Matrix Completion for Resolving Label Ambiguity

Ching-Hui Chen, Vishal M. Patel and Rama Chellappa
Department of Electrical and Computer Engineering
and the Center for Automation Research, UMIACS
University of Maryland, College Park, USA
{ching,pvishalm,rama}@umiacs.umd.edu

Abstract

In real applications, data is not always explicitly-labeled. For instance, label ambiguity exists when we associate two persons appearing in a news photo with two names provided in the caption. We propose a matrix completion-based method for predicting the actual labels from the ambiguously labeled instances, and a standard supervised classifier can learn from the disambiguated labels to classify new data. We further generalize the method to handle the labeling constraints between instances when such prior knowledge is available. Compared to existing methods, our approach achieves 2.9% improvement on the labeling accuracy of the Lost dataset and comparable performance on the Labeled Yahoo! News dataset.

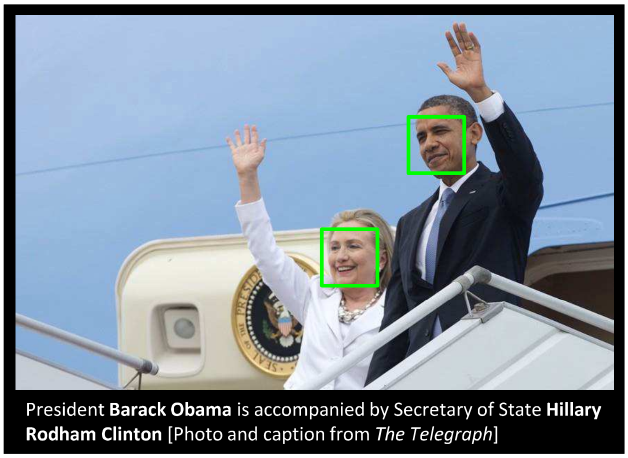

President **Barack Obama** is accompanied by Secretary of State **Hillary Rodham Clinton** [Photo and caption from *The Telegraph*]

Figure 1: The names in the captions are not explicitly associated with the face images appeared in the news photo.

1. Introduction

Learning a visual classifier requires a large amount of labeled images and videos for supervision. However, labeling images is expensive and time-consuming due to the significant amount of human efforts involved. As a result, brief descriptions such as tags, captions and screenplays accompanying the images and videos become important for training classifiers. Although such information is publicly available, it is not as explicitly labeled as human annotation. For instance, names in the caption of a news photo provide possible candidates for faces appearing in the image [2, 3, 16–18] (see Figure 1). The names in the screenplays are only weakly associated with faces in the shots [14]. The problem in which instead of a single label per instance, one is given a candidate set of labels, of which only one is correct is known as ambiguously labeled learning[1] [10, 22].

Various methods have been proposed in the literature for dealing with this ambiguously labeled learning problem. Some of these methods propose Expectation Maximization (EM)-like approaches to alternately disambiguate the labels and learn a classifier [1, 23]. Non-parametric methods have also been used to resolve the ambiguity by leveraging the inductive bias of learning methods [22]. For the ambiguously labeled training data the actual loss of mislabeling is not explicit. As a result, it is difficult to learn an effective discriminative model. Cour *et al.* [11, 12] proposed the partial 0/1 loss function for ambiguous labeling, which is a tighter upper bound for the actual loss as compared to 0/1 loss [28]. Subsequently, a discriminative classifier can be learned from the ambiguous labels by minimizing the partial 0/1 loss. Several dictionary-based methods have also been proposed in the literature for handing the partially labeled datasets [10, 26]. In particular, an EM-based dictionary learning approach was proposed in [10], where a confidence matrix and dictionary are updated in alternating iterations. Although dictionary-based methods are robust to occlusions and noise, the EM-based approach can be very sensitive to the selection of initial dictionary and also may suffer from suboptimal performance.

Luo *et al.* [25] generalize the ambiguously labeled learn-

[1] also known as partially labeled learning

ing problem addressed in [11] from single instances to a group of instances. The ambiguous loss considers the association between the group of identities and the candidate label vectors. The pairwise constraint between the instances (e.g. unique appearance of a subject) is accounted for when generating the candidate label vectors. Furthermore, Zeng *et al.* [27] use a Partial Permutation Matrix (PPM) to associate the identities in a group with the ambiguous labels. The pairwise constraint is encoded by restricting the structure of PPM. Assuming that instances of the same subject inferred by PPM can ideally form a low-rank matrix, the actual identity of an instance can be predicted by alternatively updating the low-rank subspace and PPM.

In recent years, the problem of completing a low-rank matrix with missing entries has gained a lot of attention. In particular, matrix completion methods have been shown to produce good results for multi-label image classification problems [15], [4]. In these methods, the underlying assumption is that the concatenation of feature vectors and their labels produce a low-rank matrix. Our work is motivated by these works. The proposed method, Matrix Completion for Ambiguity Resolving (MCar), takes the heterogeneous feature matrix, which is the concatenation of the label matrix and feature matrix, as input. We first show that the heterogeneous feature matrix is ideally low-rank in the absence of noise. This in turn, allows us to convert the labeling problem as a matrix completion problem. In contrast to multi-label learning, ambiguous labeling provides the clue that one of the labels in the candidate label set is the true label. This knowledge is utilized to regularize the labeling matrix in the heterogeneous feature matrix. This is essentially the main difference between our work and previously reported matrix completion techniques [15], [4]. Moreover, we generalize MCar to include the labeling constraints between the instances for practical applications. As shown by the recent success in low-rank matrix recovery [7], several prior works have developed robust methods for classification [9], [20]. The proposed method inherits the benefit of low-rank recovery and possesses the capability to resolve the label ambiguity via low-rank approximation of the heterogeneous matrix. As a result, our method is more robust compared to some of the existing discriminative ambiguous learning methods [11, 25]. The disambiguated labels from MCar are used to learn a supervised learning classifier, which can be used to classify new data.

This paper makes the following contributions:

1. We propose a matrix completion method where instances and their associated ambiguous labels are jointly considered for disambiguating class labels.

2. Our method can handle the group constraints between instances for practical applications.

3. We provide a geometric interpretation of the matrix completion framework from the perspective of recovering the potentially-separable convex hulls of each class.

