Fusing Dynamic Deep Learned Features and Handcrafted Features for Facial Expression Recognition

Xijian Fan[1*] *& Tardi Tjahjadi*[2]
[1]Department of Computer Science, Nanjing Forestry University
Nanjing, China
[2]School of Engineering, University of Warwick
Coventry, UK
[*]Corresponding Author
xijian.fan@njfu.edu.cn t.tjahjadi@warwick.ac.uk

Abstract: The automated recognition of facial expressions has been actively researched due to its wide-ranging applications. The recent advances in deep learning have improved the performance facial expression recognition (FER) methods. In this paper, we propose a framework that combines discriminative features learned using convolutional neural networks and handcrafted features that include shape- and appearance-based features to further improve the robustness and accuracy of FER. In addition, texture information is extracted from facial patches to enhance the discriminative power of the extracted textures. By encoding shape, appearance, and deep dynamic information, the proposed framework provides high performance and outperforms state-of-the-art FER methods on the CK+ dataset.

Keywords: Convolutional neural network; facial expression recognition; feature extraction.

1. Introduction

The research field of automated facial expression recognition (FER) has attracted increasing attention due to its wide-ranging applications such as healthcare monitoring, assisted driving or human computer interaction [1], [2] and [3], [40], [41]. The classical model for describing facial expressions was first proposed by Ekman et al. [4]. It includes six basic facial expressions, i.e., anger, disgust, fear, happiness, sadness and surprise. This model has inspired numerous FER methods [5], [6], [7], [8] and [9] that have mainly focused on handcrafted features to represent the different facial expressions. Nevertheless, since facial expressions are subtle, complex and variable, accurate FER still poses a challenging problem [10] and [11]. In recent years, due to the advances of deep learning methods, more and more studies [12], [13] and [14] have focused on employing deep neural networks in the recognition of facial expressions, and have demonstrated promising results. Although deep-learning-based methods are capable of extracting varied features, including high level salient information, they might overlook coarse texture and geometric properties that can be well represented using handcrafted features. Thus, handcrafted features and deep learned features are considered complementary.

Facial expressions require dynamic processes. Consequently, dynamic information may represent a facial expression more effectively than static features. Thus, for FER applications dynamic information should be captured for the entire duration of a face video. Inspired by the numerous successful applications of convolutional neural networks (CNN) in pattern recognition [15] and [16], we propose the application of CNNs to FER by learning dynamic information using texture features of deep convolutional layers. By exploiting these dynamics from the convolutional layers, the proposed deep

descriptor is better qualified for the capturing of motion patterns in subtle texture level information, and thus, in distinguishing the subtle motion difference in different facial expressions. A CNN, pre-trained on the ImageNet dataset [17], is primarily meant for the differentiation of generic objects. Thus, not all feature channels from the convolutional outputs are useful for FER applications; they might even be regarded as noise. To improve the discriminative information for FER, we introduce a constraint which enhances the channel-discriminability for the dynamic texture extracted. In addition, as handcrafted features contain valuable information such as texture and shape which describe facial expressions, we propose to combine shape and appearance handcrafted features to obtain a mixed handcrafted feature. We further fuse the mixed handcrafted feature and dynamic deep convolutional features to obtain the final characteristic feature of a facial expression. These proposed features are capable of extracting not only the texture and shape, but also the dynamic information associated with facial expressions.

The main contributions in this paper are as follows: (1) We propose to learn dynamic deep texture information from deep feature channels, which captures facial motion among different facial expressions. (2) We introduce a constraint to adaptively weight the different channels based on their discriminative ability. (3) We integrate the deep dynamic features with handcrafted features to enhance the spatial information locally.

The remainder of paper is organized as follows: Section 2 presents a brief survey of related work. The proposed FER method - including the extraction of deep dynamic discriminative features, the combined handcrafted features, the descriptor fusion, and the FER framework – is introduced in Section 3. The experimental results are presented in Section 4, Section 5 concludes the paper.

