PARTIAL FACE RECOGNITION: A SPARSE REPRESENTATION-BASED APPROACH

Luoluo Liu, Trac D. Tran, and Sang "Peter" Chin

Dept. of Electrical and Computer Engineering, Johns Hopkins Univ., Baltimore, MD 21218, USA
{lliu69,trac,schin11}@jhu.edu

ABSTRACT

Partial face recognition is a problem that often arises in practical settings and applications. We propose a sparse representation-based algorithm for this problem. Our method firstly trains a dictionary and the classifier parameters in a supervised dictionary learning framework and then aligns the partially observed test image and seeks for the sparse representation with respect to the training data alternatively to obtain its label. We also analyze the performance limit of sparse representation-based classification algorithms on partial observations. Finally, face recognition experiments on the popular AR data-set are conducted to validate the effectiveness of the proposed method.

Index Terms— Face Recognition, Image Alignment, Sparse Representation , Dictionary Learning

1. INTRODUCTION

The face recognition task attracts much attention because of its practical use for safety and surveillance purposes as well as the potential to understand how human-beings identify different people. Various approaches have been proposed. Among them, sparse representation-based classification (SRC) pioneered by J. Wright *et al.* is one of the simplest methods but still provides state-of-the-art performance [1]. The authors showed, by using an over-complete dictionary, sparse representation can be treated as a high-dimensional representation and classification is preformed on top of that. However, as with almost all of other current face recognition systems, sparse representation-based method relies highly on the success of solving sparse coding, which requires accurate alignment between training data and testing data.

In practice, much of test data are collected in uncontrolled conditions with the possibility of severe occlusion or missing information. This leads to the partial face recognition problem. In these settings, one of the most common issues is that the partial test data obviously do not align well with the training data. One may pre-process the data to achieve the same alignment. A popular approach would be detecting local features and align data by normalizing the detected features geometrically. However, this approach will fail whenever feature detection fails, which is highly likely since given partial data usually lacks these features. Another solution is accounting for every possible translations via a super-redundant dictionary, which certainly would result in exorbitant computational/memory cost.

Some of partial face recognition algorithms are based on finding transform-invariant features such as Scale-invariant feature transform (SIFT) in MKD-SRC [2], Gabor Ternary Pattern (GTP) [3] , pooled sparse codes from SIFT [4] or local patches [5], etc. Others recover the transformation and solve the sparse code simultaneously as in [6], [7], [8].

Reported success of local-feature-based holistic face recognition algorithms implies that the possibility of recognition based on a few discriminative local features. Moreover, Inspired by human-beings' ability to identify face relying on partial observations, another motivation for us to study the partial face recognition problem is the intriguing question: given limited observations, how much information is enough to reliably identify a person?

In this paper, we propose an approach to solve the aforementioned problem by alternatively finding the alignment and sparse coding. Main contributions are: *(i)* Our proposed framework make recognition task with partial observations efficiently solvable with different aligning conditions in supervised dictionary learning (SDL) framework [9]; *(ii)* Our method can be easily extend to partial observations almost in the form of any shape although for simplicity we demonstrate the one patch case; *(iii)* Sufficient conditions that lead to correctly classification on partial observations are explored.

2. METHOD

The problem is formulated as follows: given well-aligned holistic faces with labels as training data, partial face recognition task aims at classifying partially observed faces without any alignment information. Under supervised dictionary learning framework, we propose an approach that alternatively finds the sparse code as well as the best possible alignment.

We gratefully acknowledge that this work was in part supported by NSF (NSF-DMS-1222567, NSF-CCF-1117545, NSF-CCF-1422995, NSF-EECS-1443936) and AFOSR (FA9550-12-1-0136).

Notations:

Let $\mathcal{T}_1, \mathcal{T}_2$ be the sets of training samples, and the testing samples respectively, where $\forall\ \mathbf{y}_i \in \mathcal{T}_1, \mathbf{y}_i \in \mathbb{R}^m$ and $\mathbf{y}_t \in \mathcal{T}_2$ are not necessarily of the same size. Also, $\forall\ k \in \mathcal{N}^+, [k] := \{1, 2, .., k\}$. N denotes the number of different classes in the training data set. Let l_i be the label associated with $\mathbf{y}_i \in \mathcal{T}_1$, then and $l_i \in \mathcal{L} := [N]$. $\mathbf{K}$ is the label matrix of $\mathcal{T}_1$. $\mathbf{D} \in \mathbb{R}^{m\times n}$ denotes the dictionary and $\mathbf{x} \in \mathbb{R}^n$ represents sparse code and $\mathbf{W}$ provides the parameter of the classifier. Let $f(\mathbf{x}, \mathbf{W})$ represents the model of the linear classifier: $f(\mathbf{x}, \mathbf{W}) = \mathbf{W}\mathbf{x} + \mathbf{b}$ [10].

