Robust Subspace Discovery via Relaxed Rank Minimization

Xinggang Wang[1], Zhengdong Zhang[2], Yi Ma[3], Xiang Bai[1], Wenyu Liu[1], and Zhuowen Tu[4]

[1]Huazhong University of Science and Technology

[2]Massachusetts Institute of Technology

[3]Microsoft Research Asia

[4]Department of Cognitive Science, University of California, San Diego

Keywords: Subspace Learning, Low-Rank Optimization, Augmented Lagrange Multiplier, Object Discovery

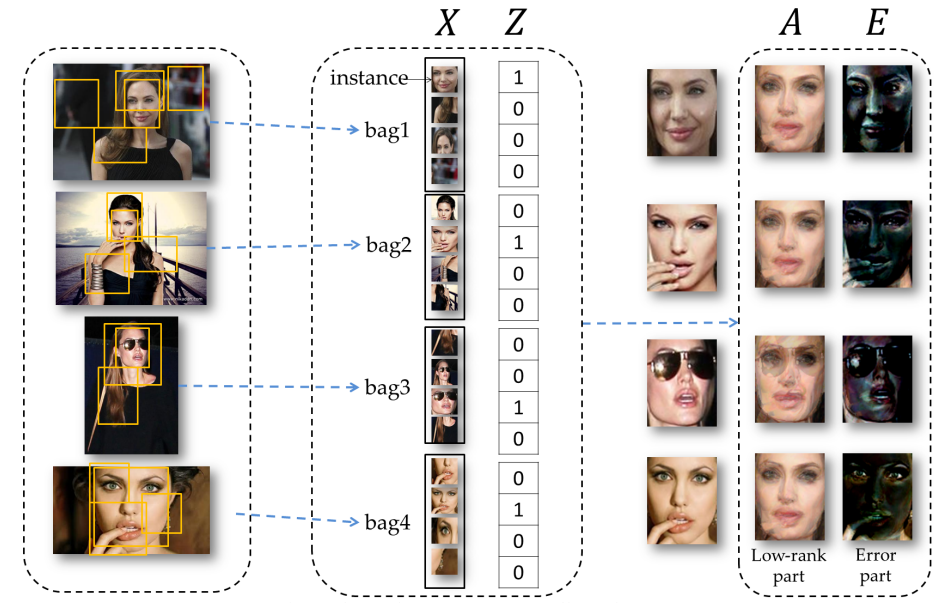
Abstract

This paper examines the problem of robust subspace discovery from input data samples (instances) in the presence of overwhelming outliers and corruptions. A typical example is the case where we are given a set of images; each image contains e.g. a

face at an unknown location of an unknown size; our goal is to identify/detect the face in the image and simultaneously learn its model. This paper explores a direction by employing a simple generative subspace model and proposes a new formulation to simultaneously infer the label information and learn the model via low-rank optimization. Solving this problem enables us to simultaneously identify the ownership of instances to the subspace and learn the corresponding subspace model. We give an efficient and effective algorithm based on the Alternating Direction Method of Multipliers (ADMM) method and provide extensive simulations and experiments to verify the effectiveness of our method. The proposed scheme can also be applied to tackle many related high-dimensional combinatorial selection problems.

1 Introduction

Subspace learning algorithms have recently been adopted for analyzing high-dimensional data in various problems (Jenatton et al., 2010; Wright et al., 2009; Wagner et al., 2009). Assuming the data are well aligned and lie in a low-dimensional linear subspace, these methods can deal with large sparse errors and learn the low-rank subspace of data. Other approaches such as (Elhamifar & Vidal, 2009; Liu et al., 2010; Luo et al., 2011; Favaro et al., 2011) have been proposed to cluster data into different subspaces. However, these methods may have difficulty in dealing with a class of unsupervised learning scenarios in which a large amount of outliers exist. In this paper, we propose a method to discover low-dimensional linear subspace from a set of data containing both inliers and a significant amount of outliers. Fig. 1 gives a typical problem setting of this paper,



(a) A collection of images containing common objects with detected salient image patches.

(b) Each patch is an instance. All patches from the same image form a bag. Each patch is associated with an indicator.

(c) The selected patches, the subspace, and sparse errors.

Figure 1: Pipeline of the object discovery task for subspace learning. Given a set of images, we first detect salient image patches (windows). All the image patches from the same image, considered as instances, form a bag. We assume the common object to appear as one instance in each bag. Our algorithm then detects and learns the subspace model for the common object while computing its residue. The symbols X, Z, A and E on the top of the figure correspond to the notations in our formulation given in Sec. 2.

as well as the pipeline of our proposed solution. Here, we are given a set of images and each image contains a common object (pattern). Our goal is to automatically identify the object and learn its subspace model.

In an abstract sense, we are given a set of data containing both inliers lying in a relative low-dimensional linear subspace and overwhelming outliers; in addition, the inliers may be corrupted by sparse errors. We make use of two constraints which have been adopted in the multiple instance learning (MIL) literature that (1) data is divided into different bags, and (2) at least one inlier exists in each bag; these two constrains

usually co-exist, e.g., as it is shown in Fig. 1(a) and (b). We may turn each image into a bag, consider image patches containing objects of the same category as inliers, and treat image patches from background or other categories as outliers. We aim to find the low-dimensional subspace and identify which data belongs to the subspace. Obviously, this problem is highly combinatorial and high-dimensional. Here we borrow the MIL concept, but assume no given negative bags in the training process, as in (Zhu et al., 2012); the original problem becomes a weakly-supervised subspace (pattern) discovery problem. We then transfer this problem into a convex optimization formulation, which can be effectively solved by the Alternating Direction Method of Multipliers (ADMM) (Gabay & Mercier, 1976; Boyd et al., 2011) method. In the proposed formulation, each instance is associated with an indicator indicating whether the instance is an inlier or an outlier; this is illustrated in Fig. 1(b); the indicators of instances are treated as latent variables; our objective function is to minimize both the rank of the subspace spanned by the selected instance and the ℓ_1 norm of the error in the selected instance; thus, by solving this optimization problem, we achieve the goal of discovering the low-dimensional subspace and identifying the instances belonging to the subspace. In Fig. 1(c), we show the discovered face subspace and errors of each face image. We deal with various object discovery tasks to demonstrate the advantage of our algorithm in the experiments. In the remainder of this section, we give the related work of our method.

Relations to Existing Work. In a nutshell, we are addressing a subspace learning problem, but a very challenging one. The existing scalable *robust subspace learning*

methods such as Robust Principal Component Analysis (RPCA) (Candes et al., 2011; Xu et al., 2012) can handle a sparse number of errors or outliers, while the dense error correction method in (Wright & Ma, 2010) can deal with dense corruptions under some restricted conditions. These do not, however, apply to our case since here the inlying instances are very few compared to the outliers and the inliers might even be partially corrupted. Nevertheless, our problem assumes an important additional structure: we know that there is at least one inlier in each set of samples. We will demonstrate that this extra information assists us in solving a seemingly impossible subspace discovery problem.

Robust Principal Component Analysis (RPCA) (Candes et al., 2011) has been successfully applied in background modeling (Candes et al., 2011), texture analysis (Z. Zhang et al., 2012) and face recognition (Jia et al., 2012). RPCA requires input data to be well-aligned, a prohibitive requirement in many real-world situations. To overcome this limitation, Robust Alignment by Sparse and Low-rank (RASL) (Peng et al., 2012) was proposed to automatically refine the alignment of input data, e.g. a set of images with a common pattern. However, RASL demands a good initialization of the common pattern on the same scale whereas here we are dealing with a much less constrained problem in which the common pattern (object) observes large scale differences at unknown locations in the images.