2. Related work

Conventional features from facial images can be divided into geometric-based and appearance-based [5] and [18] features. The geometric-based features aim to detect and locate facial landmarks or specific shapes to obtain features representing the facial geometry [19,20], and [21]. Appearance-based features are low level features, which capture texture and appearance information of the face. Gabor wavelets [19] and local binary patterns (LBPs) [22] are the two appearance-based methods that are most widely used to describe facial expressions using local appearance information. Gabor features are obtained through the convolution of a face image with a group of filters. They are robust to misalignment. Local binary patterns analyse the contrast within sub-regions of an image. In the standard configuration a pixel is compared with the eight neighbouring pixels. This yields a binary pattern of 8 bit. The LBP descriptor can be stored as a histogram. Each bin of the histogram corresponds to one binary pattern configuration that represents a facial feature. In this way, a 256-dimensional descriptor is obtained. A generalize of appearance features across different persons is not trivial. This is one of the major drawbacks of appearance-based approaches. The histogram of gradients (HOG) [23] was originally developed for object recognition and as approach for pedestrian detection. Lazebnik et al., [24] apply HOG descriptors

for face recognition and extracted HOG features from face images using a dense grid. PHOG proposed by Bosch et al., [8] is an extension of the HOG descriptor that can be used to represent the shape of a facial region.

Due to the success in the field of computer vision, deep learning methods have been applied in FER as well [12,25,14] and Li et al., 2017). Some of the studies [14] and Li et al., 2017) proposed to train an ensemble of CNNs to improve the recognition performance, while others [27] and [28] proposed to fuse deep features with handcrafted features, e.g., SIFT (Lowe et al., 2014) and HOG. Due to the dynamic nature of facial expressions, some researchers also considered video data instead of static images for FER [13] and [28]. Mohammad et al., [13] presented a network structure that includes 3D convolutional layers followed by a Long-Short Term Memory (LSTM) unit, which extracts both spatial correspondence in face images and temporal information between two consecutive image frames of a video. Hamester et al. [30] proposed a 2-channel CNN (i.e., a standard CNN and a channel that uses pre-trained parameters), which was achieved by a convolutional auto- encoder (CAE) FER. Previous research that comes closest to the FER method proposed in this paper is presented in [27] and [28]. Both approaches combine deep features and handcrafted features. However, they extract deep features from static images, thus neglecting the dynamic information in facial expressions. For this reason, we propose to extract dynamic deep convolutional features, and to combine them with two handcrafted descriptors which include appearance and shape features. The proposed method not only exploits spatial information, but also captures dynamic information.

3. Proposed FER method

The proposed framework combines learned features and handcrafted features to compose a robust and accurate method for FER.

3.1 Deep discriminative descriptor

Training samples for FER are of limited availability. For this reason, we employ a pre-trained CNN model to estimate subtle facial motions that are caused by different facial expressions. On the basis of an aligned face video, we regard each of its image frames as input for a pre-trained CNN. As a next step, we extract the feature channels (maps) of the con3-3 layer (representing texture pat- terns [31] and [32] for each frame. A set of convolutional feature channels is obtained from an input video denoted as $\{X_t\}_{t=1}^{T}$, where $X_t \in \mathbb{R}^{W\times H\times K}$, and W, H are the height and the width of a channel, $k = 1, 2, \dots, K$ is the channel index in a convolutional layer and $t = 1, 2, \dots, T$ is the frame index of a video.

According to Zeiler et al., [31] and Cimpoi et al., [32], the feature channels of convolutional layers extract fine-grained texture information. The motion patterns within the fine-grained texture information capture the dynamic information of subtle facial expression. To track this dynamic information, we use the optical flow. More specifically, we decompose each original input video is into K convolutional channel time sequences referred to as $\{C_K\}_{k=1}^{K}$, where $C_k \in \mathbb{R}^{W\times H\times T}$. For each

sequence C_k, we extract the dynamic motion of the deep convolutional texture using the optical flow approach. We assume brightness constancy and spatial smoothness and thus determine the optical flow of consecutive frames by solving the optical flow constraint equation, i.e.,

