Face recognition: Sparse Representation vs. Deep Learning

Neamah H. Alskeini, Kien Nguyen Thanh, Vinod Chandran, Wageeh Boles
Queensland University of Technology, Australia
Queensland, Australia
{neamah.alskeini, k.nguyenthanh, v.chandran, w.boles}@qut.edu.au

ABSTRACT

The pose, illumination and facial expression discrepancies between two face images are the key challenges in face recognition. The deep Convolutional Neural Networks (CNNs) and the fast Sparse Representation-based Classification (SRC) have achieved promising results in face recognition. However, CNNs require large databases and extremely expensive computations to overcome other algorithms. In this paper, we propose a novel SRC-based algorithm using test input image sets and training sub-databases, and compare its performance with CNNs. Histograms of Oriented Gradients (HOG) descriptors are used to define a new technique, named Training Image Modification (TIM), which provides image training sets with large variations of faces. The proposed algorithm divides the image training set into a number of sub-databases to address the dimensionality problem, and uses a test input image set to extract a signature from each sub-database using SRC. Each signature contains the same number of images as the test image set, although these may belong to different subjects. Considering all the sub-databases sequentially, the algorithm uses the signature of each sub-database to compute the number of images belonging to each subject. The signature that produces the Maximum Number of Images (MNI) of the same subject will have captured this subject for identification. YouTube Celebrity (YTC) and Multi-PIE databases are used in this work to evaluate the efficacy of the proposed method, which achieves high recognition rates. For relatively small databases, the proposed method is simple, scalable and stable, and it results in good face recognition rate under large face variations, as demonstrated by comparison with CNNs.

CCS Concepts

• Computing methodologies →Object recognition

Keywords

Face recognition; Deep learning; Image sets; Sparse coding; HOG descriptors

1. INTRODUCTION

Face recognition techniques work efficiently with controlled acquisition conditions and well aligned faces, but the dramatic facial appearance changes in pose, illumination and facial expression cause unreliable recognition. Although face recognition has been researched over two decades and many algorithms have been proposed, achieving a high recognition rate is still elusive [1-3].

Sparse coding or sparse representation is one of the most powerful algorithms which have been successfully used for face recognition applications [4]. A sparse representation of a test face image in terms of training data set is a promising recent direction for frontal face recognition, and is known as Sparse Representation-based Classification (SRC) [5]. The main idea of this algorithm is constructing a test face image from training samples via nonzero coefficients which fall on the correct subject [6]. In almost all face recognition systems, which use SRC, where only frontal gallery face images are used per subject, illumination and facial expression variations are regarded to provide large information. Wright et. al. [5] proposed SRC algorithm for face recognition from frontal views with varying illumination, occlusion and facial expression. They addressed the face recognition problem as multiple linear models that can be classified for one model. The tested samples can be only represented as a linear combination from training images of the same subject if a sufficient training image set is provided for each subject. To make SRC closer to the practical use, Wagner et al. [7] proposed a face recognition system which was robust to illumination variations using SRC and frontal training faces. They considered well controlled training image sets with enough variations of illumination, while testing image sets were taken under uncontrolled conditions. The proposed system was effective and efficient for face recognition under realistic conditions using only frontal face images. The main drawback of SRC is the assumption of face-accurate alignment between the test and training face images. If face images have pose variations and misalignment, these lead to a brittle inappropriate face recognition system. If a luxury acquisition system is designed for an application which demands a high face recognition rate, it is then unwise to limit a database to only frontal faces per subject [6], [8]. We seek to design a system that address this limitation.

Face image set classification uses sets or collections of face images for training and testing. Face images may have different pose and illumination even if they are captured on the same occasion [9]. Cui et al. [10] proposed an aligned image set technique to address face recognition variations, such as illumination, pose and expression. Image sets were aligned to a reference image set that is pre-structured and well defined into local linear models offline. Another study [11] developed an image set based face recognition technique using the geometric distances of closest points between two models. Image sets which are captured in different occasions for the same subject are unlikely to overlap at every point, but they should be close to each

ICGSP 2018, October 6–8, 2018, Sydney, NSW, Australia

ACM ISBN 978-1-4503-6386-0/18/10…$15.00
DOI: https://doi.org/10.1145/3282286.3282291

other at some points. Hu et al. [12] presented a method based on sparse approximated nearest points for image set classification. This method used the nearest point distance between image sets to measure the similarity by employing a sparse approximation technique. All these methods achieved considerable performance for face recognition with large variations of faces. However, the key problem of face image set classification is how to represent and model every face image set effectively because the samples within a data set are usually highly nonlinear. The existing work for face image set classification has achieved reasonably good performance, but strong assumptions were made such as appropriateness of subspace and Gaussian mixture models to represent image sets. These assumptions may not apply in several real world applications, particularly when there are large variations within image sets [9].

