Learning by Associating Ambiguously Labeled Images

Zinan Zeng[1], Shijie Xiao[2], Kui Jia[1], Tsung-Han Chan[1], Shenghua Gao[1], Dong Xu[2], Yi Ma[3]
[1]Advanced Digital Science Centers, Singapore
[2]School of Computer Engineering, Nanyang Technological University, Singapore
[3]Visual Computing Group, Microsoft Research Asia, Beijing, China
[1]{Edwin.zeng, Chris.jia, Th.chan, Shenghua.gao}@adsc.com.sg
[2]xiao0050@e.ntu.edu.sg , [2]DongXu@ntu.edu.sg , [3]yima@microsoft.com

Abstract

*We study in this paper the problem of learning classifiers from ambiguously labeled images. For instance, in the collection of new images, each image contains some samples of interest (*e.g.*, human faces), and its associated caption has labels with the true ones included, while the sample-label association is unknown. The task is to learn classifiers from these ambiguously labeled images and generalize to new images. An essential consideration here is how to make use of the information embedded in the relations between samples and labels, both within each image and across the image set. To this end, we propose a novel framework to address this problem. Our framework is motivated by the observation that samples from the same class repetitively appear in the collection of ambiguously labeled training images, while they are just ambiguously labeled in each image. If we can identify samples of the same class from each image and associate them across the image set, the matrix formed by the samples from the same class would be ideally low-rank. By leveraging such a low-rank assumption, we can simultaneously optimize a partial permutation matrix (PPM) for each image, which is formulated in order to exploit all information between samples and labels in a principled way. The obtained PPMs can be readily used to assign labels to samples in training images, and then a standard SVM classifier can be trained and used for unseen data. Experiments on benchmark datasets show the effectiveness of our proposed method.*

1. Introduction

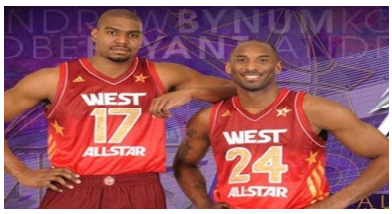
Learning classifiers for recognition purposes generally requires intensive labor work of labeling/annotating a large amount of training data. For example, in face recognition [28, 32, 13, 10], it is well known that collecting training samples with manual annotation for precise face alignment is the key to achieve high recognition accuracy. On the other hand, however, an unlimited number of images/videos with accompanying captions are freely available from the Internet, *e.g.,* images containing human faces and their associated text captions from the news websites. It becomes possible to avoid the intensive labor work if we can train good classifiers using these freely available data in the wild. Unfortunately, this is in general a difficult task. The main difficulty comes from the *ambiguous association* between samples in images and their labels in the corresponding image captions, as illustrated in Fig. 1.

Learning classifiers from the ambiguously labeled data falls in the category of ambiguous learning. The ambiguous association between samples and labels make the learning task more challenging than that in standard supervised learning. In the literature, a variety of attempts have been made to this end. For example, Multiple Instance Learning (MIL) has been proposed [1, 6, 25, 33] to learn classifiers from ambiguously labeled data, in which an image is treated as a bag, and the bag is labeled as positive if it contains at least one true positive instance, and negative otherwise. MIL essentially learns a classifier for each class of samples by iteratively estimating the instance label by some predefined losses. To explore the relations between samples and their ambiguous annotations, co-occurrence model [2, 3, 30, 20] has been proposed to infer their correspondences using the Expectation Maximization. Iterative clustering and learning approach was also proposed in [4] to assign human faces to named entities. In [12, 24], an ambiguous loss was proposed to learn a discriminant function for classification. The latent model was also explored in [31] for annotating images with unaligned object-level textual annotation.