We define a partial observation operator $(\cdot)_\Lambda, \Lambda \subseteq [m]$ as follows. Let $\mathbf{I}_\Lambda$ denote sub rows of identity matrix with row indices in the set Λ. Then for $\mathbf{y} \in \mathbb{R}^m$, $(\mathbf{y})_\Lambda := \mathbf{I}_\Lambda \mathbf{y}$. In short, Λ contains the support of image pixels that we are able to observe whereas its complement Λ^c indicates the support of missing information. For $\mathbf{D} \in \mathbb{R}^{m\times n}, \mathbf{D}_\Lambda := \mathbf{I}_\Lambda \mathbf{D}$ simply selects the corresponding rows in the observation set.

2.1. Supervised Dictionary Learning and Classification

2.1.1. Supervised Dictionary Learning

A classical approach to seek the dictionary that yields sparse representation is given as [11] [12]:

$$\min_{\mathbf{D},\mathbf{x}} (1/2) \sum_{i\in\mathcal{T}_1} \|\mathbf{y}_i - \mathbf{D}\mathbf{x}_i\|_2^2 + \rho\ \text{sparsity}(\mathbf{x}_i),\ \rho > 0, \quad (1)$$

where sparsity$(\cdot)$ denotes a regularizer that encourages sparsity where common choices are ℓ_0, ℓ_1 norms. Let $R_s(l_i, f(\mathbf{x}, \mathbf{W}))$ indicates the classification loss. By minimizing classification loss and data fidelity $\|\mathbf{y}_i - \mathbf{D}\mathbf{x}_i\|_2^2$, the supervised dictionary learning framework can be described as follows:

$$\begin{aligned}(\mathbf{D}_\star, \mathbf{x}_\star, \mathbf{W}_\star) = \arg\min_{\mathbf{D},\mathbf{x},\mathbf{W}} \sum_{i\in\mathcal{T}_1} R_s(l_i, f(\mathbf{x}_i(\mathbf{y}_i, \mathbf{D}), \mathbf{W})) \\ +\beta\|\mathbf{y}_i - \mathbf{D}\mathbf{x}_i\|_2^2 + \rho\ \text{sparsity}(\mathbf{x}_i),\ \beta > 0.\end{aligned} \quad (2)$$

For solving sparse coding, ℓ_0 minimization can be solved by Iterative Hard Thresholding (IHT) [13] whereas Orthogonal Mathching Pursuit (OMP) [14] for corresponding constraint minimization and the ℓ_1 minimization can be solved via Least Absolute Shrinkage and Selection Operator [15], or Gradient Projection for Sparse Reconstruction [16].

2.1.2. Classification

If test data $\mathbf{y}_t \in \mathcal{T}_2$ has the same alignment condition as the training data, then $\mathbf{x}_\star$ can be solved by traditional sparse coding. The technique of solving $\mathbf{x}_\star$ for partial data will be presented in Section 2.2. After obtaining $\mathbf{x}_\star$, the label assignment bases on the class that yields the minimum classification loss among all classes:

$$\hat{c}(\mathbf{y}_t) = \arg\min_{l_i\in\mathcal{L}} R_s(l_i, f(\mathbf{x}_\star(\mathbf{y}_t, \mathbf{D}_\star), \mathbf{W}_\star)). \quad (3)$$

