3D Facial Expression Recognition Based on Multi-View and Prior Knowledge Fusion

Quang Nhat Vo
Center for Machine Vision and Signal Analysis
University of Oulu
Oulu, Finland
nhat.vo@oulu.fi

Khanh Tran
Center for Machine Vision and Signal Analysis
University of Oulu
Oulu, Finland
khanh.tran@oulu.fi

Guoying Zhao*
Center for Machine Vision and Signal Analysis
University of Oulu
Oulu, Finland
guoying.zhao@oulu.fi

***Abstract*—This paper presents a novel multi-view convolutional neural network (CNN) model for 3D facial expression recognition (FER). In contrast to existing deep learning-based 3D FER approaches that mainly learn the expressions from frontal facial attribute images, the proposed model incorporates multi-view and facial prior information of the observed 3D face into the learning process. This information is jointly trained in an end-to-end manner to predict the emotion of the input 3D face model. The experiments on public 3D facial expression datasets show that training the CNN with additional information from different views and facial prior knowledge would result in learning more discriminative features as against from a single view. Our model outperforms the state-of-the-art 3D FER methods in term of recognition accuracy indicating its effectiveness. Moreover, the improvement of the proposed model is displayed more clearly in the discrimination of lowintensity facial expressions.**

Keywords*— *convolutional neural network, facial expression recognition (FER), 3D face scan

I. Introduction

The facial expression, as one of the nonverbal communications, is the movements and location of facial muscles under the skin. According to several studies [1, 2], these motions carry the emotional state of an observed person. The analysis and measurement of facial expression are essential in understanding the thinking of other people and help us delivering corresponding responses. The capability of automatic facial expression recognition (FER) leads to many exciting applications in human-computer interaction, facial animation, and medicine [3-5]. It has been an active research topic in the fields of affective computing and computer vision over the past decades.

The common data modality for most of the research in FER is 2D face images or videos [6]. Although significant improvements have been made, FER methods on 2D images are still facing challenging problems of illumination and pose variations [7]. Due to difficulties in 2D FER, the research community starts to pay attention to 3D FER. There has been an emergence of several 3D facial expression datasets [8-10] and algorithms for reconstructing the 3D face model from 2D images [11, 12]. The benefit mainly comes from the inherent characteristics of 3D face scans that make it more robust to lighting and pose variations. Besides, additional 3D geometry information may include important features for FER. Research on static 3D FER mainly focuses on how to exploit the 3D information of the 3D face scans. They can be practically divided into three main approaches as follows.

The first group extracts 3D features at landmark or patch locations. For example, Berretti et al. [13] propose to use an original solution for computing SIFT descriptors on a set of facial landmarks of depth images and classify the selected features with SVM. Li et al. [14] compute the distance, slope, and angle between facial feature points and classify the facial expressions by probabilistic Neutral Network. Regarding patch-based features, Maalej et al. [15] propose a local geometric shape analysis of facial surfaces coupled with multi-boosting and SVM for expression classification. In another work [16], the face is clustered into several regions based on their importance into the facial expression process. Then those regions are matched to reference models using Iterative Closest Points. The main drawback of these approaches is that their performance depends heavily on the accuracy of 3D facial landmark detection and region clustering, which are challenging tasks.

The second one employs the morphable models to get the one-to-one point correspondence among face scans. Methods in this group try to align an input 3D facial mesh with an intermediate 3D model. Then, extracted features could be the differences between the aligned 3D model and the intermediate 3D model [17], or the features could be the 3D information derived from the aligned 3D mesh [18], such as vertex normal, vertex coordinates, and local curvature. However, these methods require an accurate establishment of the dense correspondence among face models, which is also a complicated process.

The third one utilizes the 2D representation of the 3D face scans. The advantage of these methods is that they can reuse traditional solutions in 2D FER for 3D FER. For example, Zernike moments along with the 3D point clouds and the depth images are combined to overcome problems arising in facial expression recognition such as translation, rotation, and scaling [19]. In recent works, deep features have been applied for learning 2D images of 3D face scans and deliver promising results. Several types of 2D facial attribute maps are generated for the training. Li et al. [20] describe a 3D face scan by six types of 2D facial attribute maps. Then, these maps are jointly fed into a deep convolutional neural network (CNN) for feature learning and fusion learning. In another approach, Oyedotun et al. [21] train the CNN with RGB texture and depth images. Overall, the methods in this group can yield better results than the above two groups. However, by considering only the frontal view of the 3D face in the generation of 2D facial attribute maps, they do not fully exploit the information of the 3D model. Moreover, the facial prior knowledge which might be useful is not utilized for facial expression learning. For fulfilling the gaps of previous studies, we introduce a new 3D-FER method in this paper. The main contributions of this paper are as follow.

