Application of Sparse Coding in Face Identification

David Molin*
Nat Thavornpitak†

***Abstract*—The problem of face identification is a well studied problem. This paper proposes a face identification algorithm which represents a face image with its sparse code and uses stochastic gradient ascent to find a metric which is able to classify whether two test images depict the same face or not. The performance of this approach on the Pose, illumination and expression dataset is 94% and the performance on Labeled faces in the wild dataset is 69%.**

I. Introduction

Face Classification vs Face Verification vs Face Identification

One of the most studied problems in computer vision is face identification. Face identification is the problem which a pair of faces are given and a system determines whether or not both faces belong to the same person. Face identification is often confused with face classification and face verification(Authentication). In face classification, a test face image is given and a system determines which subjects in a dictionary a test face belongs to. Similar to face identification, face verification tries to determine whether or not two images depicts the same person. The difference is that face verification compares an unknown test face against a known face in a dictionary, while face identification compares two unknown test faces. An example of face verification is when there is a dictionary consisting of K people whose identities are known. Then there is a person who claims to be the i^{th} person in dictionary. A face verification system verifies whether a person' identity match the claimed identity[1]. Usually the images of a person in a dictionary are captured under controlled environment to ensure that different poses and different facial expressions are captured. Compared to face verification, face identification is more challenging since given test faces may belong persons who are not present in a dictionary. Therefore, a face identification system must be able to determine whether or not a pair of test images belong to the same person; although, there is no prior information about persons who are in test images.

Face Identification : Problem Specification

In this section we formally formulate the problem of face identification which we explore in this paper. The goal of face identification is to determine whether or not the given two persons have the same identify i.e. they are in fact the same person. For training data, we assume that the identity of each training face is unknown. The only label available is whether or not the two faces are the same person. In other words, a training pair consists of two image depicting two faces. We do not know which person each face belongs too. The only information, we have is whether or not the two faces are the same person. The reason why we formulate the problem this way is because we aim to develop a system that works in every scenario including in a scenario which we have very limited prior knowledge about training samples. Furthermore, we assume a test face pair consists of only two images- one for each subject. This is a challenging problem since given only one image, it is difficult to observe intra-class variation—how face of a person changes when pose and expression varies.

*dmolin@stanford.edu, †nthavorn@stanford.edu

Face Identification : Application

Face identification technology can be applied in many areas. One example of the applications is when two surveillance cameras capture images of an unknown suspect from two different crime scenes and from captured images, an identification system is used to determine whether or not the two suspects are the same person. An algorithm used to solve this problem has to be specifically designed for face identification problem. Face Verification and face recognition algorithms cannot be used since these algorithms assume that a set of deliberately-collected facial images is available for at least one of the two suspects.

Face Identification : Proposed Solution

Recently, sparse representation or sparse coding has been successfully used in face classification system. The developed system is robust to noise, illumination, slight pose difference and occlusion[2]. This suggests that sparse coding could be applied to the similar problem of face identification. In this project, we propose a face identification system based on Mahalanobis distance and sparse representation. The proposed algorithm consists of two phases, training and test phase. In the training phase a dictionary is constructed and sparse representation of all training data are calculated using the dictionary. The sparse representations of training images as well as labels of training images pairs are fed into a learning algorithm to learn a Mahalanobis matrix M which minimizes the Mahalanobis distance between images of the same person while maximizing the distance between images of different subjects. In the test phase, sparse representation of test image pairs and Mahalanobis distance between them are calculated using a dictionary D and a Mahalanobis Matrix M. Then based on, the Mahalanobis distance, a test image pair is classified as either class 0(both images depict the same subject) or class 1(each image depicts different subject). The outline of the algorithm is shown in figure 1

