Survey on Face Expression Recognition using CNN

Ankit S. Vyas
Dept. of Information Technology
Dharmsinh Desai University,
Nadiad, India
ankitvyas000@gmail.com

Harshadkumar B. Prajapati
Dept. of Information Technology
Dharmsinh Desai University,
Nadiad, India
prajapatihb.it@ddu.ac.in

Vipul K. Dabhi
Dept. of Information Technology
Dharmsinh Desai University,
Nadiad, India
vipuldabhi.it@ddu.ac.in

Abstract*— Recognition of facial expressions plays a major role in many automated system applications like robotics, education, artificial intelligence, and security. Recognizing facial expressions accurately is challenging. Approaches for solving FER (Facial Expression Recognition) problem can be categorized into 1) Static single images and 2) Image sequences. Traditionally, different techniques like Multi-layer Perceptron Model, k-Nearest Neighbours, Support Vector Machines were used by researchers for solving FER. These methods extracted features like Local Binary Patterns, Eigenfaces, Face-landmark features, and Texture features. Among all these methods, Neural Networks have gained very much popularity and they are extensively used for FER. Recently, CNNs (Convolutional Neural Networks) have gained popularity in field of deep learning because of their casual architecture and ability to provide good results without requirement of manual feature extraction from raw image data. This paper focuses on survey of various face expression recognition techniques based on CNN. It includes state-of-the-art methods suggested by different researchers. The paper also shows steps needed for usage of CNN for FER. This paper also includes analysis of CNN based approaches and issues requiring attention while choosing CNN for solving FER.

***Keywords*— Classification, Survey, FER, Face Expression Recognition, Deep Learning, CNN**

I. INTRODUCTION

Facial expressions are natural and powerful signals to interpret human's emotional states and intentions. Nowadays, everything is getting automated through computers. Facial Expression Recognition (FER) has become very popular research subject in the field of computer vision. Recognition of facial expressions can be used in robotics, neuro-marketing, academics, and more significantly in security. We can achieve much by accurately predicting facial expressions of human.

This paper focuses on a survey and analysis of two approaches for solving FER: static images based and image sequences based. Need for this survey is vital because FER is being implemented in many industrial and government sectors. Data is widely available but it needs appropriate rectification. To deal with such huge amount of data could be very time consuming using traditional feature based methods. That's why researchers prefer deep learning techniques, especially CNNs for classification of images.

In real-life scenarios, images may vary in terms of person's head poses, illumination settings, lightening conditions, and background. Expression variation and occlusion are major issues. Recently, S. Li et al. [1] presented a survey of deep learning techniques like DBN (Deep Belief Network), CNN, Auto Encoders and RNN (Recurrent Neural Network). R. Ginne et al. [2] also presented a survey on CNN based FER techniques. In both works, only few CNN based approaches are included and not discussed in depth.

This paper presents a broad survey on CNN based FER techniques. CNNs are proven very robust towards face related analysis [3]. We have done a survey on FER addressing all major problems and their available solutions like pre-processing and data augmentation. The paper discusses the mentioned challenges and highlights how those challenges are tackled by existing works. We have also shown the usage of CNN based techniques in Section II. Moreover, characteristics of CNN architecture are also analyzed to suggest an approximate architecture as per the characteristics of dataset.

This paper contains 4 Sections. Section II describes the background knowledge, usage of CNN based techniques, and issues in current FER. Section III presents a broad survey on FER methods focusing single image and image sequence, and analysis of both the methods along with CNN characteristics. Section IV presents the conclusion.

II. BACKGROUND KNOWLEDGE

A. Facial Expressions

Facial expressions are most efficient and natural way to convey emotions and intentions. S. Li et al. [4] denoted that face expressions represent intensions and emotional state of the person. There are many expressions possible, but P. Shaver et al. [5] concluded that only seven expressions are the prototypic expressions, which are: Anger, Sadness, Disgust, Happiness, Fear, Surprise, and Neutral, into which all other expressions can fall.

B. Traditional feature based approach vs CNN

Traditionally, most researchers have proposed their methods by using classifiers like MLP (Multi-layer Perceptron Model), SVM (Support Vector Machines) and k-NN (k-Nearest Neighbours). These classifiers use handcrafted features like texture and face landmark features, HoG (Histogram of Oriented Gradients), gradient feature mapping, eigen vectors, etc. These features can be extracted by techniques like Gabor filters, LBP (Local Binary Patterns), Eigen Faces, LDA (Linear Discriminant Analysis), and PCA (Principal Component Analysis). Basically, CNNs also use these features, but in their own way. The difference is, in traditional methods, we need to extract features (a.k.a *"handcrafted features"*) manually, whereas CNNs can learn such features automatically by their own.