- We propose a multi-view CNN architecture for 3D FER that jointly learns the 2D RGB texture and depth images synthesized from different views of a 3D face scan. The proposed multi-view CNN is different from existing CNN models [20, 21] that only consider the frontal view of the face. By extracting and fusing the features of facial expression in multiple views, our model can utilize more useful information and achieve a better recognition result. The target of our method is also different from other multi-view facial expression recognition works [22, 23] that more relate to the recognition of the facial expression at any viewpoint.

- The beneficial facial prior knowledge is incorporated for guiding the learning of the multi-view CNN model. By teaching the network to predict emotion-related facial areas, it can learn to extract facial-related features and reduce the nuisance factors in the input data. In contrast to Devries et al. method [24], our CNN model is trained to predict the facial maps instead of landmark locations. Furthermore, a fully convolutional network (FCN) [25] architecture is employed for the prediction. To the best of our knowledge, the fusion of multi-view deep features and facial prior knowledge have never been done before in the 3D FER.

The rest of this paper is organized as follows. The second section presents our proposed architectures and method. In the third section, our experiment setups and results are shown. Finally, the last section provides conclusions and discussions on future works.

II. Proposed Method

A. Multi-view attribute maps of a textured 3D face

There are two main approaches for modeling 3D data in the problem of 3D object recognition. The first one represents geometric 3D shapes as a probability distribution of binary variables on a 3D voxel grid [26]. The second one captures the information of 3D shapes as 2D images from different views [27]. Considering that minor and weak facial expressions may be lost in the voxel-based representation, the second approach is a better option for analyzing the emotion of the 3D face model. In this paper, one frontal view and two side views are employed to project the facial expression of a 3D face model. Besides the frontal view of the face, the expression maps from two side views of the 3D face model are acquired by performing the yaw rotation with the rotated angle of 30 and -30.

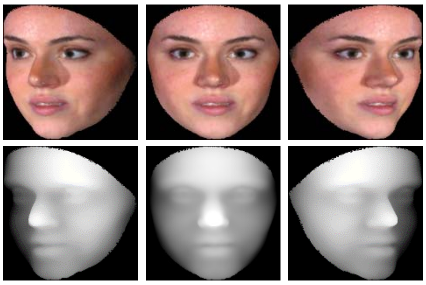


Fig. 1. Multi-view expression maps of a subject with the happy emotion in the BU-3DFE dataset.

Most of the 3D facial expression datasets such as BU-3DFE [8] and Bosphorus [28] offer both the 3D mesh and corresponding RGB value of each 3D vertex. In this paper, the selected expression maps are depth map images and RGB texture images synthesized from the 3D mesh and the related texture information. The value of the depth maps is normalized to the range of 0 to 255. All of the expression maps are fixed to the size of 244 x 244 for inputting to the proposed CNN model.

Although other feature maps such as surface normal and surface curvature [20] can also be used for learning the expression, we only employ the depth maps and RGB images due to two reasons. The first is about the training time and the required memory for storing the network parameters. Using additional feature maps will usually require a larger CNN structure in order to fit the training data. The second is based on our observation, which is also reported in the work of Oyedotun et al. [21] that presents high recognition results by training a CNN model on the only depth and RGB texture images.

Figure 1 presents a sample of multi-view expression maps extracted from a subject in the BU-3DFE dataset. This sample shows a happy emotion with low-level of expression intensity, which is quite similar to the neutral or surprise expression by looking only at the frontal view. As we can see, the projection from the side view can give us a different perspective of the expression (for example, in the cheek area near the mouth corner). It can provide more information for the learning model and improve its discrimination ability.

