2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI)
Dec. 22nd, 2017
(Iran University of Science and Technology) – Tehran, Iran

Facial Emotion Recognition using Deep Convolutional Networks

Mostafa Mohammadpour
Department of Computer Engineering
Qazvin Branch, Islamic Azad University, Qazvin, Iran
Email: mohammadpour@qiau.ac.ir

Hossein Khaliliardali
Electrical and Computer Engineering Department,
Semnan University, Semnan, Iran
Email: hossein.khaliliardali@semnan.ac.ir

Seyyed Mohammad. R Hashemi
Image Processing & Data Mining Lab
Shahrood University of Technology, Shahrood, Iran
Email: hashemi@aeuso.org

Mohammad. M AlyanNezhadi
Image Processing & Data Mining Lab
Shahrood University of Technology, Shahrood, Iran
Email: alyan.nezhadi@shahroodut.ac.ir

Abstract*—Facial emotion recognition is an emerging field which use in many nowadays application including social robots, neuromarketing and games. Non-verbal communication methods like facial expressions, eye movement and gestures are used in many application of human computer interaction, which among them facial emotion is widely used because it convey the emotional states and feelings of persons. The emotion recognition is not an easy task because there is no landmark distinction between the emotions on the face and also there are a lot of complexity and variability. In the traditional machine learning algorithm some important extracted features used for modeling the face, so, it can not achieve high accuracy rate for recognition of emotion because the features are hand-engineered and depend on prior knowledge. Convolutional neural networks (CNN) have developed in this work for recognition facial emotion expression and classify them into seven basic categories. Instead of calculating hand-engineered features, CNN calculates features by learning automatically. The novelty of the proposed method is using facial action units (AUs) of the face which first these units are recognized by CNN and incorporate to recognizing the seven basic emotion states. To evaluated the proposed model, Cohn-Kanade database is used so that the model achieves the best accuracy rate 97.01 by incorporating AU while other works in the literature used a direct CNN and achieve accuracy rate 95.75.

Keywords*—*Emotion ,Classification, Convolutional neural networks, Action Unit.

I. INTRODUCTION

The study of facial expressions comes back to Darwin's research on evolution of the species which it appeared as a shape of nonverbal communication. In his studies, the facial behavior was categorized into several groups. This type of communication is faster than verbal and it brought more advantages to the human species than others. Facial emotion stated as person's internal state, intentions and its feeling response to external stimulus [1].

Emotions are an important property of humans and are essential for effective interactions among the society. Humans communication can be either verbal of nonverbal, which it has been shown most of them refer to nonverbal communication [1]. In nonverbal communication, emotion plays effective role because it conveys humans feeling about the subject, and in the psychology research it is proven that facial expressions is more effective then spoken word in conversation[2].

Emotion recognition has intersection of several areas of computer science, cognitive science and psychology, and it can be carried out by several methods such as body language, voice intonation and electroencephalography (EEG) [3]. The easier and practical way is recognizing the emotion from facial expression. So, in the interaction environment, facial emotion recognition is more practical than recognizing emotion from EEG signal because EEG is suit for clinical application such as neurofeedback where the subjects are fixed.

In recent last twenty years, human-computer interactions (HCI) filed has been progressed and play an important role in developing computer science by creating wide variety practical application so that incorporate human beings behavior with computer devices [4]. In the robotics research, particularly humanoid robots, there is interesting to apply emotion recognition on the machine to allow a way of communication naturally [5]. These is also interest to use emotion in the HCI to carry out an efficient and intelligent interaction or communication between human and machine like human beings.

Information from facial expression distributed in different area of face and each of them has different information so that mouth and eyes include more information that cheek and forehead. There were shown on several psychological studies which culture and environment can influence the impact of emotion and the way of expressing feeling for human beings. In many of these studies shown that gender, cultural background, age have bias in expressing emotion while there is not clear evidence on importance of environment for tendency the emotion [6].

Emotion recognition methods can be divided into two main groups: First group work on static images and second one work on dynamic image sequences. In the static approaches, temporal information is not considered and they just use current image information, while in the dynamic approaches images temporal information used in order to recognize expressed emotion in frame sequences.