Robustly learning a model from noisy input data is a central problem in machine learning. In the multiple instance learning (MIL) literature (T. Dietterich et al., 1997), the input data is given in the form of bags; each positive bag contains at least one positive instance and each negative bag consists of all negative instances. MIL falls into

the class of weakly-supervised learning problems. In the MIL setting, the two central sub-tasks are to: (1) infer the missing label information, and (2) learn the correct model for the instances. The EM algorithm (Dempster et al., 1977) has been widely adopted for inferring missing labels for such MIL problems (Q. Zhang & Goldman, 2001a), and likewise for the latent SVM (Yu & Joachims, 2009). One could modify these methods; however, as we will see in our comparison, they lead to greedy iterative optimization that often produces suboptimal solutions. Recently, Lerman et al. (2012) proposed a convex optimization method called "REAPER" to learn subspace structure in datasets with large fractions of outliers. Compared to an approach like RPCA, this multiplicative approach has better thresholds for recovery in Gaussian outlier clouds. However, it is not robust to additional sparse corruption in the instances. The bMCL algorithm (Zhu et al., 2012) deals with cases of both one-class and multi-class object assumptions in object discovery with weak supervision. In a previous work, Sankaranarayanan & Davis (2012) has proposed one-class multiple instance learning (MIL) along the vein of discriminative models for target tracking. In (X. Wang et al., 2012), an EM approach of learning a low-rank subspace was proposed also for MIL with the one-class assumption. In this paper, we emphasize the task of robust subspace discovery with an explicit generative model for global optimization.

For the rest of the paper, we refer to the common pattern as an "object" and focus on the problem of object discovery. Given a set of images, our goal is to automatically discover the common object across all the images, which might appear at an unknown location with an unknown size.

Along the line of object discovery, many methods have also been proposed. In

(Russell et al., 2006), each image is treated as a bag of visual words, in which the common topics are discovered by the Latent Dirichlet Allocation. Other systems such as (Grauman & Darrell, 2006; Lee & Grauman, 2009) perform clustering on the affinity or correspondence matrix established via different approaches using different cues in the images. Although these existing methods achieve promising results on several benchmark datasets, they have notable limitations from several perspectives: 1) there often lacks a clear generative formulation and performance guarantee; 2) some systems are quite complicated with many cues, classifiers, and components involved; 3) they have a strong dependency on the discriminative models.

In contrast, this paper explores a different direction by employing a simple generative subspace model and proposes a new formulation to simultaneously infer the label information and learn the model via low-rank optimization; unlike EM-like approaches, our method is not sensitive to the initial conditions and is robust to severe corruptions. Although different from the classical robust PCA methods, our method inherits the same kind of robustness and efficiency from its convex formation and solution. Extensive simulations and experiments demonstrate the advantages of our method.

2 Formulation of Subspace Discovery

Given K bags of candidate object instances, we denote the number of instances in the kth bag as n_k. The total number of instances is $N = n_1 + \cdots + n_K$. Each instance is represented by a d-dimensional vector $x_i^{(k)} \in \mathbf{R}^d$. We may represent all the instances from one bag as columns of a matrix $X^{(k)} = \left[x_1^{(k)}, \ldots, x_{n_k}^{(k)}\right] \in \mathbf{R}^{d \times n_k}$. Furthermore,

we define $X = \left[X^{(1)}, \ldots, X^{(K)}\right] \in \mathbf{R}^{d \times N}$. By default, we assume that each bag contains at least one common object, and the rest are unrelated. To be concrete, we associate each object $x_i^{(k)}$ with a binary label $z_i^{(k)} \in \{0, 1\}$. $z_i^k = 1$ indicates that x_i^k is the common object. Similarly, we define $Z^{(k)} = \left[z_1^{(k)}, \ldots, z_{n_k}^{(k)}\right] \in \{0, 1\}^{n_k}$ and $Z = \left[Z^{(1)}, \ldots, Z^{(K)}\right]$. We assume that each bag contains at least one common object. So we have $\bigvee_{i=1}^{n_k} z_i^{(k)} = 1, \forall k \in [K]$, where $\bigvee$ is an or operator and $[K] = \{1, 2, \ldots, K\}$ is the set of positive integers less than or equal to K.

In general, different instances of the same object are highly correlated. It is reasonable to assume that such instances lie on a low-dimensional subspace $\Omega \subset \mathbf{R}^d$. This assumption can be verified empirically for real data. Fig. 4 (c) shows a comparison of the spectrum of a number of instances that are from the same object or from random image patches. Even if one applies a robust dimensionality reduction to the set of random image patches, their spectrum is still much higher than those for a common object.

However, due to many practical nuisance factors in real images, such as variation of pose, cast shadow, and occlusion, the observed instances of the common objects may no longer lie in a low-dimensional subspace. We may model all these contaminations as sparse errors added to the instances. So we could model each instance as $x = a + e$, where $a \in \Omega$ and e is a sparse vector. , and occlusion, the observed instances of the common objects may no longer lie in a low-dimensional subspace. We may model the contamination as sparse errors added to the instances. So we could model each instance as $x = a + e$, where $a \in \Omega$ and e is a sparse vector.

From the given K bags of instances $X = \left[X^{(1)}, \ldots, X^{(K)}\right]$, our goal is to find one (or more) instance from each bag so that all the selected instances form a low-

rank matrix A, subject to some sparse errors E. Or equivalently, we need to solve the following problem:

$$\begin{aligned} &\min_{A,E,Z} \operatorname{rank}(A) + \gamma \|E\|_0 \\ \text{s.t.} \quad &X\operatorname{diag}(Z) = A + E,\ \forall k \in [K]\ \textstyle\bigvee_{i=1}^{n_k} z_i^k = 1, \end{aligned} \tag{1}$$

where $\operatorname{diag}(Z)$ is an $N \times N$ block-diagonal matrix with K blocks $\left\{\operatorname{diag}\left(Z^{(k)}\right)\right\}$. To distinguish with the conventional (robust) "subspace learning" problems, we could refer to this problem as "subspace discovery".

3 Solution via Convex Relaxation

The problem in Eq. (1) is a highly combinatorial optimization problem that involves both continuous and integer variables. It is generally intractable when the dimensions d and N are large. Recent theory of RPCA (Candes et al., 2011) has suggested that rank and sparsity can be effectively minimized via their convex surrogates. So we could replace the above objective function $\operatorname{rank}(\cdot)$ with the nuclear norm $\|\cdot\|_*$ and ℓ_0 norm with ℓ_1 norm. Thus Eq. (1) is replaced with the following program:

$$\begin{aligned} &\min_{A,E,Z} \|A\|_* + \lambda \|E\|_1 \\ \text{s.t.} \quad &X\operatorname{diag}(Z) = A + E,\ \forall k \in [K]\ \textstyle\bigvee_{i=1}^{n_k} z_i^{(k)} = 1. \end{aligned} \tag{2}$$

Notice that although the objective function is now convex, the constraints on all the binary variables $z_i^{(k)}$ make this program remain intractable.

3.1 A Naive Iterative Solution

We can use a naive way (X. Wang et al., 2012) to tackle the problem in Eq. (2) by alternating between estimating Z and minimizing the objective with respect to the low-rank A and sparse E, in a spirit similar to the EM algorithm. With Z fixed, Eq. (2) becomes a convex optimization problem and can be solved by the RPCA method (Candes et al., 2011). Once the low-rank matrix A is known, one could perform ℓ_1-regression to evaluate the distance between each point and the subspace:

$$e_i^{(k)} = \min_w \|Aw - x_i^{(k)}\|_1. \tag{3}$$

Then within each bag we reassign 1 to a number of instances with errors below certain threshold and mark the rest as 0. One can iterate this process until convergence. As there are many outliers, this naive iterative method is very sensitive to initialization. So we have to run this naive method many times with random initializations and pick the best solution. This is similar to the popular RANSAC scheme for robust model estimation. Suppose there are m_k positive instances within the k-th bag, then the probability that RANSAC would succeed in selecting only the common objects is $\Pi_{k=1}^{K}(\frac{m_k}{n_k})$. Typically $\forall k, m_k/n_k \leq \frac{1}{5}$, so the probability that RANSAC succeeds vanishes exponentially as the number of objects increases. Even if the correct instances are selected, the above ℓ_1 regression does not always guarantee to work well when A contains errors. Nevertheless, with careful initialization and tuning, this method can be made to work for some relatively easy cases and datasets. It can be used as a baseline method to evaluate improved effectiveness of any new algorithm.

