Mining Discriminative Components With Low-Rank And Sparsity Constraints for Face Recognition

Qiang Zhang, Baoxin Li
Computer Science and Engineering
Arizona State University
Tempe, AZ, 85281
qzhang53, baoxin.li@asu.edu

ABSTRACT

This paper introduces a novel image decomposition approach for an ensemble of correlated images, using low-rank and sparsity constraints. Each image is decomposed as a combination of three components: one common component, one condition component, which is assumed to be a low-rank matrix, and a sparse residual. For a set of face images of N subjects, the decomposition finds N common components, one for each subject, K low-rank components, each capturing a different global condition of the set (e.g., different illumination conditions), and a sparse residual for each input image. Through this decomposition, the proposed approach recovers a clean face image (the common component) for each subject and discovers the conditions (the condition components and the sparse residuals) of the images in the set. The set of $N + K$ images containing only the common and the low-rank components form a compact and discriminative representation for the original images. We design a classifier using only these $N + K$ images. Experiments on commonly-used face data sets demonstrate the effectiveness of the approach for face recognition through comparing with the leading state-of-the-art in the literature. The experiments further show good accuracy in classifying the condition of an input image, suggesting that the components from the proposed decomposition indeed capture physically meaningful features of the input.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Feature extraction and preprocessing

Keywords

Subspace learning, Low-rank matrix, Sparse matrix, Face Recognition, Component Decomposition

WOODSTOCK '97 El Paso, Texas USA

1. INTRODUCTION

Face recognition has been an active research field for a few decades, and its challenges and importance continue to attract efforts from many researchers, resulting in many new approaches in recent years. The most recent literature may be divided into roughly two groups, where methods in the first group try to model the physical processes of image formation under different conditions (e.g., illumination, expression, pose etc.). For example, the approach of [10] models the face image under varying illumination conditions to be a linear combination of images of the same subject captured at 9 specially designed illumination conditions; the SRC algorithm of [19] further assumes that face images with illumination and expression conditions can be represented as a sparse linear combination of the training instances (i.e., the dictionary atoms). On the other hand, the second group of approaches utilizes mathematical/statistical tools to capture the latent relations among face images for classification. E.g., the SUN approach [7] uses the statistics of the human fixation of the images to recognize the face images, Volterrafaces [9] finds a latent space for face recognition, where the ratio of intra-class distance over inter-class distance is minimized. One major advantage of the techniques in the first class comes from their being generative in nature, which allows these methods to accomplish tasks like face relighting or novel pose generation in addition to recognition. The second group of methods in a sense ignores the physical property of the faces images and treats them as ordinary 2D signals.

Although the methods in the first group have the above nice property, a baseline implementation usually requires dictionaries with training images as atoms and thus may face the scalability issue in real-world applications with a huge number of subjects. Hence efforts have also been devoted to reducing the size of the dictionary while attempting to retain the level of performance of the original dictionary. Examples include those that generate more compact dictionaries through some learning procedure (e.g., [13]) and those that attempt to extract subject-specific features that are effectively used as dictionary atoms (e.g., [15]). Our approach belongs to the second group. Since the expressive power of the original dictionary-based techniques comes from largely the number of training images for each subject, a compact dictionary may suffer from degraded performance unless the reduced dictionary properly captures the conditions of the original data that are critical for a recognition task. For example, the method of [15], while shown to be effective for expression-invariant recognition, is difficulty to generalize to handle global conditions such as illumination change, which

often introduce to the data non-sparse conditions that cannot be captured by the sparsity model proposed therein.

Recognizing that non-sparse conditions such as illumination change and large occlusion are critical for face recognition, and that for a typical application we may assume only a finite number of such conditions (e.g., a relatively small number of illumination conditions or other conditions), in this paper, we propose a model for representing a set of face images by decomposing them into three components: a common component shared by images of the same subject, a low-rank component capturing non-sparse global changes, and a sparse residual component. Such a decomposition is partially inspired by the observation that the reconstruction of the image with the top few singular values and the corresponding singular vectors often capture the global information of the image, which can be represented by a low-rank matrix. To this end, a generic algorithm is proposed, with theoretic analysis on the convergence and parameter selection. The learned common and low-rank components form a compact and discriminative representation of the original set of images. A classifier is then built based on comparison of subspaces spanned by these components and by a novel image to be classified. This is very compact compared with the number of atoms in an over-determined dictionary such as that in [19]. Further, by explicitly modeling non-sparse conditions, the proposed approach is able to handle both illumination changes and large occlusions, which would fail methods like [15].

To demonstrate the effectiveness of the proposed method, we first design synthetic experiments with known ground truth to verify its key capability in recovering the underlying common, low-rank and sparse components. Then we report results on three commonly-used data sets of real face images: the Extended YaleB dataset [4], the CMU PIE dataset [17] and the AR dataset [14].The experiments show that, the proposed approach obtained better performance than the SRC algorithm [19], which utilizes a much larger dictionary, and the SUN approach [7]. The proposed approach also achieves comparable result to Volterrafaces, which is the current state-of-the-art in the literature for a few commonly-used data sets. In addition, the proposed approach can explicitly model the most important feature of the subject and the conditions in the dataset. Experiments also show that the proposed method is robust to situations where a non-trivial percentage of the training images is unavailable. Further, the capability of the proposed approach for classifying the type of condition that an input image is subject to is also demonstrated by extensive experiments. This suggests that the proposed decomposition is able to obtain physically meaningful and thus potentially discriminative components.

We introduce the proposed method in Section 2, including the proposed model, the learning algorithm and the classification method. The experiments are reported and analyzed in Section 3. We conclude in Section 4 with a summary of the work and brief discussion on future work.

In the presentation, we use upper case bold font for matrices, e.g., $\mathbf{X}$, lower case bold font for vectors, e.g., $\mathbf{x}$ and normal font for scalars, e.g., x. $\{\mathbf{X}_{i,j}\}_{i=1,j=1}^{N,M}$ denotes a set of $N \times M$ matrices, with $\mathbf{X}_{i,j}$ as its $(i,j)_{th}$ member. We assume that N is the number of the subjects, and M the number of images per subject[1]. Thus $\mathbf{X}_{i,j}$ refers to j_{th} image of the i_{th} subject. When there is no confusion, we also use $\mathbb{X}$ to denote the set $\{\mathbf{X}_{i,j}\}_{i,j=1}^{N,M}$.

[1]For simplicity, we assume that each subject has the same number of images, which can always be achieved by using some blank images, a situation the proposed method can handle.

2. PROPOSED METHOD

In this section, we first present the general formulation of the proposed model in Section 2.1, and then present our algorithm for obtaining the desired decomposition in Section 2.2 and analysis of its convergence in Sec. 2.3. With these, a face recognition algorithm is then designed in Section 2.4.

2.1 Decomposing a Face Image Set

In many applications of image and signal processing, we often consider a set of correlated signals as an ensemble. For efficient representation, a signal in the ensemble can often be viewed as a combination of a common component, which is shared among all the signals in the ensemble, and an innovation component, which is unique to this signal. Many benefits can be drawn from this decomposition of the ensemble, such as obtaining better compression rate and being able to extract more relevant features. In face recognition, all the face images, especially the subset corresponding to a subject, may be naturally viewed as forming such an ensemble of correlated signals. In a sense, a sparse-coding approach like SRC *implicitly* figures out the correlation of the images in the ensemble via the sparse coefficients under the dictionary of the training images.