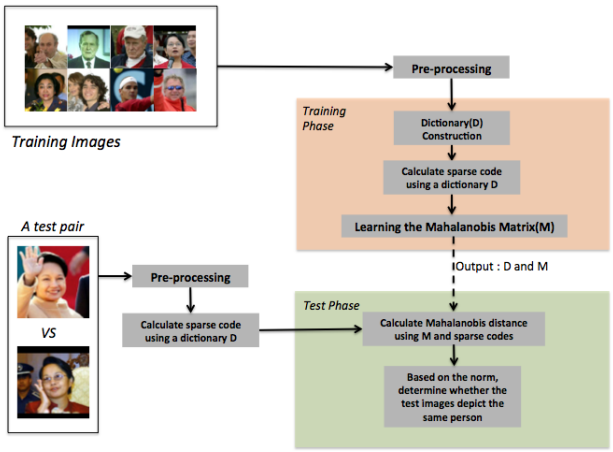

Fig. 1. Outline of the proposed algorithm

II. Related Works

A. Existing Face Identification Algorithms

Several algorithms have been proposed to solve face identification problem. Face Identification algorithms can be classified into three approaches.

The first approach is feature based classification which tries to develop features that are invariant to intra-class variation but highly discriminative among different classes. Liu et al develop a technique called Individual Eigenface Subspace Method(IESM) which is based on the assumption that Eigen subspace for images of the same person is approximately the same despite changes in illumination and expression[1]. This algorithm extracts principal components from the first image and project the second image onto extracted subspace. Then it recovers the second face from the projection. If both faces depict the same person, Eigen subspaces for both image are similar. Therefore, reconstruction of the second face will have small residue. In contrast if both images are from two different people, the residue will be large[1]. Another feature based classification algorithm is proposed by Zhou et al[3]. Zhou et al use Gabor wavelets transforms with 5 scales and 8 orientations to extract features from images and then AdaBoost is used to select subset of features that are highly discriminative[3]. Classification is based on similarity of those features. Ben-Yacoub et al extend Gabor wavelets transforms method to handle more intra-class variation by applying it with Elastic Graph Matching(EMG) [4].

Another approach is Correlation Filter approach. To determine if a pair of test faces are the same person, Savvides et al use images of the first person to construct Minimum Average Correlation Energy(MACE) Filter[5]. The second image is then passed to the filter and from the filter response, peak to side lobe ratio(PSR) is used to measure the height of the peak relative to the response in other regions. If the PSR exceeds a threshold, the pair of images is classified as depicting the same person[5].

The last approach applies learning algorithm to solve face identification problem. Chopra et al use learning algorithm to learn Similarity Metric that can be used to classify a pair of face as true match(depicting the same person) and false match(depicting two different people). The goal is to learn a mapping function which maps raw pixels to low-dimension space in which images from the same person cluster tightly together while images from different people are separated far apart. Convolutional Networks are used to learn such mapping function [6]. Guillaumin et al use Logistic Discriminant based Metric Learning to learn the metric and then use k-Nearest Neighbor to classify a test image pair [7].

B. Proposed Solution

In this project, we propose a face identification system based on Mahalanobis distance and Sparse Representation. There are three advantages of using the proposed method. First, the proposed algorithm does not rely on feature extraction. In feature extraction process, there are many parameters to be calibrated and parameters calibration requires either exhaustive search for the best parameters or humans' supervision. The algorithm in [1] requires humans to calibrate some parameters such as number of principal component retained, while method proposed in [3] requires either an exhaustive search or an algorithm to select parameters. Eliminating feature extraction process increases the speed and decreases level of human supervision. This motivates us to explore application of sparse representation. According to [2], for face images, if the size of a dictionary is large enough, a sparse representation of an image will not depend heavily on features used. Therefore, we can use raw pixels to calculate sparse representation and eliminate feature extraction process. It is also possible to improve the performance of the algorithm by using features that are invariant to pose or robust to occlusion together with sparse representation. However without those features the algorithm still can perform the task.

In addition, the proposed system is able to perform face identification by using only two test images. In contrast, correlation filter approach requires at least three images to build a MACE correlation filter[5]. Therefore, given only two images, one from each subject, the correlation approach will not be able to perform the task. The proposed method can identify a pair of faces by using only two images.

The third advantage is computational complexity and scalability. Complicated learning algorithms have high complexity; therefore, they do not scale well when dimensions of data and size of a training set grows. In addition, some learning algorithm such as the method in [7] use non-parametric learning which requires all training samples to be stored; therefore, the memory required to store this data grow with the size and dimension of data. Our proposed method uses a simple stochastic gradient ascent to learn a few parameters and after the training phase is done, training data can be discarded. The simplicity and parametric nature of the proposed algorithm allows the algorithm to be scalable to include larger training samples.

