Feature Selection Mechanism in CNNs for Facial Expression Recognition

Shuwen Zhao[1]
2111612068@zjut.edu.cn
Haibin Cai[2]
haibin.cai@port.ac.uk
Honghai Liu[2]
honghai.liu@port.ac.uk
Jianhua Zhang[1]
zjh@zjut.edu.cn
Shengyong Chen[1]
csy@zjut.edu.cn

[1] Zhejiang University of Technology
Zhejiang, China

[2] University of Portsmouth
Portsmouth, UK

Abstract

Facial Expression Recognition (FER) has been a challenging problem in computer vision for many decades, mainly due to the high-level variation of face geometry and facial appearance. In this paper, we propose a feature selection network (FSN) to automatically extract and filter facial features by embedding a feature selection mechanism inside the AlexNet. The designed feature selection mechanism effectively filters irrelevant features and emphasises correlated features according to learned feature maps. Experiment results on several databases demonstrate that the FSN outperforms the AlexNet by a large margin and achieves comparable results with the state-of-the-art methods. Furthermore, the FSN also shows improved generalisation ability over the AlexNet in the cross validation experiment of different datasets.

1 Introduction

With the developing needs of communication with machines, it becomes more important for computers to have the ability to make smarter decisions and provide a better interactive experience by understanding emotions. For example, in the application of human-computer interaction(HCI), an accurate facial expression estimate can help HCI to provide a higher quality. It can be applied in entertainment, security and education.

Commonly FER includes three steps: face detection and alignment, feature extraction, and classification. In the face detection and alignment step, an automatic face detector is used to crop faces. Then facial feature points are used to align the face image. These facial feature points can be obtained by facial landmark detector, such as Active Appearance Model (AAM)[27], Supervised Descent Method (SDM) [34], Tasks-Constrained Deep Convolutional Network (TCDCN)[25]. For feature extraction step, numerals feature vectors are

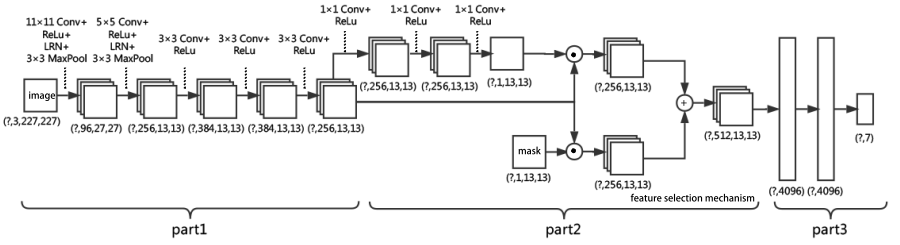


Figure 1: Feature Selection Network (FSN)

generated from the aligned face image by using the classical feature extracting methods, AU-based methods or CNNs-based methods. In the third step, of classification, the commonly used methods include Support Vector Machines (SVM) (e.g.[5, 20, 33]), Linear Discriminant Analysis (LDA), Boosted Deep Belief Networks (BDBN)[19], decision-level fusion on these classifiers [36], etc. However, because of the influence of head pose variations, illumination, occlusions, the diverse expression intensities and individual facial difference, FER in wild is a significant challenge.

The CNN-based methods have an excellent performance on recognition task related to faces. However, there still have some limitations if a classical CNN architecture is directly applied in FER datasets. First, if the architecture is simple, we will miss some crucial features. Second, if we use a very deep architecture to train the dataset, the parameters can easily fall into overfitting. Furthermore, some datasets have only a few samples and others have extremely imbalanced sample number among classes, which will influence classical FER methods. These motivate us to structure a new CNN method that could choose effective features and filter nonsense features automatically. This new CNN also uses fine-tune to avoid overfitting, and data augmentation to break the imbalance among classes at the same time.

Therefore, we propose a FSN model by embedding a feature selection mechanism inside the AlexNet. This feature selection mechanism calculates the influence proportion of each location and filters the unnecessary features in the following layer. As illustrated in Figure 1, the first part of the FSN model consists of five convolution layers. The second part is a feature selection mechanism with two branches. One branch has three convolution layers. Such a structure is first propsed by [29], and we adapt it into our FSN. Specifically, all the kernel size of convolution layer in the branch is 1, and the third layer has only one feature map for each input image. After that, a dot multiply layer is applied to filter the extracted features. On the other branch, we add another input layer, where we use the mask data as input to confine the facial region and assist feature extracting. Then we merge the features at the end of two branches. On the third part, there are three fully connected layers.

