Deformable Synthesis Model for Emotion Recognition

Diego Fabiano and Shaun Canavan
Computer Science and Engineering, University of South Florida, Tampa, Florida

Abstract— **In this paper, we propose a deformable synthesis model that can be used to synthesize data to train deep neural networks for the task of emotion recognition. This model is created through the use of 3D facial landmarks, which are then projected to the 2D image plane for training a deep network. We show that this model can accurately recognize a range of emotions that include happiness, sadness, and fear. We test the efficacy of our proposed approach on three publicly available 3D face databases, namely BU4DFE, BP4D, and BP4D+. We show that the proposed method can accurate recognize emotion when training and testing on the same database, as well as cross-database training and testing on all 3 databases. We show the proposed method results in accurate recognition of emotion using deep neural networks outperforming current state of the art on each of the tested databases.**

I. Introduction

Emotion is an important aspect of human intelligence [22] and to understand the foundation of autonomy and advance interfaces between humans and machines, we must also understand the role of emotion. Emotion recognition has real-world applications in areas such as retail shopping (e.g. customer satisfaction), medical (e.g. predicting autism), and defense (e.g. pain assessment). Considering this, there has been a great deal of research into human emotion recognition in the past decades, where many important advances have been made.

Recent works in human emotion analysis have made good use of deep neural networks as they have become increasingly popular since Hinton et al. [16] showed the use of complementary priors to make inference less difficult. Gupta et al [13] used a spatio-temporal convolutional neural network for expression recognition. They proposed a scale-invariant deep architecture for learning illumination invariant features. Using a Boosted Deep Belief Network, Liu et al. [20] trained feature learning, selection, and classifier construction iteratively in a unified loopy framework which showed an increase of the classification accuracy over state of the art. Motivated by the Generative Adversarial Model (GAN) [1], a De-expression Residue Learning [27] approach was proposed which can generate a corresponding neural expression given an arbitrary facial expression from an image. Variation in subjects can result in performance degradation for facial expression recognition. To alleviate this issue, Yang et al. [28] proposed regenerating expression from input facial images. By using a conditional GAN (cGAN) [21], they developed an identity adaptive feature space that can handle variations in subjects.

3D expression recognition has been gaining increasing attention in recent years. Li et al. [18] proposed a multimodal approach that uses 2D and 3D facial data for deep expression recognition. They represented 3D faces as size types of 2D facial attribute maps, which were then used to train a deep fusion convolutional neural network. Hariri et al. [14] investigated the application of manifold-based classification for 3D facial expression recognition. They represented the 3D facial surface by a set of covariance descriptors. Abbasnejad et al. [1], generated a synthetic dataset to train deep networks. They showed that this data can be used to train deep 3D convolutional networks to efficiently classify facial expressions.

Along with works on 2D and 3D emotion recognition, there have been many approaches to deformable models [2], [4], [6], [7]. They have successfully been used for tasks such as face recognition [17], emotional state recognition [23], and works in the medical field. In this case, Liu et al. [19] used a convolutional neural network along with 3D simplex deformable modeling for musculoskeletal magnetic resonance imaging. While these works have successfully used deformable models, their focus has mainly been on fitting unseen data and segmentation. Dong et al. [9] used a statistical model of action units (AU) for 3D facial expression recognition. They found that using global and local features of facial representations, that the AU occurrence probability can be computed from a 3D face model.

Motivated by these works and the success of deep neural networks, we propose a deformable model-based approach for emotion recognition. We refer to this approach as a Deformable Synthesis Model (DSM), which we propose to address training a deep neural network with a large-scale dataset for emotion recognition. The proposed DSM is based on a 3D statistical model of shape, which uses 3D landmarks that are then projected to the 2D image plane. To test the efficacy of our proposed approach, we create a DSM from 3 state of the art facial expression databases, namely BU4DFE [29], BP4D [30], and BP4D+ [31]. See Fig. 1 for an overview of the proposed method. The main contributions of the paper are summarized as follows:

1) We propose a Deformable Synthesis Model (DSM), for synthesizing 3D facial data, for training deep networks for emotion recognition.
2) We address the issue of a large-scale dataset that is reliable and accurate through our proposed method.
3) The proposed DSM can be used to achieve state-of-the-art results on 3 publicly available datasets.

