

Emotion Recognition Using Representative Geometric Feature Mask Based on CNN

Shaosong Lin

Department of Computing
Xi'an Jiaotong-Liverpool University
Suzhou, China
Shaosong.Lin17@student.xjtlu.edu.cn

Yong Yue*

Department of Computing
Xi'an Jiaotong-Liverpool University
Suzhou, China
Yong.Yue@xjtlu.edu.cn
*Corresponding author

Xiaozhu Zhu

Department of Computing
Xi'an Jiaotong-Liverpool University
Suzhou, China
Xiaohui.Zhu@xjtlu.edu.cn

Abstract—Emotion recognition is a growing area of facial recognition, to detect the basic emotion state of a person and then operate further analysis. For practical applications, high speed and accuracy are required as an efficient and precise system. To this end, the paper proposes an effective emotion recognition system using a representative geometric feature mask for feature extraction and a CNN model for classification. Compared with traditional emotion recognition systems, which usually extract facial key features and then convert them into mathematical information variables by equations, the system implemented in this paper extracts necessary features in facial expression through landmarks, and operates a further extraction by a transformation that converts features into a pure geometric feature mask to represent a simplified human face. Then, the mask that can be used to express the human facial emotion with fewer noise features, is input into a deep learning training CNN (Convolutional Neural Network) model. The improvement of this work is that the system combines pure geometric method to extract facial features with CNN algorithm properties in image processing, where local connectivity and shared parameter properties were fully used in further geometric feature extraction. Finally, the system achieves high accuracy and low time costs with KDEF (Karolinska Directed Emotional Faces) and CK+ (Cohn-Kanade AU-Coded Expression Database).

Keywords—Emotion recognition, Facial feature extraction, Geometric feature mask, Convolutional neural network

I. INTRODUCTION

Emotions are of importance in daily communications and they are important biological signals to show the attitude or the motivation of a person, so emotion recognition can be potentially useful in a number of applications. Emotions can be used for not only communications and management, but also public security or potential risk detection. For emotion recognition, physiological measurements such as ECG (electrocardiogram) heart rate, EOG (electrooculogram), EEG (electroencephalogram) measurements can only provide some physiological features when an irritant state occurs and equipment is also needed attaching to the person, which is not accurate enough and inconvenient for emotion recognition. In addition to physiological signal measurements, computer vision techniques can provide some solutions for human face detection and emotion classification. In recent years, with the increase of computational power and the availability of rich databases, deep learning algorithm techniques such as CNNs (convolutional neural networks), have been applied in a wide range of computer vision challengeable tasks, which include large-scale image classification systems, visual recognition, and high-dimensional shallow feature encodings, and they

have also shown a state-of-the-art accuracy level in those challenges [1].

In addition to the CNN classifier, proper methods for selecting and filtering features are also vitally important. Some traditional recognition methods aim to translate the geometric information into mathematical quantities by equations, which are abstract meanings of extracted features. However, with the relatively stronger ability of CNN algorithms on image processing, feature extraction of pure geometric shapes or textures may provide a new direction for facial emotion recognition based on CNN. The goal of this work is to propose a geometric feature extraction method for emotion recognition that can make efficient use of CNN algorithm properties in image processing. Compared with traditional works that transfer features into mathematical variables, the extraction method in this work focuses on vital human face features in facial expression such as shapes of eyes and mouth, and filters irrelevant features to reduce noise features. Then, the combination with CNN algorithm not only can ensure the local connectivity of features to remain geometric feature details, but also can maintain globality of facial features during convolution operations. Besides, the efficiency can also be improved by CNN shared parameter property that can reduce the cost of calculation. The rest of the paper will firstly review related work, discuss in detail the proposed method, and present experimental results.

II. RELATED WORK

In recent years, applications of computer vision techniques became wider and more mature, and for emotion recognition, researchers have proposed various methods to process human facial features with different data training schemes.

Some researchers transformed geometric features into corresponding mathematical variables as the model training input by specific equations. Aihua Li et al. [2] presented a novel emotion recognition method by combining SAAE (Specific Angle Abundance Entropy), improved GMRF (Gaussian Markov Random Field) and SVM (Support Vector Machine) techniques. It first used a CNN model to separate the face region into different segments and the mouth segment was selected as a factor to recognize facial expressions. Then, improved GMRF was used to enhance the image, and SAAE was applied to extract mouth shape features. When the emotion expressions were different, it showed different SAAE curves in 10 dimensions. Finally, the improved GMRF texture features combined with the SAAE features were input into the SVM model.