978-1-5386-2640-5/17/$31.00 ©2017 IEEE

Automatic emotion expression recognition include three steps: face image acquisition, feature extraction, and facial emotion expression recognition. In the optimal extracted features, within-class variations of expression should be minimum while between-class variations should be maximum. If the extracted features are not suite for task in hand and do not have enough information, even the best classifier may be unsuccessful to have best performance.

Feature extraction for emotion recognition can be divided into two approaches: Geometric feature-based methods and appearance-based methods [7]. In the first methods, location and shape of parts of the face such as eyes, mouth, eyebrows and nose are considered, while in the second methods, particular regions or whole of face are considered.

Because of differentiating expressions' feature space is a difficult problem, so expression recognition is sill a challenging task for computers. Some problems may be due to that, extracted features from two faces with equal expression may be different, while extracted features from one face with two expression may be equal, or some expression such as "fear" and "sad" are very similar.

The main contribution of the proposed is to model emotion expression on static images for recognizing seven (happy, surprise, angry, disgust, fear, sad and neutral) facial emotional states. To this end, first images have adjusted by pre-processing algorithms to have exact part of face, and the train a deep CNN for classify the emotions. Novelty of proposed method is using AUs [8] like "Lip stretcher" for recognizing the emotions. When AUs constructed by automatically learned feature by CNN, performance of emotion recognition would increase, because there is a semantic relationships among different AUs.

II. Related Work

In this section, recent approaches on facial emotion recognition which have earned a high degree of accuracy will be discussed. There are many approaches for emotion recognition which are based on hand-engineered features such as histograms of oriented gradients (HOG) features, scale invariant feature transform (SIFT) descriptors, Gabor filter, or local binary patterns (LBP).

A number of methods for emotion recognitions classify the facial image into some basic emotions like happiness, anger and sadness [9] [10] [11], while others try to classify AUs on the face in order to describe objective characteristic of the facial expression [12].

[13] proposed a novel boosted deep belief network (BDBN) in order to performing three steps of feature learning, feature selection, and classifier construction iteratively. The authors claim that their proposed method based on BDBN is able to learn a set of features which can help to characterize facial appearance for emotion expression. Experiments were conducted on the CK+ dataset and JAFFE dataset, and tested the six basic emotion. The proposed approach yields an accuracy rate of 97.70% in the CK+ database. [9] proposed a novel approach based on transformations of given image intensity into a 3D spaces, in order to be invariant to monotonic transformations. Their experiments were performed on two databases of static images static facial expression recognition sub-challenge (SFEW) and emotion recognition in the wild challenge (EmotiW 2015), and achieved a notable 40% increase in performance. [10] developed a emotion recognition system which is called extreme sparse learning (ESL), and is able to learn a dictionary of basis functions and non-linear models. The proposed network incorporate extreme learning machine (ELM) and sparse representation in order to increase accuracy of classification in noisy and imperfect images. The experiments were conducted on CK+ dataset, and achieved an accuracy of 95.33%. [14] performed a study for emotion recognition by analyzing face area such as mouth and eyes, and using principal component analysis (PCA) and neural networks. The experimental setup were conducted on JAFFE and FEEDTUM datasets. [15] proposed a method based on transfer learning of existing convolutional neural network which are fully connected layers and pretrained for emotion expression classification. Experiments were performed on CK+ and JAFFE datasets and yield training accuracy rate of 90.7% and test accuracy rate of 57.1%. [11] proposed an approach for facial expression recognition which uses particular image pre-processing techniques and CNN. By image pre-processing methods they extracted specific features of emotions on the face. The experiments were conducted on three public datasets CK+, JAFFE and BU-3DFE, and achieved 96.76% on CK+ dataset. [16] proposed an approach based on CNN which is independent of any hand-engineered features. Their network structure comprise four parts which first part is for image pre-processing, whereas the others part perform the feature extraction. After extracting features, seven expression are classified by a fully connected layer in CNN. Their proposed structure consists of 15 layers and achieved accuracy rate 99.6% and 98.63% on CK+ anf NMI datasets respectively.