C. Face expression classification using CNN

To perform FER efficiently, we concluded basic steps shown in Fig. 1.

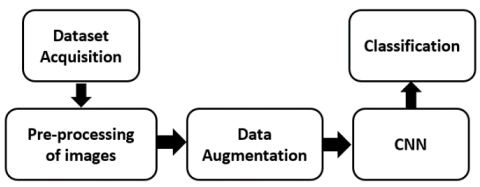

Fig. 1. Basic steps to perform CNN based approach

Each step is discussed next in more detail.

978-1-5386-9533-3/19/$31.00 ©2019 IEEE

a) Dataset Gathering

FER systems are problem specific. They highly depend on which type of images given as input. There are multiple datasets available online for various purposes like frontal face images, posed images, and spontaneous (wild setting) images. Detailed analysis of dataset is done in Section III.

b) Pre-processing of images and data augmentation

Pre-processing methods are divided into three categories: 1) Face detection, 2) Illumination normalization, and 3) Pose normalization. After pre-processing, data augmentation task is performed to obtain adequate training samples. Data augmentation is an approach to generate synthesized samples from original images. Important methods are presented in Table I.

TABLE I. PRE-PROCESSING METHODS

Pre-processing	Method	Researchers
Face Detection	Viola-Jones	[6] [7] [8] [9] [10] [11]
	Dlib	[12]
Illumination Normalization	Histogram Equalization	[6] [9] [13]
	Discrete Cosine Transform	[12] [14]
	Zero-mean Normalization	[13] [3]
Pose Normalization	Frontalization	[11] [15]
	Face Alignment	[16] [17] [18]
	Face Normalization	[19] [20]
Data Augmentation	Data Synthesizing	[21] [7] [22]

c) CNN architecture

Finally, data is fed into CNN model for classification. CNN is a class of deep neural networks. It simply involves convolution operation. Hence, it is named "Convolutional Neural Network". Mainly three layers are included in CNN: 1) Convolutional layer, 2) Pooling layer, and 3) Fully connected layer. Preceding layers extract basic shapes and succeeding layers learn more complex shapes from image. Readers can refer [23], [24] for further details on CNN. Feature maps generated by convolutional layers are decreased by using pooling layers. Fully connected layer has all the connections from previous layer. It computes matrix multiplication of those connections with corresponding biases just like artificial neural networks. We need to keep track about hyper parameters because only useful parameters and information are required to feed into CNN. Otherwise, it leads to problem of overfitting, which gives higher training accuracy, but poor test accuracy. In other words, the model learns unnecessary information and noise during training phase.

The most common issues found in frontal posed images are expression and illumination variability, whereas spontaneous images contain major problem of non-frontal head poses. Major issues and solutions proposed by different researchers are in Table II.

TABLE II. ISSUES IN FER

Issue	Solution
Expression Variation	Select sequence of images from non-peak expression to peak expressions, Select multiple images of same subject (person) with various expression intensities
Illumination Variation	Histogram Equalization, InFace Toolbox, Gamma Intensity Correction, Logarithmic Transforms
Overfitting of Model	Cross-Validation, Pruning, Data Augmentation, Regularization, Early Stopping, Dropout [25]
Uncertainty of Occlusion	Feature Reconstruction, Gradient Direction [26], Sub-region based approach, 3D data based approach
Insufficient Data	Data Augmentation
Non-frontal Face Images	Frontalization [11], Frontal Face Generation

III. SURVEY ON FER

This section presents an intensive survey on both FER approaches.

A. Survey of single image based and image sequence based approaches

Table III analyzes all major works. Widely used datasets are FER-13 [27], JAFFE [28], and CK+ [29]. Extended Cohn Kanade (CK+) dataset includes 593 image sequences of 123 subjects. Japanese Female Facial Expression (JAFFE) dataset includes 213 images of 10 Japanese women. Both datasets are laboratory controlled datasets and has posed images of seven prototypic expressions which are: Happy, Sad, Disgust, Fear, Surprise, Angry and Neutral. Moreover, CK+ contains one more "Contempt" expression, but researchers discard this expression in experiments because of evaluation purpose. FER-13 dataset contains 35887 spontaneous images collected from Google search API and seven prototypic expressions.

