FACE RECOGNITION FROM VIDEO: AN MMV RECOVERY APPROACH

A. Majumdar and R. K. Ward

Department of Electrical and Computer Engineering, University of British Columbia
{angshulm, rababw}@ece.ubc.ca

ABSTRACT

In this paper we propose a new approach to video based face recognition. Our work is based on the Sparse Classification approach which assumes that each test sample can be formed by a linear combination of the training samples of the correct class. Based on this assumption, we formulate the classification problem as one of joint sparse recovery of Multiple Measurement Vectors (MMV). This requires solving an NP hard problem. This problem has not been solved earlier; thus we derive an algorithm for solving it. The experimental evaluation is carried on the VidTIMIT database. The proposed method is compared against an HMM based method for video based face recognition and the modified Sparse Classification method. The results show that the proposed method outperforms both these methods.

Index Terms— Face recognition, joint sparsity, hard thresholding.

1. INTRODUCTION

In video based face recognition, a video sequence of each subject is collected for training. The problem is to identify the person given a test sequence. Such a situation can arise in customer identification in bank ATMs [1]. When a new customer opens a new account in the bank, a training video sequence is obtained of this person. When the same person visits an ATM at a later date, a video camera is used to shoot a test sequence. This sequence is matched against the training sequence. If the match is a success, the customer is allowed to proceed with the transaction.

In this work we focus on frontal face recognition. We extend the face recognition via the sparse classification approach [2] to a video based face recognition problem. The aforesaid work [2] addresses the problem of face recognition from 2D images. The basic assumption is that, a test sample can be formed by a linear combination of the training samples of the correct class [2].

For single images, the classification problem was formulated as an l_1-norm regularized least squares problem. The classification approach devised in [2] yielded good face recognition results. The success of this approach led to further research in this area [3, 4]. This approach is called the Sparse Classification (SC).

In [5, 6] we show that the l_1-norm regularization is not an ideal choice to classify images based on the said assumption as there are some theoretical problems associated with it. Given the assumption in [2], we have experimentally shown that slightly better results can be obtained by our Group Sparse Classifier (GSC) [5, 6].

In this work, we address the problem of face recognition from video sequences. Thus, the test sample consists of a series of frame images instead of a single image. In this work, we will show how the Sparse Classification approach can be extended to the video face recognition problem. We will formulate the classification as a Multiple Measurement Vector (MMV) recovery problem.

The MMV recovery problem requires solving an NP hard problem. Contrary to previous works [2-6], where the ideal NP hard optimization problems are approximated by their convex surrogates, in this work, we derive an algorithm to approximately solve the NP hard problem directly.

Our focus is on frontal face recognition; hence the experiments were carried out on the ViDTIMIT database [7].The Leave-One-Out (LOO) strategy was used for experimental evaluation. We compared our proposed solution with two state-of-the-art approaches [2, 8]. Our method shows considerable improvement over both.

The rest of the paper is organized into several sections. The background of the SC approach is given in Section 2. Section 3 formulates the video based face recognition problem as one of MMV recovery. The algorithm for solving the NP hard problem (arising out of our MMV recovery formulation) will be derived in Section 4. Section 5 shows the experimental results. Finally the conclusions of the work are drawn in Section 6.

2. SPARSE CLASSIFICATION

The Sparse Classification approach was first introduced in [2]. It is assumed that the new test sample of a particular class can be expressed as a linear combination of the training samples belonging to that class. For example if the test sample belongs to class k, then

$$v_{test} = \alpha_{k,1} v_{k,1} + ... + \alpha_{k,n} v_{k,n} \quad (1)$$

where $v_{k,i}$ represents the ith sample of the kth class, v_{test} is the test sample (assumed to be in the kth class) and $\alpha_{k,i}$ is a linear weight.