3.2 Relaxing Z

Instead of enforcing the variables Z to be binary $\{0, 1\}$, we relax it to have real value in $\mathbf{R}$. Also, the constraint $\bigvee_{i=1}^{n_k} z_i^{(k)} = 1$ can be relaxed with its continuous version $\mathbf{1}^T Z^{(k)} = 1$, which is linear. So the optimization problem becomes

$$\begin{aligned} &\min_{A,E,Z} \|A\|_* + \lambda \|E\|_1, \\ \text{s.t.} \quad &X\text{diag}\,(Z) = A + E, \;\; \forall k \in [K]\,, \mathbf{1}^T Z^{(k)} = 1. \end{aligned} \tag{4}$$

Although we do not explicitly require Z to be non-negative, it turns out that the optimal solution to the above program always ensures $Z^* \geq 0$, as the theorem below shows. This is due to some special nice properties of the nuclear norm and ℓ_1 norm. For our problem, this is incredibly helpful since the efficiency of the proposed algorithm based on Augmented Lagrangian Method decreases quickly as the number of constraints increases. This fact saves us from imposing N extra inequality constraints on the convex program!

Theorem 1 *If none of the columns of X is zero, the optimal solution Z^* of Eq. (4) is always non-negative.*

Proof Suppose we are given an optimal solution (A, E, Z) where Z have negative entries. Let us consider the triple $(\hat{A}, \hat{E}, \hat{Z})$ constructed in the following way:

$$\begin{aligned} \hat{Z}^{(k)} &= \tfrac{1}{1^T |Z^{(k)}|} |Z^{(k)}|, \\ \hat{A}^{(k)} &= \tfrac{1}{1^T |Z^{(k)}|} A^{(k)} \text{diag}\left(\text{sign}\left(Z^{(k)}\right)\right), \\ \hat{E}^{(k)} &= \tfrac{1}{1^T |Z^{(k)}|} E^{(k)} \text{diag}\left(\text{sign}\left(Z^{(k)}\right)\right) \end{aligned} \tag{5}$$

Since $X\mathrm{diag}\,(Z) = A+E$, obviously $X\mathrm{diag}\left(\hat{Z}\right) = \hat{A}+\hat{E}$ thus $(\hat{A}, \hat{E}, \hat{Z})$ is a feasible solution, and $\hat{Z}$ is non-negative. We will show that $\|\hat{A}\|_*+\lambda\|\hat{E}\|_1 < \|A\|_*+\lambda\|E\|_1$, thus contradicts to the fact that (A, E, Z) is optimal. Note that flipping the sign of any column of the matrix will not change the singular value of a matrix thus has no effect on the nuclear norm of it (if the svd of $W = U\Sigma V^*$, $\mathrm{diag}\,(\pm 1, \ldots, \pm 1)\, V$ is still orthogonal matrix). So if we construct another matrix A' such that $A'^{(k)} = A^{(k)}\mathrm{diag}\left(\mathrm{sign}\left(Z^{(k)}\right)\right)$, thus $\|A'\|_* = \|A\|_*$. Similarly we construct an E' and $\|E'\|_1 = \|E\|_1$. So $\hat{A}$ are $\hat{E}$ are just column-wise down-scaled version of A' and E'. Since for the k-th bag $1^T Z^{(k)} = 1$, $1^T|Z^{(k)}| > 1$ if and only if any entry of $Z^{(k)}$ is negative, otherwise $1^T|Z^{(k)}| = 1$. So the columns of A' and E' in the bags with negative $Z^{(k)}$ are down-scaled by a scalar $\alpha^k \in (0, 1)$. It can be proved that any down scaling of a non-zero column of a matrix will decrease the nuclear norm.

Lemma 2 *Given any matrix $Q \in \mathbf{R}^{m\times n}$, if $\tilde{Q}$ is Q with some column scaled by some scalar $\alpha \in (0, 1)$, then $\|\tilde{Q}\|_* < \|Q\|_*$.*

Proof:

Without loss of generality, we assume that the last column q_n get scaled. Let $Q = [Q_{n-1},\ q_n]$ and let $Q' = [Q_{n-1},\ 0]$ be the matrix by setting the last column to 0. The singular values of Q' are just the union of singular values of Q_{n-1} and an additional 0. Let $t = \min\{m, n\}$. According to [(Horn & Johnson, 2012) Theorem 7.3.9], $\sigma_1(Q) \geq \sigma_1(Q') \geq \sigma_2(Q) \geq \sigma_2(Q') \geq \ldots \geq \sigma_t(Q) \geq \sigma_t(Q') \geq 0$. So naturally $\|Q\|_* \geq \|Q'\|_*$ and the equality holds only if $\sigma_i(Q) = \sigma_i(Q'), \forall i \in [t]$, this is impossible since $\|Q\|_F^2 = \sum_i \sigma_i(Q)^2 > \|Q'\|_F^2 = \sum_i \sigma_i(Q')^2$. So we must have $\|Q\|_* > \|Q'\|_*$.

Note that $\tilde{Q} = \alpha Q + (1-\alpha)Q'$ and the nuclear norm $\|\cdot\|_*$ is convex, applying Jensen's inequality we have

$$\|\tilde{Q}\|_* \leq \alpha\|Q\|_* + (1-\alpha)\|Q'\|_* < \alpha\|Q\|_* + (1-\alpha)\|Q\|_* = \|Q\|_* \tag{6}$$

which concludes the proof. ∎

$\hat{A}$ can be viewed as a sequence of down-scaling on different columns of A, and each down-scaling will decreases the nuclear norm. The same goes for the ℓ_1 norm of the sparse error E. This shows that $\|\hat{A}\|_* + \lambda\|\hat{E}\|_1 < \|A\|_* + \lambda\|E\|_1$. This contradicts the assumption that (A, E, Z) is optimal. ∎

3.3 Solving $Eq.\ (4)$ via Alternating Direction Method of Multipliers

We apply the Alternating Direction Method of Multipliers (ADMM) method to solve Eq. (4). First write down the Augmented Lagrangian function:

$$\begin{aligned} &L(A, E, Z, Y_0, Y_1, \ldots, Y_K) \doteq \|A\|_* + \lambda\|E\|_1 \\ &+\langle Y_0, X\mathrm{diag}\,(Z) - A - E\rangle + \tfrac{\mu}{2}\|X\mathrm{diag}\,(Z) - A - E\|_F^2 \\ &+\textstyle\sum_{k=1}^{K}\left(\langle Y_k, \mathbf{1}^T Z^{(k)} - 1\rangle + \tfrac{\mu}{2}\|\mathbf{1}^T Z^{(k)} - 1\|_F^2\right). \end{aligned} \tag{7}$$

Instead of following the exact ALM procedure, we adopt the approximation scheme in (Boyd et al., 2011; Lin et al., 2010) which basically alternates the minimization with

respect to the three sets of variables in each iteration t:

$$\begin{cases} A_{t+1} = \underset{A}{\operatorname{argmin}} L(A, E_t, Z_t, Y_t, \mu_t) = \\ \underset{A}{\operatorname{argmin}} \|A\|_* + \frac{\mu_t}{2} \left\| X\operatorname{diag}(Z_t) - A - E_t + \frac{Y_{0,t}}{\mu_t} \right\|_F^2, \\ E_{t+1} = \underset{E}{\operatorname{argmin}} L(A_{t+1}, E, Z_t, Y_t, \mu_t) = \\ \underset{E}{\operatorname{argmin}} \|E\|_1 + \frac{\mu_t}{2} \left\| X\operatorname{diag}(Z_t) - A_{t+1} - E + \frac{Y_{0,t}}{\mu_t} \right\|_F^2, \\ Z_{t+1} = \underset{Z}{\operatorname{argmin}} L(A_{t+1}, E_{t+1}, Z, Y_t, \mu_t) = \\ \underset{Z}{\operatorname{argmin}} \left\| X\operatorname{diag}(Z) - A_{t+1} - E_{t+1} + \frac{Y_{0,t}}{\mu_t} \right\|_F^2 + \\ \qquad \ldots \sum_{k=1}^{K} \left\| \mathbf{1}^T Z^{(k)} - 1 + \frac{Y_{k,t}}{\mu_t} \right\|_F^2. \end{cases} \tag{8}$$

Fortunately, the above three minimization problems all have closed-form solutions. Details are given in as follows.