Feature extraction and representation are the central to the success for face recognition. Histograms of Oriented Gradients (HOG), for example, are image descriptors which are invariant to image rotation, scale and illumination, and have been used in several different tasks for computer vision [13], [14]. Dalal et al. [15] proposed these descriptors for human detection using edge orientation histograms. This algorithm divides an image into small connected cells and computes occurrences of gradient orientation in localized cells of the image. A histogram of gradient direction for the pixels within each cell is compiled, so the descriptors create a concatenation of these histograms. To improve the accuracy of the features, the local histogram is normalized using a value measured from the intensity across a large region of the image, called a block, to normalize all cells within the block. This normalization produces robust features in illumination, pose and facial expression changes which are also invariant to photometric and geometric transformations [16]. HOG descriptors provide robust features to small deformations, such as pose and illumination and they have a good discriminative ability.

The success of HOG descriptors in object detection tasks has inspired its application in face recognition. Deniz et al. [17], for example, proposed a new approach to build robust HOG descriptors by combining different scales of the descriptors and applying a linear dimensionality system to reduce face image noise. This makes the classifier less prone and more efficient to overfitting for face recognition. Li and Huo [18] also proposed a new method, called Locality Sensitive Histograms of Oriented Gradients (LSHOG) which focused on feature extraction problems to overcome the limitations of existing work. For each pixel, a histogram of gradient orientations was computed using LSHOG, and for each value of a gradient direction, a local sensitive parameter was added to decline these values exponentially with respect to the distance between pixels of the values. HOG descriptors extract robust features for face images, even under variations of pose, illumination and facial expression, and provide acceptable performance for face recognition. However, HOG descriptors extract features from a face of specific conditions, and may not be robust if the face conditions change.

Many proposed methods in the literature do not use deep learning, such as SIFT, LBP [19] and HOG [17]. This paper is concerned mainly with deep learning and the speed of SRC for face recognition. The CNNs feature extractors have the ability to compose several linear and non-linear operators of a learnable function [20], while SRC constructs sparsely a test image from training images using a linear regression function. A representative system [21] of DeepFace used deep CNNs to classify a large face dataset of 4 million examples of 4000 unique identities. The goal of face training is to minimize the distance between congruous pairs of faces of the same identity and maximise the distance between incongruous pairs to form metric learning. When DeepFace was introduced, it achieved the best performance on the Labelled Faces in the Wild (LFW) benchmark as well as the YouTube Faces in the Wild (YFW) benchmark. The DeepFace work was extended by series of DeepId papers by Sun et al. [22], [23], each of which incrementally increased the performance of face recognition on LFW and YFW. Many new ideas have been inspired over this series of papers, including using different deep CNNs architectures which are fully connected after each convolution layer, and multi-task learning over identification and verification. However, deep learning requires a large number of images, so it is unlikely to outperform other methods of face recognition if only thousands of images are used. Deep learning is also extremely computationally expensive, and some complex models require weeks to train using high end computers and expensive graphics processing units (GPUs).

In this paper, we propose a new CNN architecture and SRC-based algorithms using input image sets from YTC and Multi-PIE databases. For SRC algorithms, HOG descriptors are used to define a new training technique (Training Image Modification or TIM), which provides an image training set with large variations of faces. The proposed method divides the image training set into a number of sub-databases, and uses a test input image set to extract a signature from each sub-database using SRC. The signature that produces the Maximum Number of Images (MNI) of the same subject will have captured this subject for identification. SRC exhibits two different behaviors with sub-databases. It either returns a clear maximum number of facial images of a particular identity (this is most likely from the sub-database that contains the true identity) or returns facial images of a number of different identities.

2. THE PROPOSED ALGORITHMS

We propose two algorithms for face recognition. The first algorithm is a sparse hand crafted algorithm and the second algorithm is a face identification using deep feature learning. In this paper, many comparisons are illustrated between the two algorithms using Multi-PIE and YTC databases. Figure 1 and Figure 3 show the block diagrams of the proposed algorithms.