2.2. Alternating Alignment and Sparse Coding

Given partially observed test data $\mathbf{y}_t$, let $\mathbf{y}^h$ represent the corresponding holistic face, which is unknown (and in this scenario, the recovery is not necessary). We would like to solve for sparse code of the test data with partial constraint on the unknown observed set Λ. We know that $(\mathbf{D}_\star\mathbf{x})_\Lambda = \mathbf{I}_\Lambda(\mathbf{D}_\star\mathbf{x}) = (\mathbf{I}_\Lambda\mathbf{D}_\star)\mathbf{x} = \mathbf{D}_{\star\Lambda}\mathbf{x}$ and $\mathbf{y}_t = \mathbf{y}^h_\Lambda$. Considering minimizing the data fidelity term on Λ: $\|(\mathbf{y}^h - \mathbf{D}_\star\mathbf{x})_\Lambda\|_2^2 = \|\mathbf{y}^h_\Lambda - \mathbf{D}_{\star\Lambda}\mathbf{x}\|_2^2$, then the problem can be recast as:

$$(\mathbf{x}_\star, \Lambda_\star) = \arg\min_{\alpha\in\mathbb{R}^n, \Lambda_i\in\mathcal{S}} (1/2)\|\mathbf{y}_t - \mathbf{D}_{\star\Lambda}\mathbf{x}\|_2^2 + \rho\ \text{sparsity}(\mathbf{x}), \quad (4)$$

where $\mathcal{S}$ is the subset of all candidates that Λ could take from. (4) is solved by Alternating Minimization. Details are depicted in Algorithm 1.

However, for a learnt dictionary $\mathbf{D}_\star$, each column no longer has the physical meaning as a vector in the training data-set, so as the alignment. Updating alignment Λ seeks a optimal partial set resulting smaller representation error and sparsity level with respect to the optimal dictionary.

Algorithm 1 Partial Face Recognition based on Alternating Alignment and Sparse Coding(AASP)

1: **Require:** Training set, label set $\mathcal{T}_1, \mathbf{K}$; Testing set $\mathcal{T}_2$
2: Supervised Dictionary Learning:
   Input: $\mathcal{T}_1, \mathbf{K}$, **Output:** $(\mathbf{D}_\star, \mathbf{W}_\star)$.
3: **for** $i = 1, 2, ..|\mathcal{S}|$ **do**
4: $\mathbf{x}_{i\star} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} (1/2)\|\mathbf{y}_t - \mathbf{D}_{\star\Lambda_i}\mathbf{x}\|_2^2 + \rho\ \text{sparsity}(\mathbf{x})$,
5: **end for**
6: $\Lambda_\star = \arg\min_{\Lambda_i\in\mathcal{S}} (1/2)\|\mathbf{y}_t - \mathbf{D}_{\star\Lambda_i}\mathbf{x}_{i\star}\|_2^2 + \rho\ \text{sparsity}(\mathbf{x}_{i\star})$
7: $\mathbf{x}_\star = \arg\min_{\mathbf{x}\in\mathbb{R}^n} (1/2)\|\mathbf{y}_t - \mathbf{D}_{\star\Lambda_\star}\mathbf{x}\|_2^2 + \rho\ \text{sparsity}(\mathbf{x})$
8: **Return:** $c(\mathbf{y}_t)$: classification by (3).

2.3. Implementation Details

In our implementation, we use linear prediction with zero offset ($\mathbf{b} = \mathbf{0}$) as classifier and the squared loss as the loss function R_s. The label is represented by their binary expression $\mathbf{q}$. For $l_i = j, j \in [N]$, then $\mathbf{q}_i = \mathbf{e}_j^T = [0, 0, ..1, 0, .., 0]^T$. Therefore, $R_s(l_i, f(\mathbf{x}, \mathbf{W})) = ||\mathbf{q}_i - \mathbf{W}\mathbf{x}_i||_2^2$.

In testing, we demonstrate a simple approach to find the Λ in step 6 of Algorithm 1. A more complicated solution which extensively searching alignment uses optical flow is given in [8]. Here, Λ is initialized via mean face matching. The mean face $\mathbf{y}_m$ of the training set is calculated by taking the average of training samples. In the case when partially observed data is connected and rectangular, the partial observation operator Λ can be uniquely determined by coordinates of the top left pixel (p_i, p_j) and its width b and height h. Let B and H be the width and height of the holistic image respectively;

$\mathcal{S}_\Omega$ contains all operators in the form of $\Lambda(p_1, p_2, b, h)$ that selects a patch same size as the test, then mean face matching is simply:

$$\Lambda^\star_{\text{mean}}(p_a, p_b, b, h) = \arg\min_{\Lambda \in \mathcal{S}} \quad \|\mathbf{y}_t - (\mathbf{y}_m)_\Lambda\|^2. \quad (5)$$

To reduce the computational complexity, we restrict the possible region from $\mathcal{S}_\Omega$ to $\mathcal{S}$, which includes be all operators gives partial observations that are within a local neighbourhood of c_n pixels of $(\mathbf{y})_{\Lambda^\star_{\text{mean}}}$.