Let $\mathcal{S}_\epsilon(\cdot)$ be the following shrinkage operator.

$$\mathcal{S}_\epsilon(x) = \begin{cases} x - \varepsilon, & \text{if} \quad x > \varepsilon, \\ x + \varepsilon, & \text{if} \quad x < -\varepsilon, \\ 0, & \text{otherwise}, \end{cases} \tag{9}$$

If the svd of $X\operatorname{diag}(Z_t) - E_t + \frac{Y_{0,t}}{\mu_t} = U\Sigma V^*$, then the optimal A_{t+1} is given as $A_{t+1} = U\mathcal{S}_{\frac{1}{\mu}}(\Sigma)V^*$. For E_{t+1}, the optimal solution is $\mathcal{S}_{\frac{\lambda}{\mu_t}}\left(X\operatorname{diag}(Z_t) - A_{t+1} + \frac{Y_{0,t}}{\mu}\right)$. For Z, we can solve the original optimization via K independent ones for $Z^{(k)}$. Each sub

optimization is a typical *least square* problem for $Z^{(k)}$.

$$\begin{aligned} Z_{t+1}^{(k)} = \underset{Z^{(k)}}{\operatorname{argmin}} \left\| X^{(k)} \operatorname{diag}\left(Z^{(k)}\right) - A_{t+1}^{(k)} - E_{t+1}^{(k)} + \frac{Y_{0,t}^{(k)}}{\mu_t} \right\|_F^2 \\ \ldots + \left\| \mathbf{1}^T Z^{(k)} - 1 + \frac{Y_{k,t}}{\mu_t} \right\|_F^2 \end{aligned} \tag{10}$$

To be brief, let us denote $P^{(k)} = A_{t+1}^{(k)} + E_{t+1}^{(k)} - \mu_t^{-1} Y_{0,t}^{(k)} \in \mathbf{R}^{d \times n_k}$ and we mark the ith column of $P^{(k)}$ as $P_i^{(k)}$ and $Q^{(k)} = 1 - \mu_t^{-1} Y_{k,t} \in \mathbf{R}^1$. Furthermore, let's define $X_R^{(k)} = \begin{bmatrix} x_1^{(k)} & & \\ & \ddots & \\ & & x_{n_k}^{(k)} \end{bmatrix}$ and $P_R^{(k)} = \operatorname{vec}\left(P^{(k)}\right)$. Thus Eq. (10) can be rewritten as

$$\begin{aligned} Z_{t+1}^{(k)} &= \underset{Z^{(k)}}{\operatorname{argmin}} \left\| X_R^{(k)} Z^{(k)} - P_R^{(k)} \right\|_F^2 + \left\| \mathbf{1}^T Z^{(k)} - Q^{(k)} \right\|_F^2 \\ &= \left\| \begin{bmatrix} X_R^{(k)} \\ \mathbf{1}^T \end{bmatrix} Z_{t+1}^{(k)} - \begin{bmatrix} P_R^{(k)} \\ Q^{(k)} \end{bmatrix} \right\|_F^2 \end{aligned} \tag{11}$$

Directly applying the standard least square technique would require us to compute the pseudo-inverse of $X_R^{(k)} \in \mathbf{R}^{(dn_k+1) \times n_k}$, which is high-dimensional. So we perform a

trick so that pseudo-inverse is only calculated for a matrix in $\mathbf{R}^{n_k \times n_k}$.

$$
\begin{aligned}
Z_{t+1}^{(k)} = & \left(\begin{bmatrix} X_R^{(k)T} & \mathbf{1} \end{bmatrix} \begin{bmatrix} X_R^{(k)} \\ \mathbf{1}^T \end{bmatrix} \right)^{\dagger} \begin{bmatrix} X_R^{(k)T} & \mathbf{1} \end{bmatrix} \begin{bmatrix} P_R^{(k)} \\ Q \end{bmatrix} \\
= & ((X_R^{(k)T})X_R^{(k)} + \mathbf{1} \cdot \mathbf{1}^T)^{\dagger} (X_R^{(k)T} P_R^{(k)} + \mathbf{1} \cdot Q^{(k)}) \\
= & \begin{bmatrix} (x_1^{(k)})^T x_1^{(k)} + 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & (x_{n_k}^{(k)})^T x_{n_k}^{(k)} + 1 \end{bmatrix}^{\dagger} \\
& \cdots \begin{bmatrix} (x_1^{(k)})^T P_1 + Q^{(k)} \\ \vdots \\ (x_{n_k}^{(k)})^T P_{n_k} + Q^{(k)} \end{bmatrix}
\end{aligned}
\tag{12}
$$

After $A, E,$ and Z are updated, we only need to perform a gradient ascent on the dual variable Y_t:

$$
Y_{0,t+1} = Y_{0,t} + \mu_t \left(X \operatorname{diag}(Z_{t+1}) - A_{t+1} - E_{t+1} \right),
$$

$$
Y_{k,t+1} = Y_{k,t} + \mu_t \big(\mathbf{1}^T Z_{t+1}^{(k)} - 1 \big).
$$

And μ is also updated by $\mu_{k+1} = \rho \mu_k, \rho > 1$.

The complete algorithm is summarized in Algorithm 1 below.

The above alternating minimization process in Eq. (8) is known as Alternating Direction Method of Multipliers (ADMM) (Gabay & Mercier, 1976). A comprehensive survey of ADMM is given in (Boyd et al., 2011) and Lin et al. (2010) introduced it

Algorithm 1 : ADMM for robust subspace discovery

Input: Bags X and λ.

$Z_0 = 0, Y_0 = 0; \forall k \in [K], Y_{k,0}^{(k)} = 0; E_0 = 0; \mu_0 > 0; \rho > 1; t = 0$
while not converged **do**
  // Line 4-5 solve $A_{t+1} = \operatorname{argmin}_A L(A, E_t, Z_t, Y_t, \mu_t)$.
  $[U, \Sigma, V^*] = \text{svd}(X\text{diag}\,(Z_t) - E_t + \frac{Y_{0,t}}{\mu_t})$;
  $A_{t+1} = U\mathcal{S}_{\frac{1}{\mu}}(\Sigma)V^*$.
  // Line 7 solves $E_{t+1} = \operatorname{argmin}_E L(A_{t+1}, E, Z_t, Y_t, \mu_t)$.
  $E_{t+1} = \mathcal{S}_{\frac{\lambda}{\mu_t}}\left(X\text{diag}\,(Z_t) - A_{t+1} + \frac{Y_{0,t}}{\mu}\right)$.
  // Line 9-12 solve $Z_{t+1} = \operatorname{argmin}_Z L(A_{t+1}, E_{t+1}, Z, Y_t, \mu_t)$.
  for $k = 1 \to K$ **do**
    Obtain $Z_{t+1}^{(k)}$ via Eq. (12).
  end for
  $Z_{t+1} = [Z_{t+1}^{(1)}, \ldots, Z_{t+1}^{(K)}]$.
  // Line 14-16 update Y_{k+1} and μ_{k+1}.
  $Y_{0,t+1} = Y_{0,t} + \mu_t\,(X\text{diag}(Z_{t+1}) - A_{t+1} - E_{t+1})$.
  $Y_{k,t+1} = Y_{k,t} + \mu_t\left(\mathbf{1}^T Z_{t+1}^{(k)} - 1\right), \forall k \in [K]$.
  $\mu_{t+1} = \rho\mu_t$.
  $t \leftarrow t + 1$.
end while

Output: the converged values for (A, E, Z).

to the field of low-rank optimization. ADMM is not always guaranteed to converge to the optimal solution. If there are only two alternating terms, its convergence has been well-studied and established in (Gabay & Mercier, 1976). However, less is known for the convergence of cases where there are more than two alternating terms, despite the strong empirical observations (Z. Zhang et al., 2012). Tao & Yuan (2011) obtained convergence for a certain family of three-term alternation functions (applied to the noisy principal component pursuit problem). However, the scheme proposed in (Tao & Yuan, 2011) is different from the direct ADMM in Eq. (8), and it is also computationally heavy in practice. The convergence of the general ADMM remains an open problem although in practice a simple and fast implementation resides. Nevertheless, during the submission of this manuscript, there has been some latest development in the study of ADMM

(Shiqian Ma & Zou, 2013) that suggests one can design a convergent ADMM algorithm for the problem studied here. For instance, we could simply group the variables E and Z together and apply the proximal ADMM algorithm suggested in (Shiqian Ma & Zou, 2013), which results in slight modification to the proposed algorithm. Such proximal ADMM is guaranteed to converge. However, in practice, it might not converge faster than the proposed algorithm which exploits the natural separable structures in the augmented Lagrangian function among the three sets of variables, A, E, and Z. From our experience, the proposed algorithm works extremely well in practice and already meet our application goals.