Broadly speaking, the above methods try to learn a mapping function from training images and their associated ambiguous labels based on the following general assumptions or constraints: 1) *non-redundancy constraint* - every sample in training images belongs to a class by considering irrel-

A forceful President ***Barack Obama*** put Republican challenger ***Mitt Romney*** on the defensive on foreign policy issues on Monday night, scoring a solid victory in their third and final debate just 15 days before Election Day. [News From CNN]

President ***Barack Obama***, Italian Prime Minister ***Silvio Berlusconi***, center, and Russian President ***Dmitry Medvedev***, right, smile during a group photo at the G20 Summit in London. [News From Washington Post]



Bryant and ***Andrew Bynum*** have been named Western Conference All-Star starters at guard and center respectively. This is ***Bryant***'s 14th time starting the league's annual showcase game. All-Star nod. [News from NBA]

Figure 1. Sample photos from news websites and the corresponding text caption. In general, most of the news websites do not provide the face-name correspondence, hence it is a challenging task for the standard supervised learning method to automatically perform face recognition on such freely available data.

vant samples as *background class*, 2) *uniqueness constraint* - samples of the same class cannot simultaneously appear in an image except the background class (*e.g.,* multiple faces of the same person cannot appear in an image), and 3) *non-pairing constraint* - samples of different classes and their true labels cannot consistently appear together across the training images (*e.g.,* the faces from two subjects will not always co-occur in most of the images). With these assumptions in mind, the task of ambiguous learning is essentially to model the ambiguous relations between samples and labels both *within each image* and *across the image set*. A good approach should be able to make use of all constraints available in the sample-label relations in a principled way. Except the work in [24], however, most of existing methods can only partially achieve this.

To this end, we propose a novel framework to address the ambiguous learning problem. We are particularly interested in the problem of face recognition using ambiguously labeled images. Our framework is motivated by the observation that samples from the same class, assuming intra-class variations are reduced within a certain level, can be characterized by a low-dimensional subspace embedded in the ambient space. For example, Chen *et al.* [10] showed that face images of the same person can be represented as a low-rank matrix. Based on this low-rank assumption, our framework simultaneously optimizes a partial permutation matrix (PPM) for each of the training images by rank minimization. The PPMs are formulated so that after optimization, they can associate samples of the same classes from different images to form low-rank matrices. To address the intra-class variations, a sparse error term for each class is also introduced to achieve better robustness. The obtained PPMs can be used as indicators to assign the labels to samples in each image. Indeed, our method relies on the facts that PPMs are formulated and optimized so that the intrinsic constraints from both the intra-image and inter-image sample-label relations can be explored. For the intra-image relations, the PPM is constrained to simultaneously and exclusively assign one label to one sample in each image, where other priors could also be incorporated. For the inter-image relations, the PPMs are simultaneously optimized by rank minimization so that the aforementioned non-pairing assumption (3) in ambiguous learning can be used.

The contribution of this work is summarized as follows. In contrast to existing methods for ambiguous learning, we provide a novel perspective to address this problem by formulating it as a sample-label correspondence task with PPM optimization, where the implicit information in both the intra-image and inter-image sample-label relations can be used in a principled way. A scalable algorithm is proposed that enables our method to work on medium to large scale dataset in terms of feature dimension, number of data points and number of classes. Once the sample-label correspondences are established, standard supervised learning methods can be applied to perform the prediction on unseen data. Experiments on benchmark datasets show the effectiveness of our method.

2. Related work

Learning visual classifiers from caption-accompanying images has been an active topic in computer vision [1, 2, 20, 31, 24], of which learning face classifiers from such data is of particular interest [3, 18, 12, 24]. There are a few methods that explicitly take face-name (sample-label) correspondences into account. For example, Berg *et al,* [3] proposed a constrained mixture model to optimize the likelihood of particular face-name assignment. The work in [18] first iteratively clusters faces using EM based on face similarity and constraints from the caption. Based on these clusters, a weighted bipartite graph modelling the null assignment (*i.e.,* faces that are not assigned to any names and names that are not assigned to any faces) and caption constraints is constructed for face-name assignment. On the other hand, Support Vector Machine (SVM) based methods directly learn discriminant classifiers using the ambiguously labeled data. Cour *et al.* [12] proposed a max-margin for-

mulation by introducing an ambiguous 0/1 loss to replace the loss in the standard SVM formulation, in which they defined the ambiguous 0/1 loss as 0 if the predicted name is in the image caption, and 1 otherwise. Based on this ambiguous loss, they defined a convex loss that penalized the prediction of names as the ones not present in the caption. This formulation did not consider the uniqueness constraint, hence it generally cannot perform well for images with multiple faces. Luo *et al.* [24] extended the idea of ambiguous loss for images with multiple faces, in which they enforced the uniqueness constraint by assigning names to faces at a set level (via labeling vectors) in each image.

