2020 2nd International Conference on Artificial Intelligence and Advanced Manufacture (AIAM) | 978-1-7281-9986-3/20/$31.00 ©2020 IEEE | DOI: 10.1109/AIAM50918.2020.00024

Face Recognition Based on MTCNN and Integrated Application of FaceNet and LBP Method

Zaiye Yang
Changchun University of Science and Technology
Changchun, China
e-mail: mryangzy@outlook.com

Wei Ge
Changchun University of Science and Technology
Changchun, China
e-mail: geweiciomp@163.com

Zheng Zhang
Changchun University of Science and Technology
Changchun, China
e-mail: zz210620992@163.com

***Abstract*—Technology of face recognition has developed rapidly in the past three decades. Various face recognition methods have been proposed by a lot of research. The primary objective of the development of face recognition is improving the accuracy. Multi-task Cascaded Convolutional Networks (MTCNN) is an effective method to detect faces, which identifies the position of the face in the picture and marks five landmarks through deep Convolutional Neural Network (CNN). FaceNet is a technology of face recognition, which is also based on CNN technology, exhibit high accuracy. Local Binary Pattern (LBP) is a traditional technology of face recognition. Despite a lower accuracy than FaceNet, it has many advantages such as grayscale invariance and illumination insensitivity. In this research, we propose an enhanced model of face recognition which is based on MTCNN and integrated application of FaceNet and LBP method. The work that described in this article using LBP parallel FaceNet to improve the illumination robustness of the model only consists of MTCNN and FaceNet. Experiments show that the enhanced model is very effective in improving the illumination robustness.**

Keywords—MTCNN, FaceNet, LBP, illumination, robustness

I. INTRODUCTION

Face recognition is a widely used technology to recognize the person's identity by analyzing the image features of human faces. Face recognition is a category of computer vision. After the first attempt at face recognition in the 1970s, the technology experienced its first boom in 1988 thanks to the significant increase of computational power. In the past three decades, the development of face recognition methods has shown very impressive increase, being widely used in video surveillance and human-computer interaction. Expression, posture and lighting conditions are still theoretical challenges of face recognition technology.

Multi-task Cascaded Convolutional Networks (MTCNN) proposed by Zhang et al. [1] is a deep cascaded multi-task framework which utilize their inherent correlation to improve the performance. The network adopts a cascaded structure with three stages based on carefully designed CNNs with different functions. MTCNN predicts face and landmark location in a coarse-to-fine manner.

Local Binary Patterns (LBP) and Histogram of Oriented Gradient (HOG) are common methods of face recognition. HOG [2] maintains stability of optical and geometric deformation of image, while LBP [3] has the superiority on grayscale stability, insensitivity to illumination and so on. LBP is more often employed to get facial features of faces in lighting conditions.

The face recognition method based on LBP proposed by Ahonen et al. [4] in 2006. LBP can divide a facial image into several areas for face recognition.

The Pairwise Rotation Invariant Co-occurrence LBP (PRICoLBP) proposed by Qi et al. [5] has better rotation-invariant properties than LBP. The improved LBP maintains the same performance as LBP at the same time.

Xie et al. [6] improved the robustness of infrared face recognition through the fusion of HOG and LBP. The fusion based on extracting directional features of contours as additional information of texture features.

Deep Learning, proposed in 2006, has made breakthrough in Pattern Recognition and other areas. Deep Convolutional Neural Network (CNN), showing a higher effectiveness than the shallow structure, has been widely used in face recognition [7].

Wang et al. [8] proposed the method of face recognition based on a cascaded model of LBP and CNN, which not only reduce effectively the influence of posture change, illumination and expression but also overcome the deficiency of grayscale stability of CNN.

FaceNet based on CNN is a fairly effective method [9] proposed by Google researches in 2015. The original FaceNet provide an accuracy of up to 99.63% in the experiment with Labeled Faces in the Wild (LFW) and 95.1% in the experiment with the YouTube Faces DB.