2.1 Sparse Hand-Crafted Algorithm

This algorithm is divided to three sections, namely HOG-based Training Image Modification (TIM), Sparse Representation-based Classification (SRC) using input image sets, and computing Maximum Number of Images (MNI) from signatures.

2.1.1 HOG-based Training Image Modification (TIM)

SRC algorithm provides acceptable performance when used for face recognition; however, it drops significantly as the number of subjects in a database increases [7]. In this section, we examine how to increase the recognition rate in a database which contains a large number of subjects using a HOG-based training technique to be later used with SRC. The problem of handling a large sample size is tackled here by dividing an image training set into a number of sub-databases. Each sub-database contains q subjects with l images per subjects which show large variations of pose, illumination and facial expression. Each subject is only represented in one sub-database.

There are many benefits of dividing the training set into several sub-databases. Firstly, representing a large number of subjects in one single set or database makes face recognition more complex and computationally expensive because a lot of time is consumed on training. Moreover, our technique is flexible, as it is possible to vary the number of subjects which are represented in the sub-databases. More subjects can be added by creating extra sub-databases. Separating a database into a number of sub-databases also decreases the misalignment problem. HOG descriptors and SRC algorithm work efficiently when a small number of subjects are represented in a sub-database, even if face images are not well aligned.

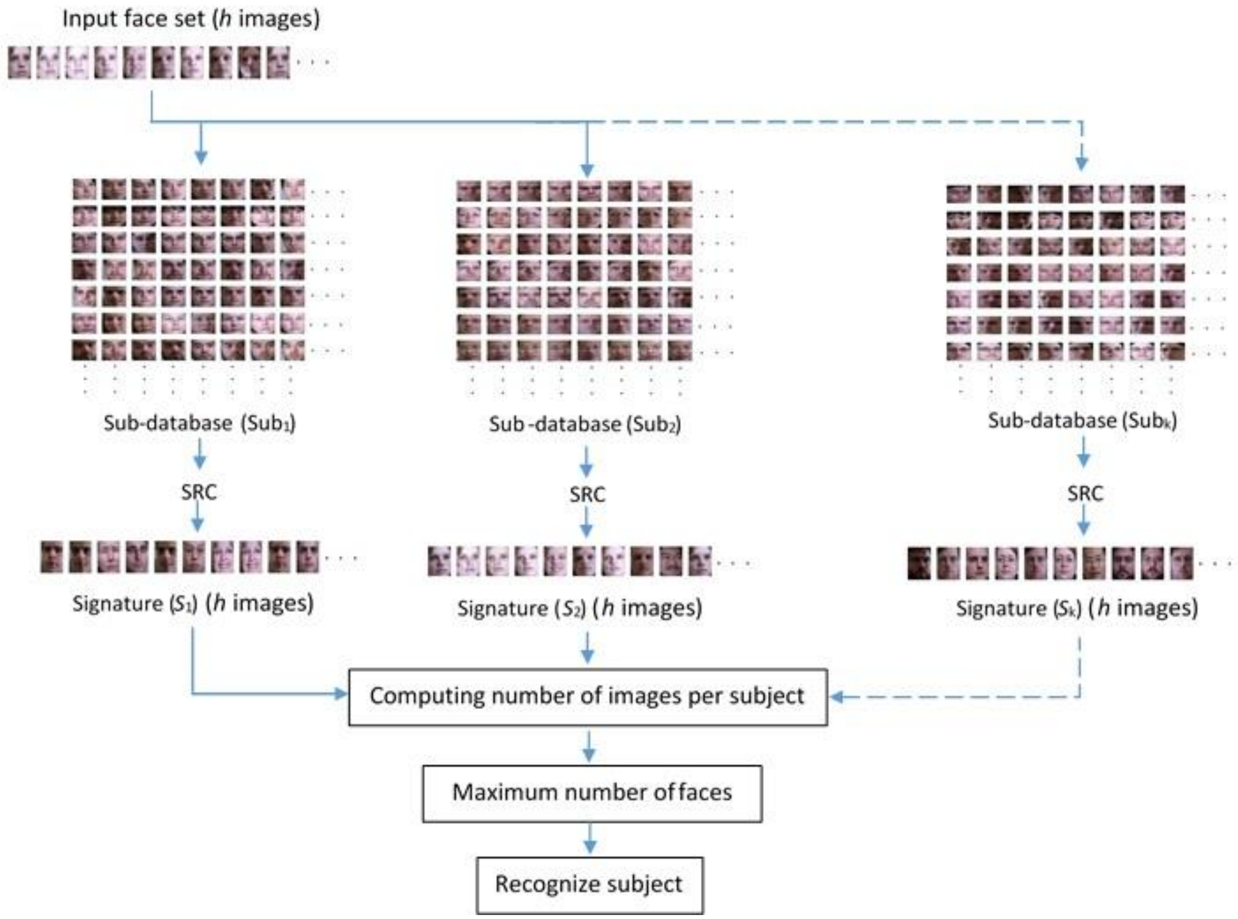
Input face set (h images)
Sub-database (Sub1)
Sub-database (Sub2)
Sub-database (Subk)
SRC
SRC
SRC
Signature (S1) (h images)
Signature (S2) (h images)
Signature (Sk) (h images)
Computing number of images per subject
Maximum number of faces
Recognize subject