Recently, low-rank property of a set of linearly correlated images shows its usefulness in many computer vision problems, such as subspace segmentation [22], face recognition [10], multi-label image classification [7], image alignment [27] and image segmentation [11]. On the other hand, PPM has been popularly used for feature point correspondence with unsupervised learning [26, 34]. Our method in this paper is essentially motivated by these pioneering works. However, with a new formulation of low-rank matrices and PPM constraints, we show that our proposed method fits well for the ambiguous learning task.

3. The proposed framework

We formally define the problem of learning from ambiguously labeled images as follows. The input is a collection of N images $\mathcal{I}_1, \cdots, \mathcal{I}_N$. Each image has different number of samples from distinctive classes, and there are $\bar{K}$ classes in total. More precisely, we assume there are K_n samples from the n^{th} image, and they are from different classes. Each sample is characterized as a d-dimensional feature vector. Hence, the n^{th} image is represented as $\mathbf{F}_n = [\mathbf{f}_n^1, \cdots, \mathbf{f}_n^{K_n}] \in \mathbb{R}^{d \times K_n}$. Associated with the n^{th} image is a binary vector $\mathbf{t}_n \in \{0,1\}^{\bar{K}}$ representing the labels appearing in the caption of the n^{th} image: $\mathbf{t}_n(i) = 1$ if the label of the i^{th} class appears in the image caption, and 0 otherwise. Given these ambiguously labeled N training images, the tasks are (1) for the n^{th} image, to assign each sample in the image a label from the caption of the image by considering the aforementioned constraints, and (2) based on the obtained sample-label assignments for all the N images, to learn classifiers and apply them to unseen data. In the following, we first introduce how the low-rank assumption of the matrix formed by the samples from the same class can be used to simultaneously optimize a set of PPMs for assigning labels to the samples in the ambiguously labeled training images.

3.1. Low-rank assumption for samples from the same class

Face images of the same individual are commonly assumed to reside in a low-dimensional subspace [28, 10]. Put it in another way, if we place sufficient face samples from the same class into a matrix, this matrix should be approximately low-rank. Denote $\bar{\mathbf{F}}_i = [\bar{\mathbf{f}}_i^1, \cdots, \bar{\mathbf{f}}_i^{n_i}] \in \mathbb{R}^{d \times n_i}$ as the matrix containing n_i samples from the i^{th} class, $i \in \{1, \ldots, \bar{K}\}$. When these samples are human faces, then $\bar{\mathbf{F}}_i$ should be approximately low-rank. However, the distribution and ground-truth labels of these n_i samples in the N training images are unknown. In our ambiguous learning tasks, we show next how this low-rank assumption can be used to seek the sample-label correspondences.

3.2. Sample-label correspondences via PPM

Given N training images, our first objective is to find the sample-label correspondences for all samples from $\bar{K}$ classes. In this work, we use partial permutation matrix (PPM) [26, 31, 34] to model such correspondences. In particular, the PPM $\mathbf{P}_n \in \bar{\mathcal{P}}_n$ for the image $\mathcal{I}_n$ is defined as:

$$\bar{\mathcal{P}}_n = \left\{ \mathbf{P}_n \in \{0,1\}^{K_n \times \bar{K}} \middle| \begin{array}{c} \mathbf{1}_{K_n}^T \mathbf{P}_n (\mathbf{1}_{\bar{K}} - \mathbf{t}_n) = 0, \\ \mathbf{P}_n \mathbf{1}_{\bar{K}} = \mathbf{1}_{K_n}, \mathbf{1}_{K_n}^T \mathbf{P}_n \leq \mathbf{1}_{\bar{K}}^T, \end{array} \right\} \quad (1)$$