3. ANALYSIS

In general, partially observed data cannot contain more information than entirely observed data. Unlike the scenario in matrix completion with randomly missing observations, the observed set is highly spatially coherent. It is not possible to recover the entire data with high accuracy and then preform classification. However, in practice, successful classification based on local discriminative features shows the possibility of accurate classification based on partial data.

3.1. Data Model

In the sparse representation based classification framework, we assume that the data model is $\mathbf{y} = \mathbf{D}_\star \mathbf{x} + \mathbf{z}$, where $\mathbf{y}, \mathbf{z} \in \mathbb{R}^m, \mathbf{x} \in \mathbb{R}^n$ and $\mathbf{x}$ is s-sparse, $s < n$ and the sparse representation $\mathbf{x}$ for data of different classes are quite distinctive to each other. Hence we might expect to find good representation $\mathbf{D}_\star$ along with a reasonable classifier parametrized by $\mathbf{W}_\star$ from the learning process. We also assume that there exists a recovery algorithm that can perfectly recover the true support set $\Lambda_\star$ of the partial data. To simplify the notations, we drop the subscript "$\star$" from $\mathbf{D}_\star, \mathbf{W}_\star, \Lambda_\star$ in this section.

Let us assume that we get the optimal dictionary $\mathbf{D}$, and test data is correctly classified in solving $\mathbf{y} = \mathbf{D}\mathbf{x}$. Now the question becomes: does it exist a partial region $\Lambda \subset [m]$ such that solving $\mathbf{x}$ via $\mathbf{y}_\Lambda = \mathbf{D}_\Lambda \mathbf{x}$ still yields the same correct classification result?

3.2. The Noiseless Case

Remark 1 *$\|\mathbf{z}\| = 0$, for $\Lambda \in [m]$, $|\Lambda| =$ rank$(\mathbf{D})$, and rows of $\mathbf{D}_\Lambda$ are linearly independent, then solving $\mathbf{y}_\Lambda = \mathbf{D}_\Lambda \mathbf{x}$ is equivalent to solving $\mathbf{y} = \mathbf{D}\mathbf{x}$.*

In solving $\mathbf{y} = \mathbf{D}\mathbf{x}$, only rank$(\mathbf{D})$ number of equations is needed. For $m > n$, rank$(\mathbf{D}) \leq n$, at most n features are needed for solving $\mathbf{y} = \mathbf{D}\mathbf{x}$. Λ can be computed by finding linearly independent rows in $\mathbf{D}$.

When $\mathbf{x}$ is fairly sparse, compressive sensing theory [17] indicates that the number of measurements needed is $\mathcal{O}(s \log(n))$, which is possibly smaller than n.

3.3. The Noisy Case

Now, consider the case when $\|\mathbf{z}\| \neq 0$, we analyze the sparse penalty being the ℓ_1 norm for its sub-differentiability. Estimation of labels can be formulated as solving the system:

$$\begin{aligned} &\text{Q(H):} \quad \arg\min_{l_i \in \mathcal{L}} R_s(l_i, f(\mathbf{x}^h_\star(\mathbf{y}, \mathbf{D}), \mathbf{W})) \text{ subject to} \\ &\mathbf{x}^h_\star = \arg\min(1/2)\|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 + \rho\|\mathbf{x}\|_1. \end{aligned} \quad (6)$$

On the other hand, similar procedure based on partial observation Λ is equivalent to solving:

$$\begin{aligned} &\text{Q(P):} \quad \arg\min_{l_i \in \mathcal{L}} R_s(l_i, f(\mathbf{x}^p_\star(\mathbf{y}_\Lambda, \mathbf{D}_\Lambda), \mathbf{W})) \text{ subject to} \\ &\mathbf{x}^p_\star = \arg\min(1/2)\|\mathbf{y}_\Lambda - \mathbf{D}_\Lambda \mathbf{x}\|_2^2 + \rho\|\mathbf{x}\|_1. \end{aligned} \quad (7)$$

As long as the solutions for Q(H) and Q(P) match, the classification based on partial observation is as accurate as the case with complete observation. Ideally, sparse codes are the same, which yields to the following sufficient condition.