4 Simulations and Experiments

In this section, we conduct simulations and experiments on both synthetic and real data for different applications for object discovery to verify the effectiveness of our method. We name the method described in Section 3.1 as Naive Iterative Method (NIM), and call the relaxed method as ADMM. In all our experiments, we set $\lambda = 1/\sqrt{d}$ where d is the dimension of instance feature.

4.1 Robust subspace learning simulation

In order to investigate the ability of the proposed ADMM method for recovering the indicators of inlier instances, in this experiment, we generate synthetic data with 50 bags; in each bag there are 10 instances which include 1 positive instance and 9 negative instances; the dimension of instance is $d = 500$. First, the positive instances are

generated by linearly combining r randomly generated $d \times 1$ vector whose entries are i.i.d. standard Gaussian, and the negative instances are independently randomly generated $d \times 1$ vector following i.i.d. normal distribution. Then, for every instance (no matter whether it is positive or negative), we normalize it to make sure its ℓ_2-norm is 1. At last, large sparse errors are added to all instances; the sparsity ratio of the error is s, and the values of the error is uniformly distributed in the range of $[-1, 1]$.

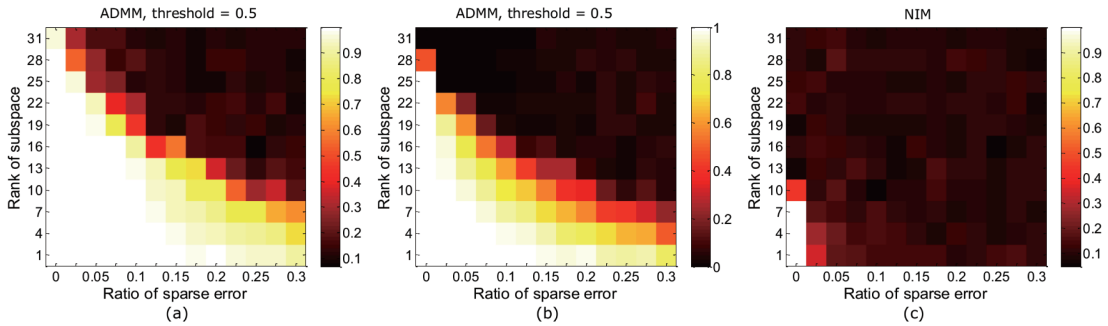
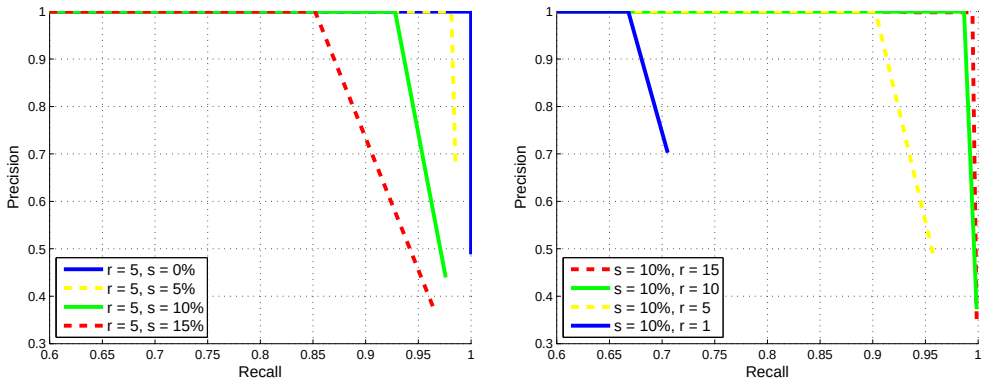


Figure 2: Accuracy of the recovered indicators when the sparsity level of error and the rank of subspace vary for ADMM at different thresholds and NIM. (a) shows the accuracy of ADMM at $\tau = 0.5$. (b) shows the accuracy of ADMM at $\tau = 0.99$. (c) shows the accuracy of NIM.

We investigate the performance when r (the rank of subspace) and s (the sparsity level the error) vary. r ranges from 1 to 31; s ranges from 0 to 0.3. For each test, we denote the ground-truth indicator vector as Z^*, the recovered indicator vector as $\hat{Z}$, the set of indexes whose corresponding values in Z^* are 1 as I^*, and the set of indexes whose corresponding values in $\hat{Z}$ are larger than a threshold $\tau \in [0, 1]$ as $\hat{I}$. *Accuracy* of the recovered indicators is defined as: $\texttt{accuracy} = \frac{\#(I^* \cap \hat{I})}{\#(I^*)}$. Given the ratio of sparsity and the rank of subspace, we run 5 random tests, and report the average accuracy of the recovered indicators for ADMM (under different τ) and NIM (randomly initialized) in Fig. 2. In Fig. 2(a), $\tau = 0.5$; 0.5 is a fair value, since there is only one

recovered indicator having the value larger than 0.5. In Fig. 2(b), $\tau = 0.99$, which is a very strict value; the accuracy matrix of the recovered indicators under $\tau = 0.99$ shows how exact of our relaxation in Sec. 3.2. The solution of NIM in Fig. 2(c) is discrete, and it is not necessary to set a threshold for the solution. Observing from the results in Fig. 2, NIM can work only when the positive instances are in a very low rank subspace in the situation of no error. No matter the threshold is 0.5 or 0.99, the working ranges of ADMM are strikingly larger than NIM. Comparing the results in Fig. 2(a) and (b), we find that it requires positive instances to be in a lower-dimensional subspace and contain less error if we want to exactly recover the indicators of them, say the indicator values of recovered instances are larger than 0.99.



(a) (b)

Figure 3: Precision-recall curves for the recovered indicator values by ADMM. (a): the rank of subspace is fixed at $r = 5$; the sparsity level of error $s = 0\%, 5\%, 10\%, 15\%$. (b): the sparsity level of error is fixed at $s = 10\%$; the rank of subspace $r = 1, 5, 10, 15$.

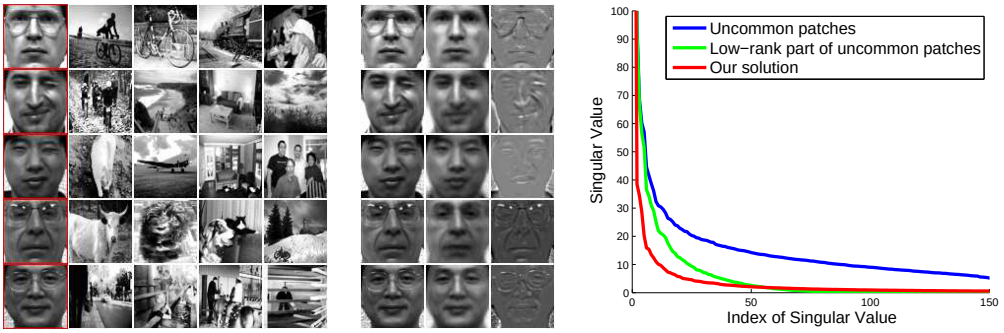
Multiple positive instances in one bag: The above simulations are focused on the situation where only one positive instance exists in each bag. Now we study how ADMM can deal with the situation where multiple instances exist in each bag. We put 3 positive

instances in each bag, which are randomly drawn from the same subspace and corrupted with large sparse errors. Thus, the three positive instances in each bag are not identical. For different values of r and s, we run ADMM for 5 times. The values of the recovered indicators are used for plotting the precision-recall curve. Results are shown in Fig. 3. Given a threshold τ, precision and recall are calculated by: $\texttt{precision} = \frac{\#(I^* \cap \hat{I})}{\#(\hat{I})}$ and $\texttt{recall} = \frac{\#(I^* \cap \hat{I})}{\#(\texttt{all positive instances})}$. As shown in Fig. 3(a), the performance of ADMM increases as the error becomes sparser; when there is no error, ADMM is able to perfectly identify all positive instances. Fig. 3(b) shows that it requires the subspace to have higher rank if more positive instances exist. When the rank of subspace is 15 and the sparsity level of error is 10%, ADMM is able to recover the indicators of 99% positive instances with 100% precision. It is observed that the current formulation for the subspace discovery problem in Eq. (4) has difficulty in dealing with multiple positive instances in some other settings.

