Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

Synthetic Sentiment Energy:
Situation Assessment in Crowds

JINJIE NI, HONGYI NIE, LUOMENG ZHANG, FENG YANG
School of Automation,
Northwestern Polytechnical University,
Shannxi, Xi'an, 710072,
Email: oliver@mail.nwpu.edu.cn
Email: hy_nie@foxmail.com
Email: p65178007@mail.nwpu.edu.cn
Email: yangfeng@nwpu.edu.cn

***Abstract*—Group Emotion Recognition has been recognized as a difficult task because of the difficulty in defining identification elements, the overlap of people, and the differences of individuals. This paper first discusses the two elements required for group emotion recognition: The facial expression on each person's face and the position of each person's face in the crowd. Then, it design Binary-Channel CNN and Yolo V3 respectively as a strategy to obtain these two kinds of information. This paper propose a Sentiment Energy Function to measure group sentiments, which consists of three components: Expression energy, Expression Change Rate Energy and Displacement energy. These three components were fused aiming to achieve the group emotion energy. Through experiment methods, it is clear that the improved Binary-Channel CNN in this paper perform good in facial expression recognition. The group emotion recognition system designed has comparatively higher robustness and adaptability. The results obtained are consistent with the results of human analysis.**

***Index Terms*—Group Emotion Recognition; Binary-Channel CNN; Yolo V3; Sentiment Energy Function**

I. INTRODUCTION

At present, the recognition of group emotions mainly depends on group behavior. When the group is in the static behavior of watching, listening, etc., group facial expression recognition is the most effective method. When the group is in the dynamic behaviors such as movement, running, and incitement, the emotion of the group can also be assessed by the magnitude of the facial or limb displacement.

Ekman and Friesen[1] have pioneered the work of modern facial expression recognition. Ekman defines six basic expressions of human beings: happy, angry, surprised, fearful, disgusted, and sad. The categories of expression identifying are determined. After that, a facial motion coding system is established, which enables researchers to divide a series of facial motion units according to the system. It relies on the relationship between facial movements and expressions and then detect details of facial expressions.

Lopes et al.[2] have done facial expression related work by using deep learning methods to recognize facial expressions. Supervised neural network learning requires a large amount of data as a support, and existing data is too limited. So they expand the sample size by data expanding methods, remove the unrelated noise in the environment by preprocessing. They finally achieved a good result.

Yu et al.[3] tried multiple CNN cascades or concatenations to recognize facial expressions, and learned the combined weights to make the fusion effect of the network better. This implementation has certain cross-database capabilities.

Hu Bin, Xia Gongcheng[4] conducted integration knowledge description and qualitative simulation studies in group behavior around 2004. Aiming at the qualitative simulation problem of group behavior change process, Qualitative Causal Reasoning and QSIM algorithm ideas were used. The knowledge description methods related to crowd behavior and group distribution are studied, including the external environment, internal cultural environment, management measures, the status of various elements in the group and the description of the relationships between the various elements. The work they have done distinguishly established a set of mechanism for behavior description and simulation.

Rassadin, Alexandr G et al.[5] did a research on group-level emotion recognition. It extracts the detected facial feature vectors using a Convolutional Neural Network trained for facial recognition tasks, rather than the traditional emotional recognition methods. A 75.4% accuracy was achieved on the validation data, which is 20% higher than prior feature based methods.The work finished by them initiates a deep learning method in group emotion recognition, and signifies a high accuracy of group sentiment analysis.

The model proposed in this paper combines Facial Expression Percentage and Body Displacement. Face localization algorithm and expression recognition algorithm were used to locate the position of each face in the crowd and recognize the facial expression presented by each face. The group emotions are more intuitively assessed by statistically analyzing the proportion of expressions, the rate at which expressions change, and the rate at which the test subjects move.

The flow chart of the Group Sentiment Assessment System is as follows:

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

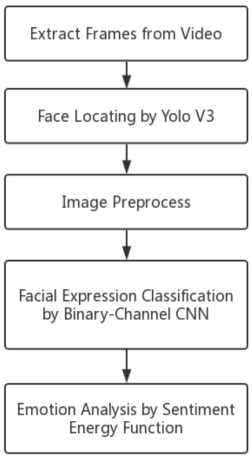

Fig. 1: Flow chart of group emotion recognition system.

