An expression recognition algorithm based on convolution neural network and RGB-D Images

HE binghua^{1*}, CHEN zengzhao¹, LI gaoyang¹, JIANG lang¹, ZHANG zhao¹ and DENG chunlin²

¹ National Engineering Research Center for E-Learning, Central China Normal University, Wuhan, 430079, China

² Information Construction Office Hubei University of Chinese Medicine, Wuhan, 430065, China

Abstract. Aiming at the problem of recognition effect is not stable when 2D facial expression recognition in the complex illumination and posture changes. A facial expression recognition algorithm based on RGB-D dynamic sequence analysis is proposed. The algorithm uses LBP features which are robust to illumination, and adds depth information to study the facial expression recognition. The algorithm firstly extracts 3D texture features of preprocessed RGB-D facial expression sequence, and then uses the CNN to train the dataset. At the same time, in order to verify the performance of the algorithm, a comprehensive facial expression library including 2D image, video and 3D depth information is constructed with the help of Intel RealSense technology. The experimental results show that the proposed algorithm has some advantages over other RGB-D facial expression recognition algorithms in training time and recognition rate, and has certain reference value for future research in facial expression recognition.

1 Introduction

Facial expression is the most effective and natural method of interaction that humans use to communicate. From the perspective of psychology, the famous psychologist Mehrabian suggested that 55% of the feelings come from facial expression when people interact with each other [1]. In the process of human-computer interaction, facial expression recognition is the basis for the machine to understand human emotions, and an effective way for humans to realize their own emotional detection. It not only can play a role in many fields of human-machine interaction, but also can provide effective analysis data for enterprise decision-making, security monitoring and auxiliary medical fields. In the context of human-computer interaction, facial expression recognition technology gains participants' real-time psychological state in a non-contact manner, attracting the attention of researchers.

In the past few decades, a large number of researches based on two-dimensional pictures have emerged and many excellent algorithms have been proposed [2],[3],[4],[5],[6]. However, due to the limitation of the two-dimensional images, the process of recognition is easy to be interfered by various factors such as posture, blocking, and illumination. In this case, this kind of information such as head deflection and

depth change are totally ignored, and the robustness and accuracy of expression recognition is affected.

In recent years, the facial expression recognition based on three-dimensional data of RGB-D (RGB-Depth) has received increasing attention from researchers in various fields. Sun Y et al. used a Hidden Markov Model to fuse two-dimensional textures and three-dimensional structural information for dynamic expression recognition [7]. Shao Jie et al. extracted four-dimensional spatio-temporal texture features as local dynamic features and three-dimensional geometric models as global static features to automatically identify the natural expression [8]. Based on the LBP-Top feature of two-dimensional gray-scale images, Sikka et al. implemented facial expression recognition by using a word bag structure [9]. These approaches verified the effectiveness of RGB-D in the field of facial expression recognition through algorithms, and proposed a new research direction for facial expression recognition.

In this paper, we propose a new facial expression recognition algorithm based on RGB-D dynamic image sequence and multi-feature fusion. First of all, in order to comprehensively obtain more details information when the expression is changing, we use a RGB-D camera to extract the depth information (Depth) while capture the RGB images. Then, on the basis of face geometry description, the features of image LBP

* Corresponding author: 1203651863@qq.com

texture, depth information and Convolutional Neural Network (CNN) convolution are fused. Finally, a Softmax classifier was constructed to get the final facial expression recognition rate. After experimental verification, the recognition effectiveness of this algorithm has been improved.

2 Images preprocessing

In order to reduce the influence of redundant information, a series of face detection and scale normalization are performed: Get an image set (RGB image and depth image both) from the beginning to the end of an expression sequence. Using the Viola-Jones algorithm to detect face regions [10]. In the experiment, the RGB images are normalized to size of 224×224 pixel, depth images are normalized to size of 120×120 pixel.

The Local Binary Pattern (LBP) feature is an operator used to describe local texture features of an image. It has rotation invariance and gray-scale invariance, and can effectively reflect the texture information of the current pixel and the region formed by its surrounding pixels[11]. Experiments show that LBP feature is one of the most effective feature descriptions in facial recognition [12].

Figure 1. Image Preprocessing

The basic idea of LBP feature is to compare the gray values of the adjacent 8 pixels in a 3×3 window with the center pixel of the window as a threshold and generate an 8-bit binary number to reflect the texture change of current region [11]. The basic LBP operator in the neighborhood space is defined as

$$L(x_c, y_c) = \sum_{p=0}^{P-1} s(i_p - i_c) 2^p \quad (3)$$

in which $s(x)$ is:

$$s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{else} \end{cases} \quad (4)$$

From the definition of LBP, we can see that the basic LBP operator can only cover an area which size needs to be given in advance, and it cannot meet the different demands for different size of area. To solve this problem, Ojala et al. improved the LBP operator by replacing the rectangle with a circle and extending the 3×3 neighborhood to any size, a LBP operator with P sample points in a circular region with radius R was proposed [13]. In this article, we use circle to get LBP features. All of the images are divided into several blocks and the histogram of each block is calculated. The LBP features vectors will be considered as the result of repeated addition by all the histograms in the circle.