Equation (1) represents the test sample by the training samples of the correct class only. It can also be represented in terms of training samples of all classes (assuming there are c classes) as

$$v_{test} = \alpha_{1,1}v_{1,1} + ... + \alpha_{1,n}v_{1,n} + ... + \alpha_{k,1}v_{k,1} + ... + \alpha_{k,n}v_{k,n} + ... + \alpha_{c,1}v_{c,1} + ... + \alpha_{c,n}v_{c,n} \quad (2)$$

In a concise matrix-vector notation (2) is expressed as,

$$v_{test} = V\alpha \quad (3)$$

$$V = \left[\underbrace{v_{1,1} \mid .. \mid v_{1,n}}_{V_1} \mid .. \underbrace{v_{c,1} \mid .. \mid v_{c,n}}_{V_c} \right], \alpha = \left[\underbrace{\alpha_{1,1}, .. \alpha_{1,n}}_{\alpha_1}, .., \underbrace{\alpha_{c,1}, .. \alpha_{c,n}}_{\alpha_c} \right]^T$$

The test sample (v_{test}) is known, and the matrix formed by stacking the training samples as columns (V) is also known. The linear weights vector (α) is unknown. In [2], the first step towards classification is the computation of the linear weights by solving the inverse problem (3). According to the assumption in [2], the vector α will be sparse, i.e. it will have zeroes everywhere except for α_k, i.e. non-zero values corresponding to the correct class (assumed to be k). Ideally, the following l_0-norm regularized least squares problem should be solved in order to recover α,

$$\hat{\alpha} = \min_{\alpha} \|v_{test} - V\alpha\|_2^2 + \lambda \|\alpha\|_0 \quad (4)$$

This is an NP hard problem. Following recent work in Compressed Sensing, [2] proposed the sparse classification approach (SC) where the NP hard l_0-norm is replaced by its tightest convex envelope, the l_1-norm. Thus the following problem is proposed instead,

$$\hat{\alpha} = \min_{\alpha} \|v_{test} - V\alpha\|_2^2 + \lambda \|\alpha\|_1 \quad (5)$$

Solving α is the first step in the SC approach. In the next step, the residual for each class is computed as follows,

$$res(i) = \|v_{test} - V_i\alpha_i\|_2, \forall i \in \{1, c\} \quad (6)$$

The test sample is assigned to the class having the lowest residual.

The term $V_i\alpha_i$ is the representative sample for the i^{th} class. The assumption is that, for the correct class (k), the representative sample will be similar to the test sample, and therefore the residual error will be the least.

3. FACE RECOGNITION FROM VIDEO: PROPOSED SOLUTION

In this work, it is assumed that there is a single training video sequence available for each person. This is a realistic assumption, since in practical situations, e.g. customer authentication in banks, the training sequence will be comprised of only one video sequence.

Each frame of the video sequence is an image that will be considered as a sample. When all the training samples are stacked as columns, the matrix V is the same as in (3). But instead of a single test sample, $\hat{v}_{test}$ will be comprised of n frames, i.e. $\hat{v}_{test} = \left[v_{test}^{(1)} \mid ... \mid v_{test}^{(n)} \right]$. Extending the assumption in [2], each frame of the test sequence is assumed to be a linear combination of the training frames i.e.

$$v_{test}^{(j)} = V\alpha_k, \forall j \in \{1, n\} \quad (7)$$

Considering all the $v_{test}^{(j)}$ in compact matrix-vector notation, (7) can be expressed as the following Multiple Measurement Vector (MMV) formulation,

$$\hat{v}_{test} = V\hat{\alpha} \quad (8)$$

where $\hat{\alpha} = \left[\alpha^{(1)} \mid ... \mid \alpha^{(n)} \right]$.

According to the assumption of SC, each of the $\alpha^{(i)}$'s will be sparse, i.e. they will have non-zero values only for the correct class. Therefore, the matrix $\hat{\alpha}$ will be row sparse, i.e. will it will have non-zero values on rows that correspond to the correct class and zeros elsewhere.

Recent works in signal processing [9-11] have shown that it is possible to solve such row sparse MMV problems by the following optimization problem,

$$\min_{\hat{\alpha}} \|\hat{v}_{test} - V\hat{\alpha}\|_F^2 + \lambda \|\hat{\alpha}\|_{2,0} \quad (9)$$

where $\|.\|_F^2$ denotes the square of Frobenius' norm and $\|\hat{\alpha}\|_{2,0} = \sum_{i=1}^{C} I(\|\hat{\alpha}_2\| > 0)$ and $I(\|\hat{\alpha}_2\| > 0) = 1$, iff $\|\hat{\alpha}_2\| > 0$.