The proposed CNN architecture has tested on two public facial expression datasets, i.e. FER2013[8] and RAF[16] datasets. Furthermore, there has cross-validation between two datasets to show the generalisation of this model. Then we compare experiment results with other methods. It has been demonstrated that our method outperforms the AlexNet[15] by a large margin and achieves comparable results with the state-of-the-art methods. Especially our FSN outperforms the AlexNet, the basis of our architecture.

2 RELATED WORK

Over the past few years, FER has been widely studied. It consists of three main steps. Because of the influence by the head pose variations, illumination, occlusions, and so on, the feature extraction step has become the most challenging one. The methods of extracting feature can be roughly divided into three categories: extracting human-designed feature by classical method, FER by action unit (AU), and CNN-based method.

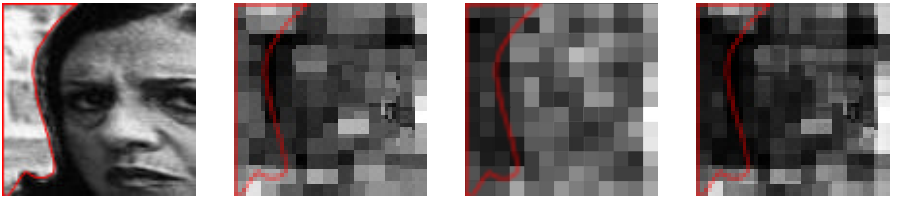
Before the popularity of deep learning, FER problem was always solved by extracting human-designed features. Those features include geometric features representing face geometry such as the shapes and locations of facial landmarks[14], motion features such as optical flow[21], Motion History Images (MHI)[30], volume LBP[37], and appearance features describing the skin texture of faces such as Gabor filters[18], Gabor wavelets [2], Haar features[32], Local Binary Patterns (LBP)[37], Scale Invariant Feature Transformation (SIFT)[17], Local Phase Quantization (LPQ)[31], Histogram of Oriented Gradients (HoG)[22], pixel intensities[23]. In addition to those features that can be directly extracted from a single image, there also have some algorithms focusing on exploiting the temporal variation in an image sequence or videos. For example, [6] proposed to learn Random Forests from heterogeneous derivative features upon pairs of images, and Local Binary Pattern from three orthogonal planes (LBP-TOP)[37] derived six feature vectors from eyes, nose and mouth components which are generated by the dynamic textures and structural shape feature descriptors to form a facial representation.

In [7], Ekman *et al.* proposed the Facial Action Coding System (FACS) which describes human facial movements by their appearance on the face. The FACS defineded many standard facial substructures called Action Units (AUs), and each AU is based on one or a few facial muscles in combinations. Thus, some works have considered classifying expressions by using AU information. The methods of AU occurrence detection include ANNs [4], AdaBoost and GentleBoost [11], SVMs [6, 20, 33]. For AU intensity detection problem, it can be approached using either a Classification-based method [3, 9, 11, 20] or a Regression-based method [8, 12].

CNN-based methods have allowed to remove or highly reduce the dependency of human-designed features. Methods that resort to the CNN architecture have also been proposed for FER. For instance, [35] used a face detection module followed by a classification module with the ensemble of multiple deep CNNs. [24] used three inception structures [26] in convolution for FER. All these methods extract feature automatically. However, many researchers are not content with the result and focus on combining the CNN with other classical algorithms such as human designed features or AUs to improve the peformance of FER. One example is [13], where appearance features and geometry features are combined using a new integration method. In another example, [10] used a jointly trained CNN for detection and intensity estimation of multiple AUs.

3 METHODOLOGY

In this section, we will first illustrate the whole structure of the network, and then introduce the algorithm of selection layer and the input data respectively in feature selection mechanism. Finally, there have some implementation details of this network.