In this work, we aim at developing a new representation of this ensemble so that the face recognition task can be better supported. In particular, considering the common challenges such as illumination conditions and large occlusions, we want to have a representation that can *explicitly* model such conditions. To this end, we propose the following decomposition of face images $\mathbf{X}_{i,j}$ in the ensemble $\mathbb{X}$ as:

$$\mathbf{X}_{i,j} = \mathbf{C}_i + \mathbf{A}_j + \mathbf{E}_{i,j},\ \forall \mathbf{X}_{i,j} \in \mathbb{X} \tag{1}$$

where $\mathbf{C}_i$ is the common part for Subject i, $\mathbf{A}_j$ is a low-rank matrix, and $\mathbf{E}_{i,j}$ is a sparse residual.

One essential difference between the proposed method and Robust PCA (RPCA [18]), is that RPCA assumes the signals are linearly dependent, with some sparsely corrupted entries in the signals. As a result, they build a big matrix with each signal as a vector. The big matrix would naturally be low-rank (because of the assumed inter-image correlation), in addition to having a sparse set of entries. On the other hand, the proposed decomposition is partially inspired by the observation that the reconstruction of the image with first few singular values and the corresponding singular vectors often capture the global information of the image [12], e.g., illumination conditions, structured patterns, which can be represented by a low-rank matrix. Here the low-rank constraint arises from certain physical conditions (rather than due to inter-image correlation), and it is imposed on each individual image. Accordingly, we represent images by matrices rather than vectors, unlike other methods like [19, 18]. With this, we can expect that:

- $\mathbf{C}_i$ is a matrix representing the common information of images for Subject i, i.e., the common components;

$\mathbf{A}_j$ is a low-rank matrix capturing the global information of the image $\mathbf{X}_{i,j}$, e.g., illumination conditions (Fig. 3), structured patterns (Fig. 1); and

$\mathbf{E}_{i,j}$ is a sparse matrix pertaining to image-specific details such as expression conditions or noise with sparse support in the images.

In this modeling, we have assumed M different low-rank matrices, which are responsible for M different global conditions such as illumination conditions or large occlusions, and they are shared among the images of different subjects. However, images of each subject do not necessarily contain all the M conditions, as we will show in Sec. 2.2.

We can also obtain a variant of the above model by considering the Retinex theory, in which image $\mathbf{I}$ can be represented as:

$$\mathbf{I}(p,q) = \mathbf{R}(p,q) \cdot \mathbf{L}(p,q) \tag{2}$$

where $\mathbf{R}(x,y)$ is the reflectance at location (x,y), which depends on the surface property, $\mathbf{L}(x,y)$ is the illumination, and $\cdot$ is element-wise product. Converting this into the logarithm domain, we have

$$log(\mathbf{I}) = log(\mathbf{R}) + log(\mathbf{L}) \tag{3}$$

The above equation indicates that we can represent the intensity of the face image as follows:

$$log(\mathbf{X}_{i,j}) = \mathbf{C}_i + \mathbf{A}_j + \mathbf{E}_{i,j},\ \forall \mathbf{X}_{i,j} \in \mathbb{X} \tag{4}$$

where $\mathbf{C}_i = log(\mathbf{R})$ captures the common property of the images for Subject i, $\mathbf{A}_j = log(\mathbf{L})$ captures the lighting conditions, and $\mathbf{E}_{i,j}$ captures the residual. This is a variant of the model in Eqn. 1, and is especially suitable for illumination-dominated datasets such as the extended YaleB dataset and the CMU-PIE dataset.

With the above decomposition, the entire dataset containing $N \times M$ images can be compactly represented by N common components and K low-rank components. If we extract the common component $\mathbf{C}_i$ for face images of Subject i under different conditions, we expect that this common component $\mathbf{C}_i$ represents the most significant feature of that subject. The set of all the learned low-rank components $\mathbb{A} = \{\mathbf{A}_j\}_{j=1}^{M}$ represents all possible global conditions of the images in the set. Hence we may use $\mathbb{A}$ and $\mathbf{C}_i$ to span the subspace of the face images for Subject i, where, in the ideal case, any face images of this subject should lie in, barring a sparse residual. This suggests that we can utilize the subspaces for face recognition by identifying which subspace a test image is more likely to lie in, which is detailed in Sec. 2.4.

2.2 An Algorithm for the Decomposition

Based on Eqn. 1, we formulate the decomposition task as the following constrained optimization problem, with an objective function derived from the requirement of decomposing a set of images into some common components, some low-rank matrices and the sparse residuals:

$$\begin{aligned} \mathbb{C}, \mathbb{A}, \mathbb{E} &= \underset{\mathbb{C},\mathbb{A},\mathbb{E}}{\operatorname{argmin}} \sum_{i,j} \|\mathbf{A}_j\|_* + \lambda_{i,j}\|\mathbf{E}_{i,j}\|_1 \\ s.t. &\quad \mathbf{X}_{i,j} = \mathbf{C}_i + \mathbf{A}_j + \mathbf{E}_{i,j},\ \forall \mathbf{X}_{i,j} \in \mathbb{X} \end{aligned} \tag{5}$$

where $\|\mathbf{A}_j\|_* = \sum_i \sigma_i(\mathbf{A}_j)$ is the nuclear norm, $\|\mathbf{E}_{i,j}\|_1 = \sum_{p,q} |\mathbf{E}_{i,j}(p,q)|$ is the ℓ_1 norm and $\mathbb{E} = \{\mathbf{E}_{i,j}\}_{i,j=1}^{N,M}$. Note that, unlike [18] where a set of images are stacked as vectors of a low-rank matrix, we do not convert the image to a vector in the decomposition stage.

To absorb the constraints into the objective function, we can reformulate Eqn. 5 with augmented Lagrange multiplier as:

$$\begin{aligned} \mathbb{C}, \mathbb{A}, \mathbb{E} &= \underset{\mathbb{C},\mathbb{A},\mathbb{E}}{\operatorname{argmin}} \sum_{i,j} \|\mathbf{A}_j\|_* + \lambda_{i,j}\|\mathbf{E}_{i,j}\|_1 \\ &+ \frac{\mu_{i,j}}{2}\|\mathbf{X}_{i,j} - \mathbf{C}_i - \mathbf{A}_j - \mathbf{E}_{i,j}\|_F^2 \\ &+ < \mathbf{Y}_{i,j}, \mathbf{X}_{i,j} - \mathbf{C}_i - \mathbf{A}_j - \mathbf{E}_{i,j} > \end{aligned} \tag{6}$$

where $\mathbf{Y}_{i,j}$ is the Lagrange multiplier, $\lambda_{i,j}$ and $\mu_{i,j}$ are scalars controlling the weight of sparsity and reconstruction error accordingly. When μ is sufficiently large, Eqn. 6 is equivalent to Eqn. 5. It is worth pointing out that, while for clarity we have written only the expression for Subject i, the optimization is actually done for the entire set of images, since the low-rank components are deemed as been shared by all images.

To solve the problem of Eqn. 6, a block coordinate descent algorithm may be designed, with each iterative step solving a convex optimization problem [3][18] for one of the unknowns. To this end, we first describe the following three sub-solutions that are needed in each iteration of such an algorithm, which correspond to solving only one of the unknowns (blocks) while fixing others.

Sub-solution 1: For finding an optimal $E_{i,j}$ in the t-th iteration, where the problem can be written as

$$\begin{aligned} \mathbf{E}_{i,j} &= \underset{\mathbf{E}_{i,j}}{\operatorname{argmin}} \lambda_{i,j}\|\mathbf{E}_{i,j}\|_1 \\ &+ \frac{\mu_{i,j}}{2}\|\mathbf{X}_{i,j}^E - \mathbf{E}_{i,j}\|_F^2 + < \mathbf{Y}_{i,j}, \mathbf{X}_{i,j}^E - \mathbf{E}_{i,j} > \end{aligned} \tag{7}$$

with $\mathbf{X}_{i,j}^E = \mathbf{X}_{i,j} - \mathbf{C}_i - \mathbf{A}_j$. So we do the following update [6]:

$$\mathbf{E}_{i,j} = S_{\frac{\lambda}{\mu_{i,j}}}(\mathbf{X}_{i,j}^E + \frac{1}{\mu_{i,j}}\mathbf{Y}_{i,j}) \tag{8}$$

where $S_\tau(\mathbf{X}) = sign(\mathbf{X}) \cdot max(0, |\mathbf{X}| - \tau)$.