4.2 Aligned face discovery among random image patches

We illustrate the effect of ADMM for object discovery by finding well aligned face images among lots of randomly selected image patches. Face images are from the Yale face dataset (Georghiades et al., 2001) which consists of 165 frontal faces of 15 individuals. Other image patches are randomly selected from the PASCAL image dataset (Everingham et al., 2011). We design bags and instances as following: the 165 face images are in 165 bags; other than the face image, in each bag, there are 9 image patches from PASCAL dataset; every image/patch is normalized to 64×64 pixels, and then vectorized to be a 4096 dimensional feature. Some of images in bags are shown in Fig. 4(a).

To evaluate the performance of this face recovery task, we get the images with the maximum indicator value in each bag, and then calculate the percentage of Yale faces among these images as the accuracy of face discovery. Because negative instances are randomly selected, we run the experiments 5 times. The average accuracy and the standard deviation of ADMM and NIM (randomly initialized) are 99.5±0.5% and 77.8±3.5% respectively. Some of the discovered faces by ADMM are shown in Fig. 4.(b). As it shows, facial expression and glasses are removed from the original images so that the repaired faces are better approximated by a low-dimensional subspace.



(a) (b) (c)

Figure 4: (a): Each row shows some sampled images in one bag. (b): Face discovery results by our algorithm. The first column shows the original patches in the bag; the second and third columns show recovered low-rank and sparse components respectively. (c): The distributions of singular values of our solution (in red), the low-rank component of uncommon patches via RPCA (Candes et al., 2011) (in green), and the original uncommon patches (in blue) in the experiment.

4.3 Object discovery on real word images

The task of object discovery is becoming a major topic in the recent years to reduce manual labeling effort to learn object class, and is a very challenging task. In this task, we are given a set of images, each containing one or more instances of the same object

class. In contrast to the fully supervised scenario, the locations of objects are not given. Different from subspace learning with simulated data, the appearance of an object varies a lot in real world images, which requires using image descriptors that are somewhat robust to substantial pose variations, e.g., HoG and LBP; moreover, location and scale of the objects are unknown that means the number of instances can rise to millions; to address this problem, we utilize an existing unsupervised salient object detection algorithm, e.g. (Feng et al., 2011), to reduce the number of instances per bag/image. The reason for us to choose the HoG and LBP descriptors for characterizing object is due to the observation that objects from the same category with the same view may not have similar color or texture. However they often have similar shapes. Both of the features, HoG and LBP, show good performance in supervised object detection (Felzenszwalb et al., 2010; Ahonen et al., 2006). The common shape structures of objects are the subspaces we want to discovery.

In the experiments of object discovery on real word images, we evaluate the proposed ADMM algorithm on four diverse datasets, which are PASCAL 2006 dataset (Everingham et al., 2006), PASCAL 2007 dataset (Everingham et al., 2007), Face Detection Data Set and Benchmark (FDDB) subset (Jain & Learned-Miller, 2010), and ETHZ Apple logo class (Ferrari et al., 2006), and compare ADMM with the state-of-the-art object discovery methods. Because different performance evaluation protocols are used, we give the experimental results for PASCAL 2006 and 2007 datasets, and for FDDB subset and ETHZ Apple logo class in two different parts.

PASCAL 2006 and 2007 datasets

The PASCAL 2006 and 2007 datasets are challenging and have been widely used as benchmarks for evaluating supervised object detection and image classification systems. For the object discovery task, we follow the protocol of (Deselaers et al., 2012). The performance is evaluated by the CorLoc measure which is the percentage of correctly localized objects, according to the PASCAL-criterion (window intersection-over-union > 0.5). Two subsets are taken from both PASCAL 2006 and 2007 datasets, which are called PASCAL06-6$\times$2, PASCAL06-all, PASCAL07-6$\times$2, and PASCAL07-all. PASCAL06-6$\times$2 contains 779 images from 12 classes/views; PASCAL06-all contains $2,184$ images from 33 classes/views; PASCAL07-6$\times$2 contains 463 images from 12 classes/views; and PASCAL07-all contains $2,047$ images from 45 classes/views. For more details about the datasets, as well as the evaluation protocol, please refer to (Deselaers et al., 2012).

Figure 5: Red rectangles: object discovery results of ADMM on the challenging PASCAL 2007. Green rectangles: annotated object ground-truth. From top to bottom: aeroplane, bicycle, bus, motorbike, plotted-plants and tv-monitors.

As mentioned previously, each image is considered as a bag, and a patch in the image detected by the salient object detector in (Feng et al., 2011) is considered as an instance. The parameter of score threshold in (Feng et al., 2011) is denoted as τ_s, which controls the number of salient objects detected. Standard HoG and LBP features are then extracted for each image patch. We let $\tau_s = 0.22$ for the PASCAL06-6×2 and PASCAL06-all datasets and use $\tau_s = 0.165$ for the PASCAL07-6×2 and PASCAL07-all datasets. We run the proposed ADMM method on these images and report the image patch with the maximum indicator value as the detected object. The results of ADMM are reported in Table. 1 and compared with the results of other methods in (Pandey & Lazebnik, 2011; Deselaers et al., 2012; Chum & Zisserman, 2007; Russell et al., 2006;

Lampert et al., 2009).

Table. 1 shows favorable results by our method compared with those by (Chum & Zisserman, 2007; Russell et al., 2006; Lampert et al., 2009). The state-of-the-art performances are reported in (Pandey & Lazebnik, 2011) and (Deselaers et al., 2012), which either uses extra bounding-box annotations or adopts complicated object models (Felzenszwalb et al., 2010). Here we study a generative model of subspace learning with a clean and effective solution. Fig. 5 shows some discovered objects on the PASCAL-all dataset.

Method	PASCAL06-6×2	PASCAL06-all	PASCAL07-6×2	PASCAL07-all
ESS (Lampert et al., 2009)	24	21	27	14
Russell et al. (2006)	28	27	22	14
Chum & Zisserman (2007)	45	34	33	19
ADMM (our method)	57	43	40	27
Deselaers et al. (2012)	**64**	**49**	50	28
Pandey & Lazebnik (2011)	N/A	N/A	**61**	**30**

Table 1: Object discovery performance evaluated by CorLoc on PASCAL 2006 and 2007 datasets.

FDDB subset and ETHZ Apple logo class

The FDDB subset contains 440 face images; the ETHZ Apple logo class contains 36 images with Apple logos. The appearance of objects and the background of the two datasets are quite diverse. In these two datasets, we only use HoG as the descriptor. Coordinating with the formulation in this paper, the low-rank term corresponds to the common shape structures of faces/apple-logos, since we use the HoG as the descriptor; the sparse error term corresponds to the occlusions and the appearance variations in

Figure 6: Face discovery results on the FDDB subset (Jain & Learned-Miller, 2010). The patches with the maximum score given by SD (Feng et al., 2011), bMCL (Zhu et al., 2012), NIM-SD (in blue) and ADMM are plotted in cyan, green, blue and red, respectively.

faces/apple-logos. We run ADMM and get the indicator value of each instance; for each image, the indicator value is normalized by dividing the maximum indicator value in the bag; the normalized indicator value is used as the score of each patch.