II. Deformable Synthesis Model

We propose a new Deformable Synthesis Model (DSM) for the synthesis of 3D facial data. The DSM allows us to create as much synthetic facial data as needed, which

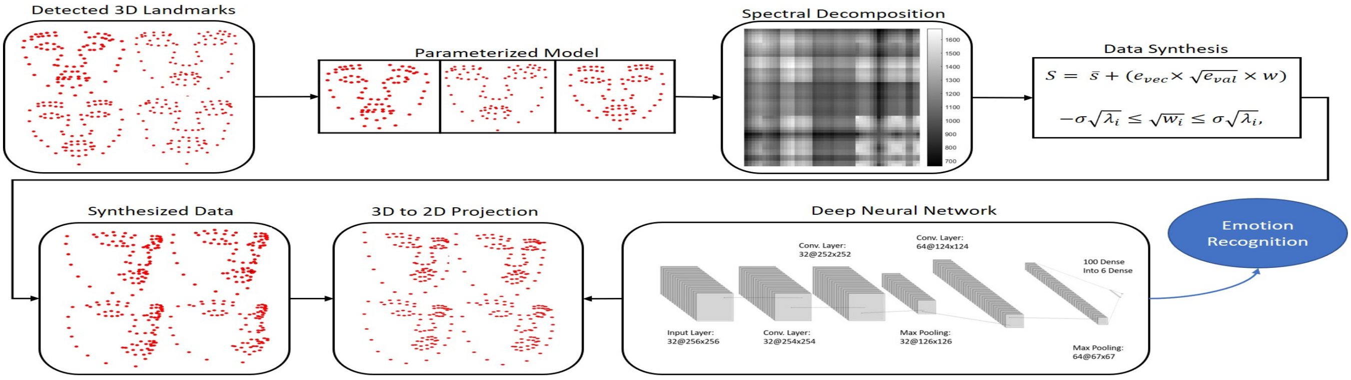

Fig. 1. Overview of the proposed method for creating a DSM and training a deep neural network for emotion recognition.

can be powerful for training deep neural networks. This model also transforms facial expressions into a new domain where the same emotion looks similar regardless of dataset it is modeled from, which allows for generalizing across databases with high recognition accuracy. Details regarding DSM creation and synthesis of data are given in the following subsections.

A. Model creation

A DSM is motivated by active shape models [6]. It is constructed from L sets of 3D facial landmarks that contain key facial features such as the nose, mouth, and eyes. Any method for detecting 3D facial landmarks can be used for creation of a DSM, however, we used a shape index-based statistical shape model (SI-SSM) [5].

Given L sets of 3D facial landmarks, where each frame of the set contains N landmarks, we split the data into classes based on each emotion that we want to model. Each class contains all subjects from the given dataset. For example, all frames that correspond to the emotion happy, are separated into their own class. This is done to ensure the DSM explicitly models each emotion independently, allowing us to synthesize the appropriate emotion data for recognition. It is important to note that each of the tested databases contains sequences of 3D data that contain both neutral and non-neutral data. The DSM contains both of these data types to ensure that it will produce more natural synthetic data.

The data is then normalized using Z-score transformation giving a mean of 0 and a standard deviation of 1. A parameterized model (S) is then created as $S = F_1, , F_N$, where F_i is the i^{th} normalized landmark and $F_i = (x_i, y_i, z_i)$. A matrix is then created where each row corresponds to one parameterized model. Spectral decomposition [15] is then performed on the transpose of the matrix to recast it in terms of its eigenvalues and corresponding eigenvectors. This is done to estimate the variance in emotion of the $N \times L_{emotion} \times 3$ feature space, where $L_{emotion}$ is the total number of faces in each emotion class. Using the eigenvalues, eigenvectors, and a vector of weights new 3D facial landmarks can be synthesized as we detail next.

B. 3D facial landmark synthesis

New 3D facial landmarks can be synthesized by a linear combination of landmarks as $S = \bar{s} + (e_{vec} \times \sqrt{e_{val}} \times w)$ where $\bar{s}$ is the average landmarks, of the modeled emotion, e_{vec} and e_{val} are the eigenvectors and eigenvalues of the spectral decomposition, and *w* is a weight vector that controls the shape of the landmarks. Given this combination, to synthesize new data, *w*, is modified within a specified limit to ensure the synthesized data retains the original shape (e.g. face). We define this limit as $-\sigma\sqrt{\lambda_i} \leq \sqrt{w_i} \leq \sigma\sqrt{\lambda_i}$, where λ_i is the i^{th} eigenvalue of the specular decomposition, w_i is the i^{th} weight in the specified range, and σ is defined as the standard deviation from the Z-score normalization (± 1). Normalizing the data, and using the standard deviation as our allowable range, allows us to capture accurate variance in emotion from our training data used to construct our DSM. To synthesize large amounts of 3D facial landmarks, we step between $[-\sigma, \sigma]$ for the desired number of sets of landmarks, where *step=2/number_of_sets, and number_of_sets = total amount of synthetic data* to create. Using this approach, we can synthesize as much data as is needed to train our deep neural networks. Synthesized data is shown in Fig. 2 below.