B. Proposed network architecture

1) Multi-view CNN for 3D facial expression recognition

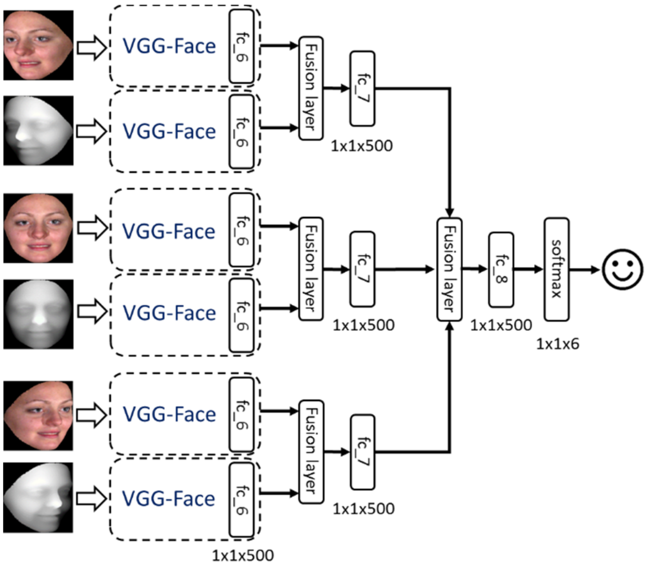


Fig. 2. Multiview CNN architecture for 3D facial expression recognition.

The proposed learning model is constructed on the hypothesis that more useful information for the emotion analysis can be obtained by observing an affective face in different viewpoints. In order to extract and take advantage of the facial information from multiple views, we design a three-stream CNN architecture for learning jointly from the depth maps and RGB texture images of three facial views. As presented in Figure 2, the network structure is formed by a set of feature extraction subnets and a feature fusion subnet. For the network training, since there are quite limited numbers of scanned 3D face meshes, convolutional layers from feature extraction subnets are initialized using pre-trained VGG-Face deep model [29]. In these subnets, there are in total six convolutional blocks for six facial attribute maps captured

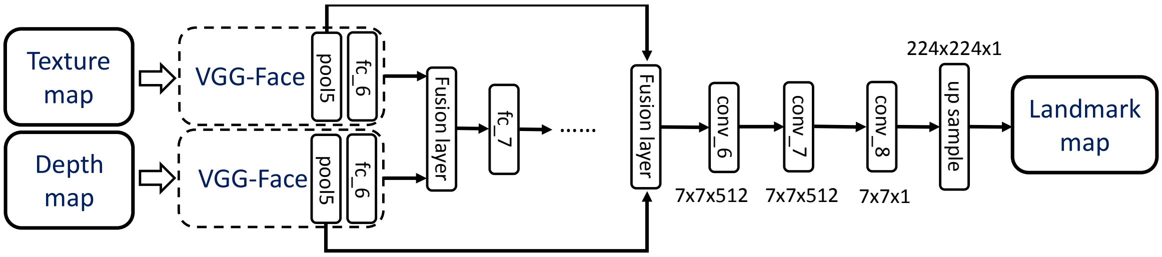


Fig. 3. The network architecture for emotion-related facial area prediction. This subnet is implemented for each view in the proposed Multi-view CNN architecture.

from three views. Each convolutional block is connected to a fully connected layer (fc_6) to outputs a feature vector of the corresponding attribute map.

The role of the feature fusion subnet is to construct a high-level feature vector for the final classification based on extracted features from each view. Instead of fusing all features into one vector from the beginning, we propose to use two levels of feature fusion. The first fusion level concatenates the feature vectors of attribute maps in the same view. Visual cues of each view are then compacted into one vector by the fc_7 layer. In the second level, the same mechanism is performed for multi-view fusion by using fc_8 layer. We observe that this hierarchical fusion strategy results in better performance compared with the straightforward concatenation of all feature vectors. It helps that the expression information of all views can be preserved and can contribute to the final classification.

Finally, the output layer's softmax generates a vector of probabilities of dimension 1 × 6 for six types of prototypical expression. The cost function for expression prediction is cross-entropy loss:

$$L_p = -\log\left(\frac{e^{f_i}}{\sum_j e^{f_j}}\right), \tag{1}$$

where f_i means the i-th element of the vector of class scores f and i is the index of the ground truth emotion class.

