Facial Data Visualization for Improved Deep Learning Based Emotion Recognition

Seung Ho Lee*

Department of Future Technology, Korea University of Technology and Education, Cheonan, Korea
E-mail: leesh903@koreatech.ac.kr

ABSTRACT

A convolutional neural network (CNN) has been widely used in facial expression recognition (FER) because it can automatically learn discriminative appearance features from an expression image. To make full use of its discriminating capability, this paper suggests a simple but effective method for CNN based FER. Specifically, instead of an original expression image that contains facial appearance only, the expression image with facial geometry visualization is used as input to CNN. In this way, geometric and appearance features could be simultaneously learned, making CNN more discriminative for FER. A simple CNN extension is also presented in this paper, aiming to utilize geometric expression change derived from an expression image sequence. Experimental results on two public datasets (CK+ and MMI) show that CNN using facial geometry visualization clearly outperforms the conventional CNN using facial appearance only.

Keywords: facial expression recognition, convolutional neural network, facial landmark points, facial geometry visualization

1. INTRODUCTION

Data science has emerged as a core area of information science in recent years. Data science uses scientific algorithms or systems to extract knowledge and insights from structured and unstructured data. Particularly, recent advances in computer vision and machine learning (e.g., deep learning) techniques can successfully reduce costs of analyzing large-scale visual data (e.g., image or video), which is one of the representative forms of unstructured data. A convolutional neural network (CNN) has been most successfully used among deep learning methods for visual data.

Face analysis can play an important role in visual data mining. Automatic face recognition (or face identification) is a mature and widely used approach, which allows face images to be organized by the identities of persons. For example, personal photos in social network services can be automatically labeled with the names of persons, which is called name tagging (Choi, Neve, Plataniotis, & Ro, 2011). Facial expression recognition (FER) has been increasingly important due to its emerging applications. Facial expression is one of the most natural and powerful tools for non-verbal human communications (Sandbach, Zafeiriou, Pantic, & Yin, 2012), which conveys emotional states, such as surprise, intention, and interest. Leveraging expressive signals would have many practical applications. For example, bunches of emotional response data of customers' faces to online video adverts can be unobtrusively collected (via smart phone camera or web camera) and analyzed for market research purposes. In addition, for improving road safety, driver state monitoring can be adopted in automobiles, which understands the driver's emotional states (e.g., anger that can negatively affect his/her driving) and moods in real time. Considering the practical importance, this paper focuses on improving deep learning (i.e., CNN) to more accurately predict the emotional state from an expression image.

A number of traditional FER methods extract hand-crafted appearance features to capture pixel intensity changes of facial expression images. These features include local binary patterns (Huang, Wang, & Ying, 2010), local phase quantization (Wang & Ying, 2012), two-dimensional principal component analysis (Yang, Zhang, Frangi, & Yang, 2004), color texture features (Lee, Kim, Ro, & Plataniotis, 2013), and so on. To classify expression features, many well-known classifiers such as the support vector machine (Bartlett et al., 2005) or sparse representation classifier (Wright, Yang, Ganesh, Sastry, & Ma, 2009) have been employed and evaluated.

More recently, due to dramatically increasing processing ability (e.g., Graphics Processing Unit processing), more powerful FER methods have been proposed based on deep learning techniques. Among various deep learning techniques, CNN (Lecun, Bottou, Bengio, & Haffner, 1998) has been perfectly designed to take expression image data as input and learn discriminative appearance features from the data.

Convolution layers in CNN automatically learn appearance features for FER. Specifically, during the training stage, the weight values of spatial filters (or kernels) are learned according to the input expression image and its ground truth (i.e., emotion class label). In the test stage, two dimensional feature maps (or activation maps) can be obtained by applying the learned spatial filters. These feature data are passed to fully connected layers in order to predict an emotion class of the input expression image.