Sub-solution 2: For finding an optimal $\mathbf{A}_k$ in the t-th iteration, where the problem can be written as

$$\begin{aligned} \mathbf{A}_j &= \underset{\mathbf{A}_j}{\operatorname{argmin}} \sum_i \|\mathbf{A}_j\|_* \\ &+ \frac{\mu_{i,j}}{2}\|\mathbf{X}_{i,j}^A - \mathbf{A}_j\|_F^2 + < \mathbf{Y}_{i,j}, \mathbf{X}_{i,j}^A - \mathbf{A}_j > \end{aligned} \tag{9}$$

We use the singular value thresholding algorithm [2, 5]:

$$\begin{aligned} \mathbf{U}\Sigma\mathbf{V}^T &\leftarrow \frac{\sum_i \mu_{i,j}\mathbf{X}_{i,j}^A + \mathbf{Y}_{i,j}}{\sum_i \mu_{i,j}} \\ \mathbf{A}_j &= \mathbf{U}S_\tau(\Sigma)\mathbf{V}^T \end{aligned}$$

with $\mathbf{X}_{i,j}^A = \mathbf{X}_{i,j} - \mathbf{C}_i - \mathbf{E}_{i,j}$ and $\tau = \frac{N}{\sum_i \mu_{i,j}}$.

Sub-solution 3: The solution to the problem of finding optimal $\mathbf{C}_i$

$$\underset{\mathbf{C}_i}{\operatorname{argmin}} \frac{\mu_{i,j}}{2} \sum_j \|\mathbf{X}_{i,j}^C - \mathbf{C}_i\|_F^2 + < \mathbf{Y}_{i,j}, \mathbf{X}_{i,j}^C - \mathbf{C}_i > \tag{10}$$

where $\mathbf{X}_{i,j}^C = \mathbf{X}_{i,j} - \mathbf{A}_j - \mathbf{E}_{i,j}$, can be obtained directly (by taking derivatives of the objective function and setting to

zero) as

$$\mathbf{C}_i = \frac{\sum_j \mathbf{Y}_{i,j} + \mu_{i,j}\mathbf{X}^C_{i,j}}{\sum_j \mu_{i,j}} \quad (11)$$

As alluded earlier, the images of any given subject may not range over all possible M conditions. This may be equivalently viewed as a problem of some images are missing for the subject. We now show how this can be addressed in a principled way. Assume that Ω is the set of (i,j) where $\mathbf{X}_{i,j}$ is available and $\bar{\Omega}$ is the complement of Ω. To deal with those missing entries, we only need to set $\mathbf{Y}_{i,j}$, $\mu_{i,j}$ and $\mathbf{X}_{i,j}$ to 0 for $(i,j) \in \bar{\Omega}$ in the initialization stage. In each iteration, we do not update $\mathbf{E}_{i,j}$ for $(i,j) \in \bar{\Omega}$. The proposed decomposition algorithm will automatically infer the missing images.

With the above preparation, we now propose the following Algorithm 1 to solve Eqn. 6:

Algorithm 1: Learning the Decomposition

Input: $\mathbb{X}$, Ω, N, M, ρ, λ and τ;
Output: $\{\mathbf{C}_i\}_{i=1}^N$, $\{\mathbf{A}_j\}_{j=1}^K$ and $\{\mathbf{E}_{i,j}\}_{i,j=1}^{N,M}$;
% Initialization
$t = 0$, $\mathbf{C}_i^0 = \mathbf{A}_j^0 = \mathbf{E}_{i,j}^0 = 0$;
$\mathbf{Y}_{i,j}^0 = \frac{\mathbf{X}_{i,j}}{\|\mathbf{X}_{i,j}\|_F}$, $\mu_{i,j}^0 = \frac{\tau}{\|\mathbf{X}_{i,j}\|_F}$ for $(i,j) \in \Omega$;
$\mathbf{Y}_{i,j}^0 = 0$, $\mu_{i,j}^0 = 0$ for $(i,j) \notin \Omega$;
while not converged do
  Solve $\mathbf{E}_{i,j}$ for $(i,j) \in \Omega$ by Sub-solution 1:
  Solve $\mathbf{A}_j$ for $j = 1, 2, ..., M$ with Sub-solution 2;
  Solve $\mathbf{C}_i$ for $i = 1, 2, ..., N$ using Sub-solution 3;
  %Update $\mathbf{Y}_{i,j}$ and $\mu_{i,j}$ for $(i,j) \in \Omega$:
  $\mathbf{Y}_{i,j}^{t+1} = \mathbf{Y}_{i,j}^t + \mu_{i,j}^t(\mathbf{X}_{i,j} - \mathbf{C}_i^{t+1} - \mathbf{A}_j^{t+1} - \mathbf{E}_{i,j}^{t+1})$;
  $\mu_{i,j}^{t+1} = \mu_{i,j}^t \rho$;
  $t = t + 1$;
end

where for convergence, we check $\frac{\sum_{i,j}\|\mathbf{X}_{i,j} - \mathbf{C}_i - \mathbf{A}_j - \mathbf{E}_{i,j}\|_F^2}{\sum_{i,j}\|\mathbf{X}_{i,j}\|_F^2}$ and if it is small enough (e.g., 10^{-6}), we terminate the algorithm. λ, τ and ρ are three parameters specified in input, which are discussed in Sec. 3.1.

2.3 Convergence of the Algorithm

The convergence property of an iterative optimization procedure like the algorithm proposed above is critical to its usefulness. The Algorithm 1 has similar convergence property as the methods described in [11], which are also augmented Lagrange multiplier based approaches. We can draw the following theorem:

Theorem 1 If $\sum_{t=1}^{\infty} \mu_{i,j}^{t+1}(\mu_{i,j}^t)^{-2} < \infty$ and $\lim_{t\to\infty} \mu_{i,j}^t(\mathbf{E}_{i,j}^{t+1} - \mathbf{E}_{i,j}^t) = 0$ $\forall i,j$, then Algorithm 1 will converge to the optimal solution for the problem of Eqn. 5.

The proof of Theorem 1 is included in the appendix.

2.4 Face Recognition Using the Decomposition

With the components in Eqn. 1 estimated from the previous algorithm, we now discuss how to classify a test image. Recognizing that the sparse residual captures only image-specific details that have not been absorbed by the common or the global condition, we discard the sparse residuals from the decomposition (training) stage and keep only the common and the low-rank components.