2) Learning with attention using facial prior knowledge

Not all information on the face is useful for emotion recognition. The facial attribute maps may contain some areas unrelated to the facial expression such as face regions near the background in the side view. According to Wegrzyn et al. [30], people were mostly relying on the eye and mouth regions when successfully recognizing an emotion. To limit the nuisance factors and guide the network to extract emotion-related features, inspired by Devries et al. [24], we incorporate the facial prior information to the training process in the manner of multi-task learning. In particular, feature extraction subnets in each view are connected to a fully convolutional network (FCN) [25] to predict a facial map representing the area of attention in input images. The used architecture is shown in Figure 3. It includes three convolutional layers for processing the feature maps of feature extraction subnets and a deconvolutional layer for generating the final binary map. The filter size of 3×3 is used for all three convolutional layers.

The ground truth for training is a binary mask image that has the same size of the input images and is created from provided 3D landmark locations. Figure 4 demonstrates some samples of the label maps from different views, which are the projection of 3D landmark points to the 2D image. We additionally perform the dilation morphology operator on the 2D binary map to highlight the projected points. The landmark points belonging to the face boundary or being occluded by facial parts are also omitted.

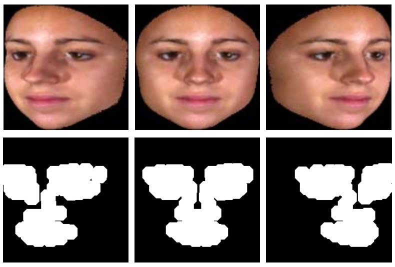


Fig. 4. Ground truth label maps from multiple views and the corresponding RGB texture images.

The objective function is the class-balanced cross-entropy loss, as follows:

$$L_l^k = -\beta \sum_{j\in Y_+} \log\left(\sigma(a_j)\right) - (1-\beta) \sum_{j\in Y_-} \log\left(1-\sigma(a_j)\right), \tag{2}$$

where k represents the k-th viewpoints, Y_+ and Y_- are the foreground and background label map, respectively, and β is a balance parameter to control the loss of the foreground/background classes, $\beta = |Y_-|/|\mathrm{Y}|$ and $1-\beta = |Y_+|/|\mathrm{Y}|$. $\sigma(a_i)$ is the sigmoid function on the activation value a_j at pixel j.

The final loss function in our proposed multi-task system, which is minimized by the standard stochastic-gradient descent, is defined as follow:

$$L = L_p + \frac{1}{N}\sum_{k=1}^{N} L_l^k, \tag{3}$$

where N is the number of viewpoints used in the training (N = 3 in our system). During the training, we set the learning rate to 1e-6, and train the network for a maximum of 1000 epochs. For the optimization algorithm, Adam optimizer [31] is employed to update the parameters of the network.

III. Experiments

A. Dataset and experimental protocol

Dataset. For validating the proposed learning system, we conduct the experiments on two public 3D FER datasets, BU-3DFE [8] and Bosphorus [28]. We apply the similar preprocessing method proposed in [20] for both two datasets. The BU-3DFE dataset contains 100 subjects (56 females and 44 males) with six types of expression (anger, disgust, fear, happiness, sadness, and surprise). Each expression has four levels of intensity, and based on that the dataset is usually divided into two subsets. The first subset (Subset I), which is the standard dataset used for 3D FER, includes expressions with two higher levels of expression intensity. The second subset (Subset II), which is seldom applied for the 3D FER, contains all four levels of intensity except the 100 neutral samples. From the first subset, we extract 1200 sets of 2D facial attribute maps (100 subjects with six prototypical expressions and two higher levels of expression intensity). Each set includes three RGB texture images and three depth images captured from three different viewpoints. We also try augmenting the data by horizontally flipping the 2D images. In the end, we created 2400 sets of 2D facial attribute maps. By doing the same process with Subset II, we gather 4800 sets of 2D facial attribute maps. With the Bosphorus dataset, only 65 subjects perform the six prototypical expressions, and each person only has one sample for each expression. However, to better partition, only 60 subjects are selected for the experiments. From this dataset, 720 sets of 2D training and testing images are generated.

Experimental protocol. Regarding the performance evaluation on Subset I, we follow the standard protocol as in the previous works [20, 21], that assigns 40 subjects to the validation and 60 subjects to the training and testing (54-versus-6-subject-partition experiments). As for Subset II and Bosphorus dataset, we apply 10-fold cross-validation training scheme as in [20] for the training and testing. The proposed CNN network is trained on a PC platform with a 3.2 GHz 2-core i7 CPU, 32 GB memory, and single NVIDIA Titan V. The system takes approximately 18 milliseconds to predict the expression for one sample.