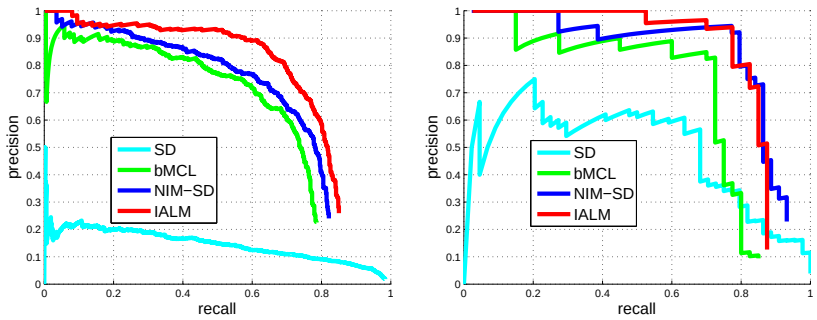
A selected patch is correct if it intersects with the ground truth object by more than half of their union (PASCAL criteria). Object discovery performance is evaluated by 1) precision-recall curves (Everingham et al., 2011), generated by varying the score threshold, 2) average precision (AP) (Everingham et al., 2011), computed by averaging multiple precisions corresponding to different recalls at regular intervals.

Method	FDDB subset	ETHZ Apple logo
SD	0.148	0.532
bMCL	0.619	0.697
NIM-SD	0.671	0.826
NIM-Rand	0.669	0.726
ADMM (our method)	**0.745**	**0.836**

Table 2: Performance comparison with APs for SD (Feng et al., 2011), bMCL (Zhu et al., 2012), NIM-SD, NIM-Rand, and ADMM on FDDB subset.

We compare ADMM with four methods: the baseline saliency detection method (SD) in (Feng et al., 2011), the state-of-the-art discriminative object discovery approach named bMCL in (Zhu et al., 2012), the naive iterative method initialized with saliency score (NIM-SD), and the naive iterative method with random initialization (NIM-Rand). The parameters of the four methods are turned to make sure they achieve their best performances. AP of NIM-Rand is the average value of 3 rounds. APs of all four methods are compared with ADMM in Table. 2 on both datasets. As we can see, ADMM significantly improves the results from the saliency detection and well outperforms all the other competing methods. The precision-recall curves of the four methods in Fig. 7 confirm this as well. SD method is a purely bottom-up approach. The other three methods make the assumption all of the input images contain a common object class of interest. The bMCL method (Zhu et al., 2012) is a discriminative method; it obtains state-of-the-art performance on image datasets with simple background, such as the SIVAL dataset (Rahmani et al., 2005). The images in the FDDB dataset are more cluttered, posing additional difficulty. Our methods, both ADMM and NIM-SD, are able to deal with cluttered background since they do not seek to discriminate the object from the background, which is an important property in tackling the problem object discovery/subspace learning. The patches with maximum scores by SD, bMCL, NIM-SD and ADMM are shown in Fig. 6.

In the experiments, we observe that there are situations in which ADMM might fail: (1) the objects are not contained in the detected salient image windows; (2) the objects observe large variation due to articulation or non-rigid transformation, which do not reside in a common low-rank space. Note that in this paper, we focus on the problem of



(a) (b)

Figure 7: Precision-recall curves of SD (Feng et al., 2011) (in cyan), bMCL (Zhu et al., 2012) (in green), NIM-SD (in blue) and ADMM (in red) in the task of object discovery in FDDB subset (Jain & Learned-Miller, 2010) (a) and ETHZ Apple logo class (Ferrari et al., 2006) (b).

subspace learning and make the assumption of the common pattern spanning a low-rank subspace.

4.4 Instance Selection for Multiple Instance Learning

In this experiment, we show how to apply the proposed ADMM to the traditional MIL problem (T. Dietterich et al., 1997). Our basic idea is to use ADMM to directly distinguish the positive instances from the negative instances in positive bags; the found positive instances together with all the negative instances from negative bags are used to train an instance-level classifier, e.g. SVM with RBF kernel, for the MIL task. In the testing stage, we use the learned instance-level SVM classifier for bag classification, based on a noisy-or model: if there exists any positive instance in a bag, the bag is identified as being positive, otherwise negative.

To use ADMM to distinguish positive instances from the negative instances, we follow the assumption that has been previously made in this paper: positive instances

Datasets	*Musk1*	*Musk2*	*Elephant*	*Fox*	*Tiger*	Average
MI-SVM	77.9	84.3	81.4	59.4	84.0	77.4
mi-SVM	87.4	83.6	82.0	58.2	78.9	78.0
MILES	86.3	87.7	-	-	-	-
EM-DD	84.8	84.9	78.3	56.1	72.1	75.2
PPMM Kernel	**95.6**	81.2	82.4	60.3	80.2	79.9
MI-CRF	87.0	78.4	85.0	65.0	79.5	79.0
ADMM (our method)	89.9±0.7	85.0±1.6	79.6±0.9	**65.4±1.2**	81.5±1.0	80.3
MIGraph	90.0±3.8	90.0±2.7	85.1±2.8	61.2±1.7	81.9±1.5	81.6
miGraph	88.9±3.3	**90.3±2.6**	**86.8±0.7**	61.6±2.8	**86.0±1.6**	**82.7**

Table 3: Performance comparison with per-class and average bag classification accuracies (%) for MI-SVM and mi-SVM in (Andrews et al., 2003), MILES (Chen et al., 2006), EM-DD (Q. Zhang & Goldman, 2001b), PPMM Kernel (H.-Y. Wang et al., 2008), MIGraph and miGraph in (Zhou et al., 2009), MI-CRF (Deselaers & Ferrari, 2010), and our method on five MIL benchmark datasets.

lie in a low-dimensional subspace. In practice, we collect all positive bags as input of ADMM algorithm in Algorithm 1 and obtain the indicator value of each instance. For each bag, the indicator value is normalized by dividing the maximum indicator value in the bag. Then, the instances whose normalized indicator values are larger than a upper threshold τ_u are labeled as positive instances; the instances whose normalized indicator values are less than a lower threshold τ_l are labeled as negative instances. In this experiment, we fix $\tau_u = 0.7$ and $\tau_l = 0.3$. The instances with normalized indicator values between 0.3 and 0.7 are omitted and not used for training the instance SVM classifier. When training the RBF kernel SVM, we adopt the LibSVM (Chang & Lin, 2011).

We evaluate the proposed method on five popular benchmark datasets, including *Musk1*, *Musk2*, *Elephant*, *Fox*, and *Tiger*. Detailed descriptions of the datasets can

be found in (T. G. Dietterich & Lathrop, 1997; Andrews et al., 2003). We compare our method with MI-SVM and mi-SVM in (Andrews et al., 2003), MILES (Chen et al., 2006), EM-DD (Q. Zhang & Goldman, 2001b), PPMM Kernel (H.-Y. Wang et al., 2008), MIGraph and miGraph in (Zhou et al., 2009), and MI-CRF (Deselaers & Ferrari, 2010) via ten times 10-fold cross validation and report the average accuracy and the standard deviation in Table. 3. Some of them were obtained in different studies and the standard deviations were not available. The average accuracy over the five tested datasets is reported in the right most column. The best performance on each compared item is noted in bold.

As shown in Table. 3, the best results are reported by MIGraph and miGraph, which exploit graph structure based on the affinities. We focus on comparing with mi-SVM which selects instance by maximizing margin between positive and negative instance under MIL condition via iterative SVM. This problem is non-convex and the optimization method of mi-SVM does not guarantee a local optima. Here, our method selects instance of a common subspace with a convex formulation and obtains promising results.

5 Conclusion

In this paper, we have proposed a robust formulation for unsupervised subspace discovery. We relax the highly combinatorial high-dimensional problem into a convex program and solve it efficiently with Augmented Lagrangian Multiplier method. Unlike the other approaches based on discriminative training, our proposed method can

discover objects of interest by utilizing the common patterns across input data. We demonstrate the evident advantage of our method over the competing algorithms in a variety of benchmark datasets. Our method suggests that an explicit low-rank sub-space assumption with a robust formulation naturally deals with a subspace discovery problem in presence of overwhelming outliers, which allows a rich emerging family of subspace learning methods to have a wider scope of applications; it enlarges the application range of the RPCA-based methods.

Acknowledgment

This work was supported by Microsoft Research Asia, NSF IIS-1216528 (IIS-1360566), NSF CAREER award IIS-0844566 (IIS-1360568), NSFC 61173120, NSFC 61222308, and Chinese Program for New Century Excellent Talents in University. Xinggang Wang was supported by Microsoft Research Asia Fellowship 2012. We thank John Wright for encouraging discussions. We thank David Wipf for valuable comments and Jun Sun for his helpful discussion on the proof of Theorem 1.