Despite successful use of CNN based FER, one limitation is that most CNN based FER methods take the original expression image as input and extract appearance features only. However, it should be noted that geometric features are also important for recognizing facial expression, as pointed out in Chen et al. (2012). For extracting geometric features, landmark points or feature points need to be detected and processed (Kotsia & Pitas, 2007). Jung, Lee, Yim, Park, & Kim (2015) have attempted to incorporate geometric information into CNN based FER. In the method in Jung et al. (2015), while CNN takes expression images, another deep neural network takes the xy coordinates of the landmark points. The outputs of the two different deep neural networks are combined by the fine-tuning method proposed in Jung et al. (2015). However, this approach has difficulty analyzing the spatial relation between two-dimensional facial appearance (i.e., expression image) and one-dimensional facial geometry (i.e., landmark points). As a result, the effectiveness of fusing geometric and appearance information could be limited due to the different domains of analysis.

To make full use of the discriminating capability of CNN, this paper suggests incorporation of facial geometry visualization. Instead of an original expression image, the expression image with visualizations of landmark detection and processing is used for CNN based FER. The visualizations on an expression image have been experimentally investigated in this paper as to whether the process is suitable for improving the discriminating power of CNN. Note that facial expression change appears in a continuous video sequence (Lee & Ro, 2016), which is a usual input format to realistic FER applications. In order to derive facial dynamics from the sequence, a simple CNN extension is also presented in this paper. More specifically, CNN takes two channel input, i.e., an expressive image with the facial geometry visualization and a non-expressive (or neutral) image with the facial geometry visualization. Experimental results on two public datasets, CK+ (Lucey et al., 2010) and MMI (Pantic,

Valstar, Rademaker, & Maat, 2005), show that CNN using facial geometry visualization clearly outperforms the conventional CNN using only facial appearance in terms of recognition accuracy. About up to 5% of improvement has been achieved for the case of one channel input (i.e., an expression image). From the CNN extension to two channel input, additional improvement has been achieved. In addition, it has also been demonstrated that the proposed method (i.e., CNN based FER using facial geometry visualization) can be comparable with some recent advances in recognition accuracy.

The rest of this paper is summarized as follows. Section 2 presents details about the proposed method. Section 3 presents experimental results to demonstrate the effectiveness of the proposed method. Conclusions are drawn in Section 4.

2. PROPOSED METHOD

This section describes a method to improve the discriminating power of CNN based FER through facial geometry visualization. Section 2.1 presents the method for static expression images. A simple extension for expression image sequences is given in Section 2.2.

2.1. Discriminative CNN with Facial Geometry Visualization

This section focuses on describing how to make use of facial geometric information for CNN based FER. From a grayscale expression image of N×N pixels, 49 landmark points
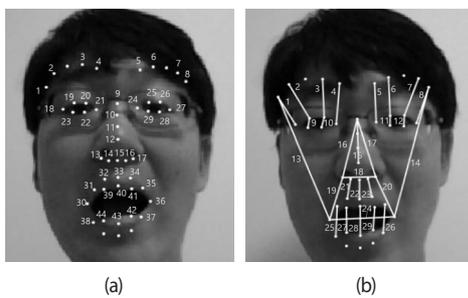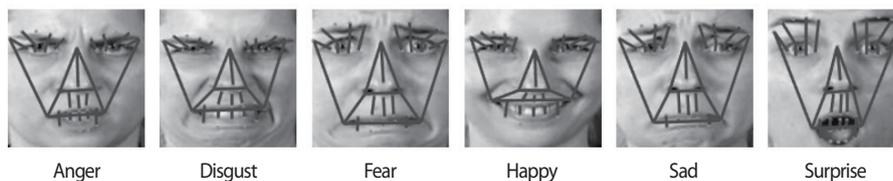
are extracted (Fig. 1a). For the automatic landmark detection, the method in Asthana, Zafeiriou, Cheng, & Pantic (2014) is adopted. To make the facial geometry more meaningful for FER, two landmark points are connected with a line as shown in Fig. 1b. For the connections, 44 out of 49 landmark points are used and a total of 29 connecting lines can be visualized. As explained in Lee and Ro (2016), the set of 29 connecting lines could be closely related to facial muscles. For example, the lengths of the connecting line No. 1-8 represent the distances between the brows and the corresponding eyelids. Hence, these connecting lines are related to Action Unit (AU) 1 (inner brow raiser), AU5 (upper lid raiser), AU7 (lid tightener), and so on (Lee & Ro, 2016).