The inner l_2-norm on the rows favors a solution that has non-zero coefficients along a row; the outer l_0-norm enforces sparsity on the number of non-zero rows. The optimization problem (9) is NP hard, and there is no algorithm to solve it even approximately. For the first time, in this work, we derive a modified iterative hard thresholding algorithm to solve it. The derivation is given in Section 4.

Once $\hat{\alpha}$ is solved, finding the class of the training sequence proceeds similar to [2]. The residual error is computed for each class,

$$res(i) = \|\hat{v}_{test} - V_i\hat{\alpha}_i\|_2, \forall i \in \{1, c\} \quad (10)$$

The class with the lowest residual error is assumed to be the class of the training sample.

4. DERIVATION OF ALGORITHM

For ease of writing, we change the notations in (9) and write it as follows,

$$\min_X J(X) : J(X) = \|Y - HX\|_F^2 + \lambda \|X\|_{2,0} \quad (11)$$

To solve the optimization problem (11), we follow the majorisation minimization (MM) approach [12]. The general MM approach is as follows:

Let *J(x)* be the (scalar) function to be minimized

1. Set *k=0* and initialize x_0.

Repeat step 2-4 until a suitable stopping criterion is met.

2. Choose $G^{(k)}(x)$ such that
   a. $G^{(k)}(x) \geq J(x)$ for all x
   b. $G^{(k)}(x_k) = J(x_k)$.
3. Set x_{k+1} as the minimizer for $G^{(k)}(x)$.

4. Set $k=k+1$, go to step 2.

Problem (11) does not have a closed form solution and therefore must be solved iteratively. At each iteration (k), we choose

$$G^{(k)}(x) = \| Y - HX \|_F^2 + (X - X^{(k)})^t (\alpha I - H^T H)(X - X^{(k)}) + \lambda \| X \|_{2,0}$$

$G^{(k)}(x)$ satisfies the condition for MM when α is greater than the maximum eigenvalue of H^TH. This guarantees stability of the algorithm. $G^{(k)}(x)$ can alternatively be expressed as,

$$G^{(k)}(x) = \alpha \| X^{(k)} + \frac{1}{\alpha} H^T (Y - HX) - X \|_2^2 + \lambda \| X \|_{2,0} + K$$

where K consists of terms independent of X.
Minimizing $G^{(k)}(x)$ is the same as minimizing the following,

$$\tilde{G}_1^{(k)}(X) = \frac{1}{2} \left\| B^{(k)} - X \right\|_F^2 + \frac{\lambda}{\alpha} \| X \|_{2,0} \qquad (12)$$

where $B^{(k)} = X^{(k)} + \frac{1}{\alpha} H^T (Y - HX^{(k)})$.

For minimizing (12) we follow the approach in [13, 14]. Each of the rows of X (denoted by $X^{r\rightarrow}$) are independent from each other. Therefore, the minimum of equation (12) can be calculated by minimizing with respect to each $X^{r\rightarrow}$ individually. To derive the minimum, we distinguish between the two cases, $\left\| X^{r\rightarrow} \right\|_2 = 0$ and $\left\| X^{r\rightarrow} \right\|_2 \neq 0$. In the first case, the row-wise cost for (12) is λ/α. In the second case the cost is $\left\| X^{r\rightarrow} \right\|_2^2 - 2X^{r\rightarrow} \bullet B^{(k)r\rightarrow}$. The minimum of which is attained at $X^{r\rightarrow} = B^{(k)r\rightarrow}$.

Comparing the cost in both cases ($\left\| X^{r\rightarrow} \right\|_2 = 0$ and $\left\| X^{r\rightarrow} \right\|_2 \neq 0$), we see that the minimum of (12) is attained when,

$$X^{(k+1)r\rightarrow} = \begin{cases} 0, & B^{(k+1)r\rightarrow} \leq \frac{\lambda}{\alpha} \\ B^{(k+1)r\rightarrow}, & B^{(k+1)r\rightarrow} > \frac{\lambda}{\alpha} \end{cases}, \forall \text{ rows (r)} \qquad (13)$$

This update is actually a modified version of the iterative hard thresholding algorithm [14].