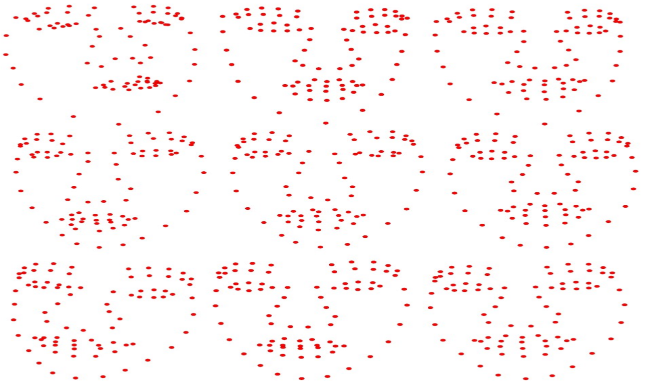

Fig. 2. Example synthesized 3D facial landmarks. Row 1: BU4DFE; Row 2: BP4D; Row 3: BP4D+.

C. Projection of 3D synthesized data to 2D image plane

Recently, Yang and Yin found that expression recognition accuracies increase when landmarks are used to generate a map which is combined with 3D depth and curvature information, that has been orthogonally projected to 2D, to create a 3-channel image, to train a CNN [26]. Motivated by this, we orthogonally project the 3D landmarks to the 2D image plane. It is important to note that the synthesized data is represented as 3D landmarks (x, y, z); once they are projected to the 2D image plane, they are represented as an image of size 256 × 256 (Fig. 3).

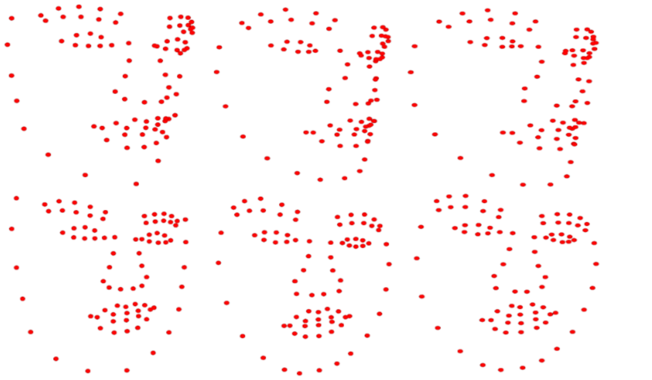

Fig. 3. Synthetic landmarks projected to 2D. From left to right: BU4DFE, BP4D, BP4D+. Top row: synthesized 3D landmarks; Bottom row: 2D projected landmarks (from 3D landmarks in same columnn in row 1.

III. 3D Facial Datasets

To test the efficacy of the proposed Deformable Synthesis Model, we tested on 3 publicly available databases; BU4DFE [29], BP4D [30], and BP4D+ [31].

A. BU4DFE

The BU4DFE dataset contains 101 subjects with six posed expressions of anger, happiness, fear, disgust, sadness, surprise, as well as neutral (no expression). Both genders are represented with 58 female and 43 male subjects and multiple racial and ethnic background with ages ranging from 18-45. Each expression is around 100 frames, and contains the 3D mesh model, 2D texture, and the corresponding annotated 3D facial landmarks. The total size of the database is over 60,000 frames of data. In the literature, it has extensively been studied for the task of emotion/expression recognition [1], [3], [8], [10], [12].

B. BP4D

The BP4D dataset was used in the Facial Expression Recognition and Analysis (FERA) challenge in 2015 [24] and 2017 [25]. It contains 41 subjects with eight dynamic expressions that that include the same six (plus neutral) that are found in the BU4DFE, plus nervousness and pain. Emotion was elicited using tasks, conducted by an interviewer, such as a sudden burst of sound (surprise), submerge hand in ice water (pain), and an unpleasant smell (disgust). The dataset contains 18 male and 23 female subjects ages 18-29 years of age, with a range of ethnicities. The dataset includes FACS action units, 2D textures and 3D models, each with the accompanying 2D and 3D annotated facial landmarks.