For a given point (x_c, y_c) , the nearest neighbor (x_p, y_p) can be calculated as follows:

$$x_p = x_c + R \cos\left(\frac{2\pi p}{P}\right) \quad (1)$$

$$y_p = y_c - R \sin\left(\frac{2\pi p}{P}\right) \quad (2)$$

in which, R is the radius of the circle and P is the number of sample point.

Figure 2, Figure 3 is a visual display of the LBP 3x3 feature extracted from this dataset.

Figure 2. Depth image and 3 x 3 LBP

Figure 3. Gray-scale image and 3x3LBP

3 Experiment

Deep learning is a very popular research method in Computer Vision and Image Processing. As a supervised method, CNN's image local perception area allows the neuron or the processing unit to access the most basic features so that it can acquire significant

features of the observation data that are invariant to translation, scaling, and rotation [14]. In the process of expression recognition, those features has great advantages in dealing with image deflection and slight changes of depth information.

3.1 CNN

CNN is a variant of the Multi-layer Perceptron (MLP), which combines the unique advantages of sparse connectivity on a fully connected basis through neurons in contiguous layers (Convolution layer, Pooling layer). The local connection pattern is used to find the spatial correlation of the input features, while reducing the number of parameters to be learned by weight sharing. For convolution kernel W , input data X , the definition of convolution is:

$$\begin{aligned} o[n] &= (X * W)(m, n) \\ &= f[m, n] * g[m, n] \\ &= \sum_{u=-\infty}^{\infty} \sum_{v=-\infty}^{\infty} f[u, v] g[m-u, n-v] \end{aligned} \quad (5)$$

In the process of CNN image processing, the convolution kernels are convolved with different positions of the image in order to obtain the output. At the same time, the sparse connection and shared-weight of the CNN model structure provide possibilities for the fusion of output, reduce the dimension of the entire experimental process.

Figure 4. CNN Structure

Figure 4 is a classic CNN expression recognition structure. The input image is preprocessed (such as gray transformation and scale normalization). After the convolution layer, pooling layer, and full connection layer, the corresponding classification results are output at the output layer (angry, disgust, fear, happiness, sadness, and surprise). ReLU in the figure

represents the activation function of the convolutional layer.

3.2 Alexnet

In order to make full use of the advantages of CNN structure and realize the integration of multidimensional features, the CNN model chosen in this paper is Alexnet. This model was proposed in ImageNet Competition 2012 and is an effective complex model for multi-feature fusion [15]. The network consists of 8 layers (excluding the input layer), including 5 convolutional layers (C1, C2, C3, C4, C5) and 3 fully-connected layers (F1, F2, F3). In addition to the image convolution, the C1, C2, and C5 also carry out the pooling operations (Max Pooling). The input layer is a pre-processed 224×224 human face pixel matrix. The convolute ional layer part is drawn up and down, representing the separation of the feature maps calculated by the current layer. Experimental parameter settings: convolution kernel size setting C1 (11×11), C2 (5×5), C3, C4, C5 convolution kernel size is set to 3×3. All pooling layers use Max pooling and the sampling window size is 3×3.

Figure 5. Algorithm of this paper

3.3 Features fusion

This paper proposes an algorithm that combines LBP features with Alexnet model. Figure 5 shows the algorithm flow of this article, includes face detection, image preprocessing, feature extraction and expression recognition.

For each expression, we extracts three kinds of features, which are the depth image LBP feature, the color image LBP feature, and the color image CNN feature. Considering that the disparities in the distribution of different types of features are relatively large, this paper use the method described in literature 16 to fuse them. In practice, firstly, the mean and variance of each of the three groups of features are normalized[16].

$$FV_i^n = \frac{FV_i - \mu_i}{\sigma_i}, i = Depth, Color, CNN \quad (6)$$

then get the fuse feature vector and put in the Softmax classifier for classification:

$$FV_{fuse} = \{FV_{Depth}^n, FV_{Color}^n, FV_{CNN}^n\} \quad (7)$$

Figure 6. Visualization of training model after 25 iterations

4 Result

Due to the lack of three-dimensional information in the existing expression database, this paper designs and implements a more complete database based on dynamic expression recognition theory and RealSense technology. It contains 7 kinds of expressions (angry, disgust, fear, happy, sad, surprised, neutral). A total of 465 groups of facial expression data (male 15, female 17) of 32 people were stored in the experiment. Each piece of data included pictures and video sequences, depth information, and data information of three-dimensional pose information. Figure 7 shows some samples in the database.