2. The Proposed Framework

2.1. Notation

We use the following notations in this paper. Upper and lower bold letters indicate matrices and vectors, respectively. The matrix element $a_{i,j}$ denotes the entity in the i^{th} row and j^{th} column of matrix $\mathbf{A}$. $\mathbf{1}_n$ represents a column vector of size $n \times 1$ consisting of 1's as its entries. $\|\cdot\|_1$ and $\|\cdot\|_0$ denote the ℓ_1 norm and ℓ_0 norm, respectively. The Frobenius norm and the nuclear norm of $\mathbf{A}$ are defined as $\|\mathbf{A}\|_F = \left(\sum_{i,j}(a_{i,j})^2\right)^{\frac{1}{2}}$ and $\|\mathbf{A}\|_* = \sum_i \sigma_i(\mathbf{A})$, respectively where σ_i is the i^{th} singular value of $\mathbf{A}$. $(\cdot)^T$ denotes transposition operation. $|S|$ returns the cardinality in set S. $\mathcal{S}_a[b] = \text{sgn}(b)\max(|b| - a, 0)$ is the shrinkage operator. The concatenation of matrix $\mathbf{A}$ and $\mathbf{B}$ is defined as $\begin{bmatrix}\mathbf{A}\\ \mathbf{B}\end{bmatrix} = [\mathbf{A}; \mathbf{B}]$.

2.2. Problem Formulation

The ambiguously labeled data is denoted as $\mathcal{L} = \{(\mathbf{x}_j, L_j),\ j = 1, 2, \ldots, N\}$, where N is the number of instances. There are c classes, and the class labels are denoted as $\mathcal{Y} = \{1, 2, \ldots, c\}$. Note that $\mathbf{x}_j$ is the feature vector of the j^{th} instance, and its ambiguous labeling set $L_j \subseteq \mathcal{Y}$ consists of the candidate labels associated with the j^{th} instance. The true label of the j^{th} instance is $l_j \in L_j$. In other words, one of the labels in L_j is the true label of $\mathbf{x}_j$. The objective is to resolve the ambiguity in $\mathcal{L}$ such that each predicted label $\hat{l}_j$ of $\mathbf{x}_j$ matches its true label l_j.

2.3. Modeling of Ambiguously Labeled Data

We interpret the ambiguous labeling set L_j with soft labeling vector $\mathbf{p}_j$, where $p_{i,j}$ indicates the probability that instance j belongs to class i. This allows us to quantitatively assign the likelihood of each class the instance belongs to if such information is provided. Given the ambiguous label of the j^{th} instance, we assign each entry of $\mathbf{p}_j$ as

$$\begin{cases} p_{i,j} = (0, 1] & \text{if } i \in L_j, \\ p_{i,j} = 0 & \text{if } i \notin L_j, \end{cases} \quad j = 1, 2, \ldots, N, \tag{1}$$

where $\sum_{i=1}^{c} p_{i,j} = 1$. Without any prior knowledge, we assume equal probability for each candidate label. Let $\mathbf{P} \in \mathbb{R}^{c \times N}$ denotes the ambiguous labeling matrix with $\mathbf{p}_j$ in its j^{th} column. With this, one can model the ambiguous labeling as

$$\mathbf{P} = \mathbf{P}^0 + \mathbf{E}_P, \tag{2}$$

where $\mathbf{P}^0$ and $\mathbf{E}_P$ denote the true labeling matrix and the labeling noise, respectively. The j^{th} column vector of $\mathbf{P}^0$ is $\mathbf{p}^0_j = \mathbf{e}_{l_j}$, where $\mathbf{e}_{l_j}$ is the canonical vector corresponding to the 1-of-K coding of its true label l_j.

Similarly, assuming that the feature vectors are corrupted by some noise or occlusion, the feature matrix $\mathbf{X}$ with $\mathbf{x}_j$ in its j^{th} column can be modeled as

$$\mathbf{X} = \mathbf{X}^0 + \mathbf{E}_X, \tag{3}$$

where $\mathbf{X} \in \mathbb{R}^{m \times N}$ consists of N feature vectors of dimension m, $\mathbf{X}^0$ represents the feature matrix in the absence of noise and $\mathbf{E}_X$ accounts for the noise. Concatenating (2) and (3), we obtain a unified model of ambiguous labels and feature vectors, which can be expressed as

$$\begin{bmatrix} \mathbf{P} \\ \mathbf{X} \end{bmatrix} = \begin{bmatrix} \mathbf{P}^0 \\ \mathbf{X}^0 \end{bmatrix} + \begin{bmatrix} \mathbf{E}_P \\ \mathbf{E}_X \end{bmatrix}. \tag{4}$$

Let

$$\mathbf{H}_{obs} = \begin{bmatrix} \mathbf{P} \\ \mathbf{X} \end{bmatrix} \text{ and } \mathbf{E} = \begin{bmatrix} \mathbf{E}_P \\ \mathbf{E}_X \end{bmatrix} \tag{5}$$

denote the heterogeneous feature matrix and its noise, respectively. If we can show that $\mathbf{H}_{obs}$ is a low-rank matrix in the absence of noise, then we can use matrix completion methods for resolving the ambiguity in labeling. In the following section, we investigate the low-rank property of $\mathbf{H}_{obs}$.

2.4. Exploiting Rank of $\mathbf{H}_{obs}$

The column vectors of $\mathbf{X}_0$ can be partitioned into sets $S_1, S_2, \ldots, S_c$ based on their true labels. We assume that the elements of S_k form a convex hull C_k of n_k vertices. It is clear that $n_k \leq |S_k|$. The representative matrix of the k^{th}class, $\mathbf{D}_k \in \mathbb{R}^{m \times n_k}$, consists of vertices of C_k as its column vectors, and each column vector is treated as a representative of the k^{th}class. Therefore, according to the definition of a convex hull, a noise-free instance $\mathbf{x}_j^0$ from class k ($\mathbf{x}_j^0 \in C_k$) can be represented as

$$\mathbf{x}_j^0 = \mathbf{D}_k \mathbf{a}_{k,j}, \text{ where } \mathbf{a}_{k,j}^T \mathbf{1}_{n_k} = 1, \mathbf{a}_{k,j} \in \mathbb{R}_+^{n_k \times 1}. \tag{6}$$

Note that $\mathbf{a}_{k,j} \in \mathbb{R}_+^{n_k \times 1}$ is the coefficient vector associated with the representative matrix of the k^{th} class. As the true label of an instance is not known in advance, we can represent $\mathbf{x}_j^0$ as

$$\begin{aligned} \mathbf{x}_j^0 &= \mathbf{D}\mathbf{q}_j, \\ \mathbf{D} &= [\mathbf{D}_1 \ \mathbf{D}_2 \ \cdots \ \mathbf{D}_c], \\ \mathbf{q}_j &= [\mathbf{a}_{1,j}^T \ \mathbf{a}_{2,j}^T \ \cdots \ \mathbf{a}_{c,j}^T]^T, \ \mathbf{q}_j^T \mathbf{1} = 1, \end{aligned} \tag{7}$$

where $\mathbf{D} \in \mathbb{R}^{m \times (\sum_{i=1}^c n_i)}$ is the collective representative matrix, and $\mathbf{q}_j \in \mathbb{R}_+^{(\sum_{i=1}^c n_i) \times 1}$ is the associated coefficient vector.