C. BP4D+

The BP4D+ multimodal spontaneous emotion corpus consists of 140 subjects with 58 males and 82 females; ages 18-66, each with highly varied emotional responses. Emotion was elicited using 10 different tasks, like the BP4D dataset. The target emotions include the eight found in the BP4D, plus skepticism, and startled (like surprise). The dataset contains the same modalities as found in the BP4D (2D, 3D, facial landmarks, action units), plus thermal images, and physiological data. The BP4D+ includes over 10TB of data and was also used in the FERA challenge 2017.

IV. Experimental Design

To evaluate the efficacy of the DSM, we trained a deep convolutional neural network with the projected 2D images.

A. Deep network architecture

Convolutional Neural Network. We fine-tuned our CNN with an initial convolutional layer containing 32 input units and a kernel size of 3x3, this is then followed by 0.2 Dropout and a second convolutional layer with the same characteristics of the previous one. Max pooling is then applied with a pool size of 2x2, a final convolutional layer is applied with 64 input units, which is then followed by max pooling and flattening. Lastly, a fully connected layer of 100 input units follows with 0.5 Dropout; and a fully connected layer for prediction. The softmax activation function was used for the last layer and the rectified linear unit was used in the inner layers. The adamax optimizer was used with a learning rate of 0.0001.

B. Training and testing data

To conduct our experiments, we used an 80/20 split of the data, with 80% used for training, and 20% used for testing. In this approach, the testing labels are not pre-used as there is a complete split between training and testing. The training and testing data were randomly selected. While conducting our experiments, we tested different random subsets of 80/20 splits and no statistically significant changes in the reported accuracies were found. The data synthesized from the proposed DSM is generated from all subjects (e.g. model is created from all subjects from each emotion), therefore we do not consider the case of separating the subjects from training and testing data. We conducted experiments where the networks are trained and tested on the same database, as well as experiments where the networks are trained on one dataset and tested on another (e.g. train on BU4DFE, and test on BP4D). There are 6 expressions shown in the BU4DFE, 8 in the BP4D, and 10 in the BP4D+. Considering this, to conduct our experiments, we only consider the emotions found in all 3 databases which are happy, sad, surprise, anger, disgust, and fear. We generated 600,000 synthetic frames for each of the tested datasets. This is approximately 10 times the size of the BU4DFE, 2 times the size of the BP4D, and 56% of the size of BP4D+ (tested data). Although, the overall size of training data decreased with the BP4D+ compared to the original dataset, we wanted a consistent number of synthesized frames across all databases.

V. Results

Using the 80/20 split of training and testing data we conducted same and cross-database experiments on BU4DFE, BP4D, and BP4D+. The CNN discussed in section 4 was trained with the 80/20 split, using the projected 2D images as input. The results can be seen in table 1.

TABLE I

Recognition accuracies of synthetic data from the DSM.

Training/Testing	**BU4DFE Score**	**BP4D**	**BP4D+**
BU4DFE	**100%**	99.75%	98.09%
BP4D	99.9%	**100%**	99.72%
BP4D+	98.87%	99.77%	**99.92%**

As can be seen in table I, the proposed DSM is able to recognize emotions across the 3 tested databases with high accuracy. An average recognition rate of 99.56% across all databases, including same and cross-database experiments. These results show the DSMs ability to generalize across multiple databases. The synthetic data created from the proposed DSM, looks similar within each emotion, but different across other emotions. The DSM also removes some of the visual anomalies that are present in the original 3D landmarks (e.g. incorrectly detected landmarks) resulting in more visually accurate 2D projections (Fig. 4). This can help explain the high emotion recognition accuracy from the proposed deformable synthesis model. The CNN can focus on the explicit shape of the face (e.g. eyes, nose, mouth).

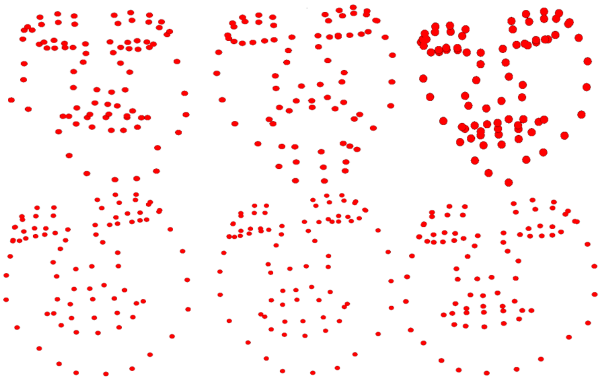

Fig. 4. Surprise expression from BP4D. Top row: original detected 3D landmarks; bottom row: DSM synthesized data.