V. Mavani et al. [30] presented a novel technique of visual saliency. Visual Saliency is a map of intensity where higher intensities show fields of maximum attention and lower intensities show minimum attention of image. K. Liu et al. [31] proposed the ensemble method which uses multiple CNNs. They created 3 subnets (3 individual CNNs) and combined their results at the succeeding layer. Z. Yu et al. [13] used pre-trained model on FER-13 dataset and fine-tuned on SFEW 2.0 dataset, finally tested on both datasets.

TABLE III. ANALYSIS OF SINGLE STATIC IMAGE BASED AND IMAGE SEQUENCE BASED APPROACHES

Researcher	Dataset	Samples	Subjects	Pre-processing Methods	Classes[a]	Model	Result
K. Liu et al. (2016) [31]	FER2013 [27]	35887	N/A	N/A	6 + 1 (HA, SU, DI, FE, AN, SA) + NEU	Ensemble CNN	65.03%
X. Zhao et al. (2017) [32]	Oulu-CASIA [33]	2880	80	Peak Gradient Suppression (PGS)	6 (HA, SU, DI, FE, AN, SA)	CNN	84.59%
	CK+ [29]	593	123				99.3%
	RaFD	1407	67				95.71%
Uçar et al. (2017) [34]	Cohn-Kanade	2105	182	N/A	6 + 1 (HA, SU, DI, FE, AN, SA) + NEU	CNN	98.70%
V. Mavani et al. (2017) [30]	CFEE [35]	1610	230	Visual saliency, Face cropping	6 + 1 (HA, SU, DI, FE, AN, SA) + NEU	CNN	74.79%
	RaFD	1407	67				95.71%
B. Yang et al. (2015) [36]	CK+ [29]	593	123	Rotation rectification, LBP	6 (HA, SU, DI, FE, AN, SA)	CNN	97.00%
	Oulu-CASIA [33]	2880	80				92.3%

A. Lopes et al. (2016) [21]	JAFFE [28]	213	10	N/A	6 + 1 (HA, SU, DI, FE, AN, SA) + NEU	CNN	53.57%
	BU-3DFE	2500	100				71.62%
	CK+ [29]	593	123				95.79%
Z. Yu et al. (2015) [13]	FER2013 [27]	35887	N/A	Face detection, Likelihood loss, Hinge loss	6 + 1 (HA, SU, DI, FE, AN, SA) + NEU	CNN	Around 80%
	SFEW 2.0	600	68				61.29%
K. Zhang et al. (2016) [37]	CK+ [29]	593	123	Dynamic features, Static features	6 (HA, SU, DI, FE, AN, SA)	PHRNN[e] + MSCNN[f]	98.50%
	Oulu-CASIA [33]	2880	80				86.25%
	MMI	1280 videos	43				81.18%
P. Hu et al. (2017) [15]	EmotiW 2017	1809	N/A	SSE[b], SDM[c], Face frontalization, DCT[d]	6 + 1 (HA, SU, DI, FE, AN, SA) + NEU	CNN	60.34%
A. Mollahosseini et al. (2016) [38]	MMI	1280 videos	43	Bidirectional warping, IntraFace	6 + 1 (HA, SU, DI, FE, AN, SA) + NEU	CNN	77.9%
	DISFA	4845 video frames	27				55.0%
	FERA	289	10				76.7%
	SFEW	663	95				47.7%
A. Fathallah et al. (2017) [39]	MUG	1462	86	N/A	6 (HA, SU, DI, FE, AN, SA)	CNN	87.65%
	RAFD	1407	67				99.33%
	CK+ [29]	593	123				99.33%
E. Ijjina et al. (2014) [26]	EURECOM kinect face dataset	N/A	52	Facial depth by kinect sensor, Gradient direction	6 (Occlusion paper, Occlusion mouth, Left profile, Open mouth, Right profile, Neutral)	CNN	87.98%
K. Shan et al. (2017) [6]	JAFFE [28]	213	10	Face detection, Histogram equalization	6 + 1 (HA, SU, DI, FE, AN, SA) + NEU	CNN	76.7442%
	CK+ [29]	593	123				80.303%
X. Chen et al. (2017) [40]	JAFFE [28]	213	10	Image normalization	6 + 1 (HA, SU, DI, FE, AN, SA) + NEU	CNN	87.735%
	CK+ [29]	593	123				99.16%
W. Li et al. (2015) [7]	CIFE	10595	N/A	Face alignment and rectification, Image cropping	6 + 1 (HA, SU, DI, FE, AN, SA) + NEU	CNN	81.5%
	CK+ [29]	593	123	Face alignment and Rectification			83%
R. Kumar et al. (2017) [9]	FER-2013 [27]	35887	N/A	Viola-Jones algorithm,	6 + 1 (HA, SU, DI, FE, AN, SA) + NEU	CNN	Around 90%
	CK+ [29]	593	123				
H. Li et al. (2017) [41]	BU-3DFE	SS[g] 1: 1200 SS 2: 2400	100 in both subsets	Nose detection, Face cropping, Re-sampling, 3D Face normalization	6 (HA, SU, DI, FE, AN, SA)	DF-CNN	SS 1-86.20% SS 2-81.33%
	Bosphorus 3D	360 (SS)	60				80.00%