Moreover, researchers also used geometric facial feature differences between emotion images and neutral images to recognize the emotion for a single person. Rathee and Ganotra [3] proposed an emotion recognition approach that modelled facial feature deformations using TPS (Thin plate spline). The approach firstly found human facial landmark locations in both neutral images and emotion images, and then computed the TPS parameters from source points in the neutral image and target points in the emotion image, where calculated TPS parameters were inputs of the SVM classifier for emotion recognition. TPS can map the non-linear deformation caused to facial muscles and its geometric significance is to generate the warp for different emotions, which refers to the neutral emotion, so the model with TPS parameters as input belongs to the emotion recognition application using geometric features. Besides, Ulukaya and Erdem [4] proposed an emotion recognition solution fusing geometric features with appearance-based features. In their work, a neutral shape dictionary was first estimated by a neutral frame and CBF (coordinate-based features) with GMM (Gaussian mixture model); based on the dictionary, the best fitting neutral face shape, which was called the baseline, was selected for each single input face image. Next, the motion vectors calculated from the baseline and the emotion image were fused with SIFT (Scale Invariant Feature Transform) descriptors as the input of SVM classifier for emotion recognition. This work is dependent on the baseline in the neutral shape dictionary to distinguish the emotion for each person, which also applied geometric feature transformation for the model training.

However, emotion recognition using the facial feature difference is not always dependable when the neutral image is not accessible. Another research direction is to operate a further feature extraction by constructing new facial geometric features based on the extracted features. Joseph and Geetha [5] propose a modified eyemap-mouthmap algorithm to extract human face features and a CNN model was applied to train the obtained features. The Viola Jones algorithm cached the eye and mouth areas in rectangular boxes and the landmark-based geometric construction of 16 triangles was operated in boxes. Moreover, Majumder et. al [6] proposed a recognition model using a 26-dimensional facial geometric feature vector as the input for the KSOM (Kohonen self-organizing map) classifier. The model used different algorithms to extract facial features such as eye, eyebrow, nose and lip respectively, and then constructed them into the 26 directional displacement features data as the classifier training input. The KSOM classifier can cluster the data and maintain the topology of input data during the extraction, which means that the abstractions of geometric human facial feature information are remained, so they implement an emotion recognition model using geometric facial features alone.

Among different classifiers such as SVM, CNN, fuzzy system, multi-SVM, CNN shows relatively high accuracy and less cost. CNN contains different layers to extract and filter features, which generally are convolutional layer, pooling layer, and fully connected layer. The convolutional layer is the fundamental building block in the CNN model, contains a set of learnable kernels that are also called filters to extract the salient data from the input data matrix [7]; the pooling layer reduces the network dimension and computation by a further extraction on the feature graph [8], and it can filter unnecessary details and remain useful information activation functions [9]; the fully connected layer, usually as the final layer of the network, implements the classifier function and it

will collect all features for classification and the number of its output neuron equals the number of recognizing classes [7]. The structure of CNN algorithm defines its unique properties in image processing, which can ensure both globality and local connectivity. And the training efficiency can also be improved by shared parameters among layers.

To summary the existing work discussed above, emotion recognition can be implemented by using transformed mathematical variables from geometric features or using purely geometric facial features alone. Most work chose eyes and mouth as important features in facial expression and operated a further extraction on those features and then input them into the training model for emotion classification. Rathee and Ganotra [3], and Joseph and Geetha [5] provided a direction for further feature extraction in this paper, which used human face key landmark locations to construct a geometric feature mask for the representation of an abstract face with fewer noise features. Due to the properties of CNN in image processing discussed above, this paper proposes an effective emotion recognition method using geometric features alone, and the method enhances key face features in facial expression such as the eyes or mouth, which is done by further processing such as landmark construction on extracted features, and then inputs them into a CNN model for classification. Details of the work are discussed in next sections.

III. METHODOLOGY

The paper proposes an effective feature extraction method using key human facial features in facial expression. The feature extraction procedure bases on the facial landmark location [3, 5], and a representative geometric feature mask is constructed as the abstraction of the human face. Then the abstracted mask is used as the input of the classifier.

