Distance-Constrained Optimization-Based Face Recognition Postprint

Authors: Zhou Shengyang, Zou Hua, Xiao Chunxia

Date: 2018-05-02T00:00:00+00:00

Abstract

To address the issues of existing face recognition methods being sensitive to factors such as face angle, expression, and pose, as well as having low accuracy, a face recognition model based on a distance-constrained optimization algorithm is proposed. This model improves upon existing face recognition methods in two aspects: a) The LBP operator is utilized to extract texture spectrum feature maps from face images, which are then fused with the R, G, B channels of the original face images; the fused image matrix serves as input to the neural network, thereby enriching the facial texture features; b) The error function is improved by employing thresholds and boundary values to constrain the distances between feature vectors, thereby constructing a new optimization objective for the model that ensures face images of the same subject have smaller Euclidean distances in the feature space, while face images of different subjects have larger Euclidean distances in the feature space. Experiments conducted on the LFW face database in unconstrained scenarios demonstrate that the model achieves an accuracy of 99.15%, can effectively improve face recognition accuracy, and exhibits excellent robustness.

Full Text

Preamble

Face Recognition Based on Distance-Constrained Optimization

Zhou Shengyang, Zou Hua, Xiao Chunxia

(School of Computer Science, Wuhan University, Wuhan 430072, China)

Abstract: Existing face recognition methods exhibit high sensitivity to facial pose, expression, and orientation, resulting in low accuracy. To address these limitations, this paper proposes a face recognition model based on distance-constrained optimization. The model introduces two key improvements: (a) It

employs the LBP operator to extract texture spectrum features from face images, which are then fused with the R, G, and B channels of the original image. This fused image matrix serves as input to the neural network, enriching facial texture features. (b) It modifies the loss function by applying threshold and margin constraints to feature vector distances, establishing a novel optimization objective that enforces small Euclidean distances between feature vectors of the same subject and large distances between those of different subjects. Experiments on the unconstrained LFW face database demonstrate that the proposed model achieves 99.15% accuracy, effectively improving face recognition performance while exhibiting strong robustness.

Keywords: face recognition; feature extraction; local binary pattern; binary loss function; residual neural network

0 Introduction

With societal development, we have entered the digital and network era where information security and confidentiality have become increasingly important. Traditional authentication methods struggle to meet modern requirements. Since the 21st century, numerous biometric recognition technologies have emerged, including fingerprint recognition, iris recognition, and DNA identification. Face recognition has found widespread application in forensics, security, and other domains due to its non-contact nature, ease of acquisition, and high reliability. Consequently, improving the efficiency and accuracy of face recognition has become a critical research topic.

In recent years, deep learning has brought breakthroughs to many computer vision problems. With the advent of big data, convolutional neural networks (CNNs) have provided new perspectives for face recognition. In practical applications, face recognition systems must correctly identify faces under unconstrained conditions involving varying expressions, poses, and illumination. The system extracts features from two face images and compares their similarity to determine whether they belong to the same individual. While traditional face recognition methods suffer from limited applicability and accuracy, deep learning-based approaches have gained prominence due to their powerful feature learning capabilities, achieving performance that even surpasses human-level recognition on many face databases.

Representative deep learning face recognition methods include ConvNet-RBM[1] and DeepFace[2]. These approaches typically train neural network models using large quantities of labeled face images with softmax classification, manually select intermediate layer features as representative face feature vectors, and finally perform face recognition through feature vector similarity measurement. Despite significant accuracy improvements, these methods exhibit three notable drawbacks: (a) Intermediate features lack direct supervision[3], as the training process does not explicitly adjust intermediate features, potentially resulting

in unrepresentative feature vectors for new images that limit algorithmic accuracy. (b) The feature vector metric must be manually selected, introducing uncertainty. (c) When training neural networks on large-scale databases, the number of softmax categories increases accordingly, requiring more intermediate neurons to maintain feature completeness. This leads to high-dimensional feature vectors and increased computational costs.

To address these issues, this paper proposes a distance-constrained optimization approach for face recognition. Unlike methods using softmax intermediate layers, our model directly employs a residual neural network to map face images into 128-dimensional feature vectors, computing Euclidean distances between feature vectors to compare face similarity. If the distance between two feature vectors falls below a given threshold, the faces are judged to belong to the same subject; otherwise, they are considered different. Compared to previous deep learning face recognition methods, this approach' s network structure remains independent of the dataset and eliminates the need for manual selection of intermediate layers and feature metrics. Additionally, drawing inspiration from LBP-based face recognition and neural network characteristics, we propose a fusion method combining LBP face feature maps with original face images for training. LBP face feature maps offer illumination invariance and can enrich geometric and texture features of face images. Using fused images as neural network input enhances model generalization and accelerates training, particularly when data volumes are relatively small. Experimental results demonstrate that our method further improves face recognition accuracy and exhibits excellent robustness for unconstrained face images.