$$\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} = 0\ , \tag{1}$$

where $\frac{\partial I}{\partial x}$, $\frac{\partial I}{\partial y}$ and $\frac{\partial I}{\partial t}$ are the spatiotemporal image brightness derivatives in the *x*, *y* and *t* dimensions, respectively. u and v are the optical flow in the horizontal and vertical dimensions, describing a local pixel translation. The Lucas-Kanade optical flow algorithm [33] is applied for our proposed FER method to estimate the optical flow vector with components u and v. We use the magnitude of the vector, i.e., $\sqrt{u^2 + v^2}$, to represent the dynamic information for computational efficiency.

Once the optical flow vectors of each video sample and for all K convolutional channel time sequences are estimated, we average the vectors to obtain a deep convolutional dynamic texture feature, which is denoted as $V \in \mathbb{R}^{K\times N}$. Given the feature set $V \in \mathbb{R}^{K\times N}$ from each input subject, which represents the facial motion of different facial expressions, we aim to learn the discriminative deep convolutional dynamic texture which integrates fine-grained dynamic information. Our experiments show that certain channels of the convolutional feature are unable to incur strong responses for all facial expressions. This could cause inaccurate motion estimation and diminish the performance. The features based on these irrelevant channels might be ambiguous for FER, thus degrading the discriminative properties of the classification approach. To this end, we introduce the channel discriminative constraint for the proposed method to adaptively weight the discriminative property of feature channels in the learning process. These weights strengthen the impact of relevant feature channels. For each input, the joint learning framework for the texture and discriminative constraint is formulated as

$$\max_{D,U_l} \frac{1}{2}\sum_{l=1}^{N} \|DV^l - U^l \mathrm{x}\|^2 \qquad s.t.\ \|D\|_2 = 1\ , \tag{2}$$

where $N = W \times H$, $V^l \in \mathbb{R}^{K\times 1}$ denotes the *l*-th column of the deep convolutional texture set V, U^l denotes the *l*-th column of the deep convolutional texture vector $U \in \mathbb{R}^{1\times N}$, and $D \in \mathbb{R}^{1\times K}$ represents the channel-discriminability constraint, which is used to weight different feature channels while extracting the texture feature. It can be shown that feature vector Eq. (2) aims to approximate the weighted aggregation of fine-grained texture features of different channels under a certain distance metric. For simplicity reasons, the Euclidean distance metric is employed in our approach.

The constraint D in Eq. (2) is trained with only one sample and is thus not capable of capturing discriminative information of all training samples. To create a more discriminative constraint vec- tor, we introduce a proper prior to learn the vector D from all training samples. In particular, we employ the metric based on the intra-class and inter-class variance to enhance the discriminability of channels. Our procedure is based on the assumption that more discriminative channels have smaller intra-class

variance but larger inter-class variance. For this purpose, we employ the Fisher criteria [ref] to measure the discriminative characteristics of each channel.

Given M training videos, there are M_i samples for each class i, where i denotes the class of a facial expression. From each video, we select the first frame for the channel-discriminability constraint learning. The feature channels are then selected using the pre-train CNN model. The related feature channels of M_i samples of class i are denoted as $\{F_1^i(k), F_2^i(k), \dots, F_{M_i}^i(k)\}$, where $F_{M_i}^i(k) \in \mathbb{R}^{N\times 1}$, $N = W \times H$.