(a) input image (b) part of the extracted features (100 features from 256) from the 5th layer (c) selection map from the 3rd layer of FSM (d) extracted features after been filtered

Figure 2: The outline in all 4 figures means the background part of the image. In figure 2(b), we can see the weight inside the outline still account for a great proportion. And the same part in figure 2(c) has a low weight. Accordingly, the result (figure 2(d)) of dot product between them shows that most of the features inside the red outline has been filtered. Moreover, such as the eyeball part and the top of the forehand in figure 2(d) also been filtered.

3.1 Network structure

We propose a CNN based on feature selection for locating facial information precisely. To obtain explicit facial details effectively and ignore the negative influence of background, we develop a feature selection mechanism. The whole network is based on the Alexnet. And the feature selection mechanism is between the convolution layer and the fully connected layer of the Alexnet.

In feature selection mechanism, one of the branches which have three convolution layers is constructed by taking the following two considerations. First, for every sample, the third layer has only one feature map which could force it locating all the effective facial features. Second, the feature maps in the third layer show the importance of each location. Different facial parts have different weights. For example, the weight in forehead is much lower than that in cheek. Therefore, it could magnify the vital part and weaken or even eliminate the insignificant part and improve generalisation ability of this model. Another branch using the face mask as input data plays an assistant role. We use the facial landmark detection method to get the facial landmarks and then make face mask figure (4). The size in the top of feature selection layer in the previous branch is 13×13, and the size of the face mask should keep the same. The features in the fifth convolution layer in part one will be filtered by both weight map and face mask. Then we combine them. The face mask could exclude all the background features which always have a strong effect on FER but have no influence on oneâĂŹs expression of emotion. Because of the limitation of datasets, it is inappropriate to classify the features filtered by the second branch directly. Most of the existing datasets have small images or even miss some parts of the face like jaw. Under these circumstances, the result of facial landmark algorithm have some bias and can not be fully trusted.

3.2 convolution layers in feature selection mechanism

As illustrated in Figure 1, the proposed feature selection mechanism includes two branches. One has three additional convolution layers with filter size 1×1 followed by a selection layer, and another one has an input layer with face mask followed by a selection layer. The former branch will be updated in each iteration.

The output of the third layer could be regarded as a weight map. It is used to estimate if the feature in this location should be taken into consideration in the final calculating. As shown in figure 2(b), one feature map without filtering includes more background parts. The background part marked by red outline is not related to facial expression, but do have the negative affect on FER. Hence, we generate a weight map to limit the influence of features in some irrelevant locations. Figure 2(d) shows the same feature map filtered by our proposed mechanism.

As shown in figure 3, given activation tensor $\mathbf{U}$ as input from the last convolution layer in the first part of FSN, where N is the number of feature channels, H and W are height and width of the tensor $\mathbf{U}$. And through three convolution layers in feature selection mechanism a matrix $\mathbf{X}$ is generated.

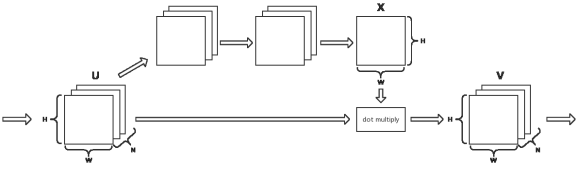


Figure 3: part of feature selection mechanism

3.2.1 forward propagation

After previous layers we could obtain one feature map $\mathbf{X}$ in each image. And then getting the selected features by element-wise multiplication of $\mathbf{X}$ with the initial activation $\mathbf{U}$:

$$\mathbf{V}_n = \mathbf{X} \odot \mathbf{U}_n \tag{1}$$

where $\mathbf{U}_n$ is the n-th feature channel of $\mathbf{U}$, and V_n corresponds to the selected feature of the same channel.

3.2.2 backward propagation

For backward propagation, gradients with respect to $\mathbf{U}$ and $\mathbf{X}$ are

$$\frac{\partial \mathbf{V}}{\partial \mathbf{U}} = \partial \mathbf{X} \tag{2}$$

and

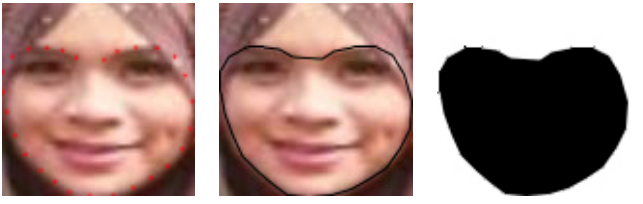
$$\frac{\partial \mathbf{V}}{\partial \mathbf{X}} = \frac{1}{N} \sum_{n=1}^{N} \partial \mathbf{U}_n \tag{3}$$

The gradient for $\mathbf{X}$ is normalised by the all number of the feature maps N since the feature filter map $\mathbf{X}$ affects all the feature maps in $\mathbf{U}$ equally.