III. METHOD

A. Overview of the Proposed Algorithm

In this project we propose to use sparse code as a representation of each image, and then the Mahalanobis distance between two sparse codes is computed to determine whether or not both images belong to the same person. The outline of the algorithm is as follows:

Training

1. Consturct a dictionary which is a matrix whose column is data from each example.
2. Calculate sparse coefficient $\hat{x}_1$ and $\hat{x}_2$ from a pair of training images y_1 and y_2
3. Learn the Mahalanobis matrix M from training data

Testing

1. Calculate sparse coefficient $\hat{x}_1$ and $\hat{x}_2$ from a pair of test images y_1 and y_2
2. Use the matrix M learned from the learning algorithm in the previous part to calculate the Mahalanobis distance between the two test images.
3. Based on the Mahalanobis distance calculated, determine whether or not a pair of test image belong to the same person(class 0) or different persons(class 1) .

B. Sparse Coding

Sparse coding is based on an assumption that if we have enough training faces in a dictionary, we can express an unknown face y as a sparse linear combination of entries in a dictionary[2]. Let A be a dictionary where every column is a representation of each face. Then a new example y can be written as

$$Ax = y \tag{1}$$

We would like find a coefficient x that satisfies the above equation. However we have many training examples i.e, the number of columns is greater than the number of rows. Therefore, the system is overdetermined. A sparse representation is a solution of $Ax = y$ where most of the elements in x are zero. To compute the spare representation, we solve the following minimization problem.

$$\hat{x} = argmin\|x\|_0 \text{ subject to } Ax = y \tag{2}$$

$\|x\|_0$ represents the ℓ^0 norm which is the number of non-zero entries of x. Due to its complexity, the ℓ^0 minimization problem cannot be solved efficiently for a large A. [8] and [9] show that a solution of the ℓ^0 minimization problem can be approximated by a solution of the ℓ^1 minimization problem with the same objective function and constraints. Therefore, the minimization problem becomes

$$\hat{x} = argmin\|x\|_1 \text{ subject to } Ax = y \tag{3}$$

In addition, due to noises, the equality constraint is relaxed to an inequality constraint.

$$\hat{x} = argmin\|x\|_1 \text{ subject to } \|Ax - y\|_2 \leq \varepsilon \tag{4}$$

where $\|x\|_2$ represents the ℓ^2 norm of x. $\hat{x}$, the solution of this problem is a sparse representation of x.

C. Occlusion Handling Technique

To deal with small occlusion, [2] recommends appending a dictionary by an identity matrix and solve the minimization problem.

$$y = \begin{bmatrix} A & I \end{bmatrix} \begin{bmatrix} x_0 \\ e_0 \end{bmatrix} = Bw \tag{5}$$

$$\hat{w} = argmin\|w\|_1 \text{ subject to } \|Bw - y\|_2 \leq \varepsilon \tag{6}$$

where $B = \begin{bmatrix} A & I \end{bmatrix}$ and $\hat{w} = \begin{bmatrix} \hat{x} \\ e_0 \end{bmatrix}$. $A\hat{x}$ is an image with occlusion removed while Ie_0 is occlusion. $\hat{x}$ is a sparse representation of y with occlusion removed.

D. Using Mahalanobis Distance in Face Identification

Let x_i and x_j be representations of face i and j then the Mahalanobis distance $d_M(x_i, x_j)$ is defined as[7]:

$$d_M(x_i, x_j) = (x_i - x_j)^T M (x_i - x_j) \tag{7}$$

Mahalanobis distance is a metric that measures how similar two sparse codes are. Since images of the same subject are more likely to have similar sparse code than images of different subjects. We expect to see small Mahalanobis distance between images of the same subject and large distance between images of different subjects. From Mahalanobis distance, we can calculate p_n which is the measure of how likely the pair of test images belong to the same subject[7].

$$p_n = \sigma(b - d_M(x_i, x_j)) \tag{8}$$

where $\sigma(z) = \frac{1}{1+exp(-z)}$ and b is a bias term[7].Then we determine whether or not a pair of test images depict the same person based on p_n. If p_n exceeds a certain threshold then an image pair is assigned to class 0(same person), otherwise it is assigned to class 1.