II. Basic Models

A. Binary-Channel CNN

This paper improves a Binary-Channel CNN[6] to identify expressions. If the original image is directly applied into the CNN training, it will extract all the feature information of the image. While in the expression recognition research, the sample facial detail information is more concerned. Some irrelevant information such as hairstyle, hair quality, background, ethnicity, photograph angles, etc. should not be counted. Therefore, instead of directly extracting the features of the gray face image using one channel model, a second channel is needed to extract the detailed face information, that is, the LBP image of faces[7], which reflects the details of the facial expression to a larger extent, such as some wrinkles that appear when people laugh, and the deformation of the corners of their mouths and cheeks. Therefore, this paper establishes a fusion model. The recognition results of both channels are weighted and fused, so that the network has both local and global information. It makes full use of information to judge facial expressions, thereby improving the recognition accuracy. The structure of Binary-Channel CNN is shown in Fig. 2.

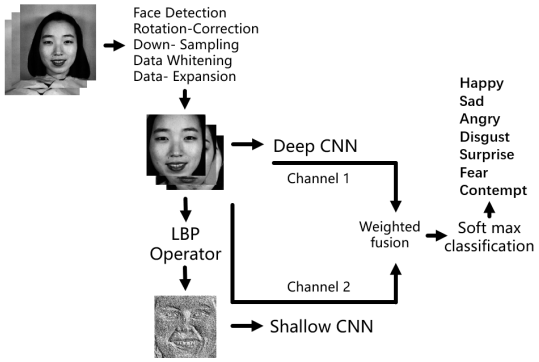

Fig. 2: Structure of Binary-Channel CNN.

The Expression Recognition Model consists of preprocessing, deep CNN based on face grayscale image, shallow CNN based on face LBP image, and a fusion layer. Before the training of the network, the face dataset needs to be preprocessed such as Face Detection, Rotation Correction, Data Whitening, Down-Sampling, Data Expanding, etc., and then use gray and its LBP image to train both CNN models, finally weighting and fusing the recognition results of the two channels based on the weights obtained by the experiment.

To improve the performance of our model, the Batch Normalization[8] (BN) layer is added to each block of the CNN. So that the input range of each layer of the neural network remains the same, thus avoiding the gradient disappearance problem and this can greatly speed up training. Since BN itself has the effect of preventing over-fitting, this paper deletes the dropout layer in original CNN theories. The depth of the two channels is deepened to four sets of hidden layers, so that more expression information can be extracted. 1st CNN channel uses gray face image for feature extraction and facial expression recognition. 2nd CNN channel uses LBP image for feature extraction and expression recognition. Each feature extraction network includes one input layer and four groups of Conv2D - Batch Normalization - MaxPooling layer, two sets of Dense layers and one output layer. The structural parameters of the two channels are the same, the input layer size is 48×48 pixels, and the number of convolution kernels in the four hidden layers are 128, 256, 256, and 256 respectively. The dimension of output feature vector is 2304.

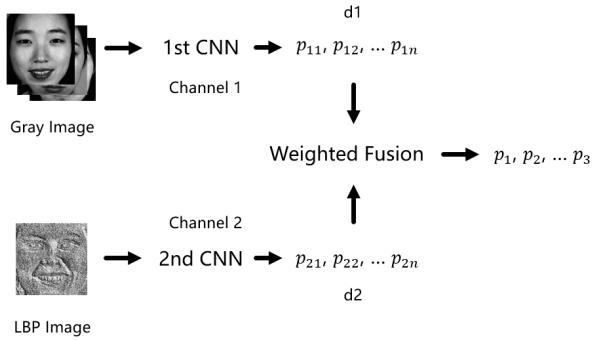

Fig. 3: Weighted Fusion Mechanism.

The weighted fusion structure is shown in the Fig. 3. $d_1 = \{p_{11}, p_{12}, \dots, p_{1n}\}$, $d_2 = \{p_{21}, p_{22}, \dots, p_{2n}\}$. p_{ij} indicates the figure of each network branch output. d_1 and d_2 respectively represent the confidence vector made by the 1st CNN and the 2nd CNN, and n represents the facial expression number.