According to (7), we can decompose $\mathbf{X}^0$ as

$$\mathbf{X}^0 = \mathbf{D}\mathbf{Q}. \tag{8}$$

The coefficient matrix $\mathbf{Q}$ in (8) is not unique as column vectors of $\mathbf{D}$ are not necessary linearly independent. However, we assume that an ideal decomposition $\mathbf{X}^0 = \mathbf{D}\mathbf{Q}^*$ satisfies the following condition

$$\begin{aligned} \mathbf{x}_j^0 = \mathbf{D}\mathbf{q}_j^*, \ \text{where } & \mathbf{a}_{k,j}^{*T} \mathbf{1}_{n_k} = 1, \ \mathbf{x}_j^0 \in S_k, \\ & \mathbf{a}_{l,j}^{*T} \mathbf{1}_{n_l} = 0, \ l \neq k, \end{aligned} \tag{9}$$

which implies that $\mathbf{x}_j^0$ is exclusively represented by $\mathbf{D}_k$ even though it is possible that it can be written as a linear combination of any other vertices from different classes.

With this, we can recover the true labels from

$$\mathbf{P}^0 = \mathbf{T}\mathbf{Q}^*, \tag{10}$$

where $\mathbf{T} = [\mathbf{e}_1 \mathbf{1}_{n_1}^T \ \mathbf{e}_2 \mathbf{1}_{n_2}^T \ \cdots \ \mathbf{e}_c \mathbf{1}_{n_c}^T]$ accumulates the coefficients associated with each matrix representative. Hence, the coefficient vector of dimension $\sum_{i=1}^c n_i$ is converted into labeling vector of dimension c. Using $\mathbf{P}^0 = \mathbf{T}\mathbf{Q}^*$ and $\mathbf{X}^0 = \mathbf{D}\mathbf{Q}^*$, we rewrite (4) as

$$\begin{bmatrix} \mathbf{P} \\ \mathbf{X} \end{bmatrix} = \begin{bmatrix} \mathbf{P}^0 \\ \mathbf{X}^0 \end{bmatrix} + \begin{bmatrix} \mathbf{E}_P \\ \mathbf{E}_X \end{bmatrix} = \begin{bmatrix} \mathbf{T} \\ \mathbf{D} \end{bmatrix} \mathbf{Q}^* + \begin{bmatrix} \mathbf{E}_P \\ \mathbf{E}_X \end{bmatrix}. \tag{11}$$

It is clear that

$$\begin{aligned} \operatorname{rank}([\mathbf{P}^0; \mathbf{X}^0]) &\leq \min\left(\operatorname{rank}([\mathbf{T}; \mathbf{D}]), \operatorname{rank}(\mathbf{Q}^*)\right) \\ &\leq \min\left(c + m, \sum_{k=1}^c n_k, N\right). \end{aligned} \tag{12}$$

Since the representatives in $\mathbf{D}$ only account for a subset of data samples, it is clear that $\sum_{k=1}^c n_k \leq N$. Therefore,

$$\operatorname{rank}([\mathbf{P}^0; \mathbf{X}^0]) \leq \min\left(c + m, \sum_{k=1}^c n_k\right). \tag{13}$$

The rank of $[\mathbf{P}^0; \mathbf{X}^0]$ is at most $\sum_{k=1}^c n_k$ if the dimension of feature vectors m is not less than the number of representatives in $\mathbf{D}$, i.e. $\sum_{k=1}^c n_k \leq m$. Hence, $[\mathbf{P}^0; \mathbf{X}^0]$ has the rank relatively smaller than N in the case of

$$N >> \min\left(c + m, \sum_{k=1}^c n_k\right).$$

Hence, we have justified the following proposition:

Proposition 1 *The heterogeneous feature matrix* $\mathbf{H}_{obs}$ *is low-rank in the absence of noise.*

Note that a similar result is also reported in [5] without the assumption of convex hull.

3. Matrix Completion for Ambiguity Resolving

According to (10), the true labeling matrix $\mathbf{P}^0$ can be recovered if $\mathbf{D}$ and $\mathbf{Q}^*$ are available. Unfortunately, obtaining $\mathbf{D}$ and $\mathbf{Q}^*$ based on the observed $\mathbf{P}$ and $\mathbf{X}$ is ill-posed. Following [15], we propose to resolve the ambiguity by recovering the underlying low-rank structure of the heterogeneous feature matrix. Hence, we transform the matrix decomposition problem to a matrix completion problem. For the ease of presentation, we start with solving a label assignment problem assuming that $\mathbf{X}$ is noise-free, i.e. $\mathbf{X} = \mathbf{X}^0$. The predicted labeling matrix $\mathbf{Y}$ can be estimated by solving the following rank minimization problem

$$\begin{aligned}
&\min_{\mathbf{Y}} \ \operatorname{rank}\left(\begin{bmatrix}\mathbf{Y}\\ \mathbf{X}^0\end{bmatrix}\right)\\
&\text{s.t.} \ \begin{bmatrix}\mathbf{Y}\\ \mathbf{X}^0\end{bmatrix} = \begin{bmatrix}\mathbf{P}\\ \mathbf{X}^0\end{bmatrix} - \begin{bmatrix}\mathbf{E}_P\\ \mathbf{0}\end{bmatrix},\\
&\qquad \mathbf{y}_j \in \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_c\}, j = 1, 2, \ldots, N,\\
&\qquad y_{i,j} = 0 \text{ if } p_{i,j} = 0.
\end{aligned} \tag{14}$$

The problem is to complete the labeling matrix $\mathbf{Y}$ via pursuing a low-rank matrix $\left[\mathbf{Y}; \mathbf{X}^0\right]$ subject to the constraints given by the ambiguous labels. The first constraint defines the feasible region of label assignment and the second constraint implies that an instance can only be labeled among its candidate labels. We cannot guarantee that the optimal solution of (14) always yields a perfect recovery of ambiguous labeling such that $\mathbf{Y}^* = \mathbf{P}^0$. Several factors contribute to our inability to resolve the ambiguity. For instance, if label 1 consistently presents in the candidate labeling set of each instance, assigning $\mathbf{e}_1$ for each column vector of $\mathbf{Y}$ yields a trivial solution. This issue is also addressed in [12], where an accurate learning from instances associated with two consistently co-occurring labels is impossible.