E. Learning algorithm to learn the Mahalanobis matrix M

[7] models p_n which is a measure of how likely that two test images belong to the same person as a function of $d_M(x_i, x_j)$.

$$p_n = \sigma(b - d_M(x_i, x_j)) \tag{9}$$

The likelihood function L can be written as

$$L = \prod_{n=1}^{N} p_n^{(1-c_n)} (1 - p_n)^{c_n} \tag{10}$$

where c_n is a class label of an image pair. The log-lokelihood function can be expressed as

$$\ell = \sum_{n=1}^{N} ((1 - c_n) log(p_n) + c_n log(1 - p_n)) \tag{11}$$

The gradient of the log-likelihood with respect to M and b is

$$\nabla_M l = \sum_{n=1}^{N} (p_n - 1 + c_n)(x_i - x_j)(x_i - x_j)^T \quad (12)$$

$$\nabla_b l = \sum_{n=1}^{N} (1 - c_n - p_n) \quad (13)$$

In this project, we use stochastic gradient ascent to find M and b that maximizes the log-likelihood function.

IV. Experiments and Results

A. Implementation

The system is implemented in MATLAB. There are two details about the implementation. First, we use ℓ^1 Minimization via Randomized First Order Algorithms to solve an optimization algorithm in (4)[10]. Secondly, to implement the face identification system, two additional tasks are performed. First all images are pre-processed by scaling, translating and rotating so that pixel locations of the two eyes are the same for all images. After pre-processing, the second additional task is to reduce dimension of images by using principal component analysis. This is done after dictionary construction. After constructing a dictionary, we calculate subspace spanned by principal components from a dictionary D. Then all training and test images are projected onto this subspace.

B. Experiments Setup

To evaluate the performance, we classify each test image pair into categories and report accuracy for each category.

A. Class 0 vs Class 1

Test case A0 : A pair of test image belong to the same person(class 0).

Test case A1 : A pair of test image belongs to different persons(class 1).

The reason why we break down test pairs into test case A0 and A1 is that we have a highly unbalanced test set. The number of test pairs that belong to two different persons are a lot greater than the number of test pairs that belong to the same person. The system can achieve high accuracy by biassing the output toward class 1, but a good system must perform identification accurately in every situation. Therefore, it is reasonable to evaluate the performance based on identification accuracy in both test case A0 and A1 instead of overall accuracy.

B. The presence of subjects in a dictionary

Let y_1 and y_2 be an image of subject A and B. We can classify a given test pair based on the presence of subject A and B in the dictionary.

Test case B1 : Both subject A and subject B are in the dictionary

Test case B2 : One of the subjects is in the dictionary

Test case B3 : None of the subjects is in the dictionary.

Test case B3 is the most challenging case since the system tries to estimate the similarity between two persons that it has never seen before. Test case B3 is also the most likely scenario to be encountered in practical application. Therefore, an efficient identification algorithm should have high performance in B3.

Fig. 2. Examples of images from Pose, Illuminaiton and Expression dataset. Each column shows images of the same person.

C. Results on Pose, Illumination and Expression(PIE) Dataset

We use two datasets to evaluate the performance of our algorithm. The first dataset is Pose, Illumination and Expression(PIE) database which consists of 66 subjects. For each subject, 21 images of different illumination and expressions and slightly different poses are collected. Each image is preprocessed as described in 4.1. To build a dictionary $numPersonD$ subjects are randomly selected and for each selected subject $numImgD$ images are randomly selected to be included in a dictionary. The entire dataset are then divided into non-overlapped training and test set. For each subject $numPersonTrain$ images are randomly selected to be in a training set while the rest are assigned to the test set. For each train and test image y, a sparse representation $\hat{x}$ is computed and normalized. All possible pairs of faces in the training set are used as training data. Similarly all possible pairs of face images in the test set are used to test the system.

Results

Parameters used in the experiment are as follow : $numPersonD$= 30 , $numImgD$= 4, $numPersonTrain$= 4.