The weighted fusion of d_1 and d_2 results in the vector $d_f = \{p_1, p_2, \dots, p_n\}$. d_f is calculated as follows:

$$d_f = \alpha \times d_1 + (1 - \alpha) \times d_2 \quad (1)$$

α is a weight, indicating the importance of the final decision from each network. The value of α is determined experimentally, and then the Softmax function is used to solve

the multi-classification problem, thereby identifying different facial expressions. The Softmax function maps the output of multiple neurons to the $(0, 1)$ interval to perform multi-classification. For the solution of the Softmax function, the Categorical Cross-Entropy[9] is used as the loss function, which is defined as follows:

$$loss = -\frac{1}{n}\sum_{x}[y \ln a + (1-y)\ln(1-a)] \tag{2}$$

y represents the expected output, and a represents the actual output of the neuron. n is the total number of outputs and x is the order number of the output. The loss can be minimized by Back Propagation based on the Gradient Descent Method[10] to obtain the network output.

B. Yolo V3

Yolo is an advanced, real-time target detection system. The main purpose of this study is to identify faces in the image in order to obtain expression information and location information later. The latest version is Yolo V3, which has very good real-time performance, and it is faster and more accurate than most detection algorithms. Considering the trade-off between the effectiveness of the target detection and the amount of computation, the pre-trained Tiny Yolo V3 model was selected. In this paper, a confidence level higher than the threshold of 0.5 is adopted, the threshold of the NMS[11] is set to 0.4. The face bounding box is set according to the face size of the training data used in the training of Binary-Channel CNN, so that the faces in the chosen frame are under similar condition to faces in training dataset, making the network do the best recognition of facial expressions.

III. Sentiment Energy Model

Sentiment Energy consists of three elements:

a) Expression Energy, that is, the energy calculated according to the energy level of each facial expression and the proportion of each facial expression in a frame image. This element can measure the group sentiment directly brought by various facial expressions.

b) Expression Change Rate Energy, which is calculated according to the change rate of each energy between frame images. This element can measure the rate of expression change in video.

c) Displacement Energy, which is calculated according to the moving speed of each face in the video. This element can measure the intensity of human displacement in the video.

Though element a and element b are normally not utilized in state-of-the-art methods related with group emotion analysis, the element c is very popular in group emotion analysis methods. Linbo Qin et al.[12] have used displacement elements to analysis group emotion. However, their methods lack precision to some extent.

In this paper, three elements are combined in order to achieve a better outcome.

$$\begin{aligned} E &= k_1 \times E_1 + k_2 \times E_2 + k_3 \times E_3 \\ &= k_1 \times \frac{\sum P_i X_i - (\sum P_i X_i)_{min}}{(\sum P_i X_i)_{max} - (\sum P_i X_i)_{min}} \\ &+ k_2 \times \frac{\sum \triangle P_i X_i - (\sum \triangle P_i X_i)_{min}}{(\sum \triangle P_i X_i)_{max} - (\sum \triangle P_i X_i)_{min}} \\ &+ k_3 \times \frac{\triangle D}{nl} \end{aligned} \tag{3}$$

The above formula is the Sentiment energy model used in this paper. The advantages of this model are as follows:

a) The proportion of facial expressions in a frame is judged by expression recognition, which is more intuitive for group sentiment analysis. The specially designed Binary-Channel CNN also provides a better recognition tool for the whole sentiment analysis system.

b) The change rate of facial expression proportion reflects the intensity of individual expression change. When the energy of expression proportion is small, as a component of emotional energy, Expression Change Rate Energy can also reflect the emotion of a group to a certain extent. This design makes the system more sensitive to group sentiment change, and also strengthens the robustness of the sentiment assessment process.

c) Displacement Energy provides an energy component measuring the speed of face displacement, which makes the intensity of face displacement also affect Sentiment Energy. This can provide another reliable criterion for emotion assessment, making it more accurate and robust.

d) The adaptive coefficients can determine the best coefficients through a self-adapting algorithm. This method makes the proportion of three elements more reasonable and closer to the actual situation, thus improving the accuracy of group emotion recognition.

A. E1: Expression Energy

Emotion	Energy Value
contempt	1
sad	2
surprised	3
happy	3
frustrated	3
scared	4
angry	5

Fig. 4: Expression energy grades.

Above is an expression energy grade table based on The Psychology of Feeling and Emotion[13]. The seven facial

expressions are graded. When doing sentiment analysis for a group, it is the facial expression of each individual in the group that can best reflect the emotional level of a group. Inspired by this, this paper introduces a component to measure the influence brought by proportions of different facial expressions.

