IIE Int'l Conference Proceedings

of

International Conference Image Processing, Computers and Industrial Engineering (ICICIE'2014)

Jan. 15-16, 2014

Kuala Lumpur (Malaysia)

Editors:

Dr. Saji Baby

Prof. Dr. P. S. Sandhu

ISBN 978-93-82242- 67-3

Organized By:

International Institute of Engineers

International Institute of Engineers
www.iieng.org

Feature Selection Method based on Sparse Representation Classification for Face Recognition

Yinn Xi Boon, and Sue Inn Ch'ng

***Abstract*—**Compressed sensing is a signal processing technique. The entity signal can be efficiently reconstructed if the sparse representation is determined. The sparse representations of all the test images are determined with respect to the training set by computing the l_1-minimization. However, sparse representation which involves high dimensional feature vector is computationally expensive. Thus, discriminative features that could perform accurately for the face recognition system under visual variations, such as illumination, expression and occlusion have to be selected carefully. In this paper, feature selection method in the application of face recognition based on sparse representation classifier (SRC) is proposed. The proposed technique first divides the images of a few subjects into chunks. Then, it selects the feature subsets based on distance based measurement, the residual, and recognition performance, the accuracy. Extensive experiments with visual variations are carried out by using ORL, AR and Yale databases.

***Keywords*—**Sparse representation, face recognition, compressed sensing, feature selection.

I. Introduction

FACE recognition has been studied extensively in machine learning, computer vision, and biometrics [1]-[3] due to the evolution of human visual system [4], and the applications of face recognition in current technologies.

An imagery data is normally of high dimension, and it is computationally expensive when processing images with high resolution. Also, when there is large scale data, system could often lead to catastrophic interference and the curse of dimensionality. Catastrophic interference here refers to the tendency to forget the previous trained data when the latest data is trained [5] whereas the curse of dimensionality refers to the phenomena occurs in high dimensional space that do not happen in low dimensional settings [6]. On the other hand, organizing and searching data depends on looking for the similar properties within the data. If the data is of high dimensions, the volume of the space will be high as well. As a result, this will cause the available data to be very sparse and dissimilar in many ways, preventing the same group data to be classified efficiently. In addition, the high dimension data will increase the computational cost and the complexity of the backend classifiers. For example the nearest neighbor classifier which considers the distance of all the data points and the neural network where the dimensions would directly increase the number of neurons and the connectionist to process the information within the network.

In pattern classification, the central question is that which are the important and informative features in recognition? Most of the time there are redundant features in the raw data that is irrelevant to the classification. Identifying those features play a key role in object based recognition and pattern classification. There are plenty of efforts in the research of feature extraction techniques where high dimensional features are transform to a lower dimensions, such as the subspace analysis and the manifold learning [7]-[8], and researchers have investigated on frequency domain extraction [9]. All these approaches are aimed to promisingly extract discriminative features for classification. It is noticed that in the presence of variations such as illumination or pose changes, the most dominant features could also degrade the recognition performance [9]. However, it is difficult to determine which feature components are bounded with specific factors [10]. Thus, choosing relevant features that contribute to high recognition rate are then considered as an important prior step to classification.

Sparse representation has been a powerful tool for signal processing applications where the entity signal can be reconstructed based on the sparse signal. It has been successfully developed in image processing applications [11] and recently to face recognition [12]-[16]. Also, sparse representation has the basic principles of working with images with much lower dimensions without significantly compromising the performance of the recognition [15]. The face recognition framework introduced by Wright et al. [13] based on compressed sensing theory have successfully shown that the introduced method could perform accurately and efficiently compared to the traditional methods [7]-[8]. Wright et al.'s idea of utilizing compressed sensing theory, namely Sparse Representation based Classification (SRC) algorithm has shown robust performance by representing the features with sparse representation that is less sensitive to outliers such as the occlusions and expression variations [15]. However, the main disadvantage is that SRC has very high computational

Yinn Xi Boon is with the Sunway University, Department of Computer Science and Networked System. Jalan Universiti, Bandar Sunway, 46150 Petaling Jaya, Selangor. (phone: +60374918622 ext. 3752; e-mail: 13005962@imail.sunway.edu.my).