Figure 7 Examples of facial expressions (from top to bottom, in order: happy, angry, disgusted, surprised, sad, fearful, neutral)

The experiment was run on the Pytorch library of the Python language. A total of 1104 facial expressions were selected for the experiment, including 3 light changes and 6 expression changes. In the experiment, the recorded expression library needs to be preprocessed, and the color image needs to be converted into a gray-scale and scale adjustment. In the experiment, 88 samples were randomly selected for the verification sample, 88 samples for the test sample, and 828 image data sets for training. Table 1 shows the

recognition effects of the six basic expressions, Table 2 shows the comparison between this paper and several other algorithms.

Table 1 Result

Label	Recognition rate
Surprise	0.93
Sadness	0.86
Fear	0.83
Angry	0.85
Happiness	0.96
Disgust	0.86

Table 2 Comparison between this paper and several other algorithms

Algorithms	Data Types	Rate (%)
Literature 7	RGB & 3D Description	90.40
Literature 8	RGB-D Image Sequence	85.04
Literature 17	3D Geometric Images	83.60
Literature 18	3D Grid Images	93.25
This Paper	RGB-D Image Sequence	87.83

In the above table, literature 7 uses a combination of HMM (Hidden Markov Model) and 3D facial expression descriptors, besides inputting raw data, the algorithm also manually obtains data, is a combination of unsupervised and supervised method. Literature 18 uses DVF(Distance Vector Fields) studies high-resolution 3D grid images. Literature 8 uses four-dimensional textures and geometric features and literature 17 uses LDA (Linear Discriminant Analysis) experimental methods. In terms of recognition results, this method is better than literature 8 and literature 17, but lower than literature 7 and literature 18, but it is better than the HMM method of document 7 and the DVF method of literature 18 in terms of training time.

5 Conclusion

This research uses the image sequence to judge the validity of the entire database. Through the experimental results, it can be found that the database recorded by RealSense camera is feasible, but the depth information of the image has not yet been fully utilized. The validity of the result proves that the three-dimensional spatial information in the database is beneficial to facial expression recognition in human-computer interaction. The method of using three-dimensional data still needs further study.

Acknowledgments

This research was partially supported by the National Science and Technology Support Program of China (No. 2015BAH33F01).

References

1. Mehrabian A. Communication without words [J]. *Psychology Today*, 1968, 2(4): 53~56.
2. 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]// *Computer Vision and Pattern Recognition Workshops*. IEEE, 2010:94-101.
3. Lyons M, Akamatsu S, Kamachi M, et al. Coding Facial Expressions with Gabor Wavelets[C]// *IEEE International Conference on Automatic Face and Gesture Recognition*, 1998. *Proceedings*. IEEE, 1998:200-205.
4. Yin L, Wei X, Sun Y, et al. A 3D facial expression database for facial behavior research[C]// *International Conference on Automatic Face and Gesture Recognition*. IEEE, 2006:211-216.
5. Phillips P J, Flynn P J, Scruggs T, et al. Overview of the Face Recognition Grand Challenge[C]// *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2005:947-954.
6. Moreno A B, Sánchez A. GavabDB: a 3D face database[J]. 2004.
7. Sun Y, Yin L J. Facial expression recognition based on 3D dynamic range model sequences[M]// *Lecture Notes in Computer Science*. Heidelberg: Springer, 2008, 5303: 58-71
8. Shao Jie, Dong Nan. Face Natural Expression Recognition Based on RGB-D Dynamic Sequence[J]. *Journal of Computer-Aided Design & Computer Graphics*, 2015, 27(5): 847-854.
9. Huang D, Shan C F, Ardabilian M, et al. Local binary patterns and its application to facial image analysis: a survey[J]. *IEEE Transactions on System, Man, and Cybernetics*, 2011, 41(6):765-781.
10. Sikka K, Wu T, Susskind J, et al. Exploring bag of words architectures in the facial expression domain[C]// *International Conference on Computer Vision*. Springer-Verlag, 2012:250-259.
11. Ahonen T, Hadid A, Pietikäinen M. Face Recognition with Local Binary Patterns[C]// *European Conference on Computer Vision*. Springer, Berlin, Heidelberg, 2004:469-481.
12. Viola P, Jones M. Rapid Object Detection using a Boosted Cascade of Simple Features[C]// *Computer Vision and Pattern Recognition*, 2001. *CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*. IEEE, 2003:I-511-I-518 vol.1.
13. Ojala T, Pietikäinen M, Mäenpää T. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns[C]// *European Conference on Computer Vision*. Springer-Verlag, 2000:404-420.
14. Ren S, Girshick R, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2017, 39(6):1137.
15. Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]// *International Conference on Neural Information Processing Systems*. Curran Associates Inc. 2012:1097-1105.
16. Y. Li, W. Zheng, Lin Nan. Facial Expression Recognition Based on AAM, CNN and LBP Features[J]. *Computer Engineering and Design*, 2017(12)
17. Wang J, Yin L, Wei X, et al. 3D Facial Expression Recognition Based on Primitive Surface Feature Distribution[C] *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2006:1399-1406.
18. Drira H, Amor B B, Daoudi M, et al. 3D dynamic expression recognition based on a novel Deformation Vector Field and Random Forest[J]. 2012:1104-1107.