A. Comparison to state of the art

We also compared our results to current state of the art on the BU4DFE, BP4D databases, BP4D+. It is important to note that while the tested databases are 4D (time-based), the proposed DSM is able to accurately recognition emotion using a single frame, which can be helpful when 4D information is not available. Comparisons to state of the art on these 3 databases can be seen in tables II, III, and IV.

Few works have used the BP4D for 3D facial expression/emotion analysis although it has been used for detecting action units [24], [25], [30]. For direct comparisons on the BP4D, we detail the results from Fabiano et al [11], as can be seen in table 3. They show that when using a random forest and raw 3D facial landmarks, high accuracy can be achieved, however, they do not show a method for generalizing across databases as our proposed DSM has the ability to do.

TABLE II

Comparisons to state of the art on BU4DFE.

Method	**Modality Score**	**Accuracy**
DSM	**Projected 3D landmarks**	**100%**
Fabiano et al. [11]	Raw 3D landmarks	99.9%
Abbasnejad et al. [1]	3D mesh models	91.22%
Fang et al. [12]	3D mesh models	91.0%
Yin et al. [29]	3D mesh models	90.44%
Yang et al. [26]	Projected 3D mesh models	75.9%

TABLE III

Comparisons to state of the art on BP4D.

Method	**Modality Score**	**Accuracy**
DSM	**Projected 3D landmarks**	**100.0%**
Fabiano et al. [11]	Raw 3D landmarks	99.69%

To the best of our knowledge, this is the first work to use BP4D+ for 3D facial expression/emotion analysis. Yang et al. [27], performed training and testing on the BP4D+, as well cross-database validation using the BP4D as training and the BP4D+ as testing, however, they used 2D images. Results of this comparison can be seen in table 4.

TABLE IV

Comparisons to state of the art on BP4D+.

Method/Database	**Modality**	**Accuracy**
DSM **Train/Test on BP4D+**	**Projected 3D landmarks**	**99.92%**
Yang et al. [27] Train/Test on BP4D+	2D images	81.39%
DSM **Train/Test on BP4D/BP4D+**	**Projected 3D landmarks**	**99.72%**
Yang et al. [27] Train/Test on BP4D/BP4D+	2D images	74.41%

To the best of our knowledge, this is the first study to perform cross-database emotion recognition across the 3 tested databases, showing accurate recognition results.

VI. Discussion

For the utility of 3D landmarks to be fully realized, being able to generalize across databases is necessary. Without generalization, the real-life applications for such algorithms are limited. We have proposed a Deformable Synthesis Model to handle this generalization, which is able to generate large amount of data to train the many weights of effective deep neural networks. We have also shown that the proposed method outperforms current state of the art on the 3 tested databases. The proposed method shows encouraging results for recognizing 6 emotions across multiple datasets which can help advance human-machine interfaces. We will next investigate the prediction of a new frame with an unknown label, as well as compare the DSM against different synthesizing methods.

References

[1] I. Abbasnejad, S. Sridharan, D. Nguyen, S. Denman, C. Fookes, and S. Luccey, Using synthetic data to improve facial expression analysis with 3D convolutional networks, *Computer Vision and Pattern Recognition*, 2017.

[2] V. Blanz and T. Vetter, A morphable model for the synthesis of 3D faces,*Computer Graphics and Interactive Techniques*, 1999.

[3] S. Canavan, Y. Sun, X. Zhang, and L. Yin, A dynamic curvature-based approach for facial activity analysis in 3D space, *Computer Vision and Pattern Recognition Workshops*, 2012.

[4] S. Canavan and L. Yin, Fitting and Tracking 3D/4D Facial data Using a Temporal Deformable Shape Model,*International Conference on Multimedia and Expo*, 2013.

[5] S. Canavan, P. Liu, X. Zhang, and L. Yin, Landmark localization on 3D/4D range data using a shape index-based statistical shape model with global and local constraints, *Computer Vision and Image Understanding*, 139, pp. 136-148, 2015.

[6] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham, Active shape models their training and application, *Computer Vision and Image Understanding*, 61(1): 38-59, 1995.

[7] T.F. Cootes, G.J. Edwards, and C.J. Taylor, Active appearance models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6): 681-685, 2001.