3.3 Input data in feature selection mechanism

First, the facial landmark detection algorithm [1] is applied in input face image to obtain landmarks precisely (figure 4(a)). Then some facial landmarks are connected in order to

form the face outline (figure 4(b)). Labelling every pixel value in the outline to be 1 and the pixel value outside to be 0 (figure 4(c)). At last, the mask image is resized to 13×13 to fit the size of the image output from the last convolution layer in feature selection mechanism. In this way, the features presenting background and making nonsense with facial expression could be filtered directly.



(a) (b) (c)

Figure 4: The process diagram of making input mask data

3.4 implement detail

We use AlexNet[15], consisting of five convolution layers and three fully connected layers, as the baseline CNN architecture. We use the pre-trained result on the ImageNet dataset to initialise the five convolution layers and fine-tune the whole network on the FER2013 dataset and RAF dataset. The input image size of our networks was 227×227 pixels, which results in an activation $\mathbf{U}$ of size 256×13×13 after the pooling layer of the 5th convolution layers.

4 EXPERIMENT

A series of experiments have been tested on the FER2013 dataset and the RAF dataset. To demonstrate the effectiveness of feature selection mechanism, our FSN is compared with the baseline AlexNet and other methods. Furthermore, there is a cross-validation between two datasets to show the generalisation ability of this model.

4.1 Pre-processing

To fit the fine-tune model which is trained in ImageNet dataset by AlexNet, face images should be resized to 256×256 and cropped to 227×227 in first image layer. Since different classes have imbalance number of images, we augment the data by flipping, rotating and cropping the original images. Finally, the classes in the same training dataset have approximately equal numbers. Furthermore, we do not use an extra alignment method because in RAF face images have already been aligned, and in FER2013 images are too small to align well.

4.2 Experimental Datasets

4.2.1 RAF Dataset

The images were collected and automatically downloaded from Flickr. Then a set of keywords was used to pick out images that were related to the six basic emotions plus the neutral

Table 1: number of images per each expression in datasets

	AN	DI	FE	HA	SA	SU	NE
RAF (train)	705	717	281	4772	1982	1290	2524
RAF (test)	162	160	74	1185	478	329	680
FER2013 (train)	3996	436	4097	7215	4830	3171	4965
FER2013 (test)	957	111	1024	1774	1247	831	1233

* AN, DI, FE, HA, Ne, SA, SU stand for Anger, Disgust, Fear, Happiness, Neutral, Sadness, Surprised respectively.

emotion. And then annotated images by 315 annotators. At last, a total of 15339 real-world facial images are presented in this dataset.

4.2.2 FER2013 Dataset

The Facial Expression Recognition 2013 (FER2013) dataset was introduced in the ICML 2013 Challenges in Representation Learning. The dataset was created using the Google image search API, and faces have been automatically registered. Most of the images are in wild settings and all images are grey with size 48×48. Faces are labelled as any of the six basic expressions as well as the neutral. The resulting database contains 28710 images for training and 7177 images for testing.

4.3 Experimental Results

There are two confusion matrices of the proposed FSN which are tested on each dataset. The diagonal of the confusion matrices represents the accuracy of each class. Specifically, there also have two cross-validation of RAF and FER2013 datasets which is used to verify that our network has a good generalisation ability. In addition to the confusion matrices, the FSN is compared with the AlexNet and other state-of-the-art methods evaluated on the two datasets (RAF, FER2013) such as DLP-CNN[16], VGG.

4.3.1 Result of RAF dataset

The confusion matrix of FSN is shown in Table II, and the comparison of each method is indicated in Table III. The FSN achieves an average FER of 72.46% on the RAF dataset. And the confusion matrix of FSN model trained by RAF and tested on FER2013 is showed in table IV.