Note that $\mathbf{Y}^* = \mathbf{P}^0$ is one of the possible optimal solutions to (14). The solution may not be unique if any one of the instances belongs to more than one convex hull, i.e. the convex hulls from different classes overlap with each other. Hence, an instance can be ideally decomposed from either one of the convex hulls without further changing the rank of $[\mathbf{Y}; \mathbf{X}^0]$. This issue is analogous to the non-separable case of linear support vector machine (SVM). Nevertheless, it is our intention to seek $\mathbf{Y} = \mathbf{P}^0$ via solving (14) bearing the premises that 1) the ambiguous labeling carries rational information, and 2) data lies in sufficiently high-dimensional space such that convex hulls of each identity are separable [13].

Figure 2 illustrates the geometric interpretation of MCar with the convex hull representation. When each element in the ambiguous labeling set is trivially treated as the true label, the convex hulls of each class are erroneously expanded and the low-rank assumption of $\left[\mathbf{Y}; \mathbf{X}^0\right]$ does not hold. MCar exploits the underlying low-rank structure of $\left[\mathbf{Y}; \mathbf{X}^0\right]$, which is equivalent to reassigning the labels for those ambiguously labeled instances such that instances of the same class cohesively form a convex hull. Hence, each over-expanded convex hull shrinks to its actual contour, and the convex hulls become potentially separable. This is essentially different from the discriminative ambiguous learning methods that construct the hyperplane between ambiguously labeled instances by minimizing the ambiguous loss.

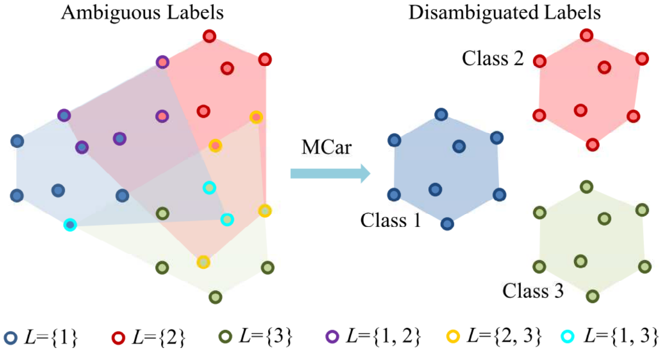


Figure 2: MCar reassigns the labels for those ambiguously labeled instances such that instances of the same subjects cohesively form potentially-separable convex hulls. The vertices of each convex hull are the representatives of each class, forming $\mathbf{D}_k$. The interior and outline of the circles are color-coded to represent three different classes and various ambiguous labels, respectively.

In the case when data is contaminated by sparse errors, the above optimization problem (14) can be reformulated as

$$\begin{aligned}
&\min_{\mathbf{Y}, \mathbf{E}_X} \ \operatorname{rank}(\mathbf{H}) + \lambda \|\mathbf{E}_X\|_0\\
&\text{s.t.} \ \mathbf{H} = \begin{bmatrix}\mathbf{Y}\\ \mathbf{Z}\end{bmatrix} = \begin{bmatrix}\mathbf{P}\\ \mathbf{X}\end{bmatrix} - \begin{bmatrix}\mathbf{E}_P\\ \mathbf{E}_X\end{bmatrix},\\
&\qquad \mathbf{y}_j \in \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_c\}, j = 1, 2, \ldots, N,\\
&\qquad y_{i,j} = 0 \text{ if } p_{i,j} = 0,
\end{aligned} \tag{15}$$

where $\mathbf{H}$ is the heterogeneous feature matrix in the absence of noise, and $\mathbf{Z}$ is the recovered feature matrix. The parameter $\lambda \in \mathbb{R}_+$ controls the rank of $\mathbf{H}$ and the sparsity of noise. The objective is to assign the predicted label $\mathbf{Y}$ and extract the sparse noise of $\mathbf{X}$ in pursuit of a low-rank $\mathbf{H}$. Figure 3 illustrates the ideal decomposition of heterogeneous feature matrix, where the underlying low-rank structure and the ambiguous labels are recovered simultaneously.

As (15) is a combinatorial optimization problem, we relax each column vector of $\mathbf{Y}$ in probability simplex in $\mathbb{R}^c$.

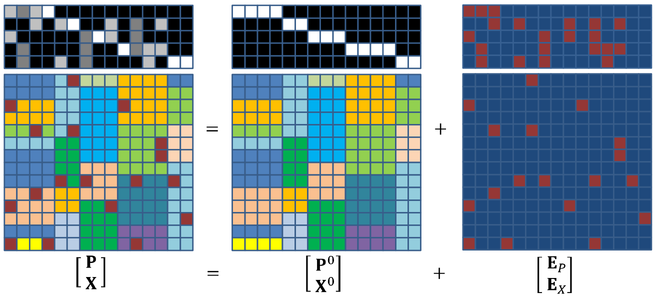

=
+
$\begin{bmatrix}\mathbf{P}\\ \mathbf{X}\end{bmatrix}$
=
$\begin{bmatrix}\mathbf{P}^0\\ \mathbf{X}^0\end{bmatrix}$
+
$\begin{bmatrix}\mathbf{E}_P\\ \mathbf{E}_X\end{bmatrix}$

Figure 3: Ideal decomposition of heterogeneous feature matrix using MCar. The underlying low-rank structure and the ambiguous labeling are recovered simultaneously.

The original formulation can be rewritten as

$$\begin{aligned} \min_{\mathbf{Y},\mathbf{E}_X} \; & \operatorname{rank}(\mathbf{H}) + \lambda\|\mathbf{E}_X\|_0 + \gamma\|\mathbf{Y}\|_0 \\ \text{s.t. } & \mathbf{H} = \begin{bmatrix}\mathbf{Y}\\ \mathbf{Z}\end{bmatrix} = \begin{bmatrix}\mathbf{P}\\ \mathbf{X}\end{bmatrix} - \begin{bmatrix}\mathbf{E}_P\\ \mathbf{E}_X\end{bmatrix}, \\ & \mathbf{1}_c^T\mathbf{Y} = \mathbf{1}_N^T, \;\; \mathbf{Y} \in \mathbb{R}_+^{c\times N}, \\ & y_{i,j} = 0 \text{ if } p_{i,j} = 0, \end{aligned} \tag{16}$$

where $\gamma \in \mathbb{R}_+$ encourages the sparsity of $\mathbf{Y}$ such that the original discrete feasible region can be well approximated. From the perspective of convex hull representation, such relaxation allows each instance to be represented from more than one set of representative matrix $\mathbf{D}_k$, while it will be penalized by the non-sparsity of $\mathbf{Y}$. Consequently, the predicted label of instance j can be obtained as

$$\hat{l}_j = \arg\max_{i\in\mathcal{Y}} \; y_{i,j}. \tag{17}$$

3.1. Optimization

The augmented Lagrangian method (ALM) has been extensively adopted for solving low-rank problems [7, 24]. In this section, we propose to incorporate the ALM with the projection step [4, 15] to solve the optimization problem for resolving the label ambiguity.