Ideally a face image from Subject i should lie in a subspace spanned by its common component $\mathbf{C}_i$ and the low-rank components $\mathbb{A}$. Therefore, we propose the following classification scheme based on comparing the distance between subspaces spanned by the training components and those spanned by replacing the training common by the test image $\mathbf{y}$. We first build the subspace $\mathbf{S}_i$ for subject i, which contains all the linear combinations of the images of Subject i under all conditions, i.e.,

$$\mathbf{S}_i = \{\mathbf{x} | \mathbf{x} = \sum_k w_k \times (\mathbf{c}_i + \mathbf{a}_j) \, \forall \mathbf{w} \in \mathbb{R}^M\} \quad (12)$$

where $\mathbf{c}_i$ and $\mathbf{a}_j$ is the vectorized form of $\mathbf{C}_i$ and $\mathbf{A}_j$ respectively. Subspace $\mathbf{S}_i$ can be sufficiently represented by a set of "basis", i.e., $\{\mathbf{c}_i + \mathbf{a}_j\}_{j=1}^M$. Accordingly, we can build the subspace $\mathbf{S}_y$ for the test image y as the follows:

$$\mathbf{S}_y = \{\mathbf{x} | \mathbf{x} = \sum_k w_k \times (\mathbf{y} + \mathbf{a}_j) \, \forall \mathbf{w} \in \mathbb{R}^M\} \quad (13)$$

Then we use the principal angles [8] between these subspace to measure their similarities. In this paper, the principal angles measure the cosine distance between the subspaces, which is calculated as $s(\mathbf{S}_i, \mathbf{S}_y) = \sum_k cos^2(\theta_k)$, where θ_k is the k_{th} principal angle between $\mathbf{S}_i$ and $\mathbf{S}_y$. The assign i as the label of $\mathbf{f}$, for which $s(\mathbf{S}_i, \mathbf{S}_y)$ is maximal.

3. EXPERIMENTAL RESULTS

Experiments have been done to evaluate the proposed model and algorithms. In this section, we report several sets of results from such experiments. First, simulations (Sec. 3.1) are employed to demonstrate the convergence and parameter selection of the proposed decomposition algorithm. Then, we show the decomposition of the images from extended YaleB dataset and also how the learned components can be used to reconstruct new images in Sec. 3.2. Finally, we demonstrate the application of the proposed method and algorithms in classification tasks, including face recognition (Sec. 3.3) and identifying the conditions of the images (Sec. 3.4). The performance of the proposed method in face recognition task is compared with that of SRC [19], Volterrafaces [9] and SUN [7] on 2 commonly used datasets, i.e., extended YaleB [10] and CMU-PIE [17].

3.1 Simulation-based Experiments

In this subsection, we use synthetic data to demonstrate the convergence of the algorithm and selection of the parameters. The common components and condition components used in this experiment are shown in Fig. 1 (b,c), where the condition components are from [16] and both components are rescaled to range $[0, 1]$. The sparse components are sampled from a uniform distribution in the range of $[0, 1]$. We use those components to generate 25 images, which are used in this experiment, as Eqn. 1.

Algorithm 1 in Sec. 2.2 requires three parameters, ρ controls the convergence speed; λ controls the sparsity of the sparse residuals; and τ is a scalar. In [11], they suggest $\rho = 1.5$, $\tau = 1.25$ and $\lambda = \frac{1}{\sqrt{m}}$ for Robust PCA, where m is the width of $\mathbf{X}_{i,j}$. We have also found that $\lambda = \frac{1}{\sqrt{m}}$ is optimal from the experiments, thus we adopt this selection in our paper. From the experiment, we found that $\tau \in [0.125, 2]$ and $\rho = 1.25$ would be an optimal choice. Fig. 1 shows an example of the recovered common components

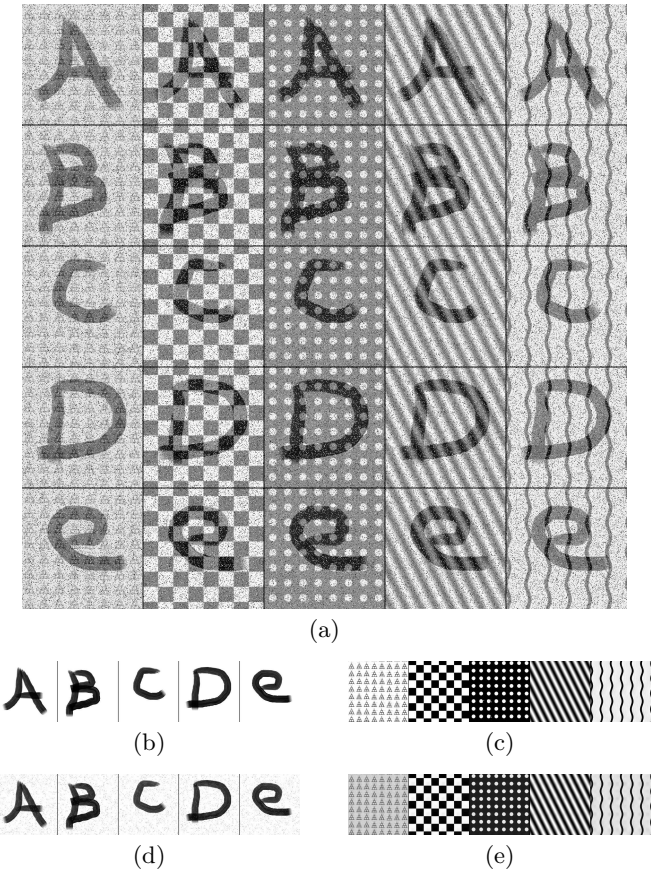

Figure 1: (a) shows the 25 images generated in the experiment, where the sparse part has 20% support of each image. (b,c) shows the ground truth of the common components and condition components accordingly. We also show the common components (d) and condition components (e) decomposed from (b) when $\rho = 1.25$ and $\tau = 2$.

(d) and condition components (e) when the sparse part has 20% support of the image.[2]

To demonstrate the robustness of the algorithm, when only part of data is available, we randomly remove 10 images from the 25 images (Fig. 2(a)) and run the algorithm with the same set of parameters. The results are shown in Fig. 2, where (b) is the recovered common components and (c) is the recovered condition components. These results suggest that the algorithm is still able to produce reasonable results even with 40% of the images missing.

3.2 Decomposing a Set of Images

In this subsection, we first demonstrate the decomposition of the set of images from Extended YaleB dataset[4]. All the 2432 images from 38 subjects under 64 illumination conditions were used. The common components and the condition components are illustrated in in Fig. 3. Comparing these with the original data, it is evident that the recovered

[2]The recovered parts are subject to a linear shift and scaling. We identify the parameters for this linear shift and scaling then map them back with those parameters.

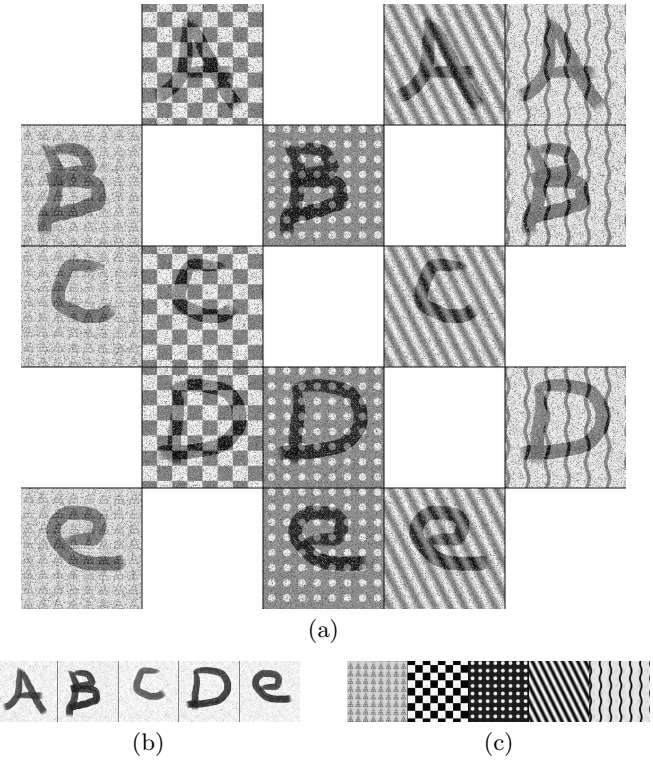

Figure 2: (a) the input data with 10 image manually removed, (b,c) is the common components and condition components decomposed from (a) accordingly.

commons are largely clean pictures of the subjects, while the condition components align well with the given illumination conditions. This experiment shows the capability of the proposed method with the Retinex model to discover the illumination conditions and the subject commons from a set of real images.