4.3.2 Result of FER2013 dataset

The confusion matrix of FSN is shown in Table V, and the comparison of each method is shown in Table VI. The FSN achieves an average FER of 67.64% on the FER2013 dataset. And the confusion matrix of FSN model trained by FER2013 and tested on RAF is showed in table VII.

Table 2: Average confusion matrix for subject-independent in RAF dataset

		predicted						
		AN	DI	FE	HA	SA	SU	NE
actual	AN	**0.728**	0.049	0.019	0.093	0.037	0.037	0.037
	DI	0.088	**0.469**	0.025	0.106	0.113	0.038	0.163
	FE	0.054	0.014	**0.568**	0.068	0.108	0.068	0.068
	HA	0.005	0.014	0.005	**0.905**	0.030	0.008	0.033
	SA	0.017	0.029	0.013	0.054	**0.816**	0.006	0.065
	SU	0.006	0.027	0.030	0.033	0.033	**0.818**	0.052
	NE	0.003	0.034	0.004	0.053	0.119	0.018	**0.769**

Table 3: Overall accuracy in RAF dataset

METHOD	Accuracy
AlexNet[15]	0.5560
VGG	0.5822
DLP-CNN(LDA)[16]	**0.7098**
FSN	**0.7246**

Table 4: RAF⟶FER2013

		predicted						
		AN	DI	FE	HA	SA	SU	NE
actual	AN	**0.247**	0.065	0.037	0.080	0.282	0.149	0.140
	DI	0.162	**0.324**	0.036	0.063	0.288	0.054	0.072
	FE	0.082	0.059	**0.059**	0.112	0.321	0.221	0.146
	HA	0.042	0.027	0.021	**0.614**	0.163	0.071	0.061
	SA	0.071	0.057	0.036	0.062	**0.435**	0.137	0.202
	SU	0.017	0.005	0.025	0.096	0.084	**0.708**	0.065
	NE	0.038	0.045	0.012	0.109	0.235	0.166	**0.394**

Table 5: Average confusion matrix for subject-independent in FER2013 dataset

		predicted						
		AN	DI	FE	HA	SA	SU	NE
actual	AN	**0.617**	0.013	0.082	0.033	0.139	0.034	0.082
	DI	0.189	**0.685**	0.018	0	0.045	0.009	0.054
	FE	0.094	0.007	**0.515**	0.019	0.178	0.102	0.087
	HA	0.026	0.002	0.017	**0.828**	0.033	0.033	0.063
	SA	0.094	0.010	0.101	0.032	**0.587**	0.024	0.152
	SU	0.012	0.005	0.046	0.022	0.040	**0.857**	0.019
	NE	0.053	0.006	0.045	0.044	0.177	0.030	**0.648**

4.3.3 Result analysis

Our proposed model has an about 16.9% higher accuracy compared with AlexNet in RAF dataset and a 6.5% higher accuracy compared with AlexNet in the FER2013 dataset. It shows that our model performs well in FER. In the method DLP-CNN, two different classifiers are used and two different results are obtained. The results of our method is better than one of them, as shown in Table III. As for cross-validation, because every dataset has

Table 6: Overall accuracy in FER2013 dataset

Method	Accuracy
AlexNet[15]	0.611
[24]	0.664
[28]	**0.693**
FSN	**0.676**

Table 7: FER2013⟶RAF

		predicted						
		AN	DI	FE	HA	SA	SU	NE
actual	AN	**0.574**	0.037	0.074	0.136	0.093	0.012	0.074
	DI	0.281	**0.113**	0.056	0.1	0.275	0.013	0.163
	FE	0.243	0	**0.270**	0.216	0.135	0.068	0.068
	HA	0.020	0.013	0.008	**0.781**	0.067	0.005	0.107
	SA	0.065	0.008	0.050	0.126	**0.575**	0.004	0.172
	SU	0.100	0.003	0.155	0.091	0.140	**0.231**	0.280
	NE	0.087	0.022	0.035	0.05	0.276	0.004	**0.525**

Table 8: comparison of across validation

accuracy / method / dataset	AlexNet	FSN
RAF⟶FER2013	0.340	**0.397**
FER2013⟶RAF	0.386	**0.456**

its peculiarity (lighting, background, expression extensions, etc.), it is difficult to perform well in cross-validation between two datasets. There was a comparison of cross-validation between AlexNet and our FSN with two datasets in table VIII. It shows that our method also has a good generalisation ability.