III. Recognition of Emotion in Deep Network

In the current study, seven states of facial emotion are recognized by deep convolutional network which it includes three steps of feature learning, selection, and classification simultaneously. Training network with more than two layers was a difficult problem in last decade that with progress of GPUs, it is possible to train neural network with more than one layers. Deep neural network has three alternating types of layers which includes convolutional, sub-sampling and fully connected layers.

A. Convolutional Neural Network

CNN comes back to 1998 which has been shown is very effective for learning feature and modeling high level of abstraction [17]. CNN includes six components: Convolutional layer, Sub-sampling layers, Rectified linear unit (ReLU), Fully connected layer, Output layer and Softmax layer [18].

Convolutional layer: Convolutional layers are determined by number of generated maps and kernel's size. The kernel is moved over the valid area of the given image (perform a convolution) for generating the map. If f_k be a filter with a kernel size $n \times m$ and is supposed to applied into the given image x, output of the layers can be calculates as follows:

$$C(x_{u,v}) = \sum_{i=-\frac{n}{2}}^{\frac{n}{2}} \sum_{i=-\frac{m}{2}}^{\frac{m}{2}} f_k(i,j) x_{u-i,v-j} \quad (1)$$

Where each CNN neuron has $n \times m$ number of input connections.

Sub-sampling layers: Sub-sampling layers in CNN reduce the map size of previous layer in order to increase the invariance of the kernels. Sub-sampling includes two types of average pooling and maximum-pooling [18]. By applying maximum function in the Max-Pooling, input value is reduced at the x_i. If m be the size of the kernel, output of max-pooling can be calculated as follows:

$$M(x_i) = Max\{x_{i+k,i+l}|\ |k| \leq \frac{m}{2}, |l| \leq \frac{m}{2}k, k \in \mathbb{N}\} \quad (2)$$

Rectified linear unit: A rectified linear unit is a activation function which it simply thresholded at zero and can be calculated as follows:

$$R(x) = max(0, x) \quad (3)$$

ReLU has advantages over tanh/sigmoid function in which it can be implemented by simple thresholding at zero, while in tanh/sigmoid there are expensive operations like exponentials. ReLU is also prevents loosing gradient error, and extremely accelerate the stochastic gradient descent convergence compared with the tanh/sigmoid functions.

Fully connected layer: Fully connected layers are similar to neurons in general neural networks which its neurons are fully connected with every neurons in the prior layer. If the x be input with size k and the number of neurons represented by l in the fully connected layer, the layer can be calculated as follows:

$$F(x) = \sigma(W * x) \quad (4)$$

Where σ is activation function.

Output Layer: The output layer represent class of the input image which its size equal to number of classes. Output vector x produce resulting class as follows:

$$C(x) = \{i|\ \exists i\ \forall j \neq i : x_j \leq x_i\} \quad (5)$$

Softmax layer: The error of the network is propagated back through a softmax layer. If N be the size of the input vector, a mapping can be calculates by softmax such that: $S(x) : \mathbb{R} \to [0, 1]^N$, and each components of the softmax layer is calculated as follows:

$$S(x)_j = \frac{x^{x_i}}{\sum_{i=0}^{N} e^{x_i}} \quad (6)$$

Where $1 \leq j \leq N$.

Learning in the CNN perform by finding the best synapses' weights of neurons. Unlike general neural network which its input consist handcrafted feature, CNN includes raw images as input [19]. In training phase, the network is feeded by training data which include grayscale images with respective labels, and a set of images are separated for validation set in order to determine the best set of weights. In test phase, the network receives a grayscale image, and outputs of the network is predicted class of given image. Figure 1 shown an example of CNN architecture. The network receives an grayscale image 24×24 as input and outputs label of each class. The class that has the highest value, is used as target in the image. The architecture of the CNN includes 2 convolutional layers, 2 pooling (sub-sampling) layers and one fully connected layer that calculate the softmax loss function and scores.