Figure 1. Block diagram of the proposed face recognition in sub-databases.

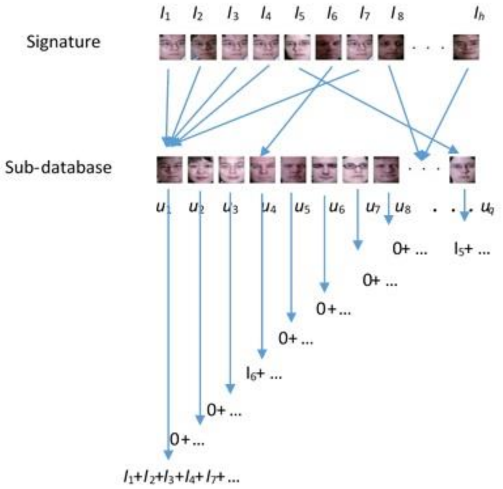

Figure 2. How to compute the total number of images per subject from a signature.

In this research, HOG descriptors are used to make sure that training images for each subject in each sub-database contain large variations. This technique will be referred to as Training Image Modification (TIM). In this technique, a number 1 of training images are randomly selected for each subject in each sub-database. The distance between the images in the training set is measured for each subject using HOG features. For each image, a vector of HOG features is extracted using a specific block size ($a \times b$ pixels of an image with $m \times n$ pixels). All the vectors are combined together and an average base vector is created for each subject. A distance with respect to the average base vector is computed for each training image. Each distance is compared with a threshold, which is the average of all distances of a subject. If a distance is less than the threshold, the image which belongs to the distance should be replaced with another image of a subject. This operation is applied for all images in training to make sure that the distance between training images is over the threshold. This means that training images will have larger variations after modifications. These descriptors then are employed to exclude the redundant faces from training images.

2.1.2 Sparse Representation-based Classification (SRC) Using Input Image Sets

The next step of the proposed algorithm consists of extracting signatures from the training sub-databases using SRC on test input image sets. SRC uses multiple registered training images for each subject. As explained in detail in [7], the images of subject i (i = 1, 2, 3, …, k) form a matrix $A = [A_1|A_2|\ldots|A_k] \epsilon R^{m\times n}$. A test image y_0 can be represented using a sparse linear combination Ax_0 from all images in the database plus a sparse error e_0 due to corrupted pixels. The sparse representation of the subject can be obtained by minimizing the l^1-norm of x and e as follows [7]:

$$min_{x,e}||x||_1 + ||e||_1 \text{ subject to } y_0 = Ax + e \quad (1)$$

If y_0 is a subject to pose, illumination and facial expression variations, the transformed image can not be sparsely represented as $y = Ax_0 + e_0$. That because the image pixels might not have an accurate correlation between test and training image sets. This is a limitation of the algorithm which leads to brittleness under illumination, pose and facial expression variations and makes it inappropriate for outside deployment settings. To deal with this problem, input image sets are used with SRC in this paper.

Traditional techniques use only one test image for face recognition. This may not be a good solution because face recognition is affected by many factors, including pose, illumination and facial expression. Instead, an input image set can be used to deal with these variations.