5 Conclusion

This paper proposed a novel CNN-based facial expression recognition method to deal with various subjects facial geometry and facial appearance. A feature selection network which aims to filter unrelated features is designed. Experimental evaluations showed that the FSN achieves a better performance over the Alexnet and can be applied to practical facial expression applications. The future direction is targeted at designing more lightweight network structure while keeping the good accuracy performance.

References

[1] Tadas Baltrusaitis, Peter Robinson, and Louis Philippe Morency. OpenFace: An open source facial behavior analysis toolkit. *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016*, 2016. doi: 10.1109/WACV.2016.7477553.

[2] Marian Stewart Bartlett, Gwen Littlewort, Mark Frank, Claudia Lainscsek, Ian Fasel,

and Javier Movellan. Fully automatic facial action recognition in spontaneous behavior. *FGR 2006: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, 2006(January):223–230, 2006. doi: 10.1109/FGR.2006.55.

[3] Marian Stewart Bartlett, Gwen C. Littlewort, Mark G. Frank, Claudia Lainscsek, Ian R. Fasel, and Javier R. Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, 1(6):22–35, 2006. ISSN 17962048. doi: 10.4304/jmm.1.6.22-35.

[4] Juliano Bazzo and Marcus Lamar. Recognizing Facial Actions Using Gabor Wavelets with Neutral Face Average Difference., 2004.

[5] S. W. Chew, P. Lucey, S. Lucey, J. Saragih, J. F. Cohn, I. Matthews, and S. Sridharan. In the pursuit of effective affective computing: The relationship between features and registration. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(4):1006–1016, 2012. ISSN 10834419. doi: 10.1109/TSMCB.2012.2194485.

[6] Arnaud Dapogny. Pairwise Conditional Random Forests for Facial Expression Recognition. *IEEE International Conference on Computer Vision*, 2015. doi: 10.1109/ICCV.2015.431.

[7] Paul Ekman and Wallace V. Friesen. Facial action coding system: A technique for the measurement of facial movement. *Consulting Psychologists Press Palo Alto*, 12, 1978.

[8] Ian Goodfellow et al. Challenges in Representation Learning: A report on three machine learning contests, 2013.

[9] Jeffrey M. Girard, Jeffrey F. Cohn, and Fernando De La Torre. Estimating smile intensity: A better way. *Pattern Recognition Letters*, 66(November 2017):13–21, 2015. ISSN 01678655. doi: 10.1016/j.patrec.2014.10.004.

[10] Amogh Gudi, H Emrah Tasli, Tim M den Uyl, and Andreas Maroulis. Deep Learning based FACS Action Unit Occurrence and Intensity Estimation. *Fg*, 06:1–5, 2015. doi: 10.1109/FG.2015.7284873.

[11] Jihun Hamm, Christian G Kohler, Ruben Gur, and Ragini Verma. Automated Facial Action Coding System for Dynamic Analysis of Facial Expressions in Neuropsychiatric Disorders. *Journal of neuroscience methods*, 200:237–256, 2011.

[12] Laszlo A. Jeni, Jeffrey M. Girard, Jeffrey F. Cohn, and Fernando De La Torre. Continuous AU intensity estimation using localized, sparse facial feature space. *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013*, (January 2014), 2013. doi: 10.1109/FG.2013.6553808.

[13] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. Joint fine-tuning in deep neural networks for facial expression recognition. *Proceedings of the IEEE International Conference on Computer Vision*, 2015 International Conference on Computer Vision, ICCV 2015:2983–2991, 2015. ISSN 15505499. doi: 10.1109/ICCV.2015.341.

[14] Hiroshi Kobayashi and Fumio Hara. Facial Interaction between Animated 3D Face Robot and Human Beings, 1997.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, pages 1097–1105, 2012. URL http://dl.acm.org/citation.cfm?id=2999134.2999257.

[16] Shan Li, Weihong Deng, and JunPing Du. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Expression Recognition. *Cvpr*, pages 1–10, 2017. ISSN 1063-6919. doi: 10.1109/CVPR.2017.277.