[a]Expression classes: Happy-HA, Surprise-SU, Disgust-DI, Fear-FE, Angry-AN, Sad-SA, Neutral-NEU [b]SSE-Supervised scoring ensemble [c]SDM-Supervised Descent Method [d]DCT-Discrete Cosine Transform [e]PHRNN-Hierarchical Bidirectional Recurrent Neural Network [f]MSCNN-Multi Signal CNN [g]SS-subset

E. Ijjina et al. [26] proposed a technique which used only a depth data of an image instead of RGB information because unlike RGB information, depth information is insensitive to illumination conditions. Depth data (bit depth) is the total number of bits for every color patch of single pixel. They used kinect depth sensor to obtain depth data.

K. Zhang et al. [37] proposed Part-based Hierarchical Bidirectional Recurrent Neural Network (PHRNN) to extract facial features from temporal sequences. Mainly two terms are involved: Spatial features and Temporal features. Spatial features are the data represented with specific location and identity whereas Temporal features refer to the data represented in some aspect of time. Therefore, their PHRNN extracts temporal features from consecutive frames and Multi-Signal Convolutional Neural Network (MSCNN) extracts spatial features from still frames to obtain the still appearance information.

B. Analysis of CNN architecture

Table IV shows an intensive analysis of the papers in order to propose a fruitful architecture for chosen problem. After analysis, we found that when we deal with posed images, CNN requires only 1 or 2 convolutional layers which are followed by pooling layers. Mostly max pooling is used by researchers. Convolutional layers require 3x3 to 7x7 kernel size (mostly odd numbers in practice). Number of kernels increase gradually from preceding to succeeding layers. All these preceding layers are followed by one or two fully connected layers, which contain neurons ranging from around 100 to 3072 (depends on problem). Convolutional layers extract useful features and max pool layer is used to reduce dimensions of feature maps generated by convolutional layer.

Spontaneous images require more number of layers to facilitate learning of complex shapes. CNN requires 2 to 6 convolutional layers followed by pooling layers. But it is not always the case to keep pooling layer right after convolutional layer. Some researchers [40], [31] used two convolutional layers consecutively which are then followed by one pooling layer. Intention behind doing so is to learn model more repeatedly and then reduce the dimensions using pooling layer. More number of layers lead to a problem of large number of hyper-parameters. Z. Yu et al. [13] used dropout technique [25] which randomly drops neurons. This randomness is helpful to reduce the risk of overfitting of model. It is used when we have high number of training

parameters. Data augmentation is also used by [21], [7], [22], [36] to minimize problem of overfitting.

We conclude the following points related to CNN architecture:

- A concise CNN with less number of layers and hyper-parameters is preferred for frontal face images.
- A complex CNN with more number of layers and parameters is preferred for spontaneous faces.
- A moderate CNN with balanced number of layers and parameters with proper pre-processing is preferred for occluded images.

IV. Conclusion

This paper focused on effectiveness of convolutional neural network for facial expression classification. FER is useful in many applications like education, HMI systems, and security. This paper conveyed useful background knowledge to understand FER domain along with different pre-processing techniques. It also presented difference between traditional feature based and CNN based approaches for solving CNN. It is concluded that traditional feature extraction based methods became time consuming for selecting appropriate features for learning of model. CNN automatically learns those features efficiently and that is why, CNN can become very convenient for real-world scenarios. We carried out an exhaustive survey and analysis of FER methods using CNN. Furthermore, we analysed CNN architectures of different works and suggested architecture as per the characteristics of dataset.