Lemma 2 *(Sufficient Condition): For Λ such that $\mathbf{x}^p_\star = \arg\min(1/2)\|\mathbf{y}_\Lambda - \mathbf{D}_\Lambda \mathbf{x}\|_2^2 + \rho\|\mathbf{x}\|_1$, if $\mathbf{x}^p_\star$ also solves $\min(1/2)\|\mathbf{y}_{\Lambda^c} - \mathbf{D}_{\Lambda^c}\mathbf{x}\|_2^2$, then classification from partially observed data is just as accurate as with complete data.*

Proof: Let $\mathbf{x}^p_\star = \arg\min(1/2)\|\mathbf{y}_\Lambda - \mathbf{D}_\Lambda \mathbf{x}\|_2^2 + \rho\|\mathbf{x}\|_1$. Let $g_{1,\Lambda} := (1/2)\|\mathbf{y}_\Lambda - \mathbf{D}_\Lambda \mathbf{x}\|_2^2 + \rho\|\mathbf{x}\|_1$, then $\mathbf{0} \in (\partial g_{1,\Lambda}/\partial \mathbf{x})|_{\mathbf{x}=\mathbf{x}^p_\star}$. Also, let $g_{\Lambda^c} := (1/2)\|\mathbf{y}_{\Lambda^c} - \mathbf{D}_{\Lambda^c}\mathbf{x}\|_2^2$.

Since $\|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 = \|\mathbf{y}_\Lambda - \mathbf{D}_\Lambda \mathbf{x}\|_2^2 + \|\mathbf{y}_{\Lambda^c} - \mathbf{D}_{\Lambda^c}\mathbf{x}\|_2^2$, $\mathbf{x}^h_\star = \arg\min \rho\|\mathbf{x}\|_1 + (1/2)\|\mathbf{y}_\Lambda - \mathbf{D}_\Lambda \mathbf{x}\|_2^2 + (1/2)\|\mathbf{y}_{\Lambda^c} - \mathbf{D}_{\Lambda^c}\mathbf{x}\|_2^2 = \arg\min g_{1,\Lambda} + g_{\Lambda^c}$. If $\mathbf{x}^p_\star$ solves $g_{1,\Lambda}$ and g_{Λ^c}, then $\mathbf{0} \in (\partial g_{1,\Lambda}/\partial \mathbf{x})|_{\mathbf{x}=\mathbf{x}^p_\star} + (\partial g_{\Lambda^c}/\partial \mathbf{x})|_{\mathbf{x}=\mathbf{x}^p_\star}$, $\mathbf{x}^h_\star = \mathbf{x}^p_\star$.

$\{\mathbf{x} : \mathbf{0} \in (\partial g_{\Lambda^c}/\partial \mathbf{x})\}$ has close form solution $\{\mathbf{D}^\dagger_{\Lambda^c}\mathbf{y} + \mathbf{v} : \mathbf{v} \in \text{Null}(\mathbf{D}_{\Lambda^c})\}$, where † denotes the pseudo-inverse. If this solution set intersects with $\{\mathbf{x} : \mathbf{x} = \arg\min g_{1,\Lambda}\}$, then such a $\mathbf{x}^p_\star = \mathbf{x}^h_\star$ exists.

If $l_\mathbf{y}$ is the true label, then $\arg\min R_s(l_i, f(\mathbf{x}(\mathbf{y}, \mathbf{D}), \mathbf{W}))$ $= \arg\min R_s(l_i, f(\mathbf{x}(\mathbf{y}_\Lambda, \mathbf{D}_\Lambda), \mathbf{W})) = l_\mathbf{y}$. □

Given $\mathbf{y}, \mathbf{D}$ and Λ, we could tell whether the set of features is good enough to return the same label as solving Q(H).

4. EXPERIMENTAL RESULTS

We perform our experiments on the cropped AR data-set [18] containing 100 subjects with half males and half females. For faster computation, we down-sample the data-set by a ratio of 0.5 and the resulting size is 82×60. Under the assumption that training data are well-aligned holistic faces, we exclude pictures taken with scarfs or sunglasses and only use types $1-7$ and types $14-23$ for each person in the data-set, which contains 4 different emotions and 4 lighting conditions.