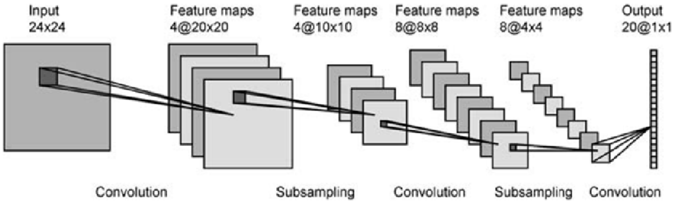


Fig. 1: An example of CNN architecture [20].

TABLE I: Seven basic emotions based on AU codes.

Emotion	Action Unites
Anger	4+5+7+23
Contempt	R12A+R14A
Disgust	9+15+16
Fear	1+2+4+5+7+20+26
Happiness	6+12
Sadness	1+4+15
Surprise	1+2+5B+26

B. Proposed Architecture

In the proposed method, in order to classify seven basic emotion states, a psychological framework which is called Facial Action Coding System (FACS) [8] is used for increasing accuracy of recognizing system, and instead of giving an image to system and classify seven emotion, it has been used coding of facial movements by AUs for classifier output, and determine final emotion state by combination of AUs.

FACS can describe emotion expression by their appearance on the face using AU which is based on anatomical movement of face muscles [8]. AU for facial expression includes 46 atomic component of facial movement and each emotions consist some AUs. FACS can be used in many approached for measuring and describing facial behaviors, and is based on actions of observable face muscles from the anatomical aspect or AUs. It also can estimate intensity of expressed emotion which can bed used in study of complex facial behavior.

There are several codes for AUs which each of them is based on facial muscles, and can be used for study of facial emotion expressions. A list of AUs and action descriptors can be found in [21], that by using them seven basic emotion can be created. Table I shown seven basic emotion which based on combination of some AUs.

In order to detect each AUs from facial expression, it is requires handcrafted high precision features to have good performance. To this end, an CNN is proposed for detecting

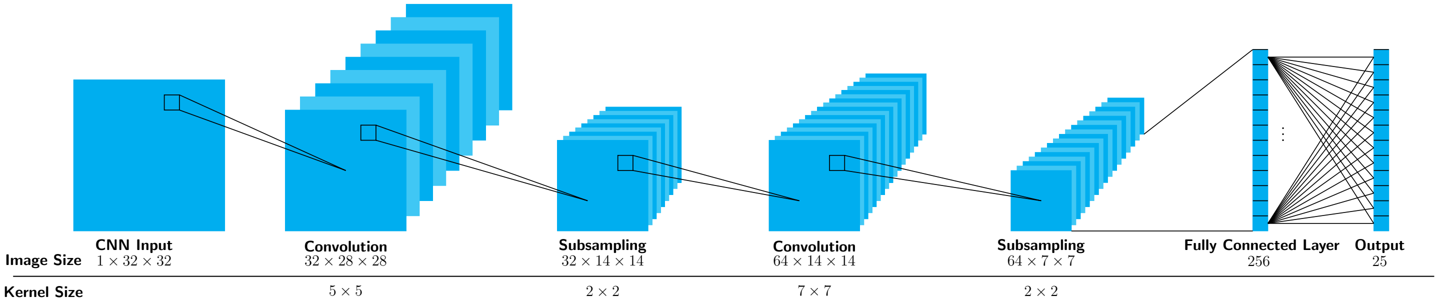


Fig. 2: Architecture of proposed CNN for detecting AUs.

AUs on the face without any hand involvement, and after detecting 25 AUs from given image, by using activation of a set of AUs the expressed emotion can be recognized.