Sue Inn Ch'ng, is with the Sunway University, Department of Computer Science and Networked System. Jalan Universiti, Bandar Sunway, 46150 Petaling Jaya, Selangor. (e-mail: sueinnc@sunway.edu.my).

cost due to its high dimensional vector representation. The l_1-regularized optimization problem utilized by [13] in SRC involves all the number of the pixels of an image resulting SRC to be computationally expensive.

Thus, in this paper, feature selection based on SRC is proposed. Firstly, the images are divided into a few chunks, and applied to SRC to find the first few chunks of features with highest accuracy and smallest residual. This is done under extensive experiments with different variations using ORL [17], AR [18] and Yale [19] databases for illumination, occlusion and expression variations. After that, the subsets of features that contributed most to the face recognition with different visual variations are selected.

The paper is organized as follows: Section 2 describes the Sparse Representation Classifier (SRC), Section 3 discusses the proposed feature selection method in detail, experiments and results are shown in Section 4, and conclusion in Section 5.

II. Sparse Representation Classifier

Generally in face recognition, it is conceived that there is a face subspace formed by a face image under visual variations. Thus, linear approaches can be used to represent these subspace analysis. The sparse representation based face recognition is basically based on this hypothesis where all the training images are used to span a face subspace [16].

This model is aimed to reconstruct an unknown test image sparsely based on the training datasets, the dictionary. Any new test image for a subject will approximately lie in the linear span of the training samples of the associated subject [16].
For example, for a database with k classes:

$$A = \{A_{1,1}, \cdots, A_{1,n_1}, \cdots, A_{k,1}, \cdots A_{k,n_k}\} \tag{1}$$

Here, $A_{i,l}$ represents the l^{th} image in the class i, and n_i is the number of images in that class. All the images are stacked together to form column vectors $\boldsymbol{v}_{i,l}$ where each column vector represents an image and is denoted as:

$$A = [v_{1,1}, \cdots, v_{1,n_1}, \cdots, v_{k,1}, \cdots, v_{k,n_k}] \in \Re^{L \times N} \tag{2}$$

Here, L is the number of pixel of an image with the size of $h \times w$, and N is the total number of the images for all k classes.

A test image, $\boldsymbol{y}$ is represented using the linear combination of the dictionary:

$$y = Ax_0 \tag{3}$$

For an ideal case, it is assumed that the face subspace of subject i is sufficient to represent a test image, $\boldsymbol{y}$ for subject i, and the coefficients $\mathbf{x}_0$ are in the form of:

$$x_0 = [0, \cdots, 0, \alpha_{i,1}, \alpha_{i,2}, \cdots, \alpha_{i,n_i}, 0, \cdots 0] \tag{4}$$

Where $\mathbf{x}_0$ should have non-zero values at the positions corresponds to all the images with the same subject as the test image, and zero values on the rest of the positions. With this, even when the test image's identity is unknown, sparsity with heuristic principle can be used to solve (3).Thus, an objective is required to measure this sparsity. It is known that the l_1-norm in the compressed sensing theory can produce the sparse solutions [18], hence, l_1-norm optimization is used:

$$\hat{x}_1 = \arg\min_x \|x\|_1 \; s.t. Ax = y \tag{5}$$

Equation (5) is used in the SRC algorithm as shown below:

Algorithm 1. Sparse Representation Classifier by [13]

1. Input: Training images, A, for k classes as shown in (2), a test image $\boldsymbol{y} \in \Re^{L \times 1}$.
2. Normalize all the columns in A to unit l_2-norm.
3. Solve the l_1-norm problem as shown in (5).
4. Compute the per-class residuals
   $r_p(y) = \|y - A\delta_p(x)\|_2$ for $p = 1, \ldots, k$
   For $i = 1, \ldots, N$,
   $$\delta_p(x) = \begin{cases} x_i & IfLabel(i)IsClass(p) \\ 0 & Otherwise \end{cases}$$
5. Output: identity($\boldsymbol{y}$) = $\arg\min_p r_p(y)$

III. Proposed Feature Selection Based on Sparse Representation Classifier

It is noticed from Algorithm 1 that a test image, $\boldsymbol{y}$ is reconstructed from the sparse coefficient $\mathbf{x}_0$, and minimum distance between the test image and the reconstructed image is computed to determine the identity of $\boldsymbol{y}$. Thus, this residual computation plays an important role in the whole classification process, and it is treated as one of the requirements in selecting the feature subsets. Besides, another requirement that is important to the classification is the feature subsets that contribute most to the recognition with high accuracy performances.