The Yolo algorithm located the position of recognizable faces in the crowd, and then these faces were input into the Binary-Channel CNN for classification to get the category of facial expression on the face of each person. After simple counting, we can know the proportion of various expressions in each frame, and then calculate this energy component. Preliminary expression of this component can be obtained as follows:

$$E_1 = \sum P_i X_i \tag{4}$$

Among them P_i is the proportion of the ith expression in this frame image, X_i is a corresponding energy grade of the ith expression.

Because the dimensions of three components E_1, E_2, E_3 are different, it is necessary to standardize each component. The range standardization method is chosen here.[14]

$$E_1 = \frac{\sum P_i X_i - (\sum P_i X_i)_{min}}{(\sum P_i X_i)_{max} - (\sum P_i X_i)_{min}} \tag{5}$$

Among them P_i is the proportion of the ith expression in this frame image, X_i is a corresponding energy grade of the ith expression.

B. E2: Expression Change Rate Energy

For a dynamic scene, people often analyze the emotions of a crowd not only limited to the analysis of a single frame. According to Cognitive neuroscience of emotional memory[15], The human brain is often more accustomed to comprehensive analysis of several coherent frames in order to achieve more accurate results.

This algorithm weighs the advantages and disadvantages and considers that the information expressed by a single frame image is limited. It only expresses the expression state of people in this frame, and human emotions can often be reflected by the change of expression (For example, the expression of panic do represent a surely happening event, but if someone encounters an emergency, they will often show some abnormal changes of expression. In some circumstances people may deliberately cover up their panic or excitement after they show ferocious expressions. In this case, only calculating the emotional energy of a single frame will be not enough. Performance of people in emergencies is closely related to the habit of individuals, and there is no inevitability.)

Therefore, this paper designs a component of frame change information before and after response: the change rate of expression proportion. Here the difference is used to represent the derivative for simple calculation purpose.

The expression of this component can be obtained preliminarily as follows:

$$E_2 = \sum \triangle P_i X_i \tag{6}$$

Among them, $\triangle \mathrm{P}_i$ is the difference of the ith expression from the former or later frames, and X_i is the corresponding energy level of the ith expression.

Because the dimensions of three components E_1, E_2, E_3 are different, it is necessary to standardize each component. The range standardization is chosen here

$$E_2 = \frac{\sum \triangle P_i X_i - (\sum \triangle P_i X_i)_{min}}{(\sum \triangle P_i X_i)_{max} - (\sum \triangle P_i X_i)_{min}} \tag{7}$$

$\triangle \mathrm{P}_i$ is the difference of the ith expression from the former or later frames, and X_i is the corresponding energy level of the ith expression.

C. E3: Displacement Energy

In the typical crowd scenes that can distinguish faces, some expressions cannot be recognized accurately because of occlusion, facial movements or some interference. But this kind of faces can be located by Yolo V3 trained on face datasets.

As mentioned above, in the analysis of group behavior, besides the expression of individual, there is another important factor: The description of human behavior. When the facial expression of a group cannot be accurately identified, its sentiment can be analyzed through its behavior. Thus, the third component is designed, and the behavior information of the group is added to the total energy. Behavior description can usually be described by syntactic structure and classified by corresponding classifiers. However, at present, the relevant datasets are very rare, so it is relatively difficult to train classifiers for behavior recognition. Moreover, this paper analyses a large sample group. It seems unrealistic to classify the behavior of each individual into a certain pattern. Therefore, the distance of individual movement between two frames is chosen to replace the description of behavior.

Expression of this component can be obtained preliminarily as follows:

$$E_2 = \triangle D \tag{8}$$

$\triangle D$ is the sum of all face displacements between two frames. Face displacement is chosen instead of human displacement because the object of emotional analysis in this paper is a group of people who can capture facial expressions. For a group of people who cannot distinguish their faces or all back-to-camera, it is out of the scope of this paper (This group of people belongs to poor quality samples, because from this group of people we can only get their location information. But in this paper we hope to get both the location and expression information, especially the expression information, because the expression information can more directly reflect emotions). Natural detection of human face position and direct detection of human body position have the same effect. Even if someone turns his back to the camera, we can still analyze the emotions of most people to represent the overall sentiment.

Because the dimensions of three components E_1, E_2, E_3 are different, it is necessary to standardize each component.