The methodology contains three procedures for emotion recognition: face alignment to find facial landmark locations for key human facial features, geometric feature mask construction to select features, and a CNN model to classify the constructed feature mask.

A. Face Alignment

Face alignment aims to detect important human facial landmarks such as the eyes, nose and mouth landmarks, and can be implemented in different ways. 3D modelling, MTCNN (multi-task CNN), and regression-based methods can all solve the face alignment task. However, to find facial landmarks in a relatively lightweight way, an entry-level alignment algorithm using the cascade of tree-based regressors [10] is used in the proposed method.

The face alignment algorithm is implemented by the cascade of regressors using the tree boosting algorithm. During the training process, the decision tree chooses the splitting node with a minimum loss and each regressor delivers updated tree information into the next cascade-level regressor.

Firstly, in the tree-based regressors, each regressor function is implemented by the node splitting in the tree structure to fit the residual targets, and selects the node with a minimum error [10]. To get a crude approximation shape of initial landmark positions Q at each node, a global similarity transform algorithm is applied, which is based on the image mean shape [10]. Then, to train the regression tree, a set of candidate splits θ are created at each node. The core of the

algorithm is to find the θ that can minimize the error E in (1), which means the difference between the training data and prediction should be as small as possible. In (2), s means the selection of the left or right node, r_i is the vector of all the residuals computed for image i and $\mu_{\theta,s}$ is the average of parent node calculated in (2).

$$E(Q, \theta) = \sum_{s \in \{l, r\}} \sum_{i \in Q_{\theta, s}} \|r_i - \mu_{\theta, s}\|^2 \quad (1)$$

$$\mu_{\theta, s} = \frac{1}{|Q_{\theta, s}|} \sum_{i \in Q_{\theta, s}} r_i, \text{ for } s \in \{l, r\} \quad (2)$$

Next, the algorithm to find human face landmarks is constructed by the cascade of regressors, where the single regressor has been described in the previous section. The structure of the cascade is shown in (3), where I stands for the input image, $\hat{S}^{(t)}$ stands for the shape vector in the current cascade level t and $r_t(\cdot, \cdot)$ stands for the regression function, which uses the gradient tree boosting algorithm with the square error loss sum [10]. In each cascade, $r_t(\cdot, \cdot)$ function predicts an update vector from image I and is added to $\hat{S}^{(t)}$ for the next cascade-level prediction.

$$\hat{S}^{(t+1)} = \hat{S}^{(t)} + r_t(I, \hat{S}^{(t)}) \quad (3)$$

The machine learning process is iterated until a sufficient level of accuracy is achieved, and then a cascade of T regressors r_0, r_1, \dots, r_{T-1} are learnt. An example is shown in Fig.1, if there are more cascades involved before the overfitting, the model will be more accurate in the recognition, which means the predicted shape vector will be nearer to the ground truth.

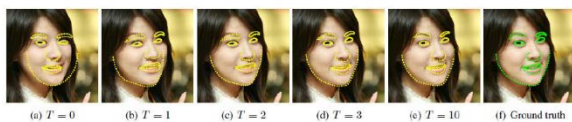


Fig. 1. The training process of the cascade [10]

B. Geometric Feature Mask

Facial features such as the eyes, mouth and eyebrow shapes are of vital importance in facial expression and emotion recognition. In the work of Aihua Li et al. [2], the geometric information was translated into mathematics information such as SAAE, which was not fully used, and the geometric information without a proper feature selection is not dependable for the classification due to noise features. However, Joseph and Geetha [5] provide a new research direction that targeted human eyes and mouth as extracted features, and then features were constructed as landmark triangle masks demonstrating good performance with high accuracy. Therefore, in this work, the geometric transformation of extracted features focuses on those key human facial features and the algorithm constructs them into a geometric feature mask for feature selection.

Firstly, 68 different facial landmarks, which are outputs of the face alignment algorithm, are extracted, and each of them is signed a specific ID. An example of extraction using the KDEF database [11] is shown in Fig. 2.