The proposed CNN architecture for detecting AUs include four parts which is shown in figure 2. The network inputs are grayscale images with of size 32×32 and comprise two convolutional layers with 32 and 64 filters, and filter sizes of 5×5 and 7×7. After each convolutional layers, a Rectified Linear Unit (ReLU) activation functions is followed. Max pooling layers are also located after convolutional layers, and a fully connected layer with 256 hidden units are placed. Finally a softmax layer with 25 neurons which indicate numbers of AUs and fully connected to prior layer is considered in order to classification of AUs. Fundamental visual features such as corners, oriented edges and shape of lips, eyes and eyebrow are extracted in the first layer. The synaptic weights among the neurons are calculated by stochastic gradient descent approach. Presence or absence of each AUs denoted by scored in the network output so that low value indicate absence of AUs, while a high value indicated presence of AUs.

IV. Experiments, results and evaluation

The proposed system was trained and tested on Extended Cohn-Kanade (CK+) dataset [22] which is a publicly available and used for facial expression studies. The dataset includes 123 subjects which they were asked to perform a series of emotion expressions, and the images were captured from front of the subjects. Dataset images are grayscale with size 640 by 480 and 8-bit precision, and for each image there is a descriptor file which contain labels for denoting AUs that presented in each images. The dataset includes images for the expressions: neutral, happy, sad, surprise, fear, anger, disgust and contempt. In order to perform a fair comparison with other proposed methods in the literature, contempt expression is omitted. Figure 3 shown some examples of images in the CK+ dataset.

In this work, some pre-processing techniques were applied to the images for extracting exact expressed emotion on the face in order to increase classification accuracy. These set of pre-processing techniques are: Image cropping, rotation correction, down-sampling, spatial normalization, intensity normalization. The pre-processing steps were implemented by C++ and OpenCV, and all others experiment were implemented by NVIDIA CUDA framework on the Ubuntu 16.04 OS. Used hardware for this work is a Intel core i5 3.3 Ghz, and a NVIDA GeForce GTX 730 with 4GB memory and 1152 CUDA Cores.

In order to train the network, cross-validation method used on all images of dataset with batch size 128 and 250 epochs. Moreover, confusion matrix is computed for showing behavior of different class of emotion. Visualization of the confusion matrix is depicted on figure 4.

As demonstrated in the figure 4, there are correlation among angry label and the neutral labels which indicate there are several images with true labels angry which the system predicted them as neutral. The network perform well in estimating happy label than other labels which indicate learning happy feature is easier that other emotions.

Table II represent comparison of average accuracy for all classes of emotion for proposed method and other state-of-the-art methods in the literature which they also used CK+ dataset for evaluation.

TABLE II: Comparison of emotion recognition algorithms for seven expressed emotions on the CK+ dataset.

Method	Description	Accuracy
[23]	SVM + Gabor filters + LBP	88.90
[24]	Local Directional Number Pattern (LDN)	89.30
[25]	LBP + SVM	91.40
[11]	Normalization+ DL	95.75
Proposed	Normalization + Action Units + DL	97.01

V. Conclusion

This paper presented a emotion recognition system with a novel approach for detecting action units (AUs) which is a coding of facial movements in psychological framework. A CNN is developed for optimal feature extraction and detecting AUs and by means of detecting seven expressed emotions. The experimental results proven that deep CNNs are able to learn characteristics of facial expression and increase facial emotion recognition accuracy.

References

[1] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80, 2001.

[2] John N Bassili. Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. *Journal of personality and social psychology*, 37(11):2049, 1979.

[3] Mostafa Mohammadpour, S M R Hashemi, and Negin Houshmand. Classification of EEG-based Emotion for BCI Applications. In *The 7th joint Conference on Artificial Intelligence & Robotics and the 9th RoboCup IranOpen International Symposium*, pages 127–131. IEEE, 2017.