The structure of MTCNN cascade FaceNet is a highly efficient method of face recognition but still improvable in the lighting environment. As a result, the parallel ensemble learning of LBP and FaceNet proposed in this research performs better at extracting features of face images. The improved structure reduces the interference of illumination and increases the accuracy of classification. Our method uses firstly MTCNN to detect face and mark five landmarks in each face, afterwards, all the data from MTCNN are fed into FaceNet and LBP at the same time. Finally, the results of the two models are weighted.

II. METHODOLOGY

A. Multi-task Cascaded Convolutional Networks(MTCNN)

MTCNN consists of three stages: [1]

In the first stage, many candidate windows produced quickly through a lightweight CNN architecture named Proposal Network (P-Net). P-Net is a fully convolutional network. Candidate windows and their bounding box

978-1-7281-9986-3/20/$31.00 ©2020 IEEE
DOI 10.1109/AIAM50918.2020.00024

regression vectors are obtained through the P-Net. Then the estimated bounding box regression vectors are used to calibrate the candidates. Afterwards, a non-maximum suppression (NMS) merges highly overlapped candidate. The structure of P-Net is shown in the figure 1. [1]

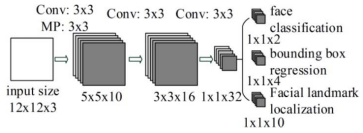

Fig. 1. The structure of P-Net

Then, all candidates obtained through P-Net are sent to Refine Network (R-Net) which will filter out lots of non-faces windows. After that, R-Net performs calibration with bounding box regression and merge candidates through NMS. The structure of R-Net is shown in the figure 2. [1]

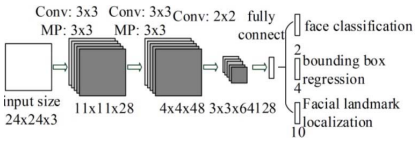

Fig. 2. The structure of R-Net

Finally, Output Network (O-Net) is a more efficient CNN to optimize the results from the R-Net and outputs five facial landmarks positions. In the last stage, O-Net outputs facial landmarks position and final bounding box. The structure of O-Net is shown in the figure3. [1]

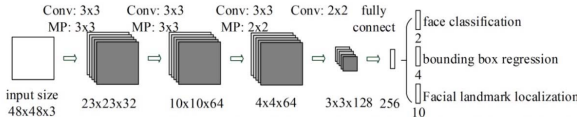

Fig. 3. The structure of O-Net

MTCNN is trained in three steps: classification, bounding box regression, and facial landmark localization. Different steps use different loss functions. The cross-entropy loss is used in the classification task which is a two-class classification problem.

$$L_i^{det} = -(y_i^{det}\log(p_i) + (1 - y_i^{det})(1 - \log(p_i))) \quad (1)$$

p_i is the probability that indicates a sample being a face and y_i^{det} is the ground-truth label.

In the task of bounding box regression we apply the Euclidean loss to each sample x_i. We predict the offset between each candidate window and their nearest ground truth.

$$L_i^{box} = \|\hat{y}_i^{box} - y_i^{box}\|_2^2 \quad (2)$$

In the formula (2), y_i^{box} represents the ground-truth coordinate. $\hat{y}_i^{box}$ represents regression target created by the P-Net. Euclidean loss is also used in the task of Facial landmarks localization. Facial landmarks location is a regression problem just like the previous task.

$$L_i^{landmark} = \|\hat{y}_i^{landmark} - y_i^{landmark}\|_2^2 \quad (3)$$

MTCNN is composed of different tasks, so different images are used to train it. Some loss functions will not be used temporarily. The overall loss function is shown in formula (4):

$$min \sum_{i=1}^{N} \sum_{j\in\{det,box,landmark\}} \alpha_j \beta_i^j L_i^j \quad (4)$$

($\alpha_{det} = 1, \alpha_{box} = 0.5, \alpha_{landmark} = 0.5$) is used in R-Net and P-Net. ($\alpha_{det} = 1, \alpha_{box} = 0.5, \alpha_{landmark} = 1$) is used in O-Net. Stochastic gradient descent is a useful method to train the CNNs. Training data come from WIDER FACE and CelebA.