Next, we randomly pick 32 illumination conditions out of the decomposed 64 conditions and the common components of Subject 1 to form a subspace as described in Eqn. 14. Then we use the proposed method to identify whether an new image is in this subspace, by reconstructing this image as the linear combination of the "basis" of this subspace, i.e., $\mathbf{c}_1 + \mathbf{a}_j$. Fig. 4 shows an example, where the new image is also picked from Subject 1; and Fig. 5 shows another example, where the new image is picked from Subject 2. These examples suggest that the learned components can be used for identifying which subject an new image belongs to. Similarly, the learned components can also be used for identifying which conditions the new image is associated with. These two scenarios are further evaluated in the following two subsections, with real face images.

3.3 Recognizing the Face Images

In this subsection, we demonstrate the performance of the proposed method in face recognition task, with the comparison to SRC, Volterraface and SUN on the extended YaleB dataset and CMU PIE dataset. As these two datasets are dominated by illumination conditions, we use the Retinex model for the proposed method, i.e., the image is converted

(a) (b)

Figure 3: The decomposition of the extended YaleB dataset. We use all the 2432 images which contain 38 subjects (b) and 64 illumination conditions (a).

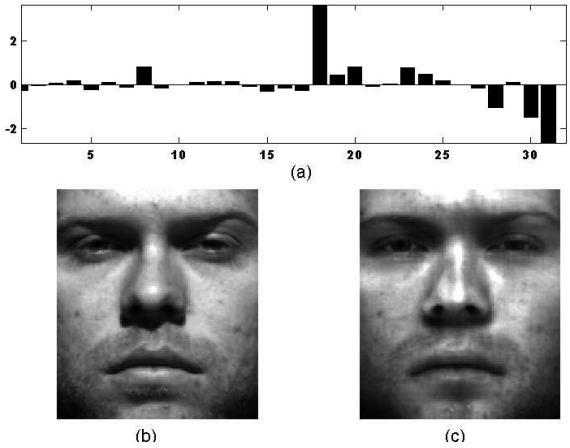

Figure 4: (a) the coefficient for the linear combination, (b) the input image, which is not observed in the images for training the 32 illumination conditions, and (c) the reconstructed image.

to logarithm. In the SRC method, we build the dictionary by containing all the training images as its columns. Since there is no code publicly available for SRC, we build our own implementation. For ℓ_1 optimization used by SRC, we used Orthonormal Matching Pursuit (OMP)[1] as the solver. We set the number of non-zero elements in the sparse coefficient (refer as K later) to be twice the number of conditions in the training data. In addition, each image is normalized to have zero mean and unit l_2 norm for SRC. For Volterrafaces and SUN, we use the author's original implementation and the provided parameters. For all the results, we present the both mean and standard deviation of the accuracies of 3 rounds of experiments.

To examine the robustness of the approaches with respect to the amount of training data, we use the following scheme. In the experiment, we only pick "#train per subject" images for each subject as the training instances, according to the randomly generated sample matrix, where some of the elements are set to 0 and the corresponding images won't be used for training.

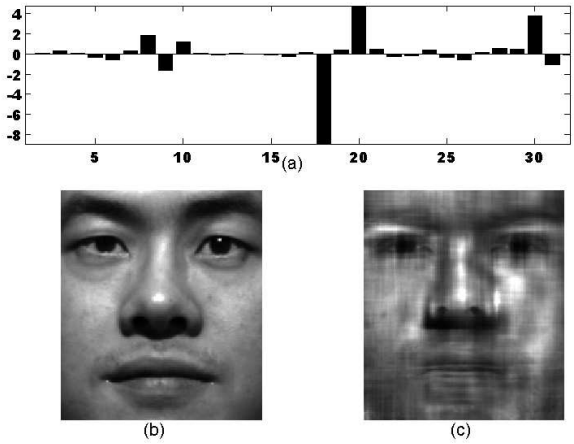

Figure 5: (a) the coefficient for the linear combination, (b) the input image and (c) the reconstructed image.

The Extended YaleB dataset [4] contains $N = 38$ subjects with 64 images for each subject, which correspond to 64 illumination conditions in the dataset. The images are resized to 48×42. The results on the extended YaleB dataset are summarized in Tab. 1. From this table, we find that the proposed approach and Volterrafaces achieve the best results; and SUN get obviously the lowest accuracy. The performance of SRC degrade dramatically as the size of dictionary (i.e., number of training instances) reduced.

The CMU PIE dataset [17] contains $N = 68$ subjects with varying poses, illuminations and expressions etc.. For all the images, we manually crop the face region, according to the eye position, then resize them to 50×35. The results are summarized in Tab. 2. In Experiment 1, all 4 methods get similar results; in Experiment 2, the proposed method and Volterrafaces get the best result; and in Experiment 3, the proposed approach gets the best result. In addition, the proposed method is more robust to the missing of training images. The performance of SRC degrades obviously as the size of dictionary reduced.

To illustrate the speed performance of the proposed approach, we compared the time required to classify one image in our approach and the SRC approach. This time was about 0.84 seconds in our method, and about 1.59 seconds in SRC. The time for the decomposition (i.e., Algorithm 1) is less than 5 minutes. The most time consuming part for the proposed approach is the singular value decomposition (SVD), which is used in computing the principle angle, so an efficient implementation of SVD can make the proposed algorithm even faster.

3.4 Identifying the Conditions

Finally, we use an experiment to show how the proposed method can be applied to to identifying the conditions the testing images are associated with. The AR dataset [14] contains $N = 100$ subjects and 26 images for each subjects. The dataset contains 2 sessions, which are taken at different times. Each session contains 13 conditions: 4 for expressions, 3 for illuminations, 3 for sun glasses and 3 for scarves.

(a) Experiment 1

#train per subject	32	24	16	8
Proposed	99.78±0.24%	99.54±0.04%	99.18±0.14%	**95.15±1.03%**
SRC	96.48±0.44%	95.29±0.52%	91.90±0.94%	78.65±1.81%
Volterrafaces	**99.95±0.06%**	**99.80±0.26%**	**99.48±0.49%**	90.22±11.84%
SUN	89.61±1.85%	87.64±2.80%	76.91±3.71%	60.17±2.09%

(b) Experiment 2

#train per subject	16	12	8	4
Proposed	**99.56±0.00%**	**99.33±0.23%**	**98.32±0.03%**	80.03±2.17%
SRC	89.14±0.00%	87.88±0.44%	81.02±0.13%	58.54±1.26%
Volterrafaces	99.25±0.34%	99.17±0.39%	96.27±4.03%	**91.03±2.43%**
SUN	79.22±0.00%	76.75±0.00%	68.86±0.00%	51.60±0.00%

Table 1: The results on extended YaleB dataset. Experiment 1: we randomly pick $M = 32$ illumination conditions for training and the remaining for testing, i.e., we will obtain $N = 38$ common components and $M = 32$ conditions by the proposed method. Experiment 2: we manually pick $M = 16$ illumination conditions for training and the remaining for testing.