In order to decouple $\mathbf{Y}$ in the first and third terms of the objective function in (16), we replace $||\mathbf{Y}||_0$ with $\|\mathbf{P} - \mathbf{E}_P\|_0$ and rewrite (16) as

$$\begin{aligned} \min_{\mathbf{Y},\mathbf{E}_X} \; & \operatorname{rank}(\mathbf{H}) + \lambda\|\mathbf{E}_X\|_0 + \gamma\|\mathbf{P} - \mathbf{E}_P\|_0 \\ \text{s.t. } & \mathbf{H} = \begin{bmatrix}\mathbf{Y}\\ \mathbf{Z}\end{bmatrix} = \begin{bmatrix}\mathbf{P}\\ \mathbf{X}\end{bmatrix} - \begin{bmatrix}\mathbf{E}_P\\ \mathbf{E}_X\end{bmatrix}, \\ & \mathbf{1}_c^T\mathbf{Y} = \mathbf{1}_N^T, \;\; \mathbf{Y} \in \mathbb{R}_+^{c\times N}, \\ & y_{i,j} = 0 \text{ if } p_{i,j} = 0. \end{aligned} \tag{18}$$

Following the procedure of ALM, we relax the first constraint in (18) and reformulate it as

$$\begin{aligned} \min_{\mathbf{H},\mathbf{E},\mathbf{\Lambda},\mu} \; & \ell(\mathbf{H},\mathbf{E},\mathbf{\Lambda},\mu) \\ \text{s.t. } & \mathbf{1}_c^T\mathbf{Y} = \mathbf{1}_N^T, \;\; \mathbf{Y} \in \mathbb{R}_+^{c\times N}, \\ & y_{i,j} = 0 \text{ if } p_{i,j} = 0, \end{aligned} \tag{19}$$

where $\mu \in \mathbb{R}_+$ and $\mathbf{\Lambda} \in \mathbb{R}^{(c+m)\times N}$. The Lagrangian is expressed as

$$\begin{aligned} \ell(\mathbf{H},\mathbf{E},\mathbf{\Lambda},\mu) = & \operatorname{rank}(\mathbf{H}) + \lambda\|\mathbf{E}_X\|_0 + \gamma\|\mathbf{P} - \mathbf{E}_P\|_0 \\ & + \langle \mathbf{\Lambda}, \mathbf{H}_{obs} - \mathbf{H} - \mathbf{E} \rangle + \frac{\mu}{2}\|\mathbf{H}_{obs} - \mathbf{H} - \mathbf{E}\|_F^2 . \end{aligned} \tag{20}$$

In order to make the optimization problem feasible, we approximate the rank with the nuclear norm and the ℓ_0 norm with the ℓ_1 norm [6]. Thus, we solve the following formulation as the convex surrogate of (19)

$$\min_{\mathbf{H},\mathbf{E},\mathbf{\Lambda},\mu} \; \ell_R(\mathbf{H},\mathbf{E},\mathbf{\Lambda},\mu) \tag{21}$$

$$\text{s.t. } \mathbf{1}_c^T\mathbf{Y} = \mathbf{1}_N^T, \;\; \mathbf{Y} \in \mathbb{R}_+^{c\times N}, \tag{22}$$

$$y_{i,j} = 0 \text{ if } p_{i,j} = 0, \tag{23}$$

where the Lagrangian is represented as

$$\begin{aligned} \ell_R(\mathbf{H},\mathbf{E},\mathbf{\Lambda},\mu) = & \|\mathbf{H}\|_* + \lambda\|\mathbf{E}_X\|_1 + \gamma\|\mathbf{P} - \mathbf{E}_P\|_1 \\ & + \langle \mathbf{\Lambda}, \mathbf{H}_{obs} - \mathbf{H} - \mathbf{E} \rangle + \frac{\mu}{2}\|\mathbf{H}_{obs} - \mathbf{H} - \mathbf{E}\|_F^2 . \end{aligned} \tag{24}$$

The ALM operates in the sense that $\mathbf{H}$, $\mathbf{E}_P$, and $\mathbf{E}_X$ can be solved alternately by fixing other variables. In each iteration, we employ a similar projection technique used in [4, 15] to enforce $\mathbf{Y}$ to be feasible. The entire procedure for solving (21) is summarized in Algorithm 1.

4. Labeling Constraints between Instances

In practical applications, several ambiguously labeled instances can appear in the same venue. As a result, pairwise relations between instances can be utilized to assist ambiguity resolving. For example, two persons in a news photo should not be identified as the same subject even though both of them are ambiguously labeled in the caption. Such prior knowledge can be easily incorporated by restricting the feasible region of the labeling matrix. Moreover, it is essential to handle the open set problem, where there are some instances whose identities never appear in the labels. These unrecognized instances can be treated as null class.

In this section, we show how MCar's formulation can be extended to associate the identities in news photos when the names are provided in the captions. We assume all the instances (face images) are collected from the K groups (photos), and G_k is the set consisting the indices of the instances (face images) appearing in the k^{th} group (photo). Note that instances (face images) from the same group (photo)

Algorithm 1 The optimization algorithm for (21)

```
Input: $\mathbf{P} \in \mathbb{R}^{c \times N}$, $\mathbf{X} \in \mathbb{R}^{m \times N}$, $\lambda$, and $\gamma$.
1:  Initialization: $\mathbf{Y} = \mathbf{0}$, $\mathbf{Z} = \mathbf{0}$, $\mu > 0, \mu_{\max} > 0$, $\rho > 1$, $\mathbf{\Lambda} = [\mathbf{\Lambda}_P; \mathbf{\Lambda}_X] = \mathbf{H}_{obs}/\|\mathbf{H}_{obs}\|_2$;
2:  while not converged do
3:      $\mathbf{E}_P = \mathbf{P} - \mathcal{S}_{\gamma\mu^{-1}}[\mathbf{Y} - \mu^{-1}\mathbf{\Lambda}_P]$;
4:      $\mathbf{E}_X = \mathcal{S}_{\lambda\mu^{-1}}[\mathbf{X} - \mathbf{Z} + \mu^{-1}\mathbf{\Lambda}_X]$;
5:      $(\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}) = \text{svd}\left(\mathbf{H}_{obs} - \mathbf{E} + \mu^{-1}\mathbf{\Lambda}\right)$;
6:      $\mathbf{H} = \mathbf{U}\mathcal{S}_{\mu^{-1}}[\mathbf{\Sigma}]\mathbf{V}^T$;
7:      $\mathbf{\Lambda} = \mathbf{\Lambda} + \mu\left(\mathbf{H}_{obs} - \mathbf{H} - \mathbf{E}\right)$;
8:      $\mu = \min(\rho\mu, \mu_{\max})$;
9:      Project $\mathbf{Y}$:
10:     ▷ Line: 11: Projection for (23)
11:     $y_{i,j} = 0$ if $p_{i,j} = 0, \forall i, j$;
12:     ▷ Line: 13-14: Projection for (22)
13:     $\mathbf{Y} = \max(\mathbf{Y}, 0)$;
14:     $\mathbf{y}_j = \mathbf{y}_j/\|\mathbf{y}_j\|_1, \forall j$;
15: end while
Output: $(\mathbf{H}, \mathbf{E})$
```