B. Local Binary Pattern (LBP)

LBP has the advantages of low time complexity, illumination insensitivity, grayscale invariance, etc. The algorithm can divide an image into several areas equally, extract the texture features of each area. After that, the LBP fuse the overall texture features [10].

The idea of LBP is comparing the gray value of the center pixel with its every pixel in neighborhood. Take Figure 4 as an example, the number of neighboring pixels is 8 and the radius is 1. Taking the gray value of center point as the threshold, comparing the threshold with the gray values of all the 8 neighboring pixels. The gray value of the pixel in neighborhood is marked as 1 if the gray value of the pixel in neighborhood is greater than the threshold; otherwise, the gray values of the neighboring pixel is marked as 0 [10].

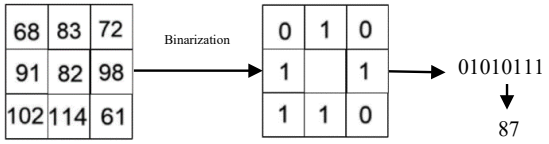

Fig. 4. Calculation Process of the LBP Algorithm

The formula of LBP can be shown in formula (5):

$$LBP_{P,R} = \sum_{i=0}^{p-1} S(g_i - g_c) \times 2^i \quad (5)$$

$$S(x) = \begin{cases} 0 & x \le 0 \\ 1 & x > 0 \end{cases} \quad (6)$$

As is shown in formula (5), $LBP_{P,R}$ represents the LBP code of center pixel. g_c represents the gray value of the center pixel and g_i represents the gray value of the ith pixel in neighborhood. R represents the radius.

An 8-bit binary number is obtained by the calculation process. The LBP code of the center point can be created by converting the 8-bit binary number to a decimal number. The image is represented in the form of a histogram by calculating the LBP code of each pixel.

Uniform pattern [11] is an extension of LBP. The number of bitwise transitions from 1 to 0 or 0 to 1 is uniformity pattern. It is called uniform if the uniformity pattern of an LBP is greater than or equal to 2. For instance, the patterns 11,111,110 (1 transitions), 11,111,100 (1 transitions) and 10,000,001 (2 transitions) are uniform and 00,1010,101 (7 transitions) and 10,1010,101 (8 transitions) are not. In uniform LBP, each uniform pattern has its separate output label, while all the non-uniform patterns have the same label. Uniform LBP is outstanding in reducing calculation time without affecting accuracy.

C. FaceNet

FaceNet is an efficient method of face recognition which based on CNN. Inception ResNet type networks can be the core architectures of FaceNet.

FaceNet contains a deep CNN structure and a batch input layer. The deep CNN followed by L2normalization to realize the embedding of human face. The triplet loss is used during training process. The structure of FaceNet is shown in the figure 5. [9]

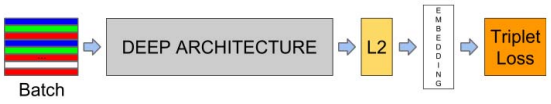

Fig. 5. FaceNet Overall Architecture Block Diagram

Positive, negative and anchor are the main elements of triplet loss training methods. The goal of triplet loss is to minimize the distance between anchors positively and maximize the distance between anchors negatively. Triplet loss is one of the best ways to get 128-dimensional vector to represent each facial features. These vectors are used to calculate the Euclidean distance. The Euclidean distances for the similar facial images would be much closer than the random non similar facial images. Triplet loss is shown in figure 6. [9]

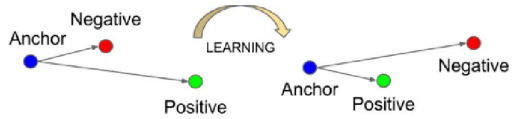

Fig. 6. Triplet Loss Training

Training data of FaceNet come from CASIA-WebFace and MS-Celeb-1M.