1 Related Work

The earliest practical face recognition algorithm was Principal Component Analysis (PCA)[4], also known as the eigenface method. This approach constructs a low-dimensional feature space from sample images, projects test images into this space to compute feature vectors, and classifies them using distance metrics. The Histogram of Oriented Gradients (HOG) method[5,6] calculates horizontal and vertical gradients, converts 2D gradient information into several undirected histogram channels, and parameterizes and cascades these features to obtain HOG feature vectors for face comparison. Reference[7] proposed a Laplacian eigenmap algorithm based on 2D kernel PCA, significantly improving face recognition accuracy. Similar methods include Local Binary Pattern (LBP)[8], Linear Discriminant Analysis (LDA)[9], and multi-feature fusion approaches[10]. These traditional methods share common limitations: they require manually designed feature extraction approaches that demand extensive parameter tuning through trial and error, struggle to capture highly nonlinear face image features, remain sensitive to facial expression and pose variations, and consequently suffer constrained recognition accuracy. Generally, traditional algorithms require image preprocessing techniques such as shadow removal[11,12] and texture upsampling[13] to enhance feature extraction.

In recent years, deep learning has achieved remarkable success in image classification and feature extraction, prompting researchers to apply these methods to face recognition with excellent results. DeepFace[2], developed by Facebook, employs softmax classification to categorize face images into 4,000 classes, using features from the penultimate fully connected layer (4,096 dimensions) as face representations. Its main advantage lies in using 3D models and local convolutional layers for face alignment and local feature extraction. After training, the model computes weighted squared distances or Euclidean distances between feature vectors for face recognition, achieving 97.35% accuracy on the LFW database.

The DeepID[14] model partitions face images into 25 regions, extracts multi-scale feature vectors, and employs a cascaded Bayesian classifier for recognition. DeepID2[15] extends this approach by jointly training with identification and verification signals. The identification signal is a standard softmax classification signal that learns inter-subject features to increase feature vector distances, while the verification signal is an image pair classification signal that learns intra-subject features to reduce feature vector distances. A joint Bayesian classifier is then trained for feature vector classification, achieving 99.15% accuracy on LFW.

Unlike DeepFace and FaceId, FaceNet[3] abandons traditional softmax training, instead using a triplet loss function to encode images into feature vectors and compute pairwise errors based on these encodings. The model trains on image pairs, designating two images from the same person as positive samples and images from different people as negative samples. FaceNet' s training objective ensures that the sum of Euclidean distance between any positive image pair and a threshold remains smaller than the distance between negative image pairs. After training, an appropriate threshold is selected. For any two images, FaceNet extracts feature vectors and compares their Euclidean distance against the threshold: distances greater than the threshold indicate different subjects, while smaller distances indicate the same subject. FaceNet' s network structure is dataset-independent, and its training process directly optimizes feature vectors, offering good interpretability and intuitiveness. However, the triplet training approach is sensitive to hard samples, and post-training threshold selection introduces uncertainty.

2 Proposed Algorithm

To further improve face recognition accuracy, this paper proposes a distance-constrained optimization model with improvements to image preprocessing and training algorithms, primarily consisting of LBP feature map fusion and binary loss optimization.

2.1 LBP Feature Map Fusion

LBP features are highly effective grayscale texture descriptors with strong discriminative power and illumination invariance, capable of describing both global and local image characteristics. Consequently, LBP features have been widely applied in pattern recognition, with Ahonen et al.[8] successfully employing them for face recognition.

The LBP operator is a local pattern-based texture operator. Its fundamental concept involves selecting P sampling points equally distant from a central pixel within a circular neighborhood of radius R, comparing them with the central pixel value, and representing the results using Boolean functions. The LBP value for texture features is calculated as:

$$LBP_{P,R}(x_c, y_c) = \sum_{i=0}^{P-1} s(g_i - g_c)2^i, \quad s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

where g_c represents the central pixel value and g_i denotes the i-th neighboring pixel value. The sampling configuration is denoted as $LBP_{P,R}$. Different neighborhood structures are illustrated in [Figure 1: see original paper].