Fig. 2 illustrates 29 connecting lines for six basic emotion classes (i.e., Anger, Disgust, Fear, Happy, Sad, and Surprise). One can see that the facial geometry visualizations (each of which is formed with the 29 connecting lines) look clearly different across the six emotion classes. It is expected that spatial relations between facial appearance and facial geometry could extract useful features for classifying emotional facial expressions. Therefore, the expression image with visualization of the 29 connecting lines is directly fed into CNN.

Fig. 3 illustrates the CNN model used in this paper. The processed expression image with facial geometry visualization is resized to 32×32 pixels. The first convolution layer employs 64 filters with sizes of 5×5. The second and third convolution layers employ 128 filters with sizes of 3×3, respectively. The three convolution layers produce 64, 128, and 128 feature maps, respectively. Next, use of a 2×2 max pooling layer is followed to reduce the spatial size of the feature maps and the computational cost of the network. Similar to conventional CNNs, a fully connected layer is included at the end of the network to classify emotion class. After the max pooling layer, each of the two-dimensional feature maps is converted into the one-dimensional feature maps which are suitable for the input to the fully connected layer. The output layer has six nodes (for six emotion classes in MMI) or seven nodes (for seven emotion classes in CK+). Through the output layer, an emotion class is predicted by finding the highest probabilistic score.

Fig. 1. Facial geometry visualization with landmark points. (a) Forty-nine landmark points detected. (b) Twenty-nine kinds of lines connecting two landmark points.

Fig. 2. Facial geometry visualization defined by 29 connecting lines.
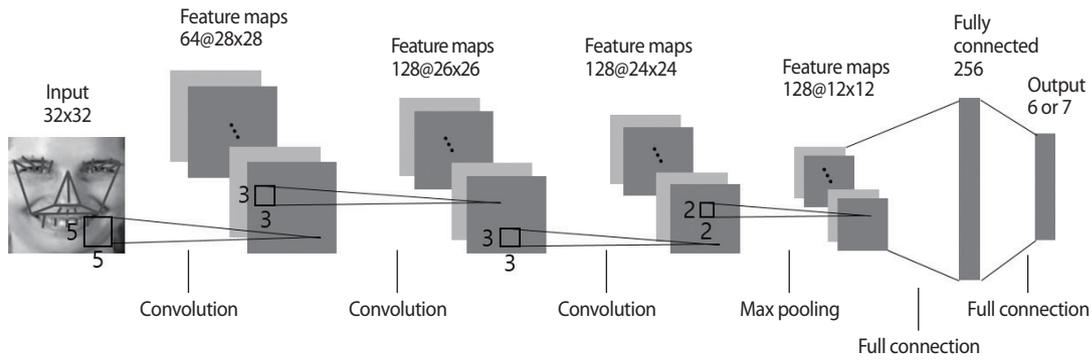
Fig. 3. Convolutional neural network model used in this paper.

2.2. Extension to Two Channel Input for Expression Image Sequence

In Section 2.1, CNN using one channel input is presented to classify a static expression image. Note that we may encounter sequences of facial expression where the face evolves from a neutral state to an emotional expressive state (Kotsia & Pitas, 2007). As extensively studied in several works (Chen et al., 2012; Kotsia & Pitas, 2007; Zafeiriou & Petrou, 2010; Donato, Stewart, Hager, Ekman, & Sejnowski, 1999), using both expressive state and neutral state can be useful to capture the expression change of a person. In this section, a simple CNN extension is described,
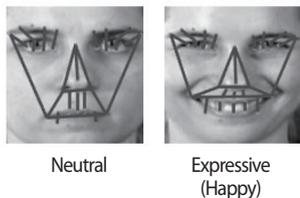
Fig. 4. Facial geometry visualizations for neutral image and expressive image.

which aims to capture more discriminative expression features from an expression image sequence.