For the k-th channel, the prior probability of class i is estimated by $P_i(k) = \frac{M_i}{\sum_{i=1}^{C} M_i}$ with class mean

$$\hat{\mu}_i(k) = \frac{1}{M_i}\sum_{j=1}^{M_i} F_j^i(k) \tag{3}$$

and gross mean

$$\hat{\mu}(k) = \sum_{i=1}^{C} P_i(k)\,\hat{\mu}_i(k). \tag{4}$$

Therefore, we estimate the sample covariance matrix $\hat{S}_i(k)$ of class i as

$$\hat{S}_i(k) = \frac{1}{M_i}\sum_{j=1}^{M_i}\left(F_j^i(k) - \hat{\mu}_i(k)\right)\left(F_j^i(k) - \hat{\mu}_i(k)\right)^T. \tag{5}$$

The intra-class scatter matrix and inter-class scatter matrix are then estimated as

$$S_w(k) = \sum_{i=1}^{C} P_i(k)\hat{S}_i(k)\ , \tag{6}$$

$$S_b(k) = \sum_{i=1}^{C} P_i(k)(\hat{\mu}_i\,(k) - \hat{\mu}(k))(\hat{\mu}_i(\mathrm{k}) - \hat{\mu}(k))^T\ . \tag{7}$$

As the final step, we consider k-th component of vector D to measure the k-th channel-discriminability, i.e.,

$$D(k) = trace\left(S_w^{-1}(k)S_b(k)\right), k = 1,2,\dots,K\ . \tag{8}$$

After measuring the discriminability of all channels, we obtain the constraint vector $D \in \mathbb{R}^{1\times K}$. Within the scope of our approach, we choose the 30 largest components in D and set the remaining components to zero, to reduce the negative effects irrelevant texture information channels may have on the classification performance. The final deep convolutional dynamic texture feature is learned as

$$U^l = DV^l, l = 1,2,\dots,N\ . \tag{9}$$

3.2 Handcrafted descriptors

A shape descriptor is a feature based on geometry, which can be defined on the basis of the locations of facial landmarks such as eyes, mouth or eyebrows. In contrast, appearance features describe facial deformations such as wrinkles or the nasolabial furrow that are caused by facial expressions. Thus, by combining shape and appearance features we capture the local facial properties and can improve the performance of a FER system.

In our proposed FER system, we start with the pre-processing of the face image using the OpenFace toolkit [34] to extract hand- crafted features. For geometric features, several metrics are extracted from

the facial points related to the eyes, the nose, and the mouth to obtain a 120-dimensional (120D) feature vectors as a shape descriptor, f_{Shape}. The metrics are estimated for each frame as follows: (1) We measure the Euclidean distances between consecutive facial points of the eyes and the associated eyebrow (20D); (2) we select the tip of the nose as well as the corners of the eyes as stable points as they do not move during facial expressions, and measure the Euclidean distances between median of these stable points and each of the facial points (49D); (3) we measure the magnitude of the angle between three consecutive points along the eyes and the associated eyebrow (18D); and (4) we measure the magnitude of the angle between three consecutive points along the mouth (16D).

The PHOG descriptor was first proposed for object classification [8]. It is inspired by HOG [23], and uses edge information in com- bination with information about the spatial layout of a local shape on the basis of image pyramids [24]. In particular, edges within an image are extracted at different levels of the image pyramid, thus at different levels of resolution. We then count the occurrences of certain gradient orientations of edges and obtain a gradient histogram in this way [8]. To construct the final PHOG descriptor, we concatenated the histograms of the individual pyramid levels.

We compute the PHOG descriptor (as illustrated in Figure 1) in every single image from a video, i.e., along the temporal axis as follows. We obtain the edge information using the Canny edge detector. The image region is then divided on the basis of gradually smaller grids. For this purpose, each cell of the previous grid is divided into two cells along each axis in the subsequent step. the grid at resolution level l has 2^l cells along each dimension.

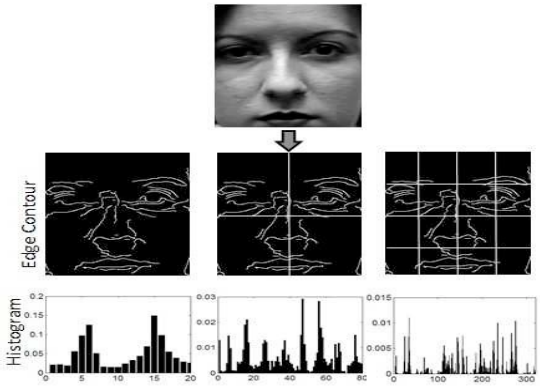

Figure 1. Generation of PHOG from a face image [9].