Here, it is normalized by dividing the maximum possible displacement:

$$E_3 = \frac{\triangle D}{nl} \tag{9}$$

Among them $\triangle D$ is the sum of all face displacements between two frames, n is the total number of successful face matching between two frames, and l is the maximum possible distance of face displacement (that is, the diagonal distance of the original video frames).

D. Adaptive Coefficients k_1,k_2,k_3

The components of three energy functions are defined above. Obviously, the total energy function should be the sum of three components.

Although three functions are range normalized and the value of the function is mapped between 0 and 1, there are two problems when three quantities are added up directly from the practical point of view:

a) The three functional components have different ability to reflect emotions of people. From the view of Emotional Psychology and Behavioral Statistics, the percentages of human emotions in each frame and the distance of face movement between frames (E_1 and E_3) should be more meaningful than expression change rates between frames (E_2).

b) The normalized values of the three function components are still quite different. From the analysis of the calculation process of E_1, E_2 and E_3, E_2 is the subtraction of two different E_1s, and its value can be less than E_1. Because the normalization of E_3 is divided by the diagonal length of the whole image of current frame (the maximum possible displacement distance), the displacement of individuals in the image is generally much smaller than the diagonal length of the whole image, so the third component, which has been normalized, is often much smaller than the other two components.

Considering the above two problems, the three components are multiplied by the corresponding weight coefficients k_1, k_2, k_3, so that the magnitude of the three components can satisfy a certain proportion. Here the weight coefficient k_i is defined:

$$\begin{aligned} k_i &= Rationalize_{ki}(k_1 \times \bar{E_1}, k_2 \times \bar{E_2}, k_3 \times \bar{E_3}) \\ &= Rationalize_{ki}(E_1', E_2', E_3') \end{aligned} \tag{10}$$

k_i is the value taken when E_1', E_2', E_3' satisfies the Rationalize function. $\bar{E_j}$ is the average energy value over the frames. Here the Rationalize function is defined as follows:

Rationalize is a function that makes the ranges of three components make sense: $E_1' : E_2' : E_3' = 2 : 1 : 2$.

In this way, an adaptive energy function is obtained.

IV. Verification and Analysis

A. Experiment Setup

1) Dataset:

CK+[16]: This dataset was jointly completed by 70 volunteers, including 6 basic expressions (happy, sad, surprised, frustrated, fear, anger and contempt). There are about 80 samples for each basic expression. After the samples are expanded, each expression will get approximately 4,000 samples.

JAFFE[17]: This dataset was jointly completed by 10 volunteers from the University of Tokyo, including 6 basic expressions and contempt, with a total of 213 samples. There are about 30 samples for each expression. After the samples are expanded, each expression will get approximately 4,000 samples.

Oulu-CASIA[18]: This dataset was jointly completed by 80 volunteers, including 6 basic expressions and contempt, with a total of 10,880 samples. There are about 1800 samples for each expression. After the samples are expanded, each expression will get approximately 5,400 samples.

2) Framework:

Keras, OpenCV, Python3.6, NVIDIA CUDA Framework 9.2

B. Details of Network Training

1) Parameter Settings:

Parameter	Value
α	0.6
Loss	Categorical Cross Entropy
Optimizer	Adam
Learning rate	0.0001
Batch size	256
Epoch	30
Performance Evaluation	Accuracy

Fig. 5: Parameter settings.

2) Training Process:

This paper judges the performance of expression recognition by accuracy, the formula:

$$accuracy = (TP + TN)/(P + N) \tag{11}$$

P represents the number of positive samples; N represents the number of negative samples; TP represents the number of target expression samples that are correctly divided; TN represents the number of non-target expression samples that are correctly divided.

$$loss = -\frac{1}{n}\sum_x [y \ln a + (1-y)\ln(1-a)] \tag{12}$$

In the training process, Categorical Cross Entropy is used as the loss of the gradient descent. In the above formula, y represents the expected output and a represents the actual output of the neuron. n is the total number of outputs and x is the order number of the output.

Figures 6 and 7 show the prediction accuracy and loss of each Epoch in the training process. It is easy to see that the training can generally stabilize after 10 $\sim$ 15 Epochs (After which Loss and Accuracy are generally unchanged).