The LBP feature extraction and computation process is shown in [Figure 2: see original paper]. LBP feature maps contain rich texture information from original face images. Traditional LBP-based face recognition methods typically concatenate local LBP features and represent them using histograms. Instead, we directly fuse LBP feature maps with original face images through channel concatenation. Xi et al.[16] demonstrated that using LBP feature maps as neural network input effectively captures facial features, converting LBP features into sparser and more effective representations. LBP feature maps offer simple and intuitive feature representation, while original face images provide richer feature information. Drawing inspiration from [16], we use LBP feature maps to construct conditional constraints for face images. The fused images possess more intuitive texture information, enabling faster convergence and higher accuracy in neural networks. The LBP feature map extraction and fusion process is illustrated in [Figure 3: see original paper].

2.2 ResNet Network Architecture

We employ a residual neural network (ResNet) as the face feature extractor to map face images into feature vectors. ResNet is an improved convolutional neural network originally proposed by He et al.[17] for ImageNet classification, achieving state-of-the-art accuracy at the time. ResNet' s advantage lies in its cross-layer connections, which strengthen feature signals and mitigate the risk of gradient explosion or vanishing in deep networks, thereby improving classification accuracy.

Similar to the ResNet-34 architecture and considering both accuracy and efficiency, our designed ResNet model comprises 20 layers. Detailed parameters are listed in . The notation follows: C for convolutional layers, P for pooling layers, RB for residual blocks (each consisting of 3 convolutional layers and 1 pooling layer), and F for fully connected layers. The complete network architecture is shown in [Figure 4: see original paper].

2.3 ResNet Model Loss Function and Training Process

The ResNet model can function as a black box for extracting representative feature vectors, mapping face images into a feature space. During initial training, each feature vector appears as a feature point distributed according to a Gaussian distribution. The key challenge is designing training rules that gradually cluster same-class points while separating different-class points for classification.

We propose an improved binary loss optimization algorithm for ResNet training. Face image pairs are defined as follows: when two images represent the same subject, they form a positive pair; when representing different subjects, they form a negative pair. The training objective requires small Euclidean distances between feature vectors of all positive pairs and large distances for negative pairs.

Based on this objective, we construct the ResNet model' s loss function. Setting a threshold τ and margin m ($m < \tau$), all positive pairs must satisfy:

$$\|x_a - x_p\|_2^2 < \tau - m$$

where x_a and x_p denote feature vectors of a positive pair. All negative pairs must satisfy:

$$\|x_a - x_n\|_2^2 > \tau + m$$

where x_a and x_n denote feature vectors of a negative pair. Based on these optimization objectives, we construct the loss function. Let Y be the pair label ($Y = 1$ for positive, $Y = -1$ for negative). The model' s loss function is:

$$D = \sum_i \max\{0, Y(\|x_1 - x_2\|_2^2 - \tau) + m\}$$

The loss function indicates that training uses image pairs. When Euclidean distances between feature vectors meet the optimization objectives, no loss is computed (loss equals 0). Otherwise, model parameters are updated using SGD. The margin m primarily serves to more clearly separate positive and negative samples during training. As shown in [Figure 5: see original paper], using margin m further reduces distances for positive samples and expands distances for

negative samples, preventing sample distances from being too close to threshold τ and creating hazardous samples.

Some methods directly minimize Euclidean distances for positive samples, such as MSE face recognition models with objective function:

$$\arg\min\{\|x_a - x_p\|_2^2\}$$

This approach minimizes L2 distance for positive samples, causing corresponding feature points to converge during training. We avoid this method for two reasons: (a) For unconstrained positive samples with significant variations in illumination, expression, and background (large intra-class variance), extracted feature vectors should maintain some differences rather than being optimized to zero distance. (b) In large databases, mislabeled negative samples frequently occur, and minimizing feature vector distances makes the model vulnerable to bad data.

Therefore, we use distance constraints where all positive samples need only be within a distance threshold rather than infinitely close, improving system robustness. Comparing our improved binary optimization objective with FaceNet' s[3] reveals that when setting $m = \alpha$, the two objectives become consistent. This shows our binary optimization objective can derive FaceNet' s triplet optimization objective, meaning our binary approach implicitly contains FaceNet' s triplet optimization target. Additionally, while FaceNet requires post-training statistical threshold selection for new samples, our method directly uses the training threshold τ, offering more intuitive operation.