Let $V = \{i_1, i_2, \ldots, i_h\}$ be an input image set with h faces. This input image set contains large variations to mimic face recognition. For each input image set, a list of images (signature) is constructed from the each training sub-database using SRC for a subject. The total number of faces (h) which are selected from each training sub-database, and called a signature (S), should be equal to the number of images in each input image set (h). Since there are k sub-databases, k signatures are collected for each input image set. In this case, all signatures should have a variety of images from different subjects except the signature obtained from the sub-database which contains the tested subject.

2.1.3 Computing Maximum Number of Images (MNI) from Signatures

Each tested subject should be represented in just one sub-database of the full training data set. In this case, SRC will take one of two different options with sub-databases. This either makes a robust decision for a maximum number of images, or a weak decision for a lower number of images for the input image set. If a tested subject is represented in a sub-database, the maximum number of face images of a signature should belong to such subject. If a tested subject is not represented in a sub-database, then SRC selects a variety of face images of different subjects from the sub-database. The number of images in a signature per subject should be lower than in the case where the subject is represented in the sub-database, as illustrated in Figure 2.

An SRC-based signature (S), as explained in the previous section, contains h images ($I_1, I_2, \ldots, I_h$) of different subjects from a sub-database. Each sub-database contains q subjects with l images per subjects. The total number of images for each subject in a signature $\{u_1, u_2, \ldots, u_q\}$ is computed based on the q subjects in the sub-database. Each of these values should be labelled for a specific subject in a sub-database. All these values are similar, except for the one which belongs to the input tested subject. This will be the maximum value over all others and will be used for face recognition. In this case, MNI from the signature can be extracted as:

$$MNI = max(u_1, u_2, \ldots, u_q, u_{q+1}, \ldots, u_{2q}, \ldots, u_j) \quad (2)$$

where j is the total number of subjects in all sub-databases. Figure 2 shows how to compute the total number of images for each subject in a sub-database. In this figure, for example, the algorithm extracts u_1 images (I_1+I_2+I_3+I_4+I_7+ …) from the signature for subject number 1 in a sub-database. This operation is repeated for all signatures to calculate the number of face images which belong to each subject in each signature.

2.2 Face Identification Using Deep Feature Learning Algorithm

The proposed algorithm learns features the variations of pose, illumination and facial expression using Convolutional Neural Networks (CNNs). The convolution layers and pooling operations in CNNs are designed to extract robust features hierarchically, from local low-level features to be globally high-level ones. Our deep CNN has three linear convolutional layers and each one is followed by non-linear layers such as Rectified non-Linear Units (ReLU) and max pooling. Soft-max layer dimensions are 249 for Multi-PIE database and 47 for YTC database at their last layers of the feature extraction cascade. The network layers to be learned are fully-connected to the convolution layers. We use ReLU for neurons after each convolution layer. Figure 3 illustrates the CNN architecture to extract deep features of a given input image size $m \times n$ pixels. Where m = 100; n = 100 for Multi-PIE database. We use 7 ×7, 5 ×5, 3 ×3 filter dimensions of convolution layers 1, 2 and 3 respectively. Also, we use 2 × 2 dimensions for all max-pooling layers. CNNs achieve good results in very large databases, however, experimentally CNNs may not able to outperform other algorithms, such as SRC for relatively small databases [21].

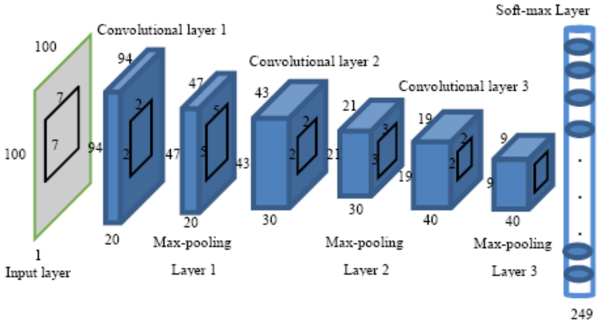

Figure 3. Block diagram of the deep learning architecture.

3. EXPERIMENTAL RESULTS

3.1 With Multi-PIE Database

We tested the proposed algorithm on Multi-PIE database for face recognition. The database includes a large number of subjects and each subject is represented by face images with different variations of pose, illumination and facial expression. These challenges are included in the testing and training images. Multi-PIE contains over 750,000 images of 337 subjects. They are imaged under 15 view points and 19 illumination conditions in up to four recording sessions. The images of this database were recorded under an improved environment, such as controlling head position and using a static background [24]. Because the database contains a large number of subjects who are affected by natural appearance variations over a period of several months, face recognition in this database is challenging. All faces were detected using Viola-Jones detector [25] and normalized to 100 × 100 pixels.