TABLE IV. Reasoning on CNN Architecture

Dataset	Researcher	CNN Architecture (Layer Name, Size of kernel, No. of kernels, Stride)	Result
CK+ [29] (Posed, Spontaneous)	A. Lopes et al. (2016) [21]	(I/P, 32x32), (Conv1, 5x5, 32), (MaxPool, 2x2, 32), (Conv2, 7x7, 64), (MaxPool, 2x2, 64), (FC1, 256)	95.79%
	X. Chen et al. (2017) [40]	(I/P, 227x227), (Conv1, 255x255, 96), (Conv2, 29x29, 128), (MaxPool, 14x14, 128), (Conv3, 12x12, 156), (Conv4, 6x6, 256), (MaxPool, 3x3, 256), (FC, 512)	99.16%
	W. Li et al. (2015) [7]	(I/P, 64x64), (Conv1, 7x7,32), MaxPool(2:1), (Conv2, 7x7, 32), (MaxPool, 2:1), (Conv3, 7x7, 64), (FC1, N/A)	83%
	R. Kumar et al. (2017) [9]	(I/P, 48x48), (Conv1, 5x5, 64), (MaxPool, 3x3, St-2), (Conv2, 5x5, 64), (MaxPool, 3x3), (Conv3, 4x4, 128), (FC, 3072)	Around 90%
	K. Shan et al. (2017) [6]	(I/P, N/A), (Conv1, 5x5, 6), (MaxPool, 2x2), (Conv2, 5x5,12), (MaxPool, 2x2), (FC1, N/A)	80.303%
	K. Zhang et al. (2016) [37]	(I/P, 64x64), (CROP, 60x60), (Conv1, 10), (MaxPool), (Conv2, 20), (MaxPool), (Conv3, 40), (MaxPool), (Conv4, 40), (MaxPool), (FC1, 80)	98.50%
	A. Fathallah et al. (2017) [39]	(I/P, 165x165), (Conv1, 4x4), (MaxPool, 2x2), (Conv2, 3x3), (MaxPool, 2x2), (Conv3, 3x3), (MaxPool, 2x2), (FC1, 160)	99.33%
Oulu-CASIA [33] (Posed)	K. Zhang et al. (2016) [37]	(I/P, 64x64), (CROP, 60x60), (Conv1, 10), (MaxPool), (Conv2, 20), (MaxPool), (Conv3, 40), (MaxPool), (Conv4, 40), (FC1, 80)	86.25%
JAFFE [28] (Posed)	A. Lopes et al. (2016) [21]	(I/P, 32x32), (Conv1, 5x5, 32), (MaxPool, 2x2, 32), (Conv2, 7x7, 64), (MaxPool, 2x2, 64), (FC1, 256)	53.57%
	Uçar et al. (2017) [34]	(I/P, N/A), (Conv1, 5x5), (MaxPool, 3x3, St-2), (Conv2, 5x5), (MaxPool, 3x3, St-1), (Conv3, 5x5), (MaxPool, 3x3, St-2), (Conv4, 2x2), (Conv5, 1x1), (FC1-N/A)	96.10%
	K. Shan et al. (2017) [6]	(I/P, N/A), (Conv1, 5x5, 6), (MaxPool, 2x2), (Conv2, 5x5,12), (MaxPool, 2x2), (FC1, N/A)	76.7442%
	X. Chen et al. (2017) [40]	(I/P, 227x227), (Conv1, 255x255, 96), (Conv2, 29x29, 128), (MaxPool, 14x14, 128), (Conv3, 12x12, 156), (Conv4, 6x6, 256), (MaxPool, 3x3, 256), (FC1, 512)	87.74%
FER-2013 [27] (Spontaneous)	Z. Yu et al. (2015) [13]	(I/P, 48x48), (Conv1, 5x5), (Stochastic Pool, 3x3, St-2), (Conv2, 3x3, 64), (Conv3, 3x3, 64), (StochasticPool, 3x3, St-2), (Conv4, 3x3, 128), (Conv5, 3x3, 128), (StochasticPool, 3x3, St-2), (FC1, 1024), (FC2, 1024)	Around 80%
	R. Kumar et al. (2017) [9]	(I/P, 48x48), (Conv1, 5x5, 64), (MaxPool, 3x3, St-2), (Conv2, 5x5, 64), (MaxPool, 3x3), (Conv3, 4x4, 128), (FC1, 3072)	Around 90%
	K. Liu et al. (2016) [31]	Subnet-1 (Conv1, 3x3, 64), (MaxPool, 2x2, St-2), (Conv2, 3x3, 128), (MaxPool, 2x2, St-2), (Conv3, 3x3, 256), (MaxPool, 2x2, St-2), (FC1, 4096), (FC2, 4096) Subnet-2 (Conv1, 3x3, 64), (MaxPool, 2x2, St-2), (Conv2, 3x3, 128), (MaxPool, 2x2, St-2), (Conv3, 3x3, 256), (Conv4, 3x3, 256), (MaxPool, 2x2, St-2), (FC1, 4096), (FC2, 4096) Subnet-3 (Conv1, 3x3, 64), (MaxPool, 2x2, St-2), (Conv2, 3x3, 128), (Conv3, 3x3, 128), (MaxPool, 2x2, St-2), (Conv4, 3x3, 256), (Conv5, 3x3, 256), (MaxPool, 2x2, St-2), (FC1, 4096), (FC2, 4096)	65.03%