After training, the ResNet model can extract features from new images for classification. During verification, we discard margin m and retain threshold τ as the Euclidean distance metric:

$$Y = \begin{cases} 1 & \text{if } \|x_1 - x_2\|_2^2 < \tau \\ -1 & \text{if } \|x_1 - x_2\|_2^2 > \tau \end{cases}$$

3 Experimental Results and Analysis

To validate our face recognition model and investigate parameter effects on recognition rate, we compare our approach with traditional PCA+LDA, CS-LBP[16], Fisher-Vector models, and deep learning-based ConvNet-RBM[1], DeepId[14], and FaceNet[3] models. Experiments run on an Intel i5 2.4 GHz CPU with 4 GB RAM, using 64-bit Ubuntu OS. Algorithms are primarily implemented in C++. Evaluation uses the Facescrube[19] and LFW[20] databases: Facescrube for parameter tuning and LFW for comparative analysis.

3.1 FaceScrube Experimental Results and Analysis

The FaceScrube[19] database contains 106,863 face images of 530 subjects, all collected in unconstrained settings with variations in illumination, expression, pose, and size. To validate model effectiveness while maintaining non-overlapping training and test sets, we select images from 450 subjects for training, 50 for validation, and 30 for testing, yielding approximately 90,000 training images, 10,000 validation images, and 6,000 test images. Using the training method described in Section 2.3, images are combined into pairs, resulting in about 1×10^7 positive pairs and 4×10^9 negative pairs.

3.1.1 Hyperparameter Selection Our model has two hyperparameters: threshold τ and margin m. Margin m is used during training, while threshold τ applies to both training and testing. Their selection determines model accuracy. For any positive pair, the Euclidean distance between feature vectors must be smaller than $\tau - m$; for negative pairs, it must exceed $\tau + m$. Facial features exhibit sparsity (small inter-class differences), so τ should not be excessively large. Conversely, margin m further reduces positive sample distances and increases negative sample distances. If m approaches τ in magnitude, it compresses all positive sample distances too small, contradicting face characteristics (large intra-class variance) and reducing accuracy.

Considering these factors, we test various τ and m combinations on FaceScrube, where $0.4 \leq \tau \leq 2.0$ and $0.05 \leq m < \tau$. Results show highest accuracy when $\tau \in (0.7, 1.0)$. The accuracy variation curve is shown in [Figure 7: see original paper]. Accuracy initially increases then decreases with τ, peaking at $\tau = 0.8, m = 0.2$. For fixed τ, excessively small or large m reduces accuracy. Overly large m compresses positive sample features into small distances, violating large intra-class variance characteristics. Overly small m fails to adequately separate positive and negative samples, potentially causing different subjects' feature vectors to have small distances (hazardous samples). Data indicates maximum accuracy of 96.68% at $\tau = 0.8, m = 0.2$, demonstrating strong performance on FaceScrube.

3.1.2 LBP Feature Map Fusion Our image preprocessing uses LBP feature map fusion to construct conditional constraints and enrich texture information. This section validates LBP feature map fusion effectiveness on FaceScrube.

We establish two models: ResFace (without LBP fusion, using raw face images as input) and LBP-ResFace (with LBP fusion). Both use $\tau = 0.8$ and $m = 0.2$. Accuracy curves across iterations are shown in [Figure 8: see original paper]. Analysis reveals that LBP-ResFace achieves higher accuracy: ResFace reaches 95.55% while LBP-ResFace attains 96.68% (1.13% improvement). Moreover, LBP-ResFace shows 2.48% higher accuracy during early training (1,000 iterations), demonstrating faster optimization. This validates the algorithm' s effectiveness.

LBP feature map fusion enriches original face image information, expresses texture features more intuitively, and accelerates neural network convergence. Different facial regions should have different convolutional weights during feature extraction, and LBP fusion provides such constraints. Additionally, LBP feature maps'illumination invariance effectively eliminates illumination and shadow effects in unconstrained scenes, improving model accuracy.

3.2 LFW Experimental Results and Analysis

The LFW[20] database, collected by the University of Massachusetts Computer Vision Laboratory, contains 13,233 face images of 5,749 subjects (4,069 with single images, 1,680 with multiple images). It serves as the academic and industrial benchmark for evaluating unconstrained face recognition algorithms. Our model is first pre-trained on FaceScrube[15] and VGG-dataface, then fine-tuned on LFW' s standard training set, and finally tested on LFW' s standard test set. To further validate LBP fusion effectiveness, we conduct two experiments: Ours-ResFace (without LBP fusion) and Ours-LBP-ResFace (with LBP fusion). Comparative results on LFW are presented in .