Note that CNN has been basically designed to take a multichannel image as input. For example, CNN can take an RGB color image and learn discriminative features from three complementary color channels. Taking the structure of CNN into consideration, two channel input of the non-expressive (or neutral) image and expressive image is applied to the CNN model in Fig. 3. The neutral and expressive images are processed to include visualization of 29 connecting lines in a similar way to Section 2.1. The visualization result is shown in Fig. 4. The difference between the two is able to encode expression change information. An extension of CNN to two channel input is illustrated in Fig. 5. From the input and the first convolutional layer in Fig. 5, one can expect that appearance and geometry changes are learned considering both neutral and expressive states of a face. It should be noted that, for consistency, two channel input of neutral and expressive images is used in both the training stage and test stage.
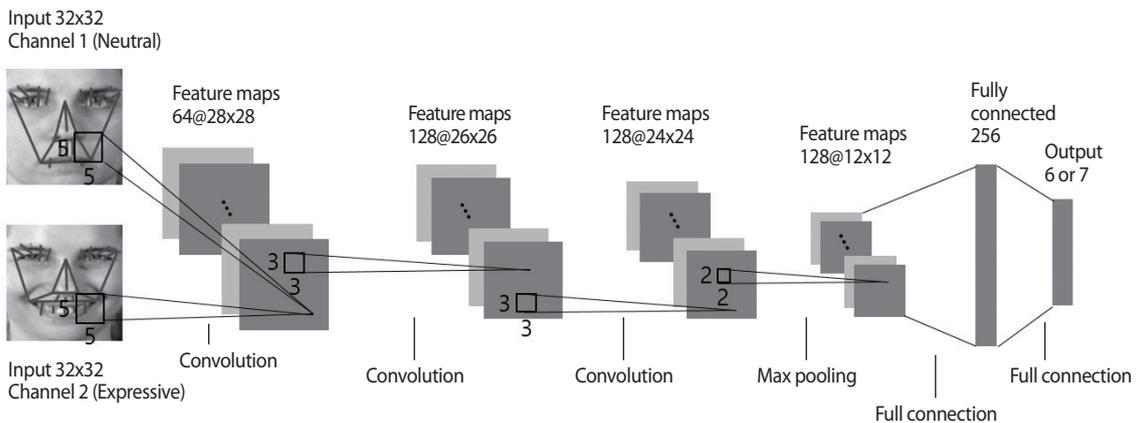
Fig. 5. Convolutional neural network extension to two channel input.

3. EXPERIMENT

3.1. Experimental Setup

The proposed method was evaluated with two public datasets, i.e., CK+ and MMI. Example expression images from the datasets are shown in Fig. 6. The dataset constructions for the experiment were performed as follows:

1. CK+ (Lucey et al., 2010): CK+ consists of 593 image sequences from 123 persons. From these, 325 video sequences of 118 persons were selected, which satisfied the criteria for one of the seven emotion classes. The selected 325 video sequences consisted of 45, 18, 58, 22, 69, 28, and 82 video sequences of Angry, Contempt, Disgust, Fear, Happy, Sadness, and Surprise, respectively. Ten-fold cross validation was used to measure recognition accuracy.

2. MMI (Pantic et al., 2005): 205 video sequences were collected from 30 persons. The dataset consisted of 31, 31, 27, 43, 32, and 41 video sequences of Anger, Disgust, Fear, Happy, Sad, and Surprise, respectively. Ten-fold cross validation was also used for this dataset.