References

[1] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *Computer Vision and Pattern Recognition,* 2018.

[2] R. Ginne and K. Jariwala, "Facial Expression Recognition using CNN: A Survey," *International Journal of Advances in Electronics and Computer Science,* vol. 5, no. 3, 2018.

[3] C. Garcia and M. Delakis, "Convolutional Face Finder: A Neural Architecture for Fast and Robust Face Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 26, no. 11, 2004.

[4] S. Z. Li and A. K. Jain, "Handbook of Face Recognition," in *Handbook of Face Recognition*, Springer, 2004.

[5] P. Shaver, J. Schwartz, D. Kirson and G. O'Connor, "Emotion Knowledge: Further Exploration of a Prototype Approach," *Personality aid Social Psychology,* vol. 52, no. 6, pp. 1061-1086, 1987.

[6] K. Shan, J. Guo, W. You, D. Lu and R. Bie, "Automatic Facial Expression Recognition Based on a Deep Convolutional-Neural-Network Structure," *IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA),* pp. 123-128, 2017.

[7] W. Li, M. Li, Z. Su and Z. Zhu, "A Deep-Learning Approach to Facial Expression Recognition with Candid Images," *14th IAPR International Conference on Machine Vision Applications (MVA),* pp. 279-282, 2015.

[8] H. W. Ng, V. D. Nguyen, V. Vonikakis and S. Winkler, "Deep Learning for Emotion Recognition on Small Datasets Using Transfer Learning," *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction,* pp. 443-449, 2015.

[9] R. Kumar, R. Kant and G. Sanyal, "Facial Emotion Analysis using Deep Convolution Neural Network," *International Conference on Signal Processing and Communication (ICSPC),* pp. 369-374, 2017.

[10] P. Viola and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001,* pp. I-I, 2001.

[11] T. Hassner, S. Harel, E. Paz and R. Enbar, "Effective face frontalization in unconstrained images," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 4295-4304, 2015.

[12] M. Shin, M. Kim and D.-S. Kwon, "Baseline CNN structure analysis for facial expression recognition," *25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN),* 2016.

[13] Z. Yu and C. Zhang, "Image based Static Facial Expression Recognition with Multiple Deep Network Learning," *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction,* pp. 435-442, 2015.

[14] C. Weilong, E. M. Joo and W. Shiqian, "Illumination compensation and normalization using logarithm and discrete cosine transform in logarithm domain," *ICARCV 2004 8th Control, Automation, Robotics and Vision Conference,* vol. 1, pp. 380-385, 2014.

[15] P. Hu, D. Cai, S. Wang, A. Yao and A. Yao, "Learning Supervised Scoring Ensemble for Emotion Recognition in the Wild," *ICMI '17,* pp. 553-560, 2017.

[16] D. Chen, S. Ren, Y. Wei, X. Cao and J. Sun, "Joint cascade face detection and alignment," *Computer Vision – ECCV,* pp. 109-122, 2014.

[17] X. Cao, Y. Wei, F. Wen and J. Sun, "Face Alignment by Explicit Shape Regression," *International Journal of Computer Vision,* vol. 107, no. 2, p. 177–190, 2014.

[18] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," *IEEE Signal Processing Letters,* vol. 23, no. 10, 2016.