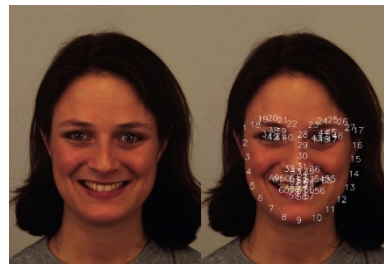


Fig. 2. 68 key points extracted (KDEF AF01HAS) [11]

Then, the further feature extraction bases on these 68 landmarks and the specific procedure is shown in Fig. 3. To reduce the influence of noise features such as haircut, environment and clothes, the image is clipped and only the face area remains. The face outline information contributes less in the recognition procedure, so key points from ID 1 to ID 17 are omitted in the geometric feature mask. Then, the remaining points are constructed as a facial expression mask shown in Fig. 3(d) [11]. The final mask contains the shape of eyebrows, eyes, nose, mouth, and upper and lower lips, which are important references in facial expression. In addition, unnecessary features like nasolabial folds are all eliminated. Therefore, as the input data of CNN model, it has been relatively detailed for vital facial features and it has reduced noise features where possible.

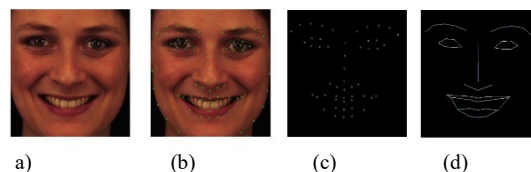


Fig. 3. Geometric feature mask construction example (KDEF AF01HAS) [11]

C. Convolutional Neural Network

Compared with traditional classifiers, the convolutional neural network (CNN) can solve problems by modelling the data into smaller pieces and combining them using deep networks [12]. The CNN model contains multiple layers, which are convolution layers for feature extraction, pooling layers for subsampling and fully connected layers for classification respectively. Moreover, during the model training, the Back Propagation (BP) algorithm is the core algorithm for parameter updating between different layers, and those convolution parameters are translation-invariant and shared among layers, and hence the network can detect features with less computation [13].

In the convolutional layer, the activation function such as sigmoid, tanh or ReLU function, is employed to decide whether an individual neuron will be activated and then work in the next layer [8], which means whether the extracted data will be saved in the feature graph map and used in the next layer.

Then, in the pooling layer, the general pooling methods are max pooling and average pooling, and there is no pooling parameter in most pooling layers. Boureau [14] has made a comparison between max pooling and average pooling. The results showed the remarkable performance of max pooling in different experiments, and thus this work applies max pooling as the pooling layer.

In addition to general CNN layers, the dropout layer, an unnecessary layer in the network but helpful, avoids the

overfitting during the model training. The dropout refers to dropping out unit nodes in the network, which are neurons, and the selection of dropout is random because it ignores the layer neurons during the training process. Then, due to the random dropout, the interaction relations of nodes are not fixed, so the possible dependence of several features on another particular feature can be reduced [15]. Moreover, the coupling between nodes is also weakened, so the effect of overfitting during the training is weakened [15].

IV. RESULTS AND EVALUATION

A. Database

Two human face datasets are used as training data in this paper, which are KDEF (Karolinska Directed Emotional Faces) [11] and CK+ (Extended Cohn-Kanade Dataset) [16, 17].

KDEF dataset contains 4900 images of 7 different emotions for human facial expressions, which are neutral, happy, angry, afraid, disgusted, sad and surprised emotions respectively [11]. The images are from 70 individuals containing 35 males and 35 females in 5 angles, which are full left profile, half left profile, straight, half right profile and full right profile [11]. In the paper, the experiment only focuses on the straight and half profile angles for emotion recognition, so only 2940 of 4900 images are used.

CK+ contains 10708 images of 7 different dynamic emotions for human facial expressions, which are angry, disgust, fear, happy, sadness, surprise and contempt emotions [16, 17]. Compared with the KDEF dataset, CK+ has more dynamic record frames of emotions, which are the emotion change from neutral to the specific facial emotion. For the model training of static emotions, the last several frames have more obvious features for the specific emotion. So, for the selection, the first emotion record frame of an individual is chosen as the neutral emotion, and the last several frames are selected as the specific emotion. However, compared with KDEF, the emotion distribution in CK+ is not uniform, so the number of each emotion is different.