The PHOG is normalized to sum to unity. As a result, *K*-vector corresponding to K histogram bins represents level 0, while level 1 is represented by a 4*K*-vector. The normalized PHOG descriptor of the entire sequence is the following vector to which we refer to as appearance descriptor, i.e.,

$$\mathrm{f}_{\mathrm{Appearance}} = \frac{\sum_k \mathrm{PHOG}_k}{k} \tag{10}$$

with dimensionality

$$\mathrm{Dim}_{xy} = K \sum_{l \in L} 4^l , \tag{11}$$

where L denotes the number of levels. We set L to 2 to avoid over fitting of the edge contours within the grids.

Bosch et al. studied which facial patches are active for certain facial expressions [8]. They observed that some facial patches show activity in all basic expressions, while some are only active for very specific expressions. The common facial patches, located below the eyes, in between the eyebrows, and around the nose and mouth corners, play vital roles in aiding FER. Naturally, patches that show high variation are more discriminative. Thus, we propose to extract the PHOG feature using these active patches to improve the recognition performance. The final PHOG features are obtained by concatenating the PHOG features from each active patch. Unlike other, related methods that concatenate all facial features, thus, generating a feature vector of high dimensionality, we only concatenate features from a few facial patches. In this way, we reduce the feature vector dimensionality while preserving the recognition accuracy.

We further combine shape features and the patch based PHOG features to obtain the final mixed handcrafted feature referred to as $f_{handcrafted}$.

3.3 Descriptor Fusion

We utilize the VGG net [42] deep convolutional network pre- trained on the ImageNet dataset to extract our deep feature as follows: Diverse fine-grained texture information is extracted from the channels of the conv3-3 layer of VGG net. The handcrafted features include shape and appearance features that are extracted using the procedure described in Section 3.2. The deep feature and the handcrafted feature are concatenated to form the fused descriptor.

3.4. Framework for Facial Expression Recognition

Our proposed FER framework consists of three steps: the pre- processing, the feature extraction and the classification. The complete framework is illustrated in Fig. 2. The pre-processing includes facial landmark detection and face alignment, to avoid artefacts due to changes in the head pose and the illumination. The OpenFace toolkit [34] is used to detect 68 facial landmarks in the first frame of the face video. These landmarks are then tracked over the subsequent frames. The locations of the selected landmarks (i.e., two eyes and nose) are used for face alignment to address in-plane facial motion.

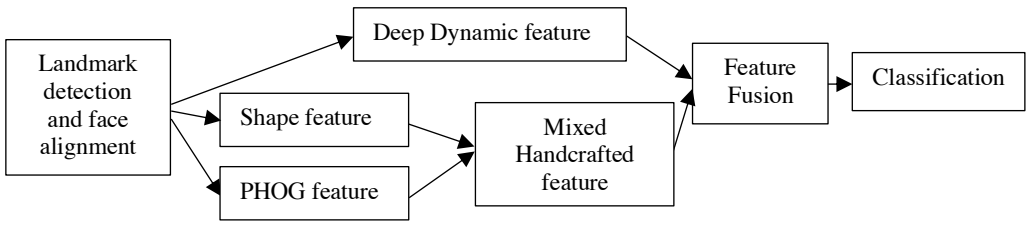

Figure 2. Framework for the facial expression recognition.

The feature extraction step involves the construction of the fused features using the handcrafted and the deep dynamic features (see Section 3.3). A support vector machine (SVM) with a lin- ear kernel is employed in the classification step.

4. Experiments

4.1 Facial expression datasets

The Extended CK+ dataset (CK+) [36] is the most popular publicly available benchmark dataset to evaluate the performance of FER methods. The CK+ dataset consists of 593 sequences that capture seven basic facial expressions that were performed by 120 subjects. In our experiments, 327 sequences of CK+ are used to evaluate the performance of the proposed FER approach.