[19] C. Zhang and Z. Zhang, "Improving multiview face detection with multi-task deep convolutional neural networks," *IEEE Winter Conference on Applications of Computer Vision,* pp. 1036-1041, 2014.

[20] V. Blanz, P. Grother, P. J. Phillips and T. Vetter, "Face Recognition based on Frontal Views generated from Non-Frontal Images," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05),* pp. 454-461, 2015.

[21] A. T. Lopes, A. F. D. S. E. de Aguiar and T. O-Santos, "Facial Expression Recognition with Convolutional Neural Networks: Coping with Few Data and the Training Sample Order," *Pattern Recognition.*

[22] S. Yang, P. Luo, C. C. Loy and X. Tang, "Faceness-Net: Face Detection through Deep Facial Part Responses," *IEEE Transactions on Pattern Analysis and Machine Intelligenc,* vol. 40, no. 8, pp. 1845 - 1859, 2017.

[23] M. A. Ponti, L. S. F. Ribeiro, T. S. Nazare, T. Bui and J. Collomosse, "Everything you wanted to know about Deep Learning for Computer Vision but were afraid to ask," *30th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T),* pp. 17-41, 2017.

[24] K. Nogueira, O. A. B. Penatti and J. A. d. Santos, "Towards Better Exploiting Convolutional Neural Networks for remote sensing scene classification," *Pattern Recognition,* vol. 61, no. C, pp. 539-556, 2016.

[25] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research,* vol. 15, no. 1, pp. 1929-1958, 2014.

[26] E. P. Ijjina and C. K. Mohan, "Facial expression recognition using kinect depth sensor and convolutional neural networks," *13th International Conference on Machine Learning and Applications,* pp. 392-396, 2014.

[27] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng and R. L. e. al., "Challenges in Representation Learning: A Report on Three Machine Learning Contests," *International Conference on Neural Information,* p. 117–124.

[28] M. J. Lyons, S. Akemastu, M. Kamachi and J. Gyoba, "Coding Facial Expressions with Gabor Wavelets," *3rd IEEE International Conference on Automatic Face and Gesture Recognition,* pp. 200-205, 1998.

[29] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih and Z. Ambadar, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," *Computer Vision and Pattern Recognition Workshops (CVPRW),* p. 94–101, 2010.

[30] M. Viraj, R. Shanmuganathan and M. K. P, "Facial Expression Recognition using Visual Saliency and Deep Learning," *IEEE International Conference on Computer Vision Workshops (ICCVW),* 2017.

[31] K. Liu, M. Zhang and Z. Pan, "Facial Expression Recognition with CNN Ensemble," *2016 International Conference on Cyberworlds (CW),* pp. 163-166, 2016.

[32] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos and S. Yan, "Peak-Piloted Deep Network for Facial Expression," *Computer Vision – ECCV 2016,* pp. 425-442, 2016.

[33] G. Zhao, X. Huang, M. Taini, S. Z. Li and M. PietikäInen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing,* vol. 29, no. 9, pp. 607-619, 2011.

[34] A. Uçar, "Deep Convolutional Neural Networks for Facial Expression Recognition," *IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA),* pp. 371-375, 2017.

[35] S. Du, A. M. Martinez and Y. Tao, "Compound facial expressions of emotion," 2014.

[36] Y. Biao, C. Jinmeng, N. Rongrong and Z. Yuyu, "Facial Expression Recognition using Weighted Mixture Deep Neural Network Based on Double-channel Facial Images," *IEEE Access,* vol. 6, pp. 4630-4640, 2018.

[37] K. Zhang, Y. Huang, Y. Du and L. Wang, "Facial Expression Recognition Based on Deep Evolutional Spatial-Temporal Networks," *IEEE Transactions on Image Processing ,* vol. 26, no. 9, pp. 4193-4203, 2016.

[38] A. Mollahosseini, D. Chan and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," *IEEE Winter Conference on Applications of Computer Vision (WACV),* pp. 1-10, 2016.

[39] A. Fathallah, L. Abdi and A. Douik, "Facial Expression Recognition via Deep Learning," *IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA),* pp. 745-750, 2017.

[40] X. Chen, X. Yang, M. Wang and J. Zou, "Convolution Neural Network for Automatic Facial Expression Recognition," *International Conference on Applied System Innovation (ICASI),* pp. 814-817, 2017.

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