[8] A. Danelakis, T. Theoharis, I. Pratikakis, and P. Perakis, An effective methodology for dynamic 3D facial expression retrieval, *Pattern Recognition*, 52, pp. 174-185, 2016.

[9] Z. Dong, X. Jia, W. Gao, and K. Wang, Training on statistical models of action units for 3D facial expression recognition, *Proceedings of Science*, 2017.

[10] H. Drira, B. Amor, M. Daoudi, A. Srivastava, and S. Berretti, 3D dynamic expression recognition based on a novel deformation vector field and random forest, *International Conference on Pattern Recognition*, 2012.

[11] D. Fabiano and S. Canavan. Spontaneous and non-spontaneous 3D facial expression recognition using a statistical model with global and local constraints, *International Conference on Image Processing*, 2018.

[12] T. Fang, X. Zhao, O. Ocegueda, S. Shah, and I. Kakadiaris, 3D/4D facial expression analysis: an advanced annotated face model approach, *Image and Vision Computing*, 30(10): 738-749, 2012.

[13] O. Gupta, D. Raviv, and R. Raskar, Illumination invariants in deep video expression recognition, *Pattern Recognition*, 76: 25-35, 2018.

[14] W. Hariri, H. Tabia, N. Farah, A. Benouareth, and D. Declercq, 3D facial expression recognition using kernel methods on Riemannian manifold, *Engineering Applications of Artificial Intelligence*, 64, pp. 25-32, 2017.

[15] T. Hawkins, Cauchy and the spectral theory of matrices, *Historia Mathematica*, 2(1): 1-29, 1975.

[16] G. Hinton, S. Osindero, and Y-W. Teh, A fast learning algorithm for deep belief nets, *Neural Computation*, 18(7): 1527-1544, 2006.

[17] I. Kakadiaris, G. Passalis, G. Toderici, M. Murtuza, Y. Lu, N. Karampatziakis, and T. Theoharis, Three-dimensional face recognition in the presence of facial expression: an annotated deformable model approach, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4): 640-649, 2007.

[18] H. Li, S. Jian, X. Zongben, and C. Liming. Multimodal 2D+3D facial expression recognition with deep fusion convolutional neural network, *IEEE Transactions on Multimedia*, 19(12):2816-2831, 2017.

[19] F. Liu, Z. Zhou, H. Jang, A. Samsonov, G. Zhao, and R. Kijowski, Deep convolutional neural network and 3D deformable approach for tissue segmentation in musculoskeletal magnetic resonance imaging, *Magnetic Resonance in Medicine*, 79(4): 2379-2391, 2018.

[20] P. Liu, S. Han, Z. Meng, and Y. Tong, Facial expression recognition via a boosted deep belief network, *Computer Vision and Pattern Recognition*, 2014.

[21] M. Mizra, and S. Osindro, Conditional generative adversarial nets, *arXiv preprint arXiv*, 1411.1784, 2014.

[22] R. W. Picard, E. Vyzas, and J. Healey, Toward machine emotional intelligence: Analysis of affective physiological state, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10):1175-1191, 2001.

[23] Y. Tie and L. Guan, A deformable 3-d facial expression model for dynamic human emotional state recognition, *IEEE Transactions on Circuits and Systems for Video Technology*, 23(1): 142-157, 2003.

[24] M. Valstar, J. Girard, et al., FERA 15 2nd facial expression recognition and analysis challenge, *Face and Gesture*, 2015.

[25] M. Valstar, E. S.-Lozano, et al., FERA 2017 Addressing head pose in the third facial expression recognition and analysis challenge, *Face and Gesture*, 2017.

[26] H. Yang, and L. Yin, CNN based 3D facial expression recognition using masking and landmark features, *Affective Computing and Intelligent Interaction*, 2017.

[27] H. Yang, U. Ciftci, and L. Yin, Facial expression recognition by de-expression residue learning, *Computer Vision and Pattern Recognition*, 2018.

[28] H. Yang, Z. Zhang, and L. Yin, Identity-adaptive facial expression recognition through expression regreneration using conditional generative adversarial networks, *Face and Gesture Recognition*, 2018.

[29] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, A high-resolution 3D dynamic facial expression database, *Face and Gesture Recognition*, 2008.

[30] X. Zhang, L. Yin, J. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. Girard, BP4D-spontaneous: A high resolution 3D dynamic facial expression database, *Image and Vision Computing*, 32(10): 692-706, 2014.

[31] Z. Zhang, J. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, Multimodal spontaneous emotion corpus for human behavior analysis, *Computer Vision and Pattern Recognition*, 2016.