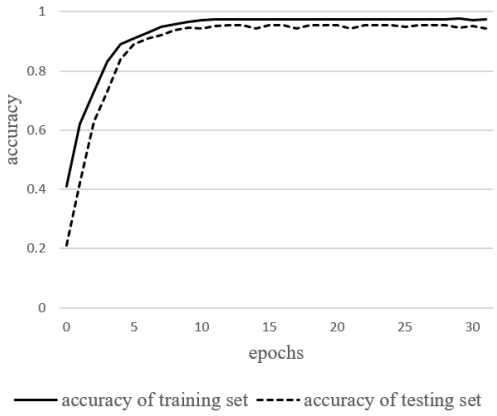

Fig. 6: Curve of accuracy.

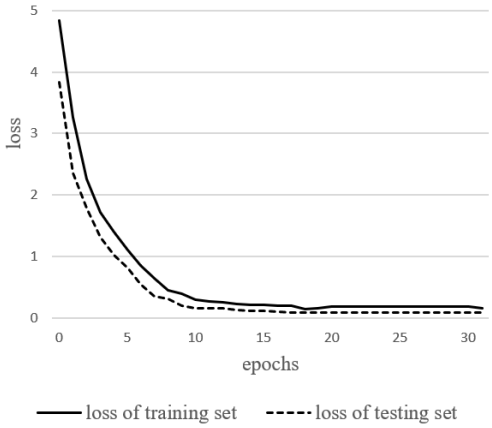

Fig. 7: Curve of loss.

C. Verification and Analysis of Binary-Channel CNN

1) Effectiveness Analysis of Binary-Channel:

In this paper, a novel neural network model, Binary-Channel Convolutional Neural Network, is proposed. The Convolutional Neural Network here consists of two channels. The input information of one channel is the preprocessed original image, and the input information of the other channel is the feature image extracted by the LBP operator. As mentioned above, the model designed in this paper has a higher recognition rate than the traditional model. The conclusions are proved by the following experimental data. The experiment was based on three expression datasets: CK+, JAFFE and Oulu-CASIA.

Respectively, we made a comparison of the recognition rates of three models: the single-channel model in which only the original gray image is input, the single-channel model in which only the LBP operator processed image is input and the binary-channel model in which two types of images are input.

As shown in Fig. 8, combined with the recognition rate of the expression, it can be seen that the binary-channel method has obvious advantages in the recognition rate of expressions, especially in comparison with the single-channel model of which only the original gray image is input. And Compared with the single-channel model of which only the LBP operator processed image is input, the recognition rate of the binary-channel method is also steadily improved. The expression recognition rate refers to the ratio of the number of correct expression recognitions to the number of total expressions.The Expression Recognition Rate is calculated as follows.

$$ERR = \frac{N_{correct\ expression\ recognitions}}{N_{total\ expressions}} \tag{13}$$

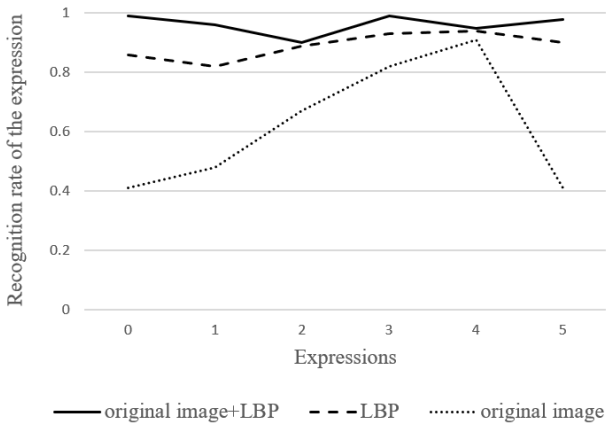

Fig. 8: Comparison of binary-channel and single-channel Expression Recognition Rate.

2) Fusion Coefficient Analysis:

To verify the effectiveness of the two channels, the fusion coefficients of the two channels need to be discussed. This experiment is still based on three expression databases: CK+, JAFFE and Oulu-CASIA. The value of each α was evaluated using the accuracy of Cross Validation. α is taken from the interval $[0, 1]$, the step size is 0.1. When α is 0, it is equivalent to the condition in which only the single original image channel is used, and when α is 1, it is equivalent to using only the single LBP image channel.