Image sets are employed as input instead of a single image because they provide large variations of pose, illumination and facial expression. In this work, image sets were used to investigate

MNI retrieved by SRC and TIM per user. The block size of HOG descriptors is 4×4 pixels. These techniques make face recognition more robust of vulnerable to outliers. For example, if some faces are invalid, that will not affect the performance of face recognition of the proposed algorithm.

3.2 Recognition with Frontal Faces in one Database

In this experiment, we first used all subjects (7 frontal images each as in [7]) from Session 1, which are 249 subjects, for training. Only random 7 natural images from folder (05_1), which has 20 images, of Multi-PIE database were used for training. The other 20 images from Sessions 2-4 were used for testing.

Table 1. Comparison of existing work and the proposed methods with frontal face.

Technique	Recognition rate (%)		
	S2	S3	S4
Nearest Subspace (NS) [7]	30.8	29.4	24.6
Nearest Neighbor (NN) [7]	26.4	24.7	21.9
LDA [7]	5.1	5.9	4.3
Local Binary Patterns (LBP) [7]	39.9	38.1	33.9
SRC	70.5	74.6	68.2
CNN	53.23	49.23	39.16
SRC+MNI+ random images	85	90.2	84.8
CNN+MNI+ random images	56.92	50.03	46.64
SRC+MNI+TIM	**95.1**	**93.4**	**93.6**
CNN+MNI+TIM	76.92	69.23	58.49

We experimented SRC algorithm and MNI technique as shown in Table 1 which explains the results of three Sessions 2-4. For both training and testing image sets, only illumination variations were represented. From the table, we see that SRC+TIM+MNI technique performed higher results than SRC+TIM, SRC+MNI and even CNN+MNI+TIM of all Sessions 2-4. MNI technique ignores invalid images which cause errors of face recognition. We also compared the proposed approach with existing work using frontal faces [7], including Nearest Neighbor (NN), Nearest Subspace (NS) [3], Linear Discriminant Analysis (LDA) [2] and Local Binary Patterns (LBP) [19]. Results are shown in Table1. These algorithms used just neutral expression in training to evaluate the performance. Their performance is low because the misalignment of faces affects them, particularly with a large number of subjects. Actually, these algorithms do not work well if faces are extracted by a face detector such as Viola-Jones. This face detector can detect frontal faces efficiently, but the detected faces have misalignment problems and these algorithms are not robust to these problems. The misalignment problem causes loss of useful information of faces. This issue produces low values of feature vectors. The low values of vectors appear clearly in a large number of subjects because they may overlap and affect face recognition. As shown in Table1, our algorithms achieved higher recognition rate than other algorithms. Using only frontal faces for recognition makes systems far for practical use and intrusive of users. Capturing only frontal faces is also not easy job, particularly if users do not pay attention to a camera.

3.3 Recognition with Pose, Illumination and Expression Variations in one Database

In order to illustrate the effectiveness of the proposed method, we validated the performance of the algorithm with variations of pose, illumination and facial expression. We used all 249 subjects which are presented in Session S1 for training sets. We used 15 random images (2 natural and 1 smile images from pose -15^{o} , 6 natural and 3 smile images from frontal faces, 2 natural and 1 smile images from pose +15^{o}) from Session 1 for training. For each subject, 25 images were randomly used for testing, which are -15^{o} pose faces, frontal faces and +15^{o} from Sessions S2-S4.

Table 2. Face recognition for pose, illumination and facial expression from one database and sub-databases.

Sessions	Techniques	One database(%)	Sub-databases(%)
Session 2	SRC	65	66.74
	CNN	60.9	-
	SRC+MNI+ random images	84.9	85.40
	CNN+MNI+ random images	68.84	-
	SRC+MNI+TIM	**90**	**91.58**
	CNN+MNI+TIM	69.23	-
Session 3	SRC	29.2	62.6
	CNN	28.84	-
	SRC+MNI+ random images	65.5	84.8
	CNN+MNI+ random images	60.76	-
	SRC+MNI+TIM	**79.8**	**86.1**
	CNN+MNI+TIM	61.53	-
Session 4	SRC	24.2	54.4
	CNN	26.33	-
	SRC+MNI+ random images	50.9	70.59
	CNN+MNI+ random images	39.23	-
	SRC+MNI+TIM	**70**	**76.4**
	CNN+MNI+TIM	53.84	-