Test Case	Accuracy
A0(same person)	92.73%
A1(diff person)	95.47%
B1(2 subjects in D)	97.47%
B2(1 subject in D)	96.49%
B3(0 subject in D)	95.37%

D. Results on Labeled Face in the Wild(LFW) Dataset

After evaluating the performance on the PIE dataset, we evaluate the performance on Labeled Face in the Wild(LFW) which is a lot more challenging due to occlusions, large variation of poses, illumination, expression and image resolution.

Fig. 3. Examples of images pairs from LFW that are correctly identified by our algorithm. The top two rows are images of the same person, while the bottom rows are images of different people.

Fig. 4. Examples of images pairs from LFW that are incorrectly identified by our algorithm. The top two rows are images of the same person, while the bottom rows are images of different people.

Unlike PIE dataset, LFW dataset consists of face images downloaded from the Internet. Since images in this dataset are captured in uncontrolled settings, faces are not centered and scales vary. Therefore, in this project, we use images preprocessed by the algorithm described by Huang et al[11] to center and scale images. LFW dataset consists of 5749 subjects and the number of total images are 13233. There are 4069 subjects of which there is only on image. Since the structure of the dataset is different from the PIE dataset. Instead of using $numPersonD$ subjects and $numImgD$ images for each subject to build a dictionary, for LFW we specify $sizeD$, a size of a dictionary. We randomly select $sizeD$ images from the dataset. All $sizeD$ images can be from the same classes or all different classes. The majority of subjects in a training set have only one sample; therefore, randomly select $numImgTrain$ images into the training set may result in having a few or none of training image of class 0(same person). Therefore, we randomly select (2/3)*$numImgTrain$ images from 1680 subjects which have two or more images samples. And select the rest of the training set randomly.

Results

Parameters used in the experiment are as follow : $sizeD$ = 200 , $numImgTrain$= 2000.

Test Case	Accuracy
A0(same person)	69.45%
A1(diff person)	69.69%
B1(2 subjects in D)	64.21%
B2(1 subject in a D)	67.89%
B3(0 subject in D)	69.71%

V. Discussion and Conclusion

A. Discussion

The proposed algorithm performs well on the PIE dataset. The overall accuracy is 95%. The algorithm can perform face identification despite variation in illumination, facial expression and slight variation of pose. Applying a learning algorithm to learn M and b increases the performance from 78% to 95%.

After evaluating the proposed algorithm on the PIE dataset. We evaluate the algorithm on a challenging LFW dataset and the overall accuracy is 69%. Applying occlusion handling technique discussed above improves the accuracy from 66% to 69%. The two major problems of the proposed algorithm are incapability to handle high variation in pose and occlusion. The first problem is that the algorithm fails to handle faces with various poses. There are two explanations for this failure. First the proposed algorithm determines the match of two test faces based on only on image for each face.Therefore, it is difficult to observe how face of the same person changes when viewpoints and poses change. The second explanation is that we use sparse codes to represent images and sparse representation does not explicitly model each component of face such as eyes and nose as separate components, whose relative locations can change when poses vary. When a given test images pair is of the same person with small variation in pose, relative locations of each face component remain approximately the same. Since sparse representation is a linear combination coefficient of a dictionary entries, it can capture distinguishing identity of faces while allowing slight shift in each face component. In contrast, if the difference in pose is extreme, areas of each face component of the first face (e.g. eyes, nose) no longer intersect areas of the

corresponding face components of the second face. This results in large Mahalanobis distance between two sparse coefficients and the algorithm incorrectly classifies the pair as faces from different people.

The second problem is occlusion. When large potion of face is occluded by sunglasses, a cap or a scarf, the algorithm cannot accurately identify the pair of images. This is because when there is a large occluded region sparse coding algorithm will focus on matching the occluded region. This results in inaccurate sparse codes and inaccurate identification result. In addition, since images are randomly selected into a dictionary, there is a possibility that there are images with occlusion in a dictionary. Therefore,occlusion also affects the performance of an algorithm on identifying an unoccluded image pair.

B. Suggested Future work

To deal with pose variation, we would suggest using other face recognition techniques to determine poses of two test images. During the training phase, M and b for each combination of poses are learned. During testing, pose of each image in the test pair is determined and M and b specifically learned for those two poses are used to calculate the distance between the two images.