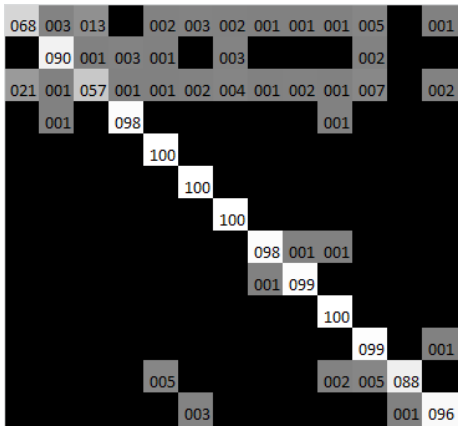

Figure 6: The confusion matrix (in percentage) of condition recognition result from the proposed method, where both axes are the condition index. The axis is index of the conditions.

In our experiments, we use one session for training and the other session for testing. The images are converted to gray scale and resized to 55×40. To recognized the associated condition, we slightly changes the formulation of the subspace:

$$\mathbf{S}_i = \{\mathbf{x} | \mathbf{x} = \sum_j w_j \times (\mathbf{a}_i + \mathbf{c}_j) \, \forall \mathbf{w} \in \mathbb{R}^N\} \quad (14)$$

$$\mathbf{S}_y = \{\mathbf{x} | \mathbf{x} = \sum_j w_j \times (\mathbf{y} + \mathbf{c}_j) \, \forall \mathbf{w} \in \mathbb{R}^N\} \quad (15)$$

where $\mathbf{S}_i$ is the subspace for condition i and $\mathbf{S}_y$ the subspace for the test image. The other settings were the same as those of previous face recognition experiments.

The proposed method achieves an accuracy of 91.77% in recognizing the conditions, with the confusion matrix given in Fig. 6, where we achieved over 96% accuracy for all but conditions 1, 2, 3 (3 expressions) and 12. This experiment again demonstrates the effectiveness of the proposed method in capturing the physical conditions in the form of low-rank components.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a novel decomposition of a set of face images of multiple subjects, each with multiple images. The decomposition finds a common image and a low-rank image for each of the subjects in the set. All the low-rank images form a set that is used to capture all possible global conditions existing in the set of images. This facilitates explicit modeling of typical challenges in face recognition, such as illumination conditions and large occlusion. Based on the decomposition, a face classifier was designed, using the decomposed components for subspace reconstruction and comparison. The classification performance shows that the proposed approach can achieve state-of-the-art performance. Experiments also showed that the proposed method is robust with missing training images, which can be an important factor to consider in a practical system. We also demonstrated with experiments that the decomposition indeed captures physically meaningful conditions, with both synthetic data and real data.

There are a few possible directions for further development of the work. In particular, the current algorithm assumes that the low-rank conditions of the training images are known and given for each of them. In practice, if the data do not have such image-level label (but still with a finite set of low-rank conditions), it is possible to expand the current algorithm by incorporating another step that attempts to estimate a mapping matrix for assigning a condition label to each image, during the optimization iteration. For example, we may define a mapping matrix Φ with $\Phi_{i,j} = k$ defining that training image $\mathbf{X}_{i,j}$ is associated with condition $\mathbf{A}_k$. Eqn. 1 suggests a constraint that we may use to solve for Φ: the optimal mapping matrix should result in the most sparsity for $\mathbf{E}_{i,j}$ or the lowest rank for $\mathbf{A}_k$, given the same reconstruction error. If we use the first criterion, the problem of finding Φ can be formulated as $\Phi = \underset{\Phi}{\operatorname{argmin}} \sum_{i,j} \|\mathbf{X}_{i,j} - \mathbf{C}_i - \mathbf{A}_{\Phi_{i,j}}\|_1$.

5. ACKNOWLEDGMENT

The work was supported in part by a grant (Grant No. 0845469) from the National Science Foundation. Any opinions, findings, and conclusions or recommendations expressed

(a) Experiment 1

#train per subject	20	15	10	5
Proposed	**100±0.00%**	**100±0.00%**	99.65±0.37%	97.49±0.21%
SRC	99.88±0.07%	99.88±0.07%	99.73±0.14%	**97.73±0.54%**
Volterrafaces	**100±0.00%**	**100±0.00%**	**100±0.00%**	95.83±4.16%
SUN	**100%**	99.84±0.11%	99.45±0.43%	95.75±0.49%

(b) Experiment 2

#train per subject	12	9	6	3
Proposed	**100±0.00%**	99.960.08%	99.17±0.15%	**94.70±0.20%**
SRC	99.91±0.16%	98.89±1.74%	96.90±3.73%	87.18±1.78%
Volterrafaces	**100±0.00%**	**100±0.00%**	**99.54±0.31%**	94.30±4.72%
SUN	**100±0.00%**	99.84±0.05%	98.53±0.29%	88.75±4.72%

(c) Experiment 3

#train per subject	40	30	20	10
Proposed	**99.98±0.03%**	**99.92±0.06%**	**99.24±0.06%**	**90.95±0.70%**
SRC	**99.98±0.03%**	99.45±0.03%	96.79±0.28%	86.98±0.16%
Volterrafaces	99.60±0.22%	98.37±0.47%	97.63±0.28%	89.72±1.45%
SUN	99.93±0.05%	99.38±0.14%	97.89±0.30%	88.29±0.02%

Table 2: The result on CMU-PIE dataset. Experiment 1: we pick the images with frontal pose (C27), which include 43 illumination conditions for each subject. We randomly pick $M = 20$ conditions for training and the remaining for testing. Experiment 2: we again only pick the image with frontal pose, but we randomly pick $M = 12$ conditions for training and the remaining for testing. Experiment 3: we use all the images from 5 near frontal poses (C05, C07, C09, C27, C29), which includes 153 conditions for each subject. We randomly pick $M = 40$ conditions for training and the remaining for testing..

in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

6. REFERENCES

[1] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: Design of dictionaries for sparse representation. *Proceedings of SPARS*, 5, 2005.

[2] J. F. Cai, E. J. Candes, and Z. Shen. A singular value thresholding algorithm for matrix completion. *preprint*, 2008.

[3] E. Candes and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 2009.

[4] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):643–660, 2001.

[5] D. Goldfarb and S. Ma. Convergence of fixed-point continuation algorithms for matrix rank minimization. *Foundations of Computational Mathematics*, pages 1–28, 2011.

[6] E. T. Hale, W. Yin, and Y. Zhang. Fixed-point continuation for ℓ_1-minimization: Methodology and convergence. *SIAM Journal on Optimization*, 19(3):1107–1130, 2008.

[7] C. Kanan and G. Cottrell. Robust classification of objects, faces, and flowers using natural image statistics. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2472 –2479, june 2010.

[8] A. Knyazev and M. Argentati. Principal angles between subspaces in an A-based scalar product: algorithms and perturbation estimates. *SIAM Journal on Scientific Computing*, 23(6):2008–2040, 2002.

[9] R. Kumar, A. Banerjee, and B. Vemuri. Volterrafaces: Discriminant analysis using volterra kernels. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 150 –155, june 2009.

[10] K. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 684–698, 2005.

[11] Z. Lin, M. Chen, L. Wu, and Y. Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *Arxiv preprint arXiv:1009.5055*, 2010.

[12] J. Liu, S. Chen, and X. Tan. Fractional order singular value decomposition representation for face recognition. *Pattern Recogn.*, 41:378–395, January 2008.

[13] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[14] A. Martinez and R. Benavente. The AR face database. Technical report, CVC Technical report, 1998.

[15] P. Nagesh and B. Li. A compressive sensing approach for expression-invariant face recognition. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1518 –1525, 2009.

[16] J. Portilla and E. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40(1):49–70, 2000.

[17] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression (PIE) database. In

Proceedings of the 5th International Conference on Automatic Face and Gesture Recognition, 2002.

[18] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma. Robust Principal Component Analysis: Exact Recovery of Corrupted Low-Rank Matrices via Convex Optimization. In *Advances in Neural Information Processing Systems 22*.