Table 2 illustrates that SRC+TIM+MNI algorithm achieves higher recognition rate than other algorithms in all Sessions S2- S4. Representing all face variations, including pose, illumination and facial expression in one large database does not solve face recognition issue because the large variations of faces causes loss useful information

3.4 Recognition with Pose, Illumination and Expression Variations in Sub-databases

To obtain high recognition rate, we separated a database into a number of sub-databases and used image sets to evaluate face recognition performance with respect to the ability and capacity of SRC algorithm. Separating the full database into a number of sub-databases, which contains a number of subjects, is investigated experimentally in MATLAB to choose the number of subjects for each sub-database. For each subject, we selected same setting of number of images of the previous one database section for testing and training, but different number of subjects in a sub-database. A curve is plotted for SRC performance as shown in Figure 4. The training images were selected for Session S1 while test images from Session S2. With the range of number of sub-databases (3- 7 sub-databases), SRC achieves the highest recognition rate (66.74%) at 50 subjects per sub-database, but SRC decreases with respect to other numbers of sub-databases. So, this number of sub-databases is selected for all other Sessions S2- S4 of our experiments.

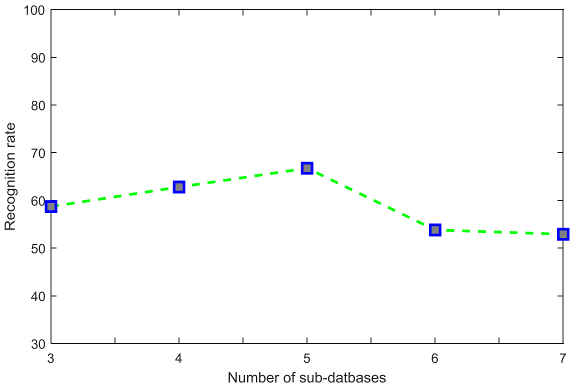

Figure 4. Recognition rate of SRC of different number of sub-databases.

In order to illustrate the effectiveness of the proposed method, we validated the performance of the algorithm with variations of pose, illumination and facial expression. We used all 249 subjects which are presented in Session S1 for the training set. For each subject in this Session, we selected as 15 images as setting in the previous one database section with variations of pose, illumination and facial expression to build the sub-databases. According to the proposed algorithm, we first constructed 5 sub-databases from Session S1 which has 249 subjects. Each sub-database contains 50 subjects except sub-database number 5, which contains 49 subjects (50 × 4 + 49 = 249). For each subject, 25 test images were randomly selected from Sessions S2-S4 to represent these variations. Table 2 shows results of the proposed methods. From this table, we can note that face recognition rates of the sub-database algorithm is the highest in all the sessions S2-S4.

3.5 With YouTube Celebrities Database

The YouTube Celebrity (YTC) Database [26] is the largest database used for face image set identification in the wild. It contains 1910 video clips of 47 politicians and celebrities. These real world videos are downloaded from YouTube and recorded at high compression rates, so they have low resolution quality. We follow the same setting in [27], [28] for the experiments and divide the database to five folds. Each fold contains 9 non-overlap video clips for each individual. Three video clips were randomly selected as a training set while the remaining six were used as test image sets. All the face images were converted to grayscale and resampled to the resolution of 30 × 30 pixels. We randomly selected 20 face images from each of the three training video clips per individual. The videos in this challenging database exhibit a wide range of appearance variations because they were captured in real life scenarios. The achieved results of our proposed method significantly outperform the existing methods to perform improvement of 9.35% over the best method as shown in Table 3. We use Viola and Jones for face detection [22]. From Table 3, we can see that our techniques are much better than the existing work. Also, SRC outperformed CNN because the limitation of the number of images per subject. SRC is much faster than CNN which takes much time for training in this database.

Table 3. YouTube face celebrities.