share the same ambiguous labels provided by their associated caption. Without loss of generality, we assume that the c^{th} class corresponds to the null class. Considering the prior knowledge, the original formulation addressed in (16) can be reformulated as

$$\min_{\mathbf{Y},\mathbf{E}_X} \ \text{rank}(\mathbf{H}) + \lambda\|\mathbf{E}_X\|_0 + \gamma\|\mathbf{Y}\|_0 \tag{25}$$

$$\text{s.t. } \mathbf{H} = \begin{bmatrix}\mathbf{Y}\\ \mathbf{Z}\end{bmatrix} = \begin{bmatrix}\mathbf{P}\\ \mathbf{X}\end{bmatrix} - \begin{bmatrix}\mathbf{E}_P\\ \mathbf{E}_X\end{bmatrix},$$

$$\mathbf{1}_c^T\mathbf{Y} = \mathbf{1}_N^T, \ \ \mathbf{Y} \in \mathbb{R}_+^{c \times N}, \tag{26}$$

$$y_{i,j} = 0 \text{ if } p_{i,j} = 0, \ \ i = 1, 2, \ldots, c-1, \tag{27}$$

$$\sum_{j \in G_k} \sum_{i=1}^{c-1} y_{i,j} \geq 1 \text{ if } \bigcup_{j \in G_k} L_j \neq \{c\}, \forall k, \tag{28}$$

$$\sum_{j \in G_k} y_{i,j} \leq 1, \ \ i = 1, 2, \ldots, c-1, \ \ \forall k. \tag{29}$$

Constraints (26) and (27) are inherited from the original formulation. The constraint in (28), assumes that there is at least one non-null identity in a photo unless all the instances in a photo are explicitly labeled as null. This constraint is enforced to avoid the trivial solution that all the instances are treated as null class. A similar constraint has been considered by [25] and [27] via restricting the candidate labeling set and confining the feasible space of PPM, respectively. The constraint in (29) enforces the uniqueness of non-null identities. Note that this framework can be easily tailored to handle other prior knowledge (e.g. must/cannot-link constraints, prior statistics) by regularizing the labeling matrix. This problem can be solved by following the similar relaxation procedures for solving (16).

5. Experimental Results

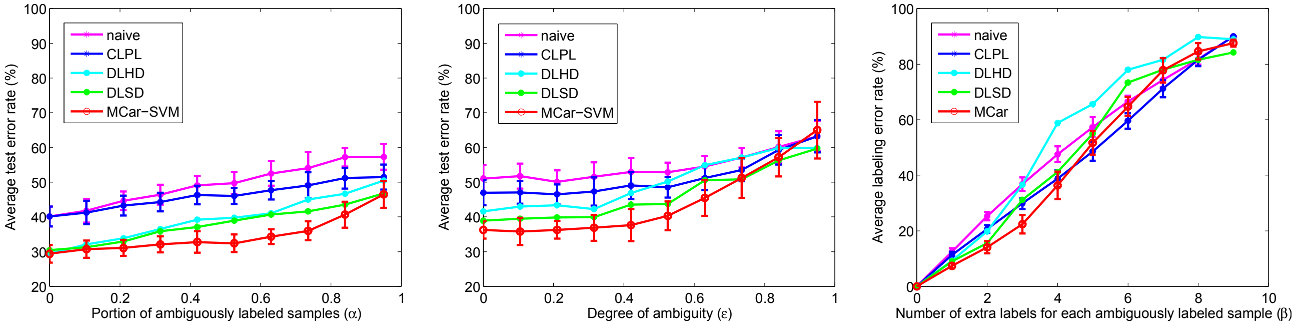
We use the Labeled Faces in the Wild (LFW) dataset [21] with synthesized ambiguous labels to evaluate the performance of our method under various controlled parameter settings. Furthermore, we use the *Lost* dataset [11] and the Labeled Yahoo! News dataset [2, 19] to demonstrate the effectiveness of our method in real-world applications. For the datasets provided with face images, we use face images in gray scale of range $[0, 1.0]$. Each instance is preprocessed with histogram equalization and converted into a column feature vector.

It is interesting to observe that (15) becomes asymptotically similar to the formulation of Robust Principle Component Analysis (RPCA) [7] as the dimension of the data feature is far greater than the number of classes. Motivated by this fact, we fix $\lambda = 1/\sqrt{\max(c+m, N)}$, which is the tradeoff parameter suggested in RPCA. In all the experiments, we use $\gamma = 2\lambda$ to encourage stronger sparsity of the labeling vector than that of the feature noise.

5.1. Experiments with the LFW Dataset

The FIW(10b) dataset [12] consists of the top 10 most frequent subjects selected from the LFW dataset [21], and the first 50 face images of each subject are used for evaluation. We use the cropped and resized face images readily provided by the authors of [12], where the face images are of 45×55 pixels. We follow the ambiguity model defined in [12] to generate ambiguous labels in the controlled experiment. Note that α denotes the number of extra labels for each instance, and β represents the portion of the ambiguously labeled data among all the instances. The degree of ambiguity ϵ indicates the maximum probability that an extra label co-occurs with a true label, over all labels and instances.

We conduct two types of controlled experiments suggested in [12]. For the *inductive* experiment, the dataset is evenly split into ambiguously labeled training set and unlabeled testing set. The proposed method, MCar-SVM, learns a multi-class linear SVM [8] with the disambiguated labels provided by MCar. The testing data is then classified with the learned classifier. For the *transductive* experiment, all the data is used as the ambiguously labeled training set. Each controlled experiment is repeated 20 times. We report the average testing (labeling) error rate for inductive (transductive) experiment, where the testing (labeling) error rate is the ratio of the number of erroneously labeled instances to the total number of instances in the testing (training) set. The standard deviations are plotted as error bars in the figures. We compare our method with several state-of-the-art ambiguous learning approaches for single instances with ambiguous labeling: CLPL [12] and DLHD/DLSD [10]. We use 'naive' [12] as the baseline method, which learns a classifier from minimizing the trivial $0/1$ loss.