Before selecting any feature subset, an image is divided into t chunks with R pixels ($R << L$). Then each chunk is rearranged into a vector form and is treated as the dictionary of (3).

The outputs of all t chunks features for all the testing images are sorted based on the residuals and classification accuracy. The first few chunks that provide minimum residuals with highest accuracy are selected as the features for the recognition.

The performance measurements of each chunk are computed as follows:

Residual = Sum of the Residuals corresponds to its Identity for all the images / Total Number of Images (6)

Accuracy (%) = (Number of Correct Identified Image / Total Number of Images) × 100% (7)

The residual is computed based on the distance between the test image and the reconstructed test image as shown in (3). Smaller residual distance indicates smaller error and thus features with small residuals is one of the requirements in selecting the features.

The algorithm for the proposed feature selection is summarized in Algorithm 2.

Algorithm 2. Proposed Feature Selection based on SRC

1. Input: Training images, $A \in \Re^{L \times N}$, and testing images, $Y \in \Re^{L \times N}$, with *t* chunks for each image.
2. SRC (Algorithm 1) for all the testing images.
3. Summed up $r_p(identity(y))$ and determine the identity for all the testing images for each chunk.
4. Sort summed up *residual* in ascending order and *Accuracy* in descending order for all *t* chunks.
5. Select first *n* chunks with low *residual* and high *Accuracy*.
6. Select these few chunks only for all the training and testing images.
7. Repeat Algorithm 1.

IV. Experimental Results and Discussion

In this section, the performance of the proposed feature selection method for the application of face recognition is evaluated. There are two experiments in this section. The first experiment shows the relationship between the residual and the accuracy for ORL database based on 5 subjects only, while the second experiment shows how the selected features based on the proposed feature selection method vary with the number of chunks chosen for full databases with visual variations.

A. Experimental Setting

For all the experiments, images are close-cropped and well-aligned. They are resized to resolution of 32×32 pixels with bicubic interpolation. After that, all the images of both training and testing sets are divided into 16 chunks due to the reason that each one of the 16 chunks are explicit and informative enough to represent the whole image for the recognition. The division of 16 chunks is shown in Fig. 1.

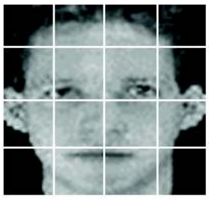

Fig. 1 An image is divided into 16 chunks

The features are selected based on the first 5 subjects from the training and testing sets of each dataset before applying the method to the whole database to reduce the computational time and find the generality of the proposed technique for each dataset. The details of the datasets are explained in Part B.

B. Databases

ORL, AR and Yale databases are used to test the robustness of the proposed method under pose changes (ORL), expression variations (AR and Yale), illumination differences (AR and Yale), and occlusion (AR). The details of each database are shown below.

1. The ORL Database of Faces

This database is used to test on pose changes images. There are 40 subjects, each with 10 different images varies from facial expressions, and pose details. The first five images are used for training and another five images are used for testing.

2. The AR Face Database

This database is used to test on expression variations, illumination changes and occluded images. There are 100 subjects, each subject comes with 26 different images captured in two sessions. The first session has 13 images, and is numbered from 1 to 13, comprising (1-4) different expressions, (5-7) illumination variations, (8-13) occlusions under illumination variations while the second session is having the same conditions as the first session. Four testing sets are carried out to investigate the robustness of the proposed feature selection method. Images 1-7 of session 1 are used for training for all the experiments. The testing images of the four testing sets are stated below:

1. AR_Expression : Images 2-4 of session 2
2. AR_Illumination : Images 5-7 of session 2
3. AR_OcclusionSunglasses : Images 8-10 of session 2
4. AR_OcclusionScarf: Images 11-13 of session 2

3. The Yale face database

This database is used to test on expression and illumination affected images. There are 15 subjects, each with 11 different images with the following conditions; with glasses, without glasses, neutral, center-light, left light, right light, sleepy, happy, sad, wink, and surprised. Two testing sets are carried out to investigate the performance of the proposed feature selection method. The neutral image and the central light image are used for training for both experiments. The conditions of the testing images of the two testing sets are stated below:

1. Yale_Expression : With glasses, without glasses, sleepy, happy, sad, wink, and surprised
2. Yale_Illumination : Left light, and right light

C. Experiment 1: The relationship between the residual and the accuracy

This experiment shows the relationship between the selection criteria – the residual and the accuracy for ORL database based on the first 5 subjects from the training and the testing sets. The residual and the accuracy are computed based on the average of the number of chunks selected. These chunks are selected based on the smallest residual and highest accuracy.