[19] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):210–227, 2008.

APPENDIX

A. PROOF OF THEOREM 1

Proposition 1 The sequences of $\tilde{Y}_{i,j}^{t+1}$, $\sum_i \hat{Y}_{i,j}^{t+1}$, $\sum_j \mathbf{Y}_{i,j}^{t+1}$ and $\dot{Y}_{i,j}^{t+1}$ are all bounded $\forall i,j$, where

$$\begin{aligned}
\mathbf{Y}_{i,j}^{t+1} &= \mu_{i,j}^t(\mathbf{X}_{i,j} - \mathbf{C}_i^{t+1} - \mathbf{A}_j^{t+1} - \mathbf{E}_{i,j}^{t+1}) + \mathbf{Y}_{i,j}^t \\
\hat{Y}_{i,j}^{t+1} &= \mu_{i,j}^t(\mathbf{X}_{i,j} - \mathbf{C}_i^{t} - \mathbf{A}_j^{t+1} - \mathbf{E}_{i,j}^{t+1} + \mathbf{Y}_{i,j}^t \\
\tilde{Y}_{i,j}^{t+1} &= \mu_{i,j}^t(\mathbf{X}_{i,j} - \mathbf{C}_i^{t} - \mathbf{A}_j^{t} - \mathbf{E}_{i,j}^{t+1}) + \mathbf{Y}_{i,j}^t \\
\dot{Y}_{i,j}^{t+1} &= \mu_{i,j}^t(\mathbf{X}_{i,j} - \dot{C}_i^{t+1} - \dot{A}_j^{t+1} - \dot{E}_{i,j}^{t+1}) + \dot{Y}_{i,j}^t
\end{aligned}$$

and $(\dot{C}^{t+1}, \dot{A}^{t+1}, \dot{E}^{t+1})$ is the optimal solution to the problem $\min_{\mathbb{C},\mathbb{A},\mathbb{E}} L(\mathbb{C},\mathbb{A},\mathbb{E},\dot{Y}^t,\mu^t)$ with $\dot{Y}^t = \{\dot{Y}_{i,j}^t\}_{i,j=1}^{N,M}$.

Proof Let's write the Lagrange function in 6 as:

$$\begin{aligned}
& L(\{\mathbf{C}_i^t\}_i, \{\mathbf{A}_j^t\}_j, \{\mathbf{E}_{i,j}^t\}_{i,j}, \{\mathbf{Y}_{i,j}^t\}_{i,j}, \{\mu^t\}_{i,j}) \\
= & \sum_{i,j} \|\mathbf{A}_j^t\|_* + \lambda_{i,j}\|\mathbf{E}_{i,j}^t\|_1 \\
+ & \frac{\mu_{i,j}^t}{2}\|\mathbf{X}_{i,j} - \mathbf{C}_i^t - \mathbf{A}_j^t - \mathbf{E}_{i,j}^t\|_F^2 \\
+ & < \mathbf{Y}_{i,j}^t, \mathbf{X}_{i,j} - \mathbf{C}_i^t - \mathbf{A}_j^t - \mathbf{E}_{i,j}^t >
\end{aligned}$$

For simplicity, we will use $L(\mathbb{C}^t,\mathbb{A}^t,\mathbb{E}^{t+1},\mathbb{Y}^t,\mu^t)$ instead of $L(\{\mathbf{C}_i^t\}_i,\{\mathbf{A}_j^t\}_j,\{\mathbf{E}_{i,j}^{t+1}\}_{i,j},\{\mathbf{Y}_{i,j}^t\}_{i,j},\{\mu^t\}_{i,j})$. The subgradient of $L(\mathbb{C}^t,\mathbb{A}^t,\mathbb{E},\mathbb{Y}^t,\mu_t)$ over $\mathbf{E}_{i,j}$ is

$$\lambda_{i,j}\partial\|\mathbf{E}_{i,j}\|_1 - \mu_{i,j}^t(\mathbf{X}_{i,j} - \mathbf{C}_i^t - \mathbf{A}_j^t - \mathbf{E}_{i,j}) - \mathbf{Y}_{i,j}^t$$

As $\mathbf{E}_{i,j}^{t+1}$ is optimal for the problem $\underset{\mathbf{E}_{i,j}}{\operatorname{argmin}} L(\mathbb{C}^t,\mathbb{A}^t,\mathbb{E},\mathbb{Y}^t,\mu_t)$

$$0 \in \lambda_{i,j}\partial\|\mathbf{E}_{i,j}^t\|_1 - \tilde{Y}_{i,j}^{t+1}$$

i.e., $\tilde{Y}_{i,j}^{t+1} \in \lambda_{i,j}\|\mathbf{E}_{i,j}^{t+1}\|_1$; and according to the Theorem 3 of [11], $\tilde{Y}_{i,j}^{t+1}$ is bounded $\forall i,j$. Similarly, we can also show that $\sum_i \hat{Y}_{i,j}^{t+1}$, $\sum_j \mathbf{Y}_{i,j}^{t+1}$ and $\dot{Y}_{i,j}^{t+1}$ are bounded $\forall i,j$.

Proposition 2 The sequences of $(\mathbb{C}^{t+1},\mathbb{A}^{t+1},\mathbb{E}^{t+1})$ is bounded.

Proof For Algorithm 1, we can find that:

$$\begin{aligned}
L(\mathbb{C}^{t+1},\mathbb{A}^{t+1},\mathbb{E}^{t+1},\mathbb{Y}^t,\mu^t) &\le L(\mathbb{C}^t,\mathbb{A}^{t+1},\mathbb{E}^{t+1},\mathbb{Y}^t,\mu^t) \\
\le L(\mathbb{C}^t,\mathbb{A}^t,\mathbb{E}^{t+1},\mathbb{Y}^t,\mu^t) &\le L(\mathbb{C}^t,\mathbb{A}^t,\mathbb{E}^t,\mathbb{Y}^t,\mu^t) \\
= L(\mathbb{C}^t,\mathbb{A}^t,\mathbb{E}^t,\mathbb{Y}^{t-1},\mu^{t-1}) &+ \sum_{i,j}\frac{\mu_{i,j}^{t-1}+\mu_{i,j}^t}{(\mu_{i,j}^t)^2}\|\mathbf{Y}_{i,j}^t - \mathbf{Y}_{i,j}^{t-1}\|_F^2
\end{aligned}$$

By boundedness of assumption that $\sum_{t=1}^{\infty}\mu_{i,j}^{t+1}(\mu_{i,j}^t)^{-2} < \infty$ and $\sum_j \mathbf{Y}_{i,j}^t$ $\forall i,j$, we have $L(\mathbb{C}^{t+1},\mathbb{A}^{t+1},\mathbb{E}^{t+1},\mathbb{Y}^t,\mu^t)$ is upper bounded. Thus $\sum_{i,j}\|\mathbf{A}_j^t\|_* + \lambda_{i,j}\|\mathbf{E}_{i,j}^t\|_1$ is bounded.

Proposition 3 The accumulation point $(\dot{C}^*,\dot{A}^*,\dot{E}^*)$ for sequences $(\dot{C}^{t+1},\dot{A}^{t+1},\dot{E}^{t+1})$ is optimal for the problem in Eqn. 5.