TABLE:2 LFW Face Recognition Accuracy Comparison

Face Recognition Model	Accuracy
PCA+LDA	
CS-LBP[17]	
Fisher-Vector Faces	
ConvNet-RBM[1]	
DeepId[13]	
FaceNet[3]	
Ours-ResFace	
Ours-LBP-ResFace	

Analysis shows that our improved binary loss function yields higher accuracy than other loss-based training methods. LBP feature map fusion significantly improves accuracy, indicating that texture-feature-fused images provide better feature representation in neural networks. Combining LBP fusion with binary loss optimization further constrains feature vector distances, extracts more effective face features, and fully demonstrates the advantages of both methods.

4 Conclusion

To enhance face recognition accuracy and robustness, this paper proposes a distance-constrained optimization model. First, LBP feature maps are fused with original images during preprocessing to add texture information. Then, a residual network encodes face images into 128-dimensional feature vectors, trained with a novel binary loss function. Experimental results demonstrate the

algorithm' s feasibility and effectiveness. The distance-constrained approach enhances feature representation capability and significantly improves recognition rates.

References

[1] Sun Y, Wang X G, Tang X O. Hybrid deep learning for face verification[C]//Proc of IEEE International Conference on Computer Vision. 2013: 1489-1496.

[2] Taigman Y, Yang M, Ranzato M A, et al. Deepface: closing the gap to human-level performance in face verification[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2014: 1701-1708.

[3] Schroff F, Kalenichenko D, Philbin J. Facenet: a unified embedding for face recognition and clustering[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2015: 815-823.

[4] Bartlett M S, Movellan J R, Sejnowski T J. Face recognition by independent component analysis[J]. IEEE Trans on neural networks, 2002, 13(6): 1450-1464.

[5] Albiol A, Monzo D, Martin A, et al. Face recognition using HOG–EBGM[J]. Pattern Recognition Letters, 2008, 29(10): 1537-1543.

[6] 万源, 李欢欢, 吴克风, 等. LBP 和 HOG 的分层特征融合的人脸识别 [J]. 计算机辅助设计与图形学学报, 2015, 27(4): 640-650.

[7] 徐梦珂, 许道云, 魏明俊. 基于 2D-KPCA 的拉普拉斯特征映射人脸识别 [J]. 计算机应用研究, 2017, 34(7): 2212-2215, 2220.

[8] Ahonen T, Hadid A, Pietikäinen M. Face recognition with local binary patterns[C]//Proc of European Conference on Computer Vision. Springer. 2004: 469-481.

[9] Simonyan K, Vedaldi A, Zisserman A. Learning local feature descriptors using convex optimisation[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2014, 36(8): 1573-1585.

[10] 杨赛, 赵春霞, 刘凡, 等. 一种基于多种特征融合的人脸识别算法 [J]. 计算机辅助设计与图形学学报, 2017, 29(9): 1667-1672.

[11] Zhang L, Zhang Q, Xiao C. Shadow remover: image shadow removal based on illumination recovering optimization[J]. IEEE Trans on Image Processing, 2015, 24(11): 4623-4636.

[12] Xiao C, Gan J. Fast image dehazing using guided joint bilateral filter[J]. The Visual Computer, 2012, 28(6-8): 713-721.

[13] Xiao C, Nie Y, Hua W, et al. Fast multi-scale joint bilateral texture upsampling[J]. The Visual Computer, 2010, 26(4): 263-275.

[14] Sun Y, Wang X, Tang X O. Deep learning face representation from predicting 10,000 classes[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2014: 1891-1898.

[15] Sun Y, Chen Y, Wang X, et al. Deep learning face representation by joint identification-verification[C]//Advances in Neural Information Processing Systems. 2014: 1988-1996.

[16] Xi Meng, Chen Liang, Polajnar D, et al. Local binary pattern network: a deep learning approach for face recognition[C]//Proc of IEEE International Conference on Image Processing. 2016: 3224-3228.

[17] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[18] Heikkilä M, Pietikäinen M, Schmid C. Description of interest regions with local binary patterns[J]. Pattern Recognition, 2009, 42(3): 425-436.

[19] Ng H W, Winkler S. A data-driven approach to cleaning large face datasets[C]//Proc of IEEE International Conference on Image Processing. 2014: 3434-3438.

[20] Labeled faces in the wild: a database for studying face recognition in unconstrained environments[R]. Amherst: University of Massachusetts, 2007.

Note: Figure translations are in progress. See original paper for figures.