[17] Zisheng Li, Jun Ichi Imai, and Masahide Kaneko. Facial-component-based bag of words and PHOG descriptor for facial expression recognition. *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, (October):1353–1358, 2009. ISSN 1062922X. doi: 10.1109/ICSMC.2009.5346254.

[18] Chengjun Liu and Harry Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 11(4):467–476, 2002. ISSN 1057-7149. doi: 10.1109/TIP.2002.999679.

[19] P Liu, S Han, Z Meng, and Y Tong. Facial Expression Recognition via a Boosted Deep Belief Network. *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, (January):1805–1812, 2014. ISSN 10636919. doi: 10.1109/CVPR.2014.233. URL http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6909629.

[20] Mohammad H. Mahoor, Steven Cadavid, Daniel S. Messinger, and Jeffrey F. Cohn. A framework for automated measurement of the intensity of non-posed facial action units. *2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009*, (June):74–80, 2009. ISSN 2160-7508. doi: 10.1109/CVPR.2009.5204259.

[21] Kenji Mase. Recognition of facial expression from optical flow. *Ieice Transactions - IEICE*, 74, 1991.

[22] S. Mohammad Mavadati, Mohammad H. Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F. Cohn. DISFA: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. ISSN 19493045. doi: 10.1109/T-AFFC.2013.4.

[23] Mohammad Reza Mohammadi, Emad Fatemizadeh, and Mohammad Mahoor. PCA-based dictionary building for accurate facial expression recognition via sparse representation. *Journal of Visual Communication and Image Representation*, 25:1082–1092, 2014.

[24] Ali Mollahosseini, David Chan, and Mohammad H. Mahoor. Going Deeper in Facial Expression Recognition using Deep Neural Networks. (November), 2015. doi: 10.1109/WACV.2016.7477450.

[25] Conference Paper, Hong Kong, Hong Kong, Hong Kong, and Hong Kong. Computer Vision âĂŞ ECCV 2014. 8690(September), 2014. ISSN 16113349. doi: 10.1007/978-3-319-10605-2. URL http://link.springer.com/10.1007/978-3-319-10605-2.

[26] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June-2015(October):1–9, 2015. ISSN 10636919. doi: 10.1109/CVPR.2015.7298594.

[27] Cootes T, Edwards G, and Taylor C J. Active Appearance Model (AAM), 1998.

[28] Yichuan Tang. Deep Learning using Linear Support Vector Machines. 2013. URL http://arxiv.org/abs/1306.0239.

[29] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christopher Bregler. Efficient Object Localization Using Convolutional Networks. (March 2015), 2014. URL http://arxiv.org/abs/1411.4280.

[30] Michel Valstar, Maja Pantic, and Ioannis Patras. Motion history for facial action detection in video. *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, 1(July 2014):635–640, 2004. ISSN 1062922X. doi: 10.1109/ICSMC.2004.1398371.

[31] Zhen Wang and Zilu Ying. Facial Expression Recognition Based on Rotation Invariant Local Phase Quantization and Sparse Representation. 2012.

[32] Jacob Whitehill and Christian W. Omlin. Haar features for FACS AU recognition. *FGR 2006: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, 2006:97–101, 2006. doi: 10.1109/FGR.2006.61.

[33] Tingfan Wu, Nicholas J. Butko, Paul Ruvolo, Jacob Whitehill, Marian S. Bartlett, and Javier R. Movellan. Multilayer architectures for facial action unit recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(4):1027–1038, 2012. ISSN 10834419. doi: 10.1109/TSMCB.2012.2195170.

[34] Xiong Xuehan and F De la Torre. Supervised Descent Method and Its Applications to Face Alignment. *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 532–539, 2013. ISSN 10636919. doi: 10.1109/CVPR.2013.75.

[35] Zhiding Yu. Image based Static Facial Expression Recognition with Multiple Deep Network Learning. *ACM on International Conference on Multimodal Interaction - ICMI*, pages 435–442, 2015. doi: 10.1145/2823327.2823341.

[36] Zhihong Zeng, Ieee Computer Society, Maja Pantic, and Senior Member. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions\n\n. 31 (1):39–58, 2009. ISSN 0162-8828. doi: 10.1109/TPAMI.2008.52.

[37] Guoying Zhao and Matti Pietikäinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007. ISSN 01628828. doi: 10.1109/TPAMI.2007.1110.