B. Implementation

Firstly, face alignment was implemented using the Dlib library, which is a modern C++ toolkit containing many machine learning algorithms and can solve real-world problems by creating complex software programs in C++. Then, the CNN framework is applied with Tensorflow 2.0, which is a frame developed by Google Brain Group. It contains rich and practical toolkits to support deep neural networks and machine learning. The frame has a high degree of flexibility that can enable user-defined data flow diagrams and also has performance optimization that can mine the computing power of personal computing equipment potential by calling both the CPU and GPU.

C. Results and Discussion

This section discusses the recognition accuracy and loss for the two databases used respectively. For each database, there is a comparison test between the original clipped human face image and geometric feature mask image as input data of the CNN model. A comparison with other related work, which is reviewed in the previous section, is also made.

1) Comparison experiments on KDEF and CK+

For each database, comparison experiments are conducted respectively. In the experiment, the input images are different, which are clipped original images such as Fig. 4(a) and geometric feature mask images such as Fig. 4(b) respectively [16,17]. The training results are shown in Table I below.

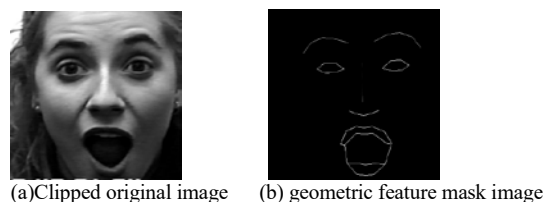


Fig. 4. Input image examples in Comparison Experiments (CK+ S014 ©Jeffrey Cohn) [16,17]

TABLE I. COMPARISON EXPERIMENT RESULTS ON KDEF AND CK+

Input	Database	Training Results		
		Accuracy	Loss	Training epoch
Clipped original image	KDEF	85.62%	0.4195	20
Geometric feature mask image	KDEF	97.72%	0.096	15
Clipped original image	CK+	90.76%	0.2760	20
Geometric feature mask image	CK+	97.84%	0.1290	15

In comparison, the model with input of the geometric feature mask images has better performance during the training, which has higher accuracy and a lower time cost. For the KDEF dataset, the test set accuracy is increased from 85.62% to 97.72%, and for the CK+ dataset, the test set accuracy is increased from 90.76% to 97.84%. Then, for the time cost, the number of training epochs is reduced from 20 to 15 epochs, which means the model needs less training time to reach the best performance before overfitting.

Experiment results show an increase in accuracy, a decrease in training cost and a higher learning efficiency during the training. This can indicate that the extraction of the facial feature mask not only amplifies weights of key facial features such as the eyes and mouth on a human face image, but also reduces unnecessary features such as haircut and facial wrinkle, which do not contribute to the emotion

expression, so the model recognition performance is enhanced.

2) Comparisons with other works

In addition, there are also comparisons among similar emotion recognition methods mentioned in the previous section. Table II summarises the performance of recognition in each related method with the database, corresponding features, the training classifier and the recognition accuracy.

In the comparison, the method proposed in the paper has a better emotion recognition performance than the most of related methods. It shows approximately 5 percent higher accuracy than the other methods. Therefore, the application of the CNN model in emotion recognition performed better than SVM models in this experiment. Moreover, both the proposed method and the work of Joseph and Geetha [5] achieved high accuracy of around 98% in the recognition, so feature

extraction using facial landmark locations can be considered as an effective method to make full use of key features in facial

expression recognition.

TABLE II. ACCURACY COMPARISONS AMONG RELATED WORKS

Method	Database	Features	Classifier	Accuracy
Aihua Li et al. [2]	JAFFE	GMRF features with SAAE	SVM	92.9%
Rathee and Ganotra [3]	CK+	TPS parameters	SVM, N-fold	90.0%
Ulukaya and Erdem [4]	CK+	Geometric + SIFT feature	SVM	90.0%
	MMI	Geometric + SIFT feature	SVM	67.0%
Majumder et. al [6]	MMI	26-dimensional geometric feature vector	multi-class SVM	92.5%
Joseph and Geetha [5]	KDEF	angle of the constructed triangles mask	CNN	98.1%
This work	KDEF	Representative geometric feature mask	CNN	97.7%
	CK+	Representative geometric feature mask	CNN	97.8%

V. CONCLUSION

This paper has proposed an effective emotion recognition system using a representative geometric feature mask based on a deep learning CNN model. The system contains three procedures to recognize the emotion: the extraction of key human face landmarks, the construction of a geometric feature mask and the classification using a CNN model. During the feature selection, important features for facial expression are enhanced while unnecessary features are filtered. Moreover, two common emotion datasets KDEF and CK+ are used as training input in experiments, and the proposed method achieved remarkable performance with a high recognition efficiency and high accuracy of about 98%; hence a timely and accurate emotion recognition system was constructed.