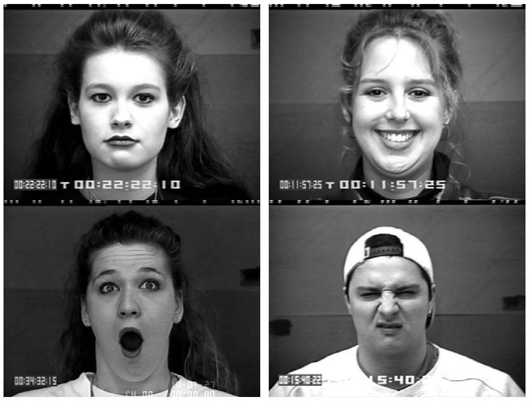
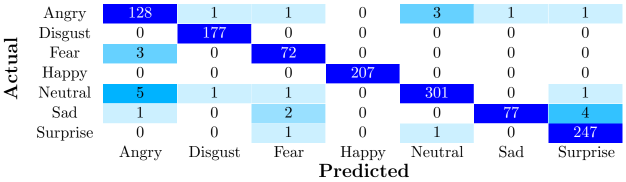


Fig. 3: Some example of images in CK+ dataset. Subjects are in emotional states neutral, happy, surprise and disgust [22].



	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	128	1	1	0	3	1	1
Disgust	0	177	0	0	0	0	0
Fear	3	0	72	0	0	0	0
Happy	0	0	0	207	0	0	0
Neutral	5	1	1	0	301	0	1
Sad	1	0	2	0	0	77	4
Surprise	0	0	1	0	1	0	247

Fig. 4: Confusion matrix of seven expressed emotion on the CK+ database.

[4] N Fragopanagos and John G Taylor. Emotion recognition in human–computer interaction. *Neural Networks*, 18(4):389–405, 2005.

[5] Li Zhang, Ming Jiang, Dewan Farid, and M Alamgir Hossain. Intelligent facial emotion recognition and semantic-based topic detection for a humanoid robot. *Expert Systems with Applications*, 40(13):5160–5168, 2013.

[6] Hillary Anger Elfenbein and Nalini Ambady. On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin*, 128(2):203, 2002.

[7] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58, 2009.

[8] Paul Ekman and Wallace Friesen. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto: Consulting Psychologists*, 1978.

[9] Gil Levi and Tal Hassner. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 503–510. ACM, 2015.

[10] Seyedehsamaneh Shojaeilangari, Wei-Yun Yau, Karthik Nandakumar, Jun Li, and Eam Khwang Teoh. Robust representation and recognition of facial emotions using extreme sparse learning. *IEEE Transactions on Image Processing*, 24(7):2140–2152, 2015.

[11] André Teixeira Lopes, Edilson de Aguiar, Alberto F De Souza, and Thiago Oliveira-Santos. Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order. *Pattern Recognition*, 61:610–628, 2017.

[12] Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):97–115, 2001.

[13] Ping Liu, Shizhong Han, Zibo Meng, and Yan Tong. Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1805–1812, 2014.

[14] Damir Filko and Goran Martinović. Emotion recognition system by a neural network based facial expression analysis. *automatika*, 54(2):263–272, 2013.

[15] Dan Duncan, Gautam Shine, and Chris English. Facial Emotion Recognition in Real Time.

[16] Peter Burkert, Felix Trier, Muhammad Zeshan Afzal, Andreas Dengel, and Marcus Liwicki. Dexpression: Deep convolutional neural network for expression recognition. *arXiv preprint arXiv:1509.05371*, 2015.

[17] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[18] Michael A Nielsen. Neural networks and deep learning, 2015.

[19] A Deshpande. A Beginner's Guide To Understanding Convolutional Neural Networks. *Retrieved March*, 31:2017, 2016.

[20] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, and Gang Wang. Recent advances in convolutional neural networks. *arXiv preprint arXiv:1512.07108*, 2015.

[21] Wallace V Friesen and Paul Ekman. EMFACS-7: Emotional facial action coding system. *Unpublished manuscript, University of California at San Francisco*, 2(36):1, 1983.

[22] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE, 2010.

[23] Thiago H H Zavaschi, Alceu S Britto, Luiz E S Oliveira, and Alessandro L Koerich. Fusion of feature sets and classifiers for facial expression recognition. *Expert Systems with Applications*, 40(2):646–655, 2013.

[24] Adin Ramirez Rivera, Jorge Rojas Castillo, and Oksam Oksam Chae. Local directional number pattern for face analysis: Face and expression recognition. *IEEE transactions on image processing*, 22(5):1740–1752, 2013.

[25] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009.