D. Improved Method

The model consists of MTCNN and FaceNet is one of the good methods for face detection and recognition. The MTCNN detects and proposes faces in the picture, and marks five key landmarks, then aligns the face image and fed it into FaceNet. FaceNet transforms the face image into a 128-dimensional vector. Finally, the vector is mapped to Euclidean space. The vectors with the closest Euclidean distance have the same label.

We add the uniform LBP model to the above model and make the LBP model parallel with FaceNet to improve the robustness in the lighting environment.

Aligning the face images that obtained by MTCNN and fed them into FaceNet and uniform LBP at the same time. FaceNet transforms the face image into a 128-dimensional vector. The LBP code of each pixel is obtained by uniform LBP. There are 59 possible values for uniform LBP code. Then the face image is divided into 7×4 blocks equally. After that, the model statistics uniform LBP code and calculate for the histogram of each block. Then the 28 histograms are connected into one. A 1652-dimensional vector is obtained through the histograms. This 1652-dimensional vector is reduced by LDA to a 128-dimensional vector.

If we want to compare the similarity between test face and train face, fed the two to FaceNet to get two 128-dimensional vectors V_{test}^{FN} and V_{train}^{FN}. At the same time, the two facial images are fed to uniform LBP and LDA and get two 128-dimensional vectors V_{test}^{LBP} and V_{train}^{LBP}. After that, calculate the Euclidean distance and weight them as shown in the following formulas:

$$dis_{FN} = \|V_{test}^{FN} - V_{train}^{FN}\|_2^2 \tag{7}$$

$$dis_{LBP} = \|V_{test}^{LBP} - V_{train}^{LBP}\|_2^2 \tag{8}$$

$$dis = \alpha_{FN} \times dis_{FN} + \alpha_{LBP} \times dis_{LBP} \tag{9}$$

The smaller dis is, the more likely that test face and train face are from the same person. The overall structure is shown in the figure 7.

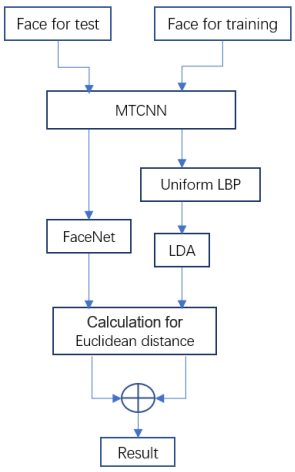

Fig. 7. Overall Structure

III. Experiments

The experiment is based on the Markus Weber's face database (MW) [12] and ORL face database (ORL) [13]. The Markus Weber's face database is collected by Markus Weber at California Institute of Technology. 450 face images, 896 $\times$ 592 pixels, JPEG format. 27 people under with different lighting/expressions/backgrounds.

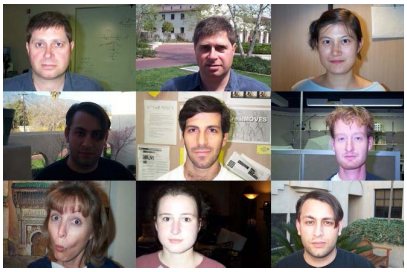

Fig. 8. Markus Weber's face database

ORL face database contains a total of 400 face photos from 40 people, 10 photos of everyone. The pictures were taken at different times, facial expressions and facial details. The background of each photo is black. All the people in the photos are front-facing.

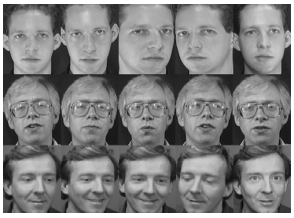

Fig. 9. ORL face database

Let's take Markus Weber's face database as an example, the data from the Markus Weber's face database contains pictures of 20 people, 20 pictures per person. Ten of them are used for learning in order to form a known face database. The other ten are used for testing to compare with the data in the known face database. The data from the ORL contains pictures of 40 people, 10 pictures per person, 5 for learning and 5 for testing.