B. Comparisions with state-of-the-arts

Results on BU-3DFE Subset I. We make the comparisons with the state-of-the-art handcrafted and deep learning based methods. We also present the result of our multi-view CNN (Multi-view CNN) with and without using the facial prior knowledge. Table I presents the average expression recognition accuracies of our model and related approaches. From Table I, we can see that, despite lacking 3D face scans with expression labels, deep learning based methods with fine-tuning strategies still outperform the traditional handcrafted-based ones. Another interesting observation is in the result of Oyedotun et al. [21]. By training only depth and RGB texture images, their network structure still shows a competitive performance compared to the method of Li et al. [20], which incorporates six types of attribute maps. Regarding our proposed method, we achieve a promising result with the multi-view CNN architecture. It shows that side views of 3D face model may provide more useful information for the FER. Table I also shows that the incorporation of face prior knowledge helps the network to learn the expression better by focusing on important regions.

Results on BU-3DFE Subset II and Bosphorus. The average accuracies of related methods are presented in Table II. They are reproduced in Li et al. [20] since these two datasets have been rarely used for the 3D FER in the literature. Similar to the conclusions obtained on the BU-3DFE subset I, the analysis of multi-view facial attribute maps can further improve the performance of the 3D FER. On the BU-3DFE Subset II, the amount of improvement in the accuracy is higher than on the BU-3DFE Subset I. An explanation for this situation is that side views for low-intensity expressions could provide more helpful information than for high-intensity expressions. In the case of the Bosphorus dataset, the amount of performance enhancement is smaller since this dataset has more difficult emotion to recognize and does not contain many samples for the training.

TABLE I. COMPARISON OF THE AVERAGE ACCURACIES WITH HANDCRAFTED AND DEEP LEARNING BASED APPROACHES ON THE BU-3DFE SUBSET I

Methods	**Feature**	**Accuracy**
Li et al. [32]	normals, curv./hist.	82.01
Zhen et al. [33]	coordinates, normals, shape index	84.50
Yang et al. [34]	depth, normals, curv./scattering	84.80
Li et al. [35]	meshHOG/SIFT meshHOS/HSOG	86.32
Li et al. [20]	depth, normal, curv., RGB, maps, deep feature	86.86
Oyedotun et al. [21]	depth, RGB, deep feature	89.31
Multi-view CNN	multiview, depth, RGB, deep feature	89.68
Multi-view CNN (with prior)	**multiview, depth, RGB, deep feature**	**91.39**

TABLE II. COMPARISON OF THE AVERAGE ACCURACIES WITH HANDCRAFTED AND DEEP LEARNING BASED APPROACHES ON THE BU-3DFE SUBSET II AND BOSPHORUS DATASETS

Methods	**BU-3DFE Subset II**	**Bosphorus**
Li et al. [35]	80.42	79.72
Yang et al. [34]	80.46	77.50
Li et al. [20]	81.33	80.00
Multi-view CNN	83.54	81.94
Multi-view CNN (with prior)	**84.30**	**82.40**

C. Network structure analysis

Multi-view CNN vs Single-view CNN. In order to demonstrate the effectiveness of using multi-view attribute maps in the 3D FER, this section compares the proposed Multi-view CNN with a CNN architecture trained on single view attribute maps (Single-view CNN). These input attribute maps are captured from the frontal view of the 3D model. Similar to the Multi-view CNN, the Single-view CNN contains two feature extraction subnets, one for the texture map and one for the depth map. Extracted feature vectors from these two subnets are then concatenated and fed to two fully connected layers, as shown in Figure 5. Finally, the softmax layer is employed to predict emotion classes.