Equations (12) and (13) suggest a compact solution for (11). This is given by the following algorithm.

Initialize: $X^{(0)} = 0$
Repeat until convergence:

$$B^{(i)} = X^{(i)} + \frac{1}{\alpha} H^T (Y - HX^{(i)})$$

$$X^{(k+1)r\rightarrow} = \begin{cases} 0, & B^{(k+1)r\rightarrow} \leq \frac{\lambda}{\alpha} \\ B^{(k+1)r\rightarrow}, & B^{(k+1)r\rightarrow} > \frac{\lambda}{\alpha} \end{cases}, \forall \text{ rows (r)}$$

4. EXPERIMANTAL RESULTS

Since our work focuses on frontal face recognition from video, we choose to use the VidTIMIT [7] database which is designed for recognition of human faces from frontal views. The VidTIMIT dataset is comprised of videos and their corresponding audio recordings for 43 people, reciting short sentences. For each person there are 13 sequences; 3 sequences contain head movements (no audio) while 10 sequences contain frontal views reciting short sentences. The recording was done in an office environment using a broadcast quality digital video camera. The video of each person is stored as a numbered sequence of JPEG images with a resolution of 512x384 pixels. A quality setting of 90% was used during the creation of the JPEG frame images.

In this work, we work with the 10 sequences containing frontal faces. Leave-One-Out cross validation (LOO) is used for evaluation. For each person, a single sequence is used for training and the remaining 9 sequences are used for testing.

We compare our proposed face recognition technique with two methods – i) Sparse Classification [2] and ii) Hidden Markov Model [8]. The Sparse Classification (SC) method was actually developed for face recognition from images (not videos) But, we have modified it for video based recognition. For each frame image of the test sequence, the SC optimization problem is solved,

$$\hat{\alpha}^{(j)} = \min_{\alpha^{(j)}} \left\| v_{test} - V\alpha^{(j)} \right\|_2^2 + \lambda \left\| \alpha^{(j)} \right\|_1$$

This optimization problem is solved via the Iterative Soft Thresholding (IST) algorithm [17]. All the $\hat{\alpha}^{(j)}$'s are stacked as columns of the matrix $\hat{\alpha}$ as in [7]. The residual error for each class is computed as:

$$res(i) = \left\| \hat{v}_{test} - V_i \hat{\alpha}_i^T \right\|_2, \forall i \in \{1, c\}$$

The test video sequence is assigned to the class having the minimum residual error. We have named this method as Modified Sparse Classification (MSC).

The difference between our proposed approach and the repeated use of the MSC on each frame of a test sequence lies in the optimization algorithm used to recover the sparse linear coefficients. The SC recovers the weights for each frame image individually, where as our method recovers the weights simultaneously for all the frames using an MMV formulation. For the theoretical differences between these two recovery approaches, the reader is referred to [10].

The details of the HMM based method are found in [8]. This method trains one HMM for each training video sequence. During testing, the likelihood score for the test sequence is computed for the HMM's trained for each class. The test sequence is assigned to the class having the maximum likelihood. The only parameter that needs to be chosen by the user is the number of hidden states. It is known that increasing the number of hidden states improves the results, but on the other hand it requires more samples

for estimation. For this work, it was found that the best results were obtained for 20 hidden states.

The original images are of very high dimensions. Dimensionality reduction has been an active area of research in face recognition for the past two decades. However, since dimensionality reduction is not the focus of our research, we employ the most often used dimensionality reduction method which is the Eigenface [15]. The Eigenspace projection from higher to lower dimension is computed from the training set. This projection is used for dimensionality reduction of both the training and the testing samples.

In Table 1, the recognition rates of the three different methods are shown; they are – i) HMM, ii) Modified SC (MSC) and iii) the Proposed. The results are shown for different lower dimensional Eigenface projections. The average recognition rates (for LOO cross validation) are shown in the following table.