TABLE I
The Residual And The Accuracy Of The Proposed Feature Selection Method For The First 5 Subjects In ORL Database

n (number of chunks)	Residual	Accuracy (%)
2	257.4	90.0
4	243.1	84.0
6	262.6	81.3
8	277.5	77.0
16	354.8	67.6

TABLE II
THE ACCURACY OF THE PROPOSED FEATURE SELECTION METHOD FOR ORL, AR AND YALE DATABASES

Method	*Dimension*	*Pose Variation*	*Expression Variation*		*Illumination Variation*		*Occlusion*	
		ORL	AR_ Expression	Yale_ Expression	AR_ Illumination	Yale_ Illumination	AR_Occlusion Sunglasses	AR_Occlusion Scarf
Proposed Technique (Algorithm 2)	128	39.0	63.3	81.0	66.7	13.3	66.0	62.3
	256	56.0	75.3	80.0	79.3	43.3	61.0	60.7
	384	68.0	73.3	90.5	69.7	46.7	53.7	55.0
	512	77.5	64.0	85.7	73.0	10.0	42.0	47.3
Original Technique [13]	1024	84.5	71.3	85.7	72.3	6.7	19.7	22.0

It is shown that the residual decreases with the increase of the accuracy, this is due to the reason that smaller residual indicates smaller error between both training and testing images, and thus contributes to higher accuracy. This experiment is conducted to the rest of the datasets in the same way.

D. Experiment 2: The performance of the proposed feature selection method with varying number of chunks

In this experiment, the performance in terms of accuracy with different dimensions of the proposed feature selection method is evaluated. Before applying the proposed method to the full database, it is first applied to the first 5 subjects for both training and testing sets in ORL, AR_Expression, AR_Illumination, AR_OcclusionSunglasses, AR_OcclusionScarf, Yale_Expression and Yale_Illumination datasets as shown in Part C to select the most discriminant features. The selected features are then applied to all the subjects for both training and testing sets for each dataset. Table II shows the result for all databases.

E. Discussion

From Table II, the performance decreases with the decrease of the dimensions selected for ORL database. The coefficients of the linear model in (3) do not contribute much to ORL database with pose details that might contain the nonlinear properties. This may cause the selected features could not recognize an identity robustly in ORL database and the accuracy degrades when the dimension is decreased.

On the other hand, for AR_Expression and AR_Illumination datasets, the overall performance for all the dimensions are high, and there is no significant improvement when the proposed feature selection method is applied. This may due to the setting of the datasets which involves the same conditions images for both training and testing sets in this case. The images with expression and illumination variations from the first session are included in the training set for all the AR datasets, and the testing set for AR_Expression and AR_Illumination are having the same conditions as the training set, but the images were from second session (taken in 14 days later). However, it is still noticed that the performance are higher compared to the method from Wright et. al. when the dimension is at 256-d for both testing sets. Even there is no significant improvement, the proposed feature selection method is still able to select the discriminative features which contributes to $\approx 70.6\%$ for both datasets.

For Yale_Expression dataset, even though there are only two normal expression images included in the training set, when a subset of 384-d is selected, the proposed method is still able to increase the accuracy up to 90.5% when the system is tested with seven images with different expressions. Whereas for Yale_Illumination variations, there is a significant improvement of 40% when a subset of 384-d are selected. Also, there is much higher accuracies for reduced dimensions. This is due to the precisely chosen features based on the sparse reprentation that do not include the feature subsets with large variation.

For the AR_OcclusionSunglasses and AR_OcclusionScarf, there is $\approx 40.3\%$ improvement when a subset of 128-d are selected for both datasets. Besides, the performance of the reduced dimensions outperform the method from [13].The sparse representation in (3) has recovered the signal even under occluded images. Thus, with this sparse representation, discriminative and informative features (mostly the features which do not include the variation part) can be selected.