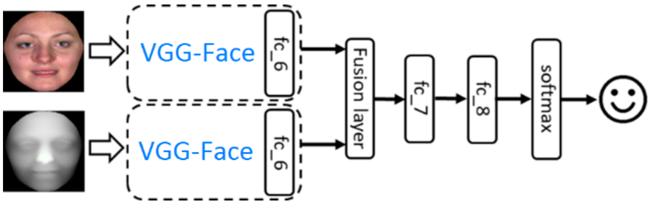
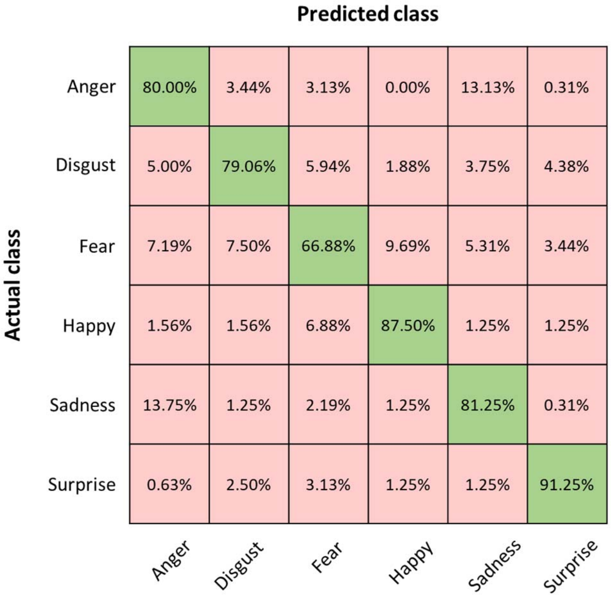


Fig. 5. The compared Single-view CNN architecture.

Insights into the recognition accuracy of our proposed Multi-view CNN versus Single-view CNN on two public 3D FER datasets are given in Table III. We can see that there is an improvement in the performance, especially on the BU-3DFE Subset II. The results of two BU-3DFE subsets demonstrate that the application of multi-view information can help to recognize samples with lower levels of expression intensity easier. We argue that the training can benefit in some way from using the multi-view information: 1) The multi-view attribute maps can provide more clues to recognize the low-intensity emotions. By considering only the frontal view, slight expressions in the 3D face models can hardly be seen on the 2D projected maps in some cases. A sample of this issue can be observed in Figure 1. We can see that the frontal view looks quite similar to the neutral or surprise expression while the emotion is expressed more clearly in the side view. 2) Some similar expressions can be distinguished more easily when looking at both the frontal view and the side view. Figure 6 presents the confusion matrix obtained for the Single-view CNN with the BU-3DFE Subset II. It shows high misclassification errors between two pairs of classes, namely (Anger, Sadness) and (Happy, Fear). As demonstrated in Figure 7, one of the reasons is due to similarities in the expression of these emotions on the frontal view that confuse the learning model. On the frontal view, happy and fear express the same movement on the face (look like smile). However, on side-view, the fear and happy expressions are quite different based on the shape of lip and fear expresses larger than happy emotion on the corner of the mouth. Additionally, the corner of the mouth in the happy is tapering than fear.

TABLE III. COMPARISON OF THE AVERAGE ACCURACIES ON DIFFERENT NETWORK ARCHITECTURES

Methods	BU-3DFE Subset I	BU-3DFE Subset II	Bosphorus
Single-view CNN	87.91	80.99	80.78
Multi-view CNN (without hierarchical fusion)	88.43	82.92	81.48
Multi-view CNN	**89.68**	**83.54**	**81.94**



Predicted class

Actual class	Anger	Disgust	Fear	Happy	Sadness	Surprise
Anger	80.00%	3.44%	3.13%	0.00%	13.13%	0.31%
Disgust	5.00%	79.06%	5.94%	1.88%	3.75%	4.38%
Fear	7.19%	7.50%	66.88%	9.69%	5.31%	3.44%
Happy	1.56%	1.56%	6.88%	87.50%	1.25%	1.25%
Sadness	13.75%	1.25%	2.19%	1.25%	81.25%	0.31%
Surprise	0.63%	2.50%	3.13%	1.25%	1.25%	91.25%

Fig. 6. Confusion matrix of recognizing six prototypic expressions on BU-3DFE Subset II by the Single-view CNN.

The effect of hierarchical fusion. To show the effectiveness of the hierarchical fusion, we compare the proposed Multi-view CNN structure with its variation that does not utilize the hierarchical fusion. Instead of using the multi-level feature concatenation, the compared network applies the straightforward concatenation of all feature vectors outputted from feature extraction subnets. The fused feature is then passed to two fully connected layers (similar to fc_7 and fc_8), and one softmax layer to predict the emotion classes. From Table III, we can see that the Multi-view CNN with the hierarchical fusion achieves better results than its variation. It can be explained that, by using the different levels of fusion, the network can create higher-level features at each view. Furthermore, the expression information of each view can be preserved until the end and contribute to the final classification. Another observation is that the use of multi-view attribute maps in both of multi-view CNN structures produces the higher performance than the Single-view CNN.