Table 1: Recognition Rates in %

Method	Number of Eigenfaces			
	20	40	60	80
HMM [9]	70.12	78.93	83.40	84.41
MSC	75.41	85.76	92.96	94.49
Proposed	78.04	90.01	94.55	97.28

The results show that both MSC and our proposed approach yield considerably better results than the HMM based technique [8]. Our proposed approach yields the best results. Our method and MSC have similar assumptions, but the methods used to recover the sparse linear coefficients are different. The theoretical difference between the two recovery approaches is discussed in [9], where it is said that, when applicable, joint MMV recovery yields better results than recovering the sparse vectors individually. The other difference between MSC and our proposed approach is in the nature of formulation of the optimization problems. MSC approximates the NP hard problem by its convex surrogate, whereas our proposed method directly solves for the NP hard problem (but only approximately).

5. CONCLUSION

A novel classification approach [2] assumes that any new test sample can be expressed as a linear combination of existing training samples belonging to the same class of the test sample. Based on this assumption, it was shown that the classification problem can be formulated as a sparse optimization problem. Their classifier was duly named the Sparse Classifier (SC).

In this paper, we address the problem of video based frontal face recognition. One can address this problem by modifying the SC and repeatedly applying it on the individual frames of the test video sequence. But this does not yield the best results as has been shown in this paper. We have formulated the video based face recognition problem as one of joint sparse MMV recovery. It yields better recognition results than the Modified SC method on the VidTIMIT database. We have also compared our work against an HMM based technique [8] for face recognition from video. The Modified SC and our proposed method yield better results than the HMM based method.

ACKNOWLEDGEMENT

This work was supported by NSERC, the Natural Sciences and Engineering Research Council of Canada and by QNRF, Qatar National Research Fund.

5. REFERENCES

[1] A. Majumdar and P. Nasiopoulos, “Frontal Face Recognition from Video”, International Symposium on Visual Computing, pp. 297–306, 2008.

[2] Y. Yang, J. Wright, Y. Ma and S. S. Sastry, “Feature Selection in Face Recognition: A Sparse Representation Perspective”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 1 (2), pp. 210-227, 2009.

[3] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, and Y. Ma, “Towards a Practical Face Recognition System: Robust Registration and Illumination via Sparse Representation”, IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[4] A. Yang, A. Ganesh, S. Sastry and Y. Ma, “Fast L1-Minimization Algorithms and an Application in Robust Face Recognition: A Review”, IEEE International Conference on Image Processing, 2010.

[5] A. Majumdar and R. K. Ward, “Classification via Group Sparsity Promoting Regularization”, IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 873-876, 2009.

[6] A. Majumdar and R. K. Ward, “Robust Classifiers for Data Reduced via Random Projections”, IEEE Trans. SMC B, Vol. 40 (5), pp. 1359 – 1371.

[7] http://itee.uq.edu.au/~conrad/vidtimit/

[8] L. Xiaoming, and C. Tsuhan, “Video-based face recognition using adaptive hidden Markov models”, IEEE CVPR, vol. 1, pp. I-340–I-345 (2003).

[9] E. van den Berg and M. P. Friedlander, “Theoretical and empirical results for recovery from multiple measurements”, IEEE Trans. Info. Theory, Vol. 56 (5), pp. 2516-2527, 2010.

[10] S. F. Cotter, B. D. Rao, K. Engang, and K. Kreutz-Delgado, “Sparse solutions to linear inverse problems with multiple measurement vectors”. IEEE Trans. Sig. Proc., Vol. 53 (7), pp. 2477-2488, 2005.

[11] J. Chen and X. Huo, “Theoretical results on sparse represenations of multiple-measurement vectors”, IEEE Trans. Sig. Proc., Vol. 54 (12), pp. 4634-4643, 2006.

[12] http://cnx.org/content/m32168/latest/

[13] T. Blumensath, M. Yaghoobi and M. E. Davies, “Iterative Hard Thresholding and L_0 Regularisation”, IEEE ICIP, pp. 877-880, 2007.

[14] T. Blumensath, and M. E. Davies, “Iterative Thresholding for Sparse Approximations”, Journal of Fourier Analysis and Applications, Vol. 14 (5), pp. 629-654, 2008.

[15] M. Turk and A. Pentland, “Eigenfaces for Recognition”, Journal of Cognitive Neuroscience, Vol.3 (1), pp.71-86, 1991.