4.2 Experiment results

To compare the performance of our proposed FER method with state-of-the-art approaches we employ a leave-one-out cross-validation scheme as follows: We select one of the video sequences for testing, and use the remaining video sequences for the training. This ensures that the testing sequence is not part of the training set, and is thus independent from the training data.

Our first set of experiments focuses on the weighting strategy used in the deep dynamic feature and investigates its effectiveness. We perform leave-one-out cross-validation 327 times, thus for all 327 video sequences in our dataset. Table 1, Table 2 and Table 3 show the recognition rates for the deep dynamic discriminative feature, the handcrafted feature and the proposed FER method with feature fusion, respectively, where the bold values denote the correct classification. The proposed framework using the feature fusion achieves better overall performance than the other methods. Both the deep dynamic discriminative feature and the handcrafted feature are outperformed by the feature fusion approach for all facial expressions except for fear. This traces back to the fusion of the deep dynamic feature and the handcrafted feature that enhances dynamic and discriminative information, while retaining texture information.

Table 1. Confusion matrix of the deep dynamic feature on classification of seven facial expressions of the CK+ dataset.

	A	**D**	**F**	**H**	**Sa**	**Su**	**C**
Anger(A)	**88.9**	0	2.2	0	6.7	0	2.2
Disgust(D)	1.7	**94.9**	0	0	0	1.7	1.7
Fear(F)	4.0	0	**88.0**	0	4.0	0	4.0
Happiness (Sa)	1.4	1.4	0	**97.1**	0	0	0
Sadness (Sa)	3.6	3.6	3.6	0	**85.7**	0	3.6
Surprise (Su)	0	0	0	0	0	**100**	0
Contempt (C)	11.1	11.1	0	0	0	0	**77.8**

Table 2. Confusion matrix of the handcrafted feature on classification of seven facial expressions of the CK+ dataset.

	A	**D**	**F**	**H**	**Sa**	**Su**	**C**
Anger (A)	**82.2**	2.2	6.7	0	6.7	0	2.2
Disgust (D)	0	**84.7**	5.2	0	0	0	0
Fear (F)	0	4.0	**84.0**	8.0	8.0	0	0
Happiness(Sa)	0	1.4	1.4	**95.7**	0	0	1.4
Sadness (Sa)	7.2	7.2	3.6	0	**78.6**	0	3.6
Surprise (Su)	0	0	1.2	2.4	2.4	**94.0**	0
Contempt (C)	16.7	11.1	5.6	0	5.6	0	**61.1**

Table 3. Confusion matrix of the propose feature on classification of seven facial expressions of the CK+ dataset.

	A	**D**	**F**	**H**	**Sa**	**Su**	**C**
Anger (A)	**91.1**	0	2.2	0	2.2	4.4	0
Disgust (D)	0	**100**	0	0	0	0	0
Fear(F)	0	4.0	**88.0**	4.0	0	0	4.0
Happiness(Sa)	0	1.4	2.9	**95.7**	0	0	0
Sadness (Sa)	7.1	0	0	3.6	**89.3**	0	0
Surprise (Su)	0	0	0	0	0	**100**	0
Contempt (C)	5.6	0	5.6	0	5.6	0	**83.3**

To compare with other methods, we utilized the same leave-one-out cross-validation strategy as it is used in the work of Eskil and Benli [37], Lucey et al. [36], and Li and Lam [38]. We compare the proposed framework with these three methods. The proposed framework for all seven facial expressions in average achieves a recognition rate of 92.5% and outperforms the other methods. We thus conclude that the proposed method that involves the fusion of deep dynamic features and handcrafted features improves the recognition performance (see Table 4).

Table 4. Comparative evaluation of the proposed method with 3 methods using leave-subject-out cross-validation.