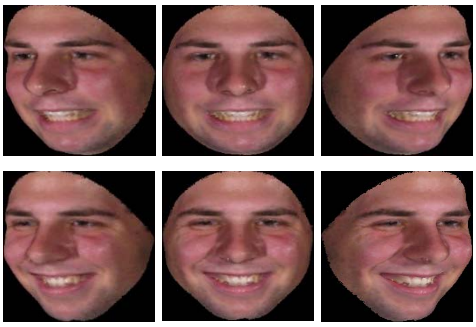


Fig. 7. Example of fear emotion (top) and happy emotion (bottom) in multi-view. On the frontal view, these two emotions look similar to smile, while they are different in side-view.

IV. CONCLUSION

In this paper, we proposed a novel CNN model for 3D FER by jointly training 2D facial attribute maps from different views and incorporating the facial prior knowledge to guide the learning process. Following the experimental results on two public datasets, our method presents promising results compared with the state-of-the-art studies. Nevertheless, there are still rooms for improvements and exploration in future work. In the next research, the importance of each view for the recognition will be studied. We are also going to extend our method to 4D data (3D and temporal) and investigate the data augmentation.

ACKNOWLEDGEMENT

This work was supported by Infotech Oulu, the National Natural Science Foundation of China (No. 61772419), Tekes Fidipro Program (No. 1849/31/2015), Business Finland Project (No. 3116/31/2017), and Academy of Finland ICT 2023 project (313600). Moreover, the authors wish to acknowledge CSC-IT Center for Science, Finland, for computational resources. Furthermore, we gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPU used for this research.

REFERENCES

[1] A.J. Fridlund, Human facial expression (1 ed.), Academic Press, 1994.

[2] J. A. Russell and J. M. Fernandez Dols, *The psychology of facial expression* (*1 ed.*). Cambridge University Press, 1997.

[3] R.A. Calix, S.A. Mallepudi, B. Chen, and G.M. Knapp, "Emotion recognition in text for 3-d facial expression rendering," *IEEE Transaction on Multimedia*, vol. 12, no. 6, pp. 544–551, Oct. 2010.

[4] M. Yeasin, B. Bullot, and R. Sharma, "Recognition of facial expressions and measurement of levels of interest from video," *IEEE Transaction on Multimedia*, vol. 8, no. 3, pp. 500–508, Jun. 2006.

[5] G. Muhammad, M. Alsulaiman, S.U. Amin, A. Ghoneim, and M.F. Alhamid, "A Facial-Expression Monitoring System for Improved Healthcare in Smart Cities," *IEEE Access*, vol. 5, pp. 10871–10881, 2017.

[6] C.A. Corneanu, M. Oliu, J.F. Cohn, and S. Escalera, "Survey on RGB, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1548–1568, Aug. 2016.

[7] M. Pantic and L.J.M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Transactions on Pattern*

Analysis and Machine Intelligence, vol. 22, no. 12, pp. 1424–1445, Dec. 2000.

[8] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, “A 3D facial expression database for facial behavior research,” *IEEE International Conference on Automatic Face Gesture Recognition*, Apr. 2006, pp. 211–216.

[9] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, “A highresolution 3D dynamic facial expression database,” *IEEE International Conference on Automatic Face and Gesture Recognition*, Sept. 2008, pp. 1–6.

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

[11] J. Booth, E. Antonakos, S. Ploumpis, G. Trigeorgis, Y. Panagakis, and S. Zafeiriou, “3D Face Morphable Models In-the-Wild,” *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5464-5473.

[12] A.T. Tran, T. Hassner, I. Masi, E. Paz, Y. Nirkin, and G. Medioni, “Extreme 3D Face Reconstruction: Seeing Through Occlusions,” *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3935-3944.

[13] S. Berretti, A.D. Bimbo, and P. Pala, “A Set of Selected SIFT Features for 3D Facial Expression Recognition,” *IEEE International Conference on Pattern Recognition*, 2010, pp. 4125–4128.

[14] X. Li, Q. Ruan, Y. Ming, “3D Facial Expression Recognition Based on Basic Geometric Features,” *IEEE International Conference on Signal Processing Proceedings*, 2010, pp. 1366–1369.