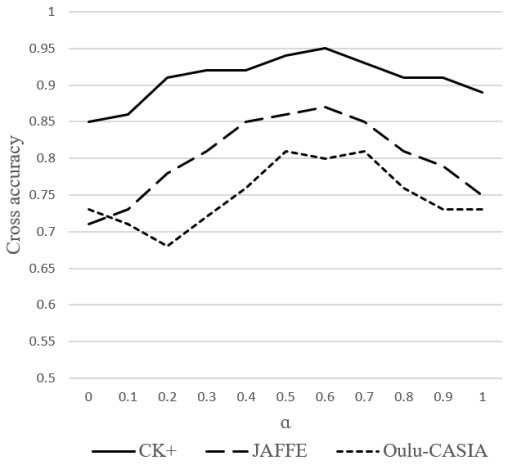

Fig. 9: Comparison of algorithm precision under different fusion weights.

It can be seen from Fig. 9 that the accuracy of the LBP image for expression recognition is higher than that of the original image, and the recognition accuracy is high when α is 0.5, 0.6, 0.7. In CK+ and JAFFE, the recognition accuracy is the highest when α is 0.6, so 0.6 is the most suitable in reality. It can also be seen that the contribution of the LBP image to the recognition accuracy is higher than that of the gray image.

D. Group Emotion Recognition and Analysis

1) Preparatory Work:

Before the group sentiment analysis, the image of each frame is extracted by video processing. Then the Yolo V3 model is used to locate the position of each face in the picture. In this way, the target for analysis is obtained: The gray image of face and the position of each person in each frame. Then, through the Rotation Transformation, Histogram Equalization, Image Normalization and other pre-processing operations, the processed image matrix is input into the Binary-Channel CNN for expression classification. In this way, the expression and position of each face in each frame image can be obtained.

2) Analysis of Emotion Based on Sentiment Energy:

According to the designed sentiment function, three kinds of energy are calculated by the classification information of expression and the face position information in each frame: E_1: the energy generated by the expression proportion, E_2: the energy generated by the changes in the proportion of expressions, E_3: the energy generated by the face displacement. The total energy E can be obtained by weighting the three energies according to the adaptive coefficients.

Fig. 10: Locating the faces in experimental video.

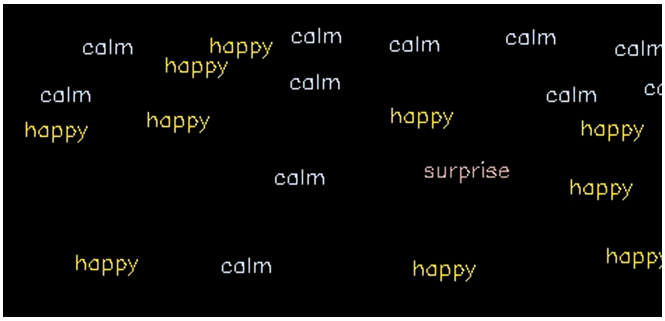

Fig. 11: Classifying the facial expression of each face.

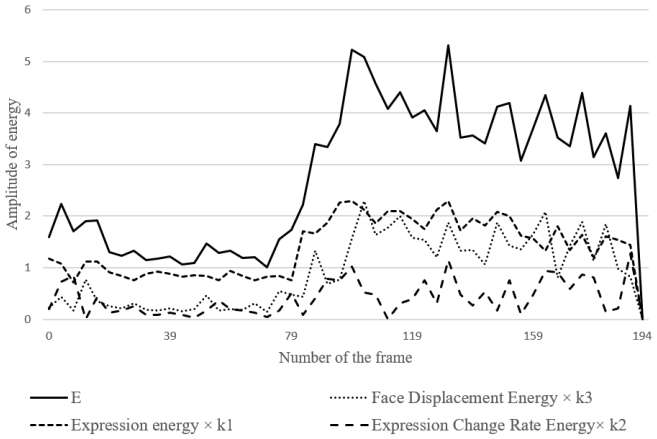

Fig. 12: Emotion energy trend of experimental video.

The experimental target is a video of the audience in a theater watching a movie or drama, which is downloaded from the internet. Related videos and codes can be found in attached files or in the link[19]. The changes in the three Sentiment Energies of the entire video over time and the changes in total energy over time are visualized, as shown in Fig. 12.