(a) (b) (c)

Figure 4: Performance comparisons on the FIW(10b) dataset. (a) $\alpha \in [0, 0.95]$, $\beta = 2$, *inductive* experiment. (b) $\alpha = 1.0$, $\beta = 1$, $\epsilon \in [1/(c-1), 1]$, *inductive* experiment. (c) $\alpha = 1.0$, $\beta \in [0, 1, \ldots, 9]$, *transductive* experiment.

Figure 4a and 4b show the results of the inductive experiments. Figure 4a shows that the proposed MCar method consistently outperforms all the other methods especially when half of the instances are ambiguously labeled. In Figure 4b, MCar outperforms other methods over various degree of ambiguity except in the cases that $\epsilon > 0.7$. This shows that MCar yields improved performance in mild degree of ambiguity, but it becomes susceptible to high degree of ambiguity. An explanation is that both the true label and the extra labels of a subject will result in low-rank component of the labeling matrix when they are likely to co-occur in high degree of ambiguity. Consequently, separating the true label from the extra labels in MCar becomes challenging. In Figure 4c, MCar outperforms the other approaches only when the number of extra labels is less than 5 in the transductive experiment. This shows that MCar cannot operate when the labeling is severely cluttered such that the low-rank approximation of heterogeneous feature fails.

Figure 5 shows the intermediate result of low-rank decomposition of the feature matrix using MCar. Note that variations due to illumination, occlusions (e.g. glasses, hand), and expressions are suppressed such that the low-rank component of a subject is preserved. In contrast to MCar, the discriminative methods (e.g. naive, CLPL) can be susceptible to such variations. Furthermore, it also demonstrates the robustness of our methods even though the face images are not perfectly aligned. The proposed method outperforms the dictionary-based method for all cases except when there is severe ambiguity. Note the low-rank approximation of MCar operates on the feature matrix and ambiguous labeling matrix as a whole by concatenating them such that the actual labels and the low-rank component of feature matrix are recovered simultaneously. This essentially demonstrates the advantage of the proposed method over the DLHD/DLSD methods that iteratively alternate between confidence and dictionary update.

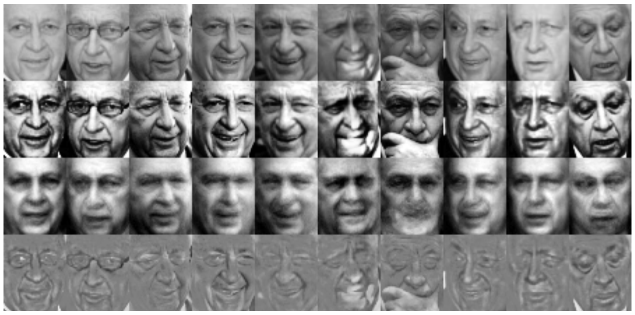


Figure 5: A subset of images from FIW(10b) demonstrates the low-rank decomposition of feature matrix in MCar: the original face images, histogram-equalized images $\mathbf{X}$, low-rank component $\mathbf{Z}$, and noisy component $\mathbf{E}_X$, from the first row to the forth row, respectively.

Method	naive	CLPL [12]	MMS [25]	MCar
Error Rate	18.6%	12.6%	11.4%	8.5%

Table 1: Labeling error rates for *Lost* $(16, 8)$ dataset (available at http://www.timotheecour.com/tv_data/tv_data.html).

5.2. Lost Dataset

The *Lost* dataset consists of the face images and the ambiguous labels automatically extracted using the screenplays provided in the TV series *Lost*. We use the *Lost* $(16, 8)$ dataset released by the authors of [11] for evaluation. The *Lost* $(16, 8)$ dataset consists of 1122 registered face images from 8 episodes, and the size of each is 60×90 pixels. The labeling covers 16 subjects, but only 14 of them appear in the dataset. Using the ambiguous labels provided by [11], we compare our method with the performance of 'naive', CLPL , and MMS [25]. No labeling constraint between instances is considered in this experiment. Results

Method	CL-SVM	MIMLSVM	MMS [25]	LR-SVM [27]	MCar-SVM
Error Rate	23.1 % ± 0.6	25.3 % ± 0.3	14.3 % ± 0.5	19.2 % ± 0.4	14.5 % ± 0.4

Table 2: Average testing error rates for the Labeled Yahoo! News dataset (available at http://lear.inrialpes.fr/data).

are shown in Table 1. It can be seen from this table that MCar outperforms CLPL and MMS by 4.1% and 2.9%, respectively. This shows that MCar better resolves the ambiguity and handles variations of instances in the TV series as compared to discriminative methods. Note that the performance of MMS is close to that of CLPL since the ambiguous loss functions of both methods become similar when no labeling constraint between the instances is considered.

5.3. Labeled Yahoo! News Dataset

The Labeled Yahoo! News dataset contains fully annotated faces in the images with names in the captions. It consists of 31147 detected faces from 20071 images. We use the precomputed SIFT feature of dimension 4992 extracted from that face images provided by Guillaumin *et al.* [19]. Following the protocol suggested in [25], we retain the 214 subjects with at least 20 occurrences in the captions. The remaining face images and names are treated as the additional null class. We conduct experiments on 5 training/testing splits by randomly selecting 80% of images and their associated captions as training set, and the rest are used as testing set. In each split, we also maintain the ratio between the number of training and testing instances from each subject.

The baseline approaches are CL-SVM and MIMLSVM [29], where their implementation details are provided in [25]. We compare with two state-of-the-art methods: MMS [25] and LR-SVM [27], which are based on discriminative model and low-rank framework, respectively. Both of these consider the labeling constraints between instances. We resolve the ambiguity for the labels in the training set using (25) and train a multi-class linear SVM [8] to classify the testing data. Our MCar-SVM algorithm exhibits a slightly 0.2% higher error rate as compared to MMS. One explanation is that the low-rank approximation for a class of insufficient instances is not quite effective such that some of the labels of those classes in the training data are erroneously labeled. This results in the performance degradation in the learned classifier. This issue is also pointed out by [27] as evidenced by the fact that the number of instances per class in this dataset ranges from 2 to 1168 with mean and standard deviations equal to 41.5 and 90.5, respectively.

Compared to the LR-SVM method, the MCar-SVM algorithm demonstrates 4.7% improvement on the testing accuracy. Since MCar assigns the labels across all the instances via low-rank approximation of heterogeneous feature matrix, it is more effective than the LR-SVM method, which updates the PPM and the low-rank subspace of each class alternately. However, LR-SVM still possesses its own advantage in large datasets in terms of the scalability.