In the experiment, we randomly partition the data-set: half for training and remaining for testing. To test our algorithm, we generate 3 specific patterns (Type 1-3) with fixed

coordinates and 3 random patterns with controlled sizes (Type 4-6) in which the width b and height h are randomly generated from a predefined range: $b \in [r_1 B, r_2 B], h \in [r_1 H, r_2 H]$ and (r_1, r_2) are $(0.5, 0.9)$, $(0.3, 0.8)$ and $(0.2, 0.5)$ respectively. Examples are shown in Fig. 1.

Different dictionary learning algorithms are adopted in training. We apply online dictionary learning (ODL) [19], and label consistent K-SVD dictionary learning (LC-KSVD) [20] for constraint ℓ_1, ℓ_0 minimization respectively. We compare our algorithm with supervised dictionary learning methods using the same mean-face matching initialization (denoted by MF-Match) and one of the state-of-the-art feature-based partial face recognition algorithms: MKD-SRC [2].

In our experiments, all dictionary learning algorithms share the same parameter setting: $\rho = 0.1$, $\beta = 1$, dictionary size $n = 700$, and number of iterations is 10. The sparsity level of OMP is set to be 30 and the neighborhood size c_n is 5. Table 1 lists the averaged accuracy among all classes. Cumulative Matching Curves for two hard cases (Type 3, 6 with small observation sets) are depicted in Fig. 2 and Fig. 3.

Our algorithm performs better than others in both relatively easy cases (Type 1, 4) and hard cases (Type 3, 6). In SDL with MF Match and AASP, ℓ_1 and ℓ_0 minimization algorithms perform similarly. For holistic face recognition (Type 0), two methods coincidence since $\mathcal{S}_\Omega$ contains only one element: the whole observation set. All algorithms perform reasonably well when size of the observation set is large and when decreasing, they all suffer. By Section 3, larger size of Λ leads to tighter relaxation of the original constraint in (6), and therefore one would expect better result. Ours is relatively robust against the decreasing size of Λ and especially will not be effected that much when observation contains only part of major local areas such as eyes used for feature extraction as feature based methods.

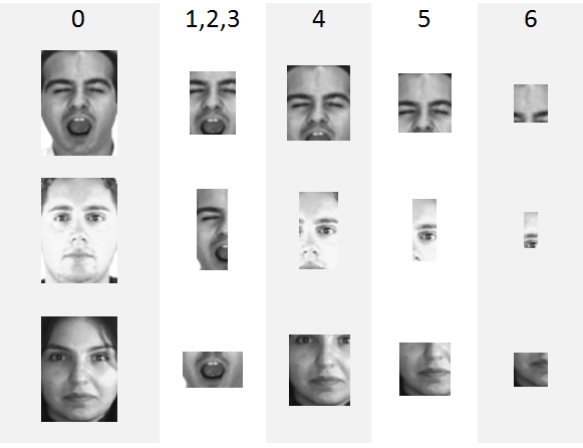

Fig. 1: Examples: Col. 1: holistic faces; Col. 2 (top to bottom): fixed patterns 1 (a large partial block), 2 (left face), 3 (mouth and chin); Col. 3, 4, 5 : random patterns 4, 5, 6 .
Size of partial faces among types: $0 > 1 > 2 > 3, 0 > 4 > 5 > 6$