[15] A. Maalej, B. Amor, M. Daoudi, A. Srivastava, and S. Berretti, “Local 3D shape analysis for facial expression recognition,” *IEEE International Conference on Pattern Recognition*, 2010, pp. 4129–4132.

[16] P. Lemaire, B. Ben Amor, M. Ardabilian, L. Chen, and M. Daoudi, “Fully automatic 3D facial expression recognition using a region-based approach,” *ACM Workshop on Human Gesture and Behavior Understanding*, 2011, pp. 53–58.

[17] B. Gong, Y. Wang, J. Liu, and X. Tang, “Automatic facial expression recognition on a single 3D face by exploring shape deformation,” *Proceedings of the Seventeen ACM International Conference on Multimedia*, 2009, pp. 569–572.

[18] O. Ocegueda, T. Fang, S. Shah, and I. Kakadiaris, “Expressive maps for 3D facial expression recognition,” *ICCV Workshops*, 2011, pp. 1270–1275.

[19] N. Vretos, N. Nikolaidis, I. Pitas, “3D facial expression recognition using Zernike moments on depth images,” *IEEE International Conference on Image Processing*, 2011, pp. 773-776.

[20] H. Li, J. Sun, Z. Xu, and L. Chen, “Multimodal 2D+3D Facial Expression Recognition With Deep Fusion Convolutional Neural Network,” *IEEE Transactions on Multimedia*, vol. 19, no. 12, pp. 2816–2831, 2017.

[21] O.K. Oyedotun, G.G. Demisse, A.E.R. Shabayek, D. Aouada, and B.E. Ottersten, “Facial Expression Recognition via Joint Deep Learning of RGB-Depth Map Latent Representations,” *ICCV Workshops*, 2017, pp. 3161–3168.

[22] Y. Hu, Z. Zeng, L. Yin, X. Wei, X. Zhou, and T.S. Huang, “Multi-view facial expression recognition,” *IEEE International Conference on Automatic Face and Gesture Recognition*, 2008, pp. 1–6.

[23] X. Huang, G. Zhao, and M. Pietikäinen, ”Emotion recognition from facial images with arbitrary views,” *British Machine Vision Conference*, 2013, pp. 76.1–76.11.

[24] T. Devries, K. Biswaranjan, and G.W. Taylor, “Multi-task learning of facial landmarks and expression,” *Canadian Conference on Computer and Robot Vision* (*CRV*) , 2014, pp. 98-103.

[25] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431-3440.

[26] N. Sedaghat, M. Zolfaghari, E. Amiri, and T. Brox, “Orientation-boosted voxel nets for 3D object recognition,” *arXiv* preprint arXiv:1604.03351, 2016.

[27] H. Su, S. Maji, E. Kalogerakis, E.L. Miller, “Multi-view Convolutional Neural Networks for 3D Shape Recognition,” *IEEE International Conference on Computer Vision*, 2015, pp. 945-953.

[28] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, “Bosphorus database for 3D face analysis,” *European Workshop on Biometrics and Identity Management*, 2008, pp. 47–56.

[29] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep Face Recognition,” *British Machine Vision Conference*, 2015, pp. 41.1–41.12.

[30] M. Wegrzyn, M. Vogt, B. Kireclioglu, J. Schneider, and J. Kissler, “Mapping the emotional face. How individual face parts contribute to successful emotion recognition,” *PloS one*, 12(5), pp. e0177239, 2017.

[31] D.P. Kingma, J. Ba, “Adam: A method for stochastic optimization,” *arXiv* preprint arXiv:1412.6980, 2014.

[32] H. Li, J.-M. Morvan, and L. Chen, “3D facial expression recognition based on histograms of surface differential quantities,” in *International Conference on Advanced Concepts for Intelligent Vision Systems*, 2011, pp. 483–494.

[33] Q. Zhen, D. Huang, Y. Wang, and L. Chen, “Muscular movement model based automatic 3D facial expression recognition,” *International Conference on MultiMedia Modeling*, 2015, pp. 522–533.

[34] X. Yang, D. Huang, Y. Wang, and L. Chen, “Automatic 3D facial expression recognition using geometric scattering representation,” *IEEE International Conference on Automatic Face and Gesture Recognition*, 2015, vol. 1, pp. 1–6.

[35] 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*, vol. 140, pp. 83-92, 2015.