For both datasets, from each video sequence, a pair of neutral frame and peak expression frames (i.e., most expressive frame) was manually selected by the author. From each frame, facial region was detected by using Viola Jones algorithm (Viola & Jones, 2004). The detected facial region was aligned based on two eye locations and cropped resulting in an expression image. To compute two eye locations, 49 facial landmark points were detected using the method in Asthana et al. (2014). The coordinates of the left eye and the right eye were obtained by averaging those of the facial landmark no. 18-23 and no. 24-29 (Fig. 1a), respectively. Forty-four out of 49 landmark points were used to visualize facial geometry with 29 connecting lines. The expression images with visualization of 29 connecting lines were resized to 32×32 pixels.

The CNN models in Fig. 3 and Fig. 5 were implemented using Python ver. 3.5.4 and an open-source deep learning library called Keras. As activation functions, Rectified Linear Unit (ReLU) and Softmax were adopted for the convolutional layers and the output layer, respectively. In order to learn the CNN model, the number of epochs and batch size were set to 45 and 30, respectively. In addition, Adam was selected as optimizer and its learning rate was set to 0.3.

3.2. Experimental Results

In this section, the effectiveness of using facial geometry visualization (i.e., visualization of 29 connecting lines on expression image) was investigated. For this purpose, four

different FER methods were defined as follows:

1. Peak without facial geometry: An expression image with the highest expression intensity in each video sequence was used as input for CNN based FER. The expression image did not include the visualization of 29 connecting lines.

2. Peak with facial geometry (proposed): An expression image with the highest expression intensity in each video sequence was used as input for CNN based FER. The expression image included the visualization of 29 connecting lines.

3. Peak/neutral without facial geometry: A pair of peak expression image and neutral expression image was used as input for CNN based FER. The expression images did not include the visualizations of 29 connecting lines.

4. Peak/neutral with facial geometry (proposed): A pair of peak expression image and neutral expression image was used as input for CNN based FER. The expression images included the visualizations of 29 connecting lines.

Table 1 shows comparisons of the four FER methods on CK+. From the comparison results, two observations can be made. First, compared to using only a peak expression image, using a pair of peak expression and neutral expression images yields improved recognition accuracies. This is mainly because a neutral state can be useful to capture the pure expression change (neutral to expressive) present in an expression image sequence. Second, regardless of using a neutral image, including the visualization of 29 connecting lines on an expression image is clearly better in recognition accuracy than using the expression image without the visualization. By incorporating facial geometry visualization, about 3% to 5% of improvements are achieved for CNN based FER. Table 2 shows the confusion matrix for 'Peak/Neutral with facial geometry' on CK+. It is shown that Fear and Sad expressions are often misclassified as Surprise and Angry, respectively.

Table 3 shows comparisons results of the four FER methods on MMI. Note that MMI is much more difficult to analyze than CK+ for the following reasons. First, different persons make the same emotional expression differently as shown in Fig. 6b. Second, some persons have accessories such as glasses or head cloths (Fig. 6b). Although the average recognition accuracy in Table 3 is relatively low (70.47%), visible improvements are made by using facial geometry visualization, while achieving up to 74.11%. Table 4 shows the confusion matrix for 'peak/neutral with facial geometry' on MMI. It is observed that recognizing Fear is very difficult as it is often confused with Surprise, similar to the case in CK+. It is also observed that Anger and Disgust are confused with each other.

Fig. 6. Datasets used in experiment. (a) CK+. (b) MMI.