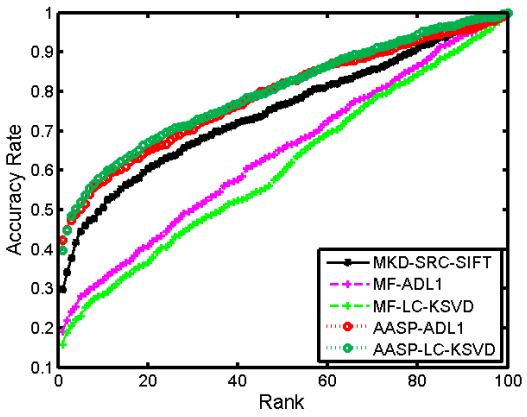

Fig. 2: Type 3: Cumulative Match Score Curves

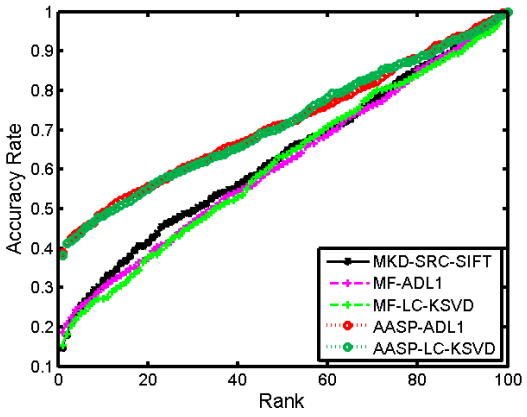

Fig. 3: Type 6: Cumulative Match Score Curves

Types	MKD -SRC [2]	MF-Match -ODL	MF-Match -LC-KSVD	AASP -ODL	AASP -LC-KSVD
0	0.93	0.96	**0.97**	0.96	**0.97**
1	0.66	0.63	0.64	**0.90**	0.87
2	0.58	0.31	0.37	0.76	**0.79**
3	0.29	0.19	0.18	**0.42**	0.40
4	0.68	0.60	0.63	**0.83**	**0.83**
5	0.47	0.41	0.41	**0.65**	**0.65**
6	0.15	0.19	0.15	**0.39**	0.38

Table 1: Rank-1 accuracy with various partial patterns

5. CONCLUSIONS AND FUTURE WORK

We develop a sparse representation-based classification algorithm called AASP to solve the partial face recognition problem. Experiments shows that our proposed method performs better than other comparison methods especially in the case of severe information missing. In the future, we plan to extend our method to handle various other types of corruptions and explore the lower bound of size of partial data that still provides correct classification in the noisy case.

6. REFERENCES

[1] J. Wright, A. Y Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, 2009.

[2] S. Liao and A. K. Jain, "Partial face recognition: An alignment free approach," in *Biometrics (IJCB), 2011 International Joint Conf. on*. IEEE, 2011, pp. 1–8.

[3] S. Liao, A. K Jain, and S. Z Li, "Partial face recognition: Alignment-free approach," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 5, pp. 1193–1205, 2013.

[4] J. Yang, K. Yu, and T. Huang, "Supervised translation-invariant sparse coding," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conf. on*. IEEE, 2010, pp. 3517–3524.

[5] Y. Chen, T. T. Do, and T. D. Tran, "Robust face recognition using locally adaptive sparse representation," in *Image Processing (ICIP), 2010 17th IEEE Conf. on*. IEEE, 2010, pp. 1657–1660.

[6] 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," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, vol. 34, no. 2, pp. 372–386, 2012.

[7] J. Huang, X. Huang, and D. Metaxas, "Simultaneous image transformation and sparse representation recovery," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conf. on*. IEEE, 2008, pp. 1–8.

[8] X. Sun, N. M. Nasrabadi, and T. D. Tran, "Sparse coding with fast image alignment via large displacement optical flow," *arXiv; International Conference on Acoustics, Speech, and Signal Processing(ICASSP), 2016 IEEE Conf. on (to be appear)*.

[9] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. R. Bach, "Supervised dictionary learning," pp. 1033–1040, 2009.

[10] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. R. Bach, "Supervised dictionary learning," in *Advances in neural information processing systems*, 2009, pp. 1033–1040.

[11] K. Engan, S. O. Aase, and J. H. Husøy, "Frame based signal compression using method of optimal directions (mod)," in *Circuits and Systems, 1999. ISCAS'99. Proceedings of the 1999 IEEE International Symposium on*. IEEE, 1999, vol. 4, pp. 1–4.

[12] M. Elad, M. Aharon, and A. M. Bruckstein, "The k-SVD: An algorithm for designing of overcomplete dictionaries for sparse representations," *IEEE Trans. Image Process*, vol. 15, no. 12, pp. 3736–3745, 2006.

[13] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 265–274, 2009.

[14] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *Information Theory, IEEE Trans. on*, vol. 53, no. 12, pp. 4655–4666, 2007.

[15] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[16] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 1, no. 4, pp. 586–597, 2007.

[17] E.J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *Information Theory, IEEE Trans. on*, vol. 52, no. 2, pp. 489–509, 2006.

[18] A. M. Martínez and A. C. Kak, "PCA versus LDA," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, vol. 23, no. 2, pp. 228–233, 2001.

[19] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *The Journal of Machine Learning Research*, vol. 11, pp. 19–60, 2010.

[20] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent K-SVD," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conf. on*. IEEE, 2011, pp. 1697–1704.