We observe the change of accuracy as the weight value α_{FN} and α_{LBP} in the formula (9) change, by taking the value of α_{LBP} from 0 to 1 in 0.1 intervals. At the same time $\alpha_{LBP} = 1 - \alpha_{FN}$. We apply the above steps to experiment with Markus Weber's face database and ORL respectively. Among them, Markus Weber's face database is a face library with strong light interference, while ORL does not.

All the results is shown in the table I.

TABLE I. ACCURACY OF RECOGNITION

α_{LBP}		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
α_{FN}		1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
Accuracy	***MW***	98%	98%	98%	98.5%	98%	98%	98%	96.5%	94.5%	92%	80%
	ORL	99.5%	99.5%	99.5%	99.5%	99.5%	99.5%	99.5%	99.5%	97.5%	91%	85%

It can be seen from the experimental data corresponding to Markus Weber's face database that the addition of uniform LBP improves the accuracy of recognition. The model has the highest recognition accuracy when the weight of uniform LBP is 0.3. When using the Markus Weber's face database, the accuracy of the improved model is increased by 0.005 compared to the model with only FaceNet. When using ORL, as the weight of uniform LBP increases, the recognition accuracy has decreased. It shows that the parallel connection of uniform LBP and FaceNet improve the recognition accuracy only in the case of light interference. Otherwise, LBP drags down the accuracy of FaceNet.

IV. CONCLUSION

This study aims to improve the robustness of illumination on the basis of MTCNN and FaceNet. The parallel connection of LBP and FaceNet improve the face recognition accuracy because LBP can reduce the interference of illumination on facial features. Markus Weber's face database and ORL was utilized to experiment with the enhanced model. The accuracy of the model is significantly upgraded on the database with light interference. Therefore, the improved method proposed in this article has better illumination robustness.

ACKNOWLEDGEMENT

The research is supported by Scientific Research Planed Project of The Education Department of Jilin Province (No. JJKH20200782KJ)

REFERENCES

[1] Zhang K, Zhang Z, Li Z, Qiao Y (2016) Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett, vol. 23, no. 10, pp. 1499–1503.

[2] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005, pp. 886–893.

[3] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. vol. 24, no. 7, pp. 24, 971–987. 2002.

[4] T. Ahonen, A. Hadid, M. Pietikainen, Face Description with Local Binary Patterns: application to Face Recognition, IEEE Trans. Pattern Anal. Mach. Intell. vol. 28, no. 12, pp. 2037–2041. 2006.

[5] X.B. Qi, R.Xiao and C.G. Li, "Pairwise rotation invariant co-occurrence local binary pattern," IEEE Transactions on Pattern Analysis and Machine Intelligence. vol. 36, no. 11, pp. 2199, 2014.

[6] Z.H. Xie, P. Jiang and S. Zhang, Fusion of LBP and HOG using multiple kernel learning for infrared face recognition, 16th International Conference on Computer and Information Science (2017), 81–84.

[7] M Fatahi, M Ahmadi,A Ahmadi, et al. Towards an spiking deep belief network for face recognition application. Computer and Knowledge.

[8] M. Wang, Z. Wang, J. Li, Deep convolutional neural network applies to face recognition in small and medium databases, in: 4th International Conference on Systems and Informatics (ICSAI), pp. 1368–1372. 2017.

[9] F. Schroff, D. Kalenichenko and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815-823, 2015.

[10] Jialin Tang, Qinglang Su, Binghua Su, Parallel ensemble learning of convolutional neural networks and local, COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 197 (2020) 105622.

[11] Ojala T, Pietikainen M, Maenpaa T (2002) Multiresolution grayscale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell vol. 24, no. 7, pp. 971–987.

[12] Markus Weber's face database, http://www.vision. caltech.edu/archive. htmlJune 2007.

[13] F. Samaria, Andy Harter. (1994, December). ORL face database. Available: http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html.