Another interesting research is on how to handle occluded images when they are either in a test set or in a dictionary. The technique presented in section 3.3 alleviates the effect of occlusion, but does not resolve it. One possible way to solve occlusion problem in test images is to apply the technique used in [2]. To apply this technique, images are divided into non-overlapped section and a dictionary for each section is built. M and b for each section are learned. During testing image pairs are divided and each section of images is tested separately. For each section, the algorithm decides whether or not the same person is depicted in both images and also evaluates the confidence that the decision is correct. Results from all sections are combined based on the confidence. The assumption is that if a region is occluded, it is likely that the classification result from that section is inaccurate. It is also likely that the confidence measure from that section is low. Since the result from that section is given little weight, the final result is still correct. In order to handle occlusion which may occur in faces presented in a dictionary, the occlusion extraction technique based on Low-rank matrix decomposition introduced by Chen et al in [12]can be used.

C. Conclusion

In this project we have shown that sparse representation can be used to solve the problem of face identification. The proposed algorithm can perform the task under unrestricted settings. No prior information about persons in test faces is assumed. The number of images used in testing is limited to one image for each person, and there may be images with occlusion in a dictionary due to random selection of dictionary entries. In addition, the proposed algorithm requires low level of human supervision. The trade-off is the accuracy. The performance on the LFW dataset which has very large intra-class variations is 69% while the performance on the PIE dataset is 95%. Although the algorithm does not perform well when head poses variation is large, we have shown that sparse representation is a potential solution to face identification problem.

D. Acknowledgement

We would like to thank Prof. Silvio Savarese and CS231A teaching assistants especially Jiayuan Ma, our project TA, for their advices on this project.

REFERENCES

[1] X. Liu, T. Chen, B.V.K. Vijaya Kumar, "Face Authentication for Multiple Subjects Using Eigenflow". Ac- cepted for publication, Pattern Recognition, special is- sue on Biometrics (2001).

[2] Wright, J.; Yang, A.Y.; Ganesh, A.; Sastry, S.S.; Yi Ma, "Robust Face Recognition via Sparse Representation," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.31, no.2, pp.210,227, Feb. 2009

[3] Mian Zhou; Hong Wei, "Face Verification Using GaborWavelets and AdaBoost," Pattern Recognition, 2006. ICPR 2006. 18th International Conference on , vol.1, no., pp.404,407, 0-0 0

[4] Ben-Yacoub, S.; Abdeljaoued, Y.; Mayoraz, E., "Fusion of face and speech data for person identity verification," Neural Networks, IEEE Transactions on , vol.10, no.5, pp.1065,1074, Sep 1999

[5] Savvides, Marios, BVK Vijaya Kumar, and Pradeep Khosla. "Face verification using correlation filters." 3rd IEEE Automatic Identification Advanced Technologies (2002): 56-61.

[6] Chopra, Sumit, Raia Hadsell, and Yann LeCun. "Learning a similarity metric discriminatively, with application to face verification." Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Vol. 1. IEEE, 2005.

[7] Guillaumin, M.; Verbeek, J.; Schmid, C., "Is that you? Metric learning approaches for face identification," Computer Vision, 2009 IEEE 12th International Conference on , vol., no., pp.498,505, Sept. 29 2009-Oct. 2 2009

[8] D. Donoho, "For Most Large Underdetermined Systems of Linear Equations the Minimal l1-Norm Solution Is Also the Sparsest Solution," Comm. Pure and Applied Math., vol. 59, no. 6, pp. 797- 829, 2006.

[9] E. Cande's and T. Tao, "Near-Optimal Signal Recovery from Random Projections: Universal Encoding Strategies" IEEE Trans. Information Theory, vol. 52, no. 12, pp. 5406-5425, 2006.

[10] Juditsky, A., Kilin?c Karzan, F., Nemirovski, A. (2011), "L1 Minimization via Randomized First Order Algorithms"

[11] Huang, Gary B., et al. "Learning to Align from Scratch." NIPS. 2012.

[12] Chih-Fan Chen; Chia-Po Wei; Wang, Y.-C.F., "Low-rank matrix recovery with structural incoherence for robust face recognition," Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on , vol., no., pp.2618,2625, 16-21 June 2012