Compared with related methods that transform features into mathematical variables by equations, this work improved by using purely geometric features as the training model input and then combining it efficiently with CNN algorithm. The geometric information is not replaced by the mathematics information like SAAE, while geometric properties remain such as facial shape and location relationship, and then the method enhanced them during the feature extraction process. Besides, its combination with CNN model not only shows the feasibility of feature selection, but also makes full use of the CNN properties to enhance the local spatial correlation by local connectivity property, and the globality also remains in image processing. This work may be extended for computer vision with deep learning to further improve facial expression recognition performance.

ACKNOWLEDGMENT

This work is under partial support from an XJTLU Key Programme Special Fund (KSF-A-19).

REFERENCES

[1] A. Agrawal and N. Mittal, "Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy," *The Visual Computer*. vol.36, pp. 405–412, Jan. 2019.

[2] A. Li et al., "A Facial Expression Recognition Model Based on Texture and Shape Features," *Traitement du Signal*. vol.37, no.4, pp.627-632, Aug.2020.

[3] N. Rathee and D. Ganotra, "Modelling Facial Features for Emotion Recognition and Synthesis," *IETE Journal of Research*. vol. 63, no. 6,

pp. 845-852. Nov. 2017.

[4] S. Ulukaya and C.E. Erdem, "Gaussian mixture model based estimation of the neutral face shape for emotion recognition," *Digital Signal Processing*. vol. 32, pp. 11-23. Sept. 2014.

[5] A. Joseph and P. Geetha, "Facial emotion detection using modified eyemap-mouthmap algorithm on an enhanced image and classification with tensorflow," *Visual Computer*. vol.36 Issue 3, pp.529-539. 11p. Mar. 2020

[6] A. Majumder et al., "Emotion recognition from geometric facial features using self-organizing map," *Pattern Recognition*, vol.47, no. 3, pp. 1282-1293, Mar. 2014.

[7] T. Bwzdan and N. Bacanin, "Convolutional Neural Network Layers and Architectures," in *Sinteza 2019*. Jan.2019.

[8] S. M. Razak and B. Jafarpour, "Convolutional neural networks (CNN) for feature-based model calibration under uncertain geologic scenarios," *Computational Geosciences*. vol.24, pp.1625-1649, Jun.2020.

[9] Z. Zeng et al., "CNN Model Design of Gesture Recognition Based on Tensorflow Framework," in *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, Chengdu, China. Mar.15-17, 2019.

[10] V.Kazemi and J.Sullivan, "One Millisecond Face Alignment with an Ensemble of Regression Trees," in *Computer Vision and Pattern*. Columbus, Ohio, USA. Jun.2014.

[11] E.Lundqvist, D., Flykt, A., & Öhman, A. (1998). *The Karolinska Directed Emotional Faces - KDEF*, CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, ISBN 91-630-7164-9.

[12] S. Gollapudi, "Deep Learning for Computer Vision," in *Learn Computer Vision Using OpenCV*, ed., Hyderabad, Telangana, India: W.Spahr et al., 2019, ch.3, pp. 51-70.

[13] S. Jaiswal and G. C.Nandi, "Robust real-time emotion detection system using CNN architecture," *Neural Computing and Application*. vol.32, pp.11253–11262, Oct.2019.

[14] Y. Boureau et al., "Learning mid-level features for recognition" in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, 2010.vol.1, pp.2559-2566.

[15] J. Yang and G. Yang, "Modified Convolutional Neural Network Based on Dropout and the Stochastic Gradient Descent Optimizer," *Algorithms*. vol.11, no.3, pp.28. 15p, Mar. 2018.

[16] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete expression dataset for action unit and emotion-specified expression. *Proceedings of the Third International Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB 2010)*, San Francisco, USA, 94-101.

[17] Kanade, T., Cohn, J. F., & Tian, Y. (2000). Comprehensive database for facial expression analysis. *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG'00)*, Grenoble, France, 46-53.