5.4. Convergence

The convergence of Algorithm 1 is currently not theoretically guaranteed but observed empirically. Figure 6 shows the objective value with iterations for Algorithm 1 evaluated on the *Lost* dataset. It can be seen that Algorithm 1 converges in a few iterations. To gain more insight on the convergence of these methods, one may need to investigate the projections onto the convex sets are non-expansive [5].

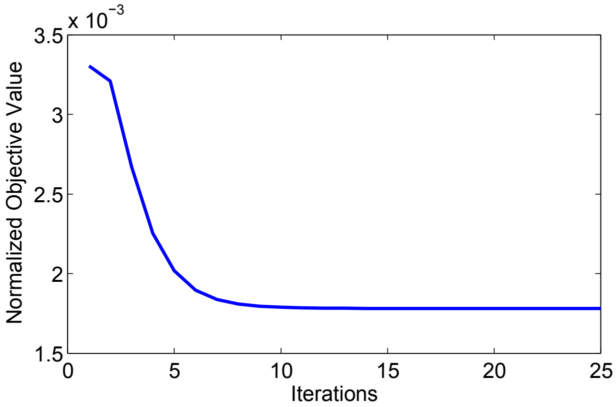


Figure 6: The optimization algorithm converges as the number of iterations increases. The objective value is normalized by the number of entries in $\mathbf{H}_{obs}$.

6. Conclusions

We have introduced a novel matrix completion framework for resolving the ambiguity of labels. In contrast to existing iterative alternating approaches, the proposed MCar method ensures all the instances and their associated ambiguous labels are utilized as a whole for resolving the ambiguity. Since MCar is capable of discovering the underlying low-rank structure of subjects, it is robust to within-subject variations. Hence, MCar can serve as the counterpart of discriminative ambiguous learning methods. As demonstrated by the experiments on synthesized ambiguous labels and two datasets collected from real world, MCar consistently resolves the ambiguity when single instances or group of instances are ambiguously labeled as compared to some of the previously proposed methods.

7. Acknowledgments

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2014-14071600012. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

[1] C. Ambroise, T. Denoeux, G. Govaert, and P. Smets. Learning from an imprecise teacher: Probabilistic and evidential approaches. *Applied Stochastic Models and Data Analysis, volume 1*, pages 100–105, 2001. 1

[2] T. L. Berg, A. C. Berg, J. Edwards, and D. A. Forsyth. Who's in the picture? In *NIPS*, 2004. 1, 6

[3] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y.-W. Teh, E. Learned-Miller, and D. A. Forsyth. Names and faces in the news. In *CVPR*, 2004. 1

[4] R. S. Cabral, F. D. la Torre, J. P. Costeira, and A. Bernardino. Matrix completion for multi-label image classification. In *NIPS*, 2011. 2, 5

[5] R. S. Cabral, F. D. la Torre, J. P. Costeira, and A. Bernardino. Matrix completion for weakly-supervised multi-label image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014. 3, 8

[6] E. Candes and B. Recht. Exact low-rank matrix completion via convex optimization. In *Allerton Conference on Communication, Control, and Computing*, 2008. 5

[7] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, Jun. 2011. 2, 5, 6

[8] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. 6, 8

[9] C.-F. Chen, C.-P. Wei, and Y.-C. F. Wang. Low-rank matrix recovery with structural incoherence for robust face recognition. In *CVPR*, 2012. 2

[10] Y.-C. Chen, V. M. Patel, J. K. Pillai, R. Chellappa, and P. J. Phillips. Dictionary learning from ambiguously labeled data. In *CVPR*, 2013. 1, 6

[11] T. Cour, B. Sapp, C. Jordan, and B. Taskar. Learning from ambiguously labeled images. In *CVPR*, 2009. 1, 2, 6, 7

[12] T. Cour, B. Sapp, and B. Taskar. Learning from partial labels. *J. Mach. Learn. Res.*, 12:1501–1536, 2011. 1, 4, 6, 7

[13] T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3):326–334, Jun. 1965. 4

[14] M. Everingham, J. Sivic, and A. Zisserman. Hello! My name is... Buffy - Automatic naming of characters in TV video. In *BMVC*, 2006. 1

[15] A. B. Goldberg, X. Zhu, B. Recht, J.-M. Xu, and R. D. Nowak. Transduction with matrix completion: Three birds with one stone. In *NIPS*, 2010. 2, 4, 5

[16] V. Govindaraju, D. B. Sher, R. K. Srihari, and S. N. Srihari. Locating human faces in newspaper photographs. In *CVPR*, 1989. 1

[17] V. Govindaraju, S. N. Srihari, and D. B. Sher. A computational model for face location. In *ICCV*, 1990. 1

[18] V. Govindaraju, S. N. Srihari, and D. B. Sher. A computational model for face location based on cognitive principles. In *AAAI*, 1992. 1

[19] M. Guillaumin, J. Verbeek, and C. Schmid. Multiple instance metric learning from automatically labeled bags of faces. In *ECCV*, 2010. 6, 8

[20] D. Huang, R. S. Cabral, and F. D. la Torre. Robust regression. In *ECCV*, 2012. 2

[21] G. B. Huang, V. Jain, and E. Learned-Miller. Unsupervised joint alignment of complex images. In *ICCV*, 2007. 6

[22] E. Hüllermeier and J. Beringer. Learning from ambiguously labeled examples. In *Intell. Data Anal.*, 2006. 1

[23] R. Jin and Z. Ghahramani. Learning with multiple labels. In *NIPS*, 2002. 1

[24] Z. Lin, M. Chen, L. Wu, and Y. Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *(UIUC Technical Report UILU-ENG-09-2215, November 2009)*. 5

[25] J. Luo and F. Orabona. Learning from candidate labeling sets. In *NIPS*, 2010. 1, 2, 6, 7, 8

[26] A. Shrivastava, J. K. Pillai, V. M. Patel, and R. Chellappa. Learning discriminative dictionaries with partially labeled data. In *ICIP*, 2012. 1

[27] Z. Zeng, S. Xiao, K. Jia, T.-H. Chan, S. Gao, D. Xu, and Y. Ma. Learning by associating ambiguously labeled images. In *CVPR*, 2013. 2, 6, 8

[28] T. Zhang. Statistical analysis of some multi-category large margin cassification methods. *J. Mach. Learn. Res.*, 5:1225–1251, 2004. 1

[29] Z. Zhou and M. Zhang. Multi-instance multilabel learning with application to scene classification. In *NIPS*, 2006. 8