Proof For $(\dot{C}^{t+1},\dot{A}^{t+1},\dot{E}^{t+1})$, we have the following:

$$\begin{aligned}
& L(\dot{C}^{t+1},\dot{A}^{t+1},\dot{E}^{t+1},\dot{Y}^t,\mu^t) = \min_{\mathbb{C},\mathbb{A},\mathbb{E}} L(\mathbb{C},\mathbb{A},\mathbb{E},\dot{Y}^t,\mu^t) \\
\le & \min_{\mathbf{C}_i+\mathbf{A}_j+\mathbf{E}_{i,j}=\mathbf{X}_{i,j},\forall(i,j)} L(\mathbb{C},\mathbb{A},\mathbb{E},\dot{Y}^t,\mu^t) \\
\le & \min_{\mathbf{C}_i+\mathbf{A}_j+\mathbf{E}_{i,j}=\mathbf{X}_{i,j},\forall(i,j)} \sum_{i,j}\|\mathbf{A}_j\|_* + \lambda_{i,j}\|\mathbf{E}_{i,j}\|_1 \\
= & f^*
\end{aligned}$$

We also have:

$$\begin{aligned}
& \sum_{i,j}\|\dot{A}_j^{t+1}\|_* + \lambda_{i,j}\|\dot{E}_{i,j}^{t+1}\|_1 \\
= & L(\dot{C}^{t+1},\dot{A}^{t+1},\dot{E}^{t+1},\dot{Y}^t,\mu^t) - \sum_{i,j}\frac{\|\dot{Y}_{i,j}^t - \dot{Y}_{i,j}^{t-1}\|_F^2}{2\mu_{i,j}^t} \\
\le & f^* - \sum_{i,j}\frac{\|\dot{Y}_{i,j}^t - \dot{Y}_{i,j}^{t-1}\|_F^2}{2\mu_{i,j}^t} = f^* + O(\sum_{i,j}(\mu_{i,j}^t)^{-1})
\end{aligned}$$

where we use the knowledge that $\dot{Y}_{i,j}^{t+1}$ is bounded $\forall i,j$. Take $t\to\infty$, we have $\sum_{i,j}\|\dot{A}_j^*\|_* + \lambda_{i,j}\|\dot{E}_{i,j}^*\|_1 = f^*$. Using $(\dot{Y}_{i,j}^t - \dot{Y}_{i,j}^{t-1}) = \mu_{i,j}^{t-1}(\dot{X}_{i,j} - \dot{C}_i^{t-1} - \dot{A}_j^{t-1} - \dot{E}_{i,j}^{t-1})$ and boundedness of $\dot{Y}_{i,j}^{t+1}$ $\forall i,j$, we also have $\mathbf{X}_{i,j} - \dot{C}_i^* - \dot{A}_j^* - \dot{E}_{i,j}^* = 0$ $\forall i,j$. Thus $(\dot{C}^*,\dot{A}^*,\dot{E}^*)$ is the optimal solution for Eqn. 5.

By $\mathbf{X}_{i,j} - \mathbf{C}_i^{t+1} - \mathbf{A}_j^{t+1} - \mathbf{E}_{i,j}^{t+1} = \mu_{i,j}^t(\mathbf{Y}_{i,j}^{t+1} - \mathbf{Y}_{i,j}^t)$ and boundedness of $\mathbf{Y}_{i,j}^t$, we have $\lim_{t\to\infty}\mathbf{C}_i^{t+1} + \mathbf{A}_j^{t+1} + \mathbf{E}_{i,j}^{t+1} = \mathbf{X}_{i,j}$ $\forall i,j$, i.e., $(\mathbb{C}^{t+1},\mathbb{A}^{t+1},\mathbb{E}^{t+1})$ approaches to a feasible solution. In addition, we have

$$\|\sum_i \mathbf{A}_j^{t+1} - \mathbf{A}_j^t\|_F = \|\sum_i(\mu_{i,j}^t)^{-1}(\hat{Y}_{i,j}^{t+1} - \tilde{Y}_{i,j}^{t+1})\|_F$$

With the assumption $\sum_{t=1}^{\infty}(\mu_{i,j}^t)^{-1} < \infty$, boundedness of $\sum_i \hat{Y}_{i,j}^{t+1}$ and $\tilde{Y}_{i,j}^t$, $\mathbf{A}_j^{t+1}$ has a limit $\mathbf{A}_j^*$. Similarly:

$$\|\sum_j \mathbf{A}_j^{t+1} - \mathbf{A}_j^t + \mathbf{C}_i^{t+1} - \mathbf{C}_i^t\|_F = \|\sum_j(\mu_{i,j}^t)^{-1}(\mathbf{Y}_{i,j}^{t+1} - \tilde{Y}_{i,j}^{t+1})\|_F$$

Thus $\lim_{t\to\infty}\sum_j \mathbf{A}_j^{t+1} - \mathbf{A}_j^t + \mathbf{C}_i^{t+1} - \mathbf{C}_i^t = 0$. Since $\mathbf{A}_j^{t+1}$ has limit $\mathbf{A}_j^*$, then $\mathbf{C}_i^{t+1}$ has limit $\mathbf{C}_i^*$, then $\mathbf{E}_{i,j}^{t+1}$ has limit $\mathbf{X}_{i,j} - \mathbf{A}_j^* - \mathbf{C}_i^*$. So $(\mathbb{C}^*,\mathbb{A}^*,\mathbb{E}^*)$ is a feasible solution.

Considering the subgradients and the optimality of $\mathbf{E}_{i,j}^{t+1}$ and $\mathbf{A}_j^{t+1}$, we have $\tilde{Y}_{i,j}^{t+1} \in \partial\|\mathbf{E}_{i,j}^{t+1}\|_1$ and $\sum_i \hat{Y}_{i,j}^{t+1} \in \partial\|\mathbf{A}_j^{t+1}\|_*$. According to the property of subgradients:

$$\begin{aligned}
& (\sum_{i,j}\|\mathbf{A}_j^{t+1}\|_* + \lambda_{i,j}\|\mathbf{E}_{i,j}^{t+1}\|_1) - (\sum_{i,j}\|\dot{A}_j^{t+1}\|_* + \lambda_{i,j}\|\dot{E}_{i,j}^{t+1}\|_1) \\
\le & \sum_{i,j} - < \hat{Y}_{i,j}^{t+1}, \dot{A}_j^{t+1} - \mathbf{A}_j^{t+1} > - < \tilde{Y}_{i,j}^{t+1}, \dot{E}_{i,j}^{t+1} - \mathbf{E}_{i,j}^{t+1} > \\
= & \sum_{i,j} -\mu_{i,j}^t < \mathbf{E}_{i,j}^{t+1} - \mathbf{E}_{i,j}^t, \dot{A}_j^{t+1} - \mathbf{A}_j^{t+1} > \\
- & \frac{< \tilde{Y}_{i,j}^{t+1}, \tilde{Y}_{i,j}^{t+1} - \tilde{Y}_{i,j}^t >}{\mu_{i,j}^t} - \frac{< \tilde{Y}_{i,j}^{t+1}, \dot{Y}_{i,j}^{t+1} - \dot{Y}_{i,j}^t >}{\mu_{i,j}^t}
\end{aligned}$$

By Proposition 1 and 2 that $\mathbf{Y}_{i,j}^{t+1}$, $\dot{Y}_{i,j}^{t+1}$ are bounded; by Proposition 3 that $\sum_{i,j}\|\dot{A}_j^*\|_* + \lambda_{i,j}\|\dot{E}_{i,j}^*\|_1 = f^*$; and by assumption $\lim_{t\to\infty}\mu_{i,j}^t(\mathbf{E}_{i,j}^{t+1} - \mathbf{E}_{i,j}^t) = 0$, we have $\sum_{i,j}\|\mathbf{A}_j^*\|_* + \lambda_{i,j}\|\mathbf{E}_{i,j}^*\|_1 = f^*$. That is $(\mathbb{C}^*,\mathbb{A}^*,\mathbb{E}^*)$ is optimal for the problem in Eqn. 5. This completes the proof of Theorem 1.