Study	**Methododogy**	**Recognition Rate**
Esil and Benli (2014)	SVM	76.8
	Adaboost	76.3
Lucey et al., (2010)	SVM(shape)	50.4
	SVM(Appearance)	66.7
	SVM(Combined)	83.3
Li and Lam (2015)	DBM + SVM	86.8
Proposed method with feature fusion	SVM	**92.5**

5. Conclusion

This paper proposes the fusion of deep dynamic convolutional features and handcrafted features to obtain a novel feature for FER. The FER framework on the basis of this fused feature achieves a higher recognition rate than three classical methods and deep-learning based methods on the CK+ dataset. The proposed framework thus proves suitable for an application in FER. Investigating the generalization ability and robustness of the proposed feature by employing more datasets will be part of our future work.

Acknowledgement

The work is supported by National Science Foundation of China (Grant No. 61902187), Jiangsu Province Innovative and Entrepreneurial Talent Project and Nanjing Forestry University Start-up Foundation for Research.

6. References

[1] A.A. Salah, N. Sebe, T. Gevers, Communication and automatic interpretation of affect from facial expressions, affective computing and interaction: psychological, Cognitive Neurosci. Perspect. (2010) 157–183.
[2] D. Nguyen et al., Deep spatio-temporal features for multimodal emotion recognition, in: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2017, pp. 1215–1223.
[3] Z. Zeng et al., A survey of affect recognition methods: Audio, visual, and spontaneous expressions, IEEE Trans. Pattern Anal. Mach. Intell. 31 (1) (2008) 39–58.
[4] P. Ekman, W. Friesen, Constants across cultures in the face and emotion, J. Pers. Soc. Psychol. 17 (2) (1971) 124–129.
[5] Y.-I. Tian, T. Kanade, J. Cohn, Recognizing action units for facial expression analysis, IEEE Trans. Pattern Anal. Machine Intell. 23 (2) (2002) 97–115.
[6] C. Shan, S. Gong, P. McOwan, Facial expression recognition based on local binary patterns: a comprehensive study, Image Vis. Comput. 27 (2009) 803– 816.
[7] Dalal, N., and Bill T., Histograms of oriented gradients for human detection. 2005.
[8] A. Bosch et al., Representing shape with a spatial pyramid kernel, in: Proceedings of the International Conference on Image and Video Retrieval, 2007, pp. 401–408.
[9] X. Fan, X. Yang, Q. Ye, Y. Yang, A discriminative dynamic framework for facial expression recognition in video sequences, J. Vis. Commun. Image Represent 56 (2018) 182–187.
[10] M. Pantic, L.J.M. Rothkrantz, Automatic analysis of facial expressions: the state of the art, IEEE Trans. Pattern Anal. Machine Intelligence 22 (12) (2000) 1424– 1445.
[11] B. Fasel, J. Luettin, Automatic facial expression analysis: a survey, Pattern Recogn. 36 (1) (2003) 259–275.
[12] E. Barsoum, C. Zhang, C.C. Ferrer, Z. Zhang, Training deep networks for facial expression recognition with crowd-sourced label distribution, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016, pp. 279–283.
[13] M. Mohammad, H. Behzad, Facial expression recognition using enhanced deep 3D convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 30–40.
[14] Bo-Kyeong Kim et al., Hierarchical committee of deep convolutional neural networks for robust facial expression recognition, J. Multimodal User Interf. 10 (2) (2016) 173–189.