Table 1. Comparison results on CK+

Facial expression recognition method	Recognition rate (%)
Peak without facial geometry	85.87
Peak with facial geometry	88.34
Peak/neutral without facial geometry	87.74
Peak/neutral with facial geometry	92.63

Table 2. Confusion matrix on CK+ obtained by using 'peak/neutral with facial geometry'

Actual	Predicted						
	Angry	Contempt	Disgust	Fear	Happy	Sad	Surprise
Angry	**91.11**	0.00	4.44	2.22	0.00	2.22	0.00
Contempt	5.56	**88.89**	0.00	0.00	0.00	5.56	0.00
Disgust	1.72	0.00	**96.55**	0.00	1.72	0.00	0.00
Fear	0.00	0.00	0.00	**76.00**	4.00	8.00	12.00
Happy	0.00	0.00	0.00	0.00	**100.00**	0.00	0.00
Sad	10.71	0.00	0.00	7.14	0.00	**82.14**	0.00
Surprise	1.22	0.00	0.00	1.22	0.00	3.66	**93.90**

Table 3. Comparison results on MMI

Facial expression recognition method	Recognition rate (%)
Peak without facial geometry	67.77
Peak with facial geometry	69.29
Peak/neutral without facial geometry	70.72
Peak/neutral with facial geometry	74.11

Table 4. Confusion matrix on MMI obtained by using 'Peak/neutral with facial geometry'

Actual	Predicted					
	Anger	Disgust	Fear	Happy	Sad	Surprise
Anger	**67.74**	19.35	3.23	0.00	6.45	3.23
Disgust	19.35	**58.06**	6.45	6.45	6.45	3.23
Fear	7.41	3.70	**55.56**	3.70	7.41	22.22
Happy	4.65	0.00	2.33	**90.70**	2.33	0.00
Sad	9.38	6.25	6.25	3.13	**71.88**	3.13
Surprise	4.88	0.00	7.32	0.00	0.00	**87.80**

Table 5. Comparisons with recent advances in FER on MMI

FER method	Cross validation	Recognition rate (%)
Proposed	10-fold	74.11
GAN (Yang et al., 2018)	10-fold	73.23
Identity-Aware CNN (Meng et al., 2017)	10-fold	71.55
Partial matching (Lee & Ro, 2016)	10-fold	72.10
DTAGN-Joint (Jung et al., 2015)	10-fold	70.24
HMM (Weng, Wang, & Ji, 2013)	15-fold	51.50
ITBN (Weng et al., 2013)	15-fold	59.70

FER, acial expression recognition; GAN, generative adversarial network; CNN, convolutional neural network; DTAGN, deep temporal appearance-geometry network; HMM, hidden Markov model; ITBN, interval temporal Bayesian network.

3.3. Comparisons with Recent Advances

In this section, the proposed method (CNN based FER using facial geometry visualization) was compared with some recent advances in FER on MMI dataset under similar testing protocols (e.g., recognition of six emotion classes under 10- or 15-fold cross validation). Table 5 shows the comparison results. It is observed that the deep learning based methods (Jung et al., 2015; Yang, Ciftci, & Yin, 2018; Meng, Liu, Cai, Han, & Tong, 2017) achieve relatively high recognition accuracies ranging from 70.24 to 74.11. It should be noted that the proposed method is comparable with the DTAGN-Joint method (Jung et al., 2015) that also uses geometric information for deep neural networks. In the method (Jung et al., 2015), the detected facial landmark points are converted into one-dimensional data and fed into a deep neural network (fully connected layer). On the other hand, the proposed method directly processes the facial geometry and its original appearance as a single two-dimensional image data without any data conversion such as vectorization. Thus, the features learned via the convolutional layers could be two-dimensional and thus are more straightforward for visual emotion classification.

4. CONCLUSION

The ability to accurately recognize and interpret a person's facial expressions is a key to practical data science applications. For FER, CNN has been widely adopted because it can automatically learn discriminative appearance features from an expression image. For more accurate FER, this paper proposes a simple but effective method to incorporate geometric information into CNN. In the proposed method, instead of an original expression image that contains facial appearance only, the expression image with facial geometry visualization is used

as input to CNN. Spatial relation between facial appearance and facial geometry could make the learned expression features more discriminative. For future work, the various visualization methods of facial geometry will be studied for further improving the proposed approach.

Note that the proposed method is very simple and easy to implement because it does not need to change the structure of conventional CNN models. Thus, it is believed that it could be practically used for various applications including emotion mining from online video adverts or driver state monitoring for road safety.