References

Ahonen, T., Hadid, A., & Pietikainen, M. (2006). Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(12), 2037–2041.

Andrews, S., Tsochantaridis, I., & Hofmann, T. (2003). Support vector machines for multiple-instance learning. In *Advances in neural information processing systems*

(pp. 561–568). MIT Press.

Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, *3*(1), 1–122.

Candes, E., Li, X., Ma, Y., & Wright, J. (2011, May). Robust principal component analysis? *Journal of the ACM*, *58*(3), 1–37.

Chang, C.-C., & Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *2*(3), 1–27.

Chen, Y., Bi, J., & Wang, J. Z. (2006). Miles: Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(12), 1931–1947.

Chum, O., & Zisserman, A. (2007). An exemplar model for learning object classes. In *Ieee conference on computer vision and pattern recognition* (pp. 1–8).

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *J. Royal Statist. Soc. Series B*, *39*(1), 1–38.

Deselaers, T., Alexe, B., & Ferrari, V. (2012). Weakly supervised localization and learning with generic knowledge. *International Journal of Computer Vision*, *100*(3), 275–293.

Deselaers, T., & Ferrari, V. (2010). A conditional random field for multiple-instance learning. In *Proceedings of the 26th international conference on machine learning* (p. 287-294).

Dietterich, T., Lathrop, R., & Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, *89*(1-2), 31–71.

Dietterich, T. G., & Lathrop, R. H. (1997). Solving the multiple-instance problem with axis-parallel rectangles. *Artificial Intelligence*, *89*, 31–71.

Elhamifar, E., & Vidal, R. (2009). Sparse subspace clustering. In *Ieee conference on computer vision and pattern recognition* (p. 2790-2797).

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2007). *The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.* http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2011). *The PASCAL Visual Object Classes Challenge (VOC) Results.* http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index .html.

Everingham, M., Zisserman, A., Williams, C. K. I., & Van Gool, L. (2006). *The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results.* http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf.

Favaro, P., Vidal, R., & Ravichandran, A. (2011). A closed form solution to robust subspace estimation and clustering. In *Ieee conference on computer vision and pattern recognition* (p. 1801-1807).

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(9), 1627–1645.

Feng, J., Wei, Y., Tao, L., Zhang, C., & Sun, J. (2011). Salient object detection by composition. In *International conference on computer vision* (pp. 1028–1035).

Ferrari, V., Tuytelaars, T., & Van Gool, L. (2006). Object detection by contour segment networks. In *European conference on computer vision* (p. 14-28).

Gabay, D., & Mercier, B. (1976). A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, *2*(1), 17–40.

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

Grauman, K., & Darrell, T. (2006). Unsupervised learning of categories from sets of partially matching image features. In *Ieee conference on computer vision and pattern recognition* (p. 19-25).

Horn, R. A., & Johnson, C. R. (2012). *Matrix analysis*. Cambridge University Press.

Jain, V., & Learned-Miller, E. (2010). *Fddb: A benchmark for face detection in unconstrained settings* (Tech. Rep. No. UM-CS-2010-009). University of Massachusetts, Amherst.

Jenatton, R., Obozinski, G., & Bach, F. (2010). Structured sparse principal component analysis. In *International conference on artificial intelligence and statistics* (p. 366-373).

Jia, K., Chan, T.-H., & Ma, Y. (2012). Robust and practical face recognition via structured sparsity. In *Eccv* (pp. 331–344). Springer.

Lampert, C., Blaschko, M., & Hofmann, T. (2009). Efficient subwindow search: a branch and bound framework for object localization. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 2129-2142.

Lee, Y., & Grauman, K. (2009). Foreground focus: Unsupervised learning from partially matching images. *International Journal of Computer Vision*, *85*(2), 143–166.

Lerman, G., McCoy, M. B., Tropp, J. A., & Zhang, T. (2012). Robust computation of linear models, or how to find a needle in a haystack. *CoRR*, *abs/1202.4044*.

Lin, Z., Chen, M., & Ma, Y. (2010). The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *Arxiv preprint arXiv:1009.5055*.

Liu, G., Lin, Z., & Yu, Y. (2010). Robust subspace segmentation by low-rank representation. In *Proceedings of the 26th international conference on machine learning* (p. 663-670).

Luo, D., Nie, F., Ding, C., & Huang, H. (2011). Multi-subspace representation and discovery. *Machine Learning and Knowledge Discovery in Databases*, 405–420.

Pandey, M., & Lazebnik, S. (2011). Scene recognition and weakly supervised object localization with deformable part-based models. In *Ieee international conference on computer vision* (pp. 1307–1314).

Peng, Y., Ganesh, A., Wright, J., Xu, W., & Ma, Y. (2012). Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2233–2246.

Rahmani, R., Goldman, S. A., Zhang, H., Krettek, J., & Fritts, J. E. (2005). Localized content based image retrieval. In *Acm sigmm international workshop on multimedia information retrieval* (pp. 227–236).

Russell, B., Freeman, W., Efros, A., Sivic, J., & Zisserman, A. (2006). Using multiple segmentations to discover objects and their extent in image collections. In *Ieee conference on computer vision and pattern recognition* (p. 1605-1614).

Sankaranarayanan, K., & Davis, J. (2012). One-class multiple instance learning and applications to target tracking. In *Proceedings of the asian conference on computer vision* (p. 126-139).

Shiqian Ma, L. X., & Zou, H. (2013). Alternating direction methods for latent variable gaussian graphical model selection. *Neural Computation*, *25*(8), 2172-2198.

Tao, M., & Yuan, X. (2011). Recovering low-rank and sparse components of matrices from incomplete and noisy observations. *SIAM Journal on Optimization*, *21*(1), 57–81.

Wagner, A., Wright, J., Ganesh, A., Zhou, Z., & Ma, Y. (2009). Towards a practical face recognition system: Robust registration and illumination via sparse representation. In *Ieee conference on computer vision and pattern recognition* (p. 597-604).

Wang, H.-Y., Yang, Q., & Zha, H. (2008). Adaptive p-posterior mixture-model kernels for multiple instance learning. In *Proceedings of the 25th annual international conference on machine learning* (pp. 1136–1143).

Wang, X., Zhang, Z., Ma, Y., Bai, X., Liu, W., & Tu, Z. (2012). One-class multiple instance learning via robust pca for common object discovery. In *Asian conference on computer vision* (p. 246-258).

Wright, J., & Ma, Y. (2010). Dense error correction via ℓ^1-minimization. *IEEE Transactions on Information Theory*, *56*(7), 3540-3560.

Wright, J., Yang, A., Ganesh, A., Sastry, S., & Ma, Y. (2009, February). Robust face recognition via sparse representation. *IEEE Transactions Pattern Analysis and Machine Intelligence*, *31*(2), 210-227.

Xu, H., Sanghavi, S., & Caramanis, C. (2012). Robust PCA via outlier pursuit. *IEEE Transactions on Information Theory*, *58*(5), 3047-3064.

Yu, C., & Joachims, T. (2009). Learning structural svms with latent variables. In *Proceedings of the 26th annual international conference on machine learning* (pp. 1169–1176).

Zhang, Q., & Goldman, S. A. (2001a). Em-dd: An improved multiple-instance learning technique. In *Advances in neural information processing systems* (p. 1073-1080).

Zhang, Q., & Goldman, S. A. (2001b). Em-dd: An improved multiple-instance learning technique. In *Advances in neural information processing systems* (pp. 1073–1080). MIT Press.

Zhang, Z., Ganesh, A., Liang, X., & Ma, Y. (2012). Tilt: transform invariant low-rank textures. *International Journal of Computer Vision*, *99*(1), 1–24.

Zhou, Z.-H., Sun, Y.-Y., & Li, Y.-F. (2009). Multi-instance learning by treating instances as non-iid samples. In *Proceedings of the 26th annual international conference on machine learning* (pp. 1249–1256).

Zhu, J., Wu, J., Wei, Y., Chang, E., & Tu, Z. (2012). Unsupervised object class discovery via saliency-guided multiple class learning. In *Ieee conference on computer vision and pattern recognition* (p. 3218-3225).