[15] A. Krizhevsky, I. Sutskever, G. Hinton, ImageNet classification with deep convolutional neural, Networks. NIPS vol (2012) 25.
[16] K. He, X. Zhang, S. Ren, S. Jian, Deep residual learning for image recognition, IEEE Conference on Computer Vision & Pattern Recognition, 2016.
[17] J. Deng, W. Dong, R. Socher, L.J. Li, F.F. Li, ImageNet: a Large-Scale Hierarchical Image Database, 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2009, 20-25 June.
[18] Tian, Y.-I., Kanade, T., and Cohn, J., Recognizing action units for facial expression analysis. IEEE Transactions Pattern Analysis and Machine Intelligence 2002; 23(2), pp. 97-115.
[19] Brais Martinez et al., Automatic analysis of facial actions: a survey, IEEE Trans. Affective Comput. (2017), July 25.
[20] Z. Zhang, M. Lyons, M. Schuster, S. Akamatsu, Comparison between geometry based and gabor-wavelets-based facial expression recognition using multilayer, Proc. Int. Conf. Automatic Face Gesture Recognit. (1998) 454–459.
[21] Y. Zhang, Q. Ji, Active and dynamic information fusion for facial expression understanding from image sequences, IEEE Trans Pattern. Anal. Mach. Intell. 27 (5) (2005) 699–714.
[22] M. Pantic, I. Patras, Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences, IEEE Trans. Syst Man Cybernet. 36 (2) (2006) 433–449.
[23] G. Zhan, M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expression, IEEE Trans Pattern. Anal. Mach. Intell. 29 (6) (2007) 915–928.
[24] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, IEEE Conf. Comput. Vision Pattern Recognit. 1 (2005) 886–893.
[25] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene Categories, IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit. 2 (2006) 2169–2178.
[26] H. Ding, S.K. Zhou, R. Chellappa, Facenet2expnet: regularizing a deep face recognition net for expression recognition, in: In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG), 2017, pp. 118–126.
[27] Li, D. and Wen, G., MRMR-based ensemble pruning for facial expression recognition. Multimedia Tools and Applications 2018; 77(12), pp.15251-15272.
[28] T. Connie, M. Al-Shabi, W.P. Cheah, M. Goh, Facial expression recognition using a hybrid CNN–SIFT aggregator, in: In International Workshop on Multi- disciplinary Trends in Artificial Intelligence, 2017, pp. 139–149, November.
[29] H. Kaya, F. Gürpınar, A.A. Salah, Video-based emotion recognition in the wild using deep transfer learning and score fusion, Image Vision Comput. 65 (2017) 66–75.
[30] Lowe, D.G., Distinctive image features from scale-invariant keypoints. International journal of computer vision 2004; 60(2), pp. 91-110.
[31] D. Hamester, P. Barros, S. Wermter, Face expression recognition with a 2- channel convolutional neural network, in: In 2015 international joint conference on neural networks, 2015, pp. 1–8.
[32] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European conference on computer vision, Springer, Cham, 2014, pp. 818– 833.
[33] M. Cimpoi, S. Maji, A. Vedaldi, Deep filter banks for texture recognition and segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3828–3836.
[34] Lucas, B D, Kanade, T. An iterative image registration technique with an application to stereo vision. 1981.
[35] T. Baltrusaitis et al., Constrained local neural fields for robust facial landmark detection in the wild, Proc. IEEE ICCV Workshops (2013) 354–361.
[36] Hsu, Chih-Wei, Lin,Chih-Jen., A Comparison of Methods for Multiclass Support Vector Machines. IEEE Transactions on Neural Networks 2002; 13(2), pp. 415-425.

[37] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, 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, 2010, pp. 94–101.
[38] M.T. Eskil, K. Benli, Facial expression recognition based on anatomy, Comput. Vis. Image Underst. 119 (2014) 1–14.
[39] J. Li, E.Y. Lam, Facial expression recognition using deep neural networks, in: In 2015 IEEE International Conference on Imaging Systems and Techniques (IST), 2015, pp. 1–6.
[40] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, Comput. Sci. (2014).
[41] B. Sriman, L. Schomaker, Multi-script text versus non-text classification of regions in scene images, J. Vis. Commun. Image Represent. 62 (2019) 23–42.

Xijian Fan received B.Sc. in Information and Communication Technology from Nanjing University of Posts and Telecommunications and M.Sc. in Computer Information and Science from Hohai University in 2008 and 2012, respectively. He received Ph.D. in School of Engineering from University of Warwick, U.K. His research interests include image processing, computer vision and biomedical engineering.

Tardi Tjahjadi received B.Sc. in Mechanical Engineering from University College London in 1980, and M.Sc. in Management Sciences in 1981 and Ph.D. in Total Technology in 1984 from UMIST, U.K. He has been an associate professor at Warwick University since 2000. His research interests include image processing and computer vision.