Overall, it can be noticed that with the proposed feature selection method, the images for all the databases with variations have improved their performance in terms of accuracy in reduced dimensions especially when a subset of 256-d or 384-d is selected. Although for AR_Occlusion testing sets, the accuracy is the best when a subset of 128-d is selected, the accuracy at 256-d is still considered robust. However, the system is no longer promising for pose variations images. Since SRC is of linear model, pose variations may more likely to depend on the nonlinear methods such as kernel that would transform the nonlinearity in the face structures to make it linearly separable [15].

The accuracy improved in the reduced dimensions due to the reason that the redundant features has been eliminated, and the proposed feature selection method has selected only the relevant features that is decisive enough for face recognition. There will be large differences for the images with large variability even when they are of same subject. The training images would not cover the whole space of all the possibilities [15], so the distance of the test image and the train images are unlikely to be closed in this case. Thus, SRC based on compressed sensing theory has reconstructed the images even with the small number of measurement with its non-zero values that concentrates on the training images with the same subject as the test image, and this makes the proposed feature selection method based on SRC a robust feature selection technique when its criteria are used to select the relevant feature subsets..

V. CONCLUSION

One of the challenges in face recognition is the large variations that caused by the environments, such as occlusion and expression variations. Thus, discriminative features which could differentiate an image better from the rest of the features which are not only redundant but degrade the classification performance should be carefully selected. By using the proposed feature selection method based on Sparse Representation Classifier, features can be reconstructed even with sparse measurements. The reconstructed features with minimum residual and high accuracy are then selected as the relevant features. From the experiments, the system performance could be further improved with visual variations due to the sparsity that is properly harnessed and the important criteria included when selecting the discriminative features.

REFERENCES

[1] Y. Su, S. Shan, X. L.Chen, and W. Gao, W., "Hierarchical ensemble of global and local classifiers for face recognition," *IEEE Computer Vision*, *11th international conference*, pp. 1-8, Oct 2009.
[2] W.Y. Zhao, R. Chellppa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *in ACM Computing Survey*, vol. 35, issue. 4, pp. 399-258, New York, Dec 2003.
[3] W.C. Zhang, S. Shan, W. Gao, X. L. Chen, "Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation," *IEEE in Computer Vision, 10th IEEE Conference*, vol. 1. Pp. 786-791, Oct 2005.
[4] P. Sinha, B. Balas, Y. Ostrovsky, and R. Rusell, "Face recognition by humans: Nineteen results all computer vision researchers should know about," *Proceeding of the IEEE*, vol. 94, no. 11, pp. 1948-1962, Nov 2006.
[5] R.M. French, "Catastrophic forgetting in connectionist networks," in *Trends in Cognitive Sciences*, vol. 3, No. 4, April 1999.
[6] H. H. Zhang, "Lecture 10: Curse of dimensionality," 2013.
[7] M. Turk, and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71-86, 1991.
[8] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, "Face recognition using Laplacianfaces*," IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328-340, March 2005.
[9] N. Liu, and H. Wang, "Improving predictive accuracy by evolving feature selection for face recognition, in *IEICE Electronics Express*, no. 1-6, 2008.
[10] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Trans. Knowledge Data Engineering*, vol. 17, Issue. 4, pp. 491-502, April 2005.
[11] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Robust sparse coding for face recognition," in *CVPR*, 2011.
[12] J. Wright and Y. Ma, " Dense error correction via l1 minimization," *IEEE Trans. on Information Theory*, pp. 3033-3036, April 2009.
[13] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE in PAMI*, vol. 31, issue. 2, pp. 210-227, Feb 2009.
[14] M. Yang, and L. Zhang, "Gabor feature based sparse representation for face recognition with gabor occlusion dictionary," in *ECCV*, 2010.
[15] A. Y. Yang, J. Wright, Y. Ma, and S. Sastry, "Feature selection in sparse representation: A sparse representation prespective," in *CiteSeerX*, 2007
[16] H. Qiu, D.Pham, S. Venkatesh, J. Lai, and W. Liu, "Innovative sparse representation algorithms for robust face recognition," in *ICIC*, vol. 7, no. 10, pp. 5645-5667, Oct 2011.
[17] The ORL database of faces, AT&T Laboratories Cambridge.
[18] A. Martinez, and R. Benavente, "The AR face database," CVC Technical Report No. 24, Tech. Rep., 1998.
[19] The yale face database, CVC, 1997.