Technique	Recognition rate (%)
SSDML [29]	67.00
DRM-WV [27]	72.23
DRM-PWV [27]	72.55
SRC	64.30
CNN	62.95
SRC+MNI+ random images	80.16
CNN+MNI + random images	73.33
SRC+MNI+TIM	**81.90**
CNN+MNI + TIM	74.21

4. CONCLUSION

In this research, we proposed two algorithms to improve face recognition performance. The first algorithm uses Sparse Representation-based Classification (SRC), Training Image Modification (TIM), Histograms of Oriented Gradients (HOG) descriptors, Maximum Number of Images (MNI) from sub-databases. The second algorithm is Convolution Neural Networks (CNNs). For the SRC algorithm, the full database is divided into a number of sub-databases to deal with the dimensionality problem. For each input image set which has large variations of pose, illumination and facial expression, SRC is used to capture a signature from each sub-database. Depending on MNI of signatures, this algorithm can recognize a tested subject from sub-databases. According to the experiments performed on Multi- PIE and YTC databases, this algorithm achieved considerable results under large face variations. It is robust to outliers and misalignment. We also noted that SRC outperforms the proposed CNN architecture because deep learning needs very large databases to achieve good results.

For future work, we will further improve face recognition by combining the locality of CNNs and linearity of SRC in large face variations. Investigating errors of face recognition in pose, illumination and facial expression is still an open problem.

5. REFERENCES

[1] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[2] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711– 720, 1997.

[3] K.-C. Lee, J. Ho, and D. J. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):684–698, 2005.

[4] J. Yang, L. Luo, J. Qian, Y. Tai, F. Zhang, and Y. Xu. Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(1):156–171, 2017.

[5] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009.

[6] L. Zhang, M. Yang, and X. Feng. Sparse representation or collaborative representation: Which helps face recognition?

IEEE International Conference on Computer Vision (ICCV), pages 471–478, 2011.

[7] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma. Toward a practical face recognition system: Robust alignment and illumination by sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(2):372–386, 2012.

[8] C. Ding and D. Tao. A comprehensive survey on pose-invariant face recognition. *ACM Transactions on Intelligent Systems and Technology (TIST),* 7(3):1–40, 2016.

[9] J. Lu, G. Wang, W. Deng, P. Moulin, and J. Zhou. Multi-manifold deep metric learning for image set classification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1137– 1145, 2015.

[10] Z. Cui, S. Shan, H. Zhang, S. Lao, and X. Chen. Image sets alignment for video-based face recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pages 2626–2633, 2012.

[11] H. Cevikalp and B. Triggs. Face recognition based on image sets*. IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pages 71–86, 2010.

[12] Y. Hu, A. S. Mian, and R. Owens. Sparse approximated nearest points for image set classification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pages 121–128, 2011.

[13] P. Tian and Dong. A review on image feature extraction and representation techniques. *International journal of multimedia and ubiquitous engineering*, 8(4):385–396, 2013.

[14] T. Watanabe, S. Ito, and K. Yokoi. Co-occurrence histograms of oriented gradients for pedestrian detection. *Advances in Image and Video Technology*, pages 37–47, 2009.

[15] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR),* 1:886–893, 2005.

[16] M. Pedersoli, J. Gonz`alez, B. Chakraborty, and J. J. Villanueva. Enhancing real-time human detection based on histograms of oriented gradients. *Computer Recognition Systems 2,* pages 739–746, 2007.

[17] O. D ́eniz, G. Bueno, J. Salido, and F. De la Torre. Face recognition using histograms of oriented gradients. *Pattern Recognition Letters,* 32(12):1598–1603, 2011.

[18] B. Li and G. Huo. Face recognition using histograms of oriented gradients. *Optik-international journal for light and electron optics*, 127(6):3489–3494, 2016.

[19] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 28(12):2037–2041, 2006.

[20] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. *BMVC*, 1(3):1–12, 2015.

[21] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.

[22] Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873,* pages 1–5, 2015.

[23] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2892–2900, 2015.

[24] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. *Image and Vision Computing*, 28(5):807–813, 2010.

[25] P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[26] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pages 1–8, 2008.

[27] M. Hayat, M. Bennamoun, and S. An. Deep reconstruction models for image set classification. *IEEE transactions on Pattern Analysis and Machine Intelligence*, pages 713–727, 2015.

[28] Y. Hu, A. S. Mian, and R. Owens. Face recognition using sparse approximated nearest points between image sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1992–2004, 2012.

[29] P. Zhu, L. Zhang, W. Zuo, and D. Zhang. From point to set: Extend the learning of distance metrics. *IEEE International Conference on Computer Vision (ICCV),* pages 2664–2671, 2013.