The three components of the energy function and the total energy change over time can be clearly seen in the graph. Through analysis, it can be inferred that an event before or after 80 frames triggers the emotions of the group, so that the Sentiment Energy of the group increases significantly. Correspondingly, it can be seen that the three components of the Sentiment Energy also rise before or after 80 frames, so it is determined that people in the video have a significant emotional change due to some external interference before or after the 80 frames. By watching the video, we do see that the expressions of the audiences were calm at first, but almost everyone around the 80 frames became excited at the wonderful performance on the stage or screen. Their expressions were mostly happy and the amplitude of their head swaying also increased. Therefore, it can be concluded that the results evaluated by the model are consistent with the results of the human brain analysis.

V. Conclusions

This paper establishes a neural network structure and proposes a new energy model for group emotion assessment. For the establishment of Binary-Channel CNN, the main purpose is to increase the ability of the network in extracting facial expression features, and try to ignore some factors that are not related to expressions such as hair, background, clothing and so on. Through experiments, it can be seen that the expression recognition accuracy of the proposed model can reach more than 93% on the overall dataset. This shows that the ability of Binary-Channel CNN to recognize expressions is very prominent. The Sentiment Energy model designed in this paper mainly focuses on the facial expressions, expression changes and face displacement in the video to comprehensively evaluate the magnitude of Sentiment Energy. From the

theoretical and experimental results, the energy model can accurately reflect the emotion changes of the group.

VI. Reference

[1] Ekman P , Friesen W V , O"Sullivan M , et al. Universals and cultural differences in the judgments of facial expressions of emotion.[J]. Journal of Personality and Social Psychology, 1987, 53(4):712-717.

[2] Lopes A T , Aguiar E D , Souza A F D , et al. Facial Expression Recognition with Convolutional Neural Networks: Coping with Few Data and the Training Sample Order[J]. Pattern Recognition, 2016, 61:610-628.

[3] Yu Z . Image based Static Facial Expression Recognition with Multiple Deep Network Learning[C]// Acm on International Conference on Multimodal Interaction. ACM, 2015.

[4] Bin Hu, Gongcheng Xia. Qualitative Simulation of Crowd Behavior Based on Integrated Causal Reasoning and QSIM[J]. Industrial Engineering and Management, 2004, 9(3): 32-36.

[5] Rassadin A G , Gruzdev A S , Savchenko A V . Group-level Emotion Recognition using Transfer Learning from Face Identification[J]. 2017, 11(1):56-64.

[6] Cao Jinmeng, Ni Rongrong, Yang Wei. Binary-Channel Convolutional Neural Networks[J]. Journal of Nanjing Normal University(Engineering Technology), 2018, 18(03): 7-15.

[7] XIAOYANG TAN, BILL TRIGGS. Fusing Gabor and LBP Feature Sets for Kernel-Based Face Recognition[C]. 2017:235-249.

[8] Santurkar S , Tsipras D , Ilyas A , et al. How Does Batch Normalization Help Optimization?[J]. 2018, 11(6):212-220.

[9] Rusiecki A . Trimmed categorical cross-entropy for deep learning with label noise[J]. Electronics Letters, 2019, 55(6):319-320

[10] Shor N Z . The rate of convergence of the generalized gradient descent method[J]. Cybernetics, 1968, 4(3):79-80.

[11] Neubeck A , Gool L J V . Efficient Non-Maximum Suppression[C]// 18th International Conference on Pattern Recognition (ICPR 2006), 20-24 August 2006, Hong Kong, China. IEEE Computer Society, 2006.

[12] Linbo Qin, Wenshi Xiong, Wenjun Zhou, Shanshan Xiong, Xiaohong Wu . Crowd emotion recognition based on multi-stream CNN-LSTM networks[J]. Application Research of Computers, 2018, 35(12): 3828-3831.

[13] Ruckmick C A . The Psychology of Feeling and Emotion[J]. McGraw-Hill, 1936, 4(1): 32-40.

[14] Hofmann B , Yamamoto M . Convergence rates for Tikhonov regularization based on range inclusions[J]. Inverse Problems, 2015, 21(3):805-820.

[15] Labar K S , Cabeza R . Cognitive neuroscience of emotional memory[J]. Nature Reviews Neuroscience, 2006, 7(1):54-64.

[16] Lucey P , Cohn J F , Kanade T , et al. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression[C]// 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops. IEEE, 2010.

[17] http://www.kasrl.org/jaffe.html.

[18] https://www.oulu.fi/cmvs/node/41316.

[19] https://pan.baidu.com/s/1A9ECx7_duzmVqcHlpQobOQ