REFERENCES

Asthana, A., Zafeiriou, S., Cheng, S., & Pantic, M. (2014). Incremental face alignment in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 23-28 June (pp. 1859-1866).

Bartlett, M. S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., & Movellan, J. (2005). Recognizing facial expression: Machine learning and application to spontaneous behavior. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 20-25 June (pp. 568-573). Piscataway: IEEE.

Chen, J., Chen, D., Gong, Y., Yu, M., Zhang, K., & Wang, L. (2012). Facial expression recognition using geometric and

appearance features. In *Proceedings of the 4th International Conference Internet Multimedia Computing and Service*, Wuhan, China (pp. 29-33). New York: ACM.

Choi, J. Y., Neve, W. D., Plataniotis, K. N., & Ro, Y. M. (2011). Collaborative face recognition for improved face annotation in personal photo collections shared on online social networks. *IEEE Transactions on Multimedia*, 13(1), 14-28.

Donato, G., Stewart, M. S., Hager, J. C., Ekman, P., & Sejnowski, T. J. (1999). Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10), 974-989.

Huang, M. W., Wang, Z. W., & Ying, Z. L. (2010). A new method for facial expression recognition based on sparse representation plus LBP. In *2010 3rd International Congress on Image and Signal Processing*, Yantai, China, 16-18 October (Vol. 4, pp. 1750-1754). Piscataway: IEEE.

Jung, H., Lee, S., Yim, J., Park, S., & Kim, J. (2015). Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, 7-13 December (pp. 2983-2991).

Kotsia, I., & Pitas, I. (2007). Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Transactions on Image Processing*, 16(1), 172-187.

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.

Lee, S. H., Kim, H., Ro, Y. M., & Plataniotis, K. N. (2013). Using color texture sparsity for facial expression recognition. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Shanghai, China, 22-26 April (pp. 1-6). Piscataway: IEEE.

Lee, S. H., & Ro, Y. M. (2016). Partial matching of facial expression sequence using over-complete transition dictionary for emotion recognition. *IEEE Transactions on Affective Computing*, 7(4), 389-408.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, San Francisco, CA, USA, 13-18 June (pp. 94-101). Piscataway: IEEE.

Meng, Z., Liu, P., Cai, J., Han, S., & Tong, Y. (2017). Identity-aware convolutional neural network for facial expression recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition*, Washington, DC, USA, 30 May-3 June (pp. 558-565). Piscataway: IEEE.

Pantic, M., Valstar, M., Rademaker, R., & Maat, L. (2005). Web-based database for facial expression analysis. In *2005 IEEE International Conference on Multimedia and Expo*, Amsterdam, Netherlands, 6 July (p. 5). Piscataway: IEEE.

Sandbach, G., Zafeiriou, S., Pantic, M., & Yin, L. (2012). Static and dynamic 3D facial expression recognition: A comprehensive survey. *Image and Vision Computing*, 30(10), 683-697.

Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2), 137-154.

Wang, Z., & Ying, Z. (2012). Facial expression recognition based on local phase quantization and sparse representation. In *2012 8th International Conference on Natural Computation*, Chongqing, China, 29-31 May (pp. 222-225). Piscataway: IEEE.

Weng, Z., Wang, S., & Ji, Q. (2013). Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA, 23-28 June (pp. 3422-3429).

Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., & Ma, Y. (2009). Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2), 210-227.

Yang, H., Ciftci, U., & Yin, L. (2018). Facial expression recognition by de-expression residue learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 18-23 June (pp. 2168-2177).

Yang, J., Zhang, D. D., Frangi, A. F., & Yang, J. Y. (2004). Two-dimensional PCA: A new approach to appearance-based face representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(1), 131-137.

Zafeiriou, S., & Petrou, M. (2010). Sparse representation for facial expressions recognition via l1 optimization. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, San Francisco, CA, USA, 13-18 June (pp. 32-39). Piscataway: IEEE.