm is invariant to pose and illumination variations due to the solely use of 3D geometric facial models without any texture information, and has been tested on two public 3D facial expression databases: BU-4DFE and BU-3DFE. The results show: (I) CNN model benefits from the masking; (II) the combination of landmark features and CNN features can further improve the 3D FER accuracy.

Our future work will extend the current work to make it more robust for unseen subjects for facial expression recognition. We will test the algorithm using the more challenging spontaneous expression dataset - BP4D database, and also address the issue of 3D AU detection and 3D AU intensity estimation using the newly developed multi-modal dataset - BP4D+ database.

Acknowledgements

The material is based upon the work supported in part by the National Science Foundation under grants CNS-1629898 and CNS-1205664.

References

[1] 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. PAMI*, 38(8):1548–1568, 2016.

[2] A. Danelakis, T. Theoharis, and I. Pratikakis. Geotopo: dynamic 3d facial expression retrieval using topological and geometric information. In *Proceedings of the 7th Eurographics Workshop on 3D Object Retrieval*, pages 1–8. Eurographics Association, 2014.

[3] A. Dapogny, K. Bailly, and S. Dubuisson. Dynamic facial expression recognition by joint static and multi-time gap transition classification. In *InAutomatic Face & Gesture Recognition'15*, volume 1, pages 1–6. IEEE, 2015.

[4] A. Dapogny, K. Bailly, and S. Dubuisson. Pairwise conditional random forests for facial expression recognition. In *InProceedings of the IEEE international conference on computer vision*, pages 3783–3791, 2015.

[5] H. Ding, S. K. Zhou, and R. Chellappa. Facenet2expnet: Regularizing a deep face recognition net for expression recognition. *arXiv preprint arXiv:1609.06591*, 2016.

[6] R. Girshick. Fast r-cnn. In *InProceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[7] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli. Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 73–80. ACM, 2015.

[8] R. e. a. Irani. Spatiotemporal analysis of rgb-dt facial images for multimodal pain level recognition. In *InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 88–95, 2015.

[9] T. Jost and H. Hügli. Fast icp algorithms for shape registration. In *Joint Pattern Recognition Symposium*, pages 91–99. Springer, 2002.

[10] P. Khorrami, T. Paine, and T. Huang. Do deep neural networks learn facial action units when doing expression recognition? In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 19–27, 2015.

[11] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[13] V. Le, H. Tang, and T. S. Huang. Expression recognition from 3d dynamic faces using robust spatio-temporal shape features. In *InAutomatic Face & Gesture Recognition'11*, pages 414–421. IEEE, 2011.

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

[15] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos. Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order. *Pattern Recognition*, 61:610–628, 2017.

[16] Z. Meng and at al. Identity-aware convolutional neural network for facial expression recognition. 2017.

[17] A. Mollahosseini, D. Chan, and M. H. Mahoor. Going deeper in facial expression recognition using deep neural networks. In *Applications of Computer Vision (WACV), 2016*, pages 1–10. IEEE, 2016.

[18] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *ACM on multimodal interaction*, pages 443–449. ACM, 2015.

[19] C. Pramerdorfer and M. Kampel. Facial expression recognition using convolutional neural networks: State of the art. *arXiv preprint arXiv:1612.02903*, 2016.

[20] S. Ramanathan, A. Kassim, Y. Venkatesh, and W. S. Wah. Human facial expression recognition using a 3d morphable model. In *Image Processing, 2006 IEEE International Conference on*, pages 661–664. IEEE, 2006.

[21] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[22] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Trans. PAMI*, 37(6):1113–1133, 2015.

[23] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[24] N. Sheng, Y. Cai, C. Zhan, C. Qiu, Y. Cui, and X. Gao. 3d facial expression recognition using distance features and lbp features based on automatically detected keypoints. In *Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), International Congress on*, pages 396–401. IEEE, 2016.

[25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[26] Y. Tang. Deep learning using support vector machines. *CoRR, abs/1306.0239*, 2, 2013.

[27] F. Tombari, S. Salti, and L. Di Stefano. Unique signatures of histograms for local surface description. In *ECCV*, pages 356–369. Springer, 2010.

[28] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1912–1920, 2015.

[29] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale. A high-resolution 3d dynamic facial expression database. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE, 2008.

[30] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. In *Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on*, pages 211–216. IEEE, 2006.

[31] G. Zamzmi, C.-Y. Pai, D. Goldgof, R. Kasturi, T. Ashmeade, and Y. Sun. An approach for automated multimodal analysis of infants' pain. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 4148–4153. IEEE, 2016.

[32] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

[33] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3438–3446, 2016.

[34] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan. Peak-piloted deep network for facial expression recognition. In *ECCV*, pages 425–442. Springer, 2016.