where $\{0,1\}^{K_n \times \bar{K}}$ denotes a $K_n \times \bar{K}$ binary matrix and $\mathbf{1}_c$ (*resp.* $\mathbf{0}_c$) denotes a column vector of length c with all entries as 1 (*resp.* 0). The first row in (1) enforces that only labels appearing in the caption can be assigned to samples in the image $\mathcal{I}_n$. The second row in (1) is designed to satisfy the non-redundancy and uniqueness constraints. The PPMs $\{\mathbf{P}_n \in \bar{\mathcal{P}}_n\}_{n=1}^N$ for all the N images are similarly defined. Note that PPM has been used in [31] for ambiguous learning. However, their work did not enforce the uniqueness constraint when using PMMs. Given $\{\mathbf{F}_n\}_{n=1}^N$, there exist the PPMs such that samples of the same class can be identified and columnly corresponded in $\{\mathbf{F}_n\mathbf{P}_n \in \mathbb{R}^{d \times \bar{K}}\}_{n=1}^N$, or equivalently, the $\bar{K}$ sub matrices $\mathbf{L}_1, ..., \mathbf{L}_{\bar{K}}$ of size $\mathbb{R}^{d \times N}$ concatenated in

$$\begin{aligned} \mathcal{F}(\{\mathbf{P}_n\}_{n=1}^N) &\triangleq [\mathrm{vec}(\mathbf{F}_1\mathbf{P}_1)|\cdots|\mathrm{vec}(\mathbf{F}_N\mathbf{P}_N)] \in \mathbb{R}^{d\bar{K} \times N} \\ &= [\mathbf{L}_1^T, ..., \mathbf{L}_{\bar{K}}^T]^T \end{aligned}$$

are rank deficient, where $\mathrm{vec}(\cdot)$ is an operator that vectorizes a matrix by concatenating its column vectors. Based on our low-rank assumption for samples from the same class, the sample-label correspondence problem can be formulated as the following problem:

$$\min_{\substack{\{\mathbf{L}_i\}_{i=1}^{\bar{K}} \\ \{\mathbf{P}_n \in \bar{\mathcal{P}}_n\}_{n=1}^N}} \sum_{i=1}^{\bar{K}} \mathrm{rank}(\mathbf{L}_i) \ \text{ s.t. } \mathcal{F}(\{\mathbf{P}_n\}_{n=1}^N) = [\mathbf{L}_1^T, ..., \mathbf{L}_{\bar{K}}^T]^T.$$

Considering intra-class variations and inevitable data noise or corruption, the above low-rank assumption is likely to be violated. To improve the robustness, we introduce a sparse matrix to model those data variations as sparse errors. In addition, we relax the non-convex function $\mathrm{rank}(\cdot)$

by its convex surrogate, the nuclear norm [17], resulting in the optimization problem as follows:

$$\min_{\substack{\{\mathbf{L}_i,\mathbf{E}_i\}_{i=1}^{\bar{K}}\\ \{\mathbf{P}_n\in\bar{\mathcal{P}}_n\}_{n=1}^{N}}} \sum_{i=1}^{\bar{K}} \|\mathbf{L}_i\|_* + \lambda\|\mathbf{E}_i\|_1,$$
$$\text{s.t. } \mathcal{F}(\{\mathbf{P}_n\}_{n=1}^{N}) = [(\mathbf{L}_1+\mathbf{E}_1)^T, ..., (\mathbf{L}_{\bar{K}}+\mathbf{E}_{\bar{K}})^T]^T,$$

where $\|\cdot\|_1$ is the l_1 norm and $\lambda > 0$ is a trade-off parameter that balances the low-rank and sparse terms.

3.3. Modeling for the background samples

In practical ambiguously labeled images from Internet, there are many irrelevant or background samples that co-occur with the samples we are interested in. In line with the convention in ambiguous learning [24, 18], we call these background samples as samples of *null class*. Without loss of generality, we let the $\bar{K}^{th}$ class be the null class. Note that enforcing low-rank and sparse constraints on samples of the null class is inappropriate. In addition, there might be no true labels appearing in image captions for samples from the null class, we again take the convention to set $\mathbf{t}_n(\bar{K}) = 0$. Moreover, to avoid the trivial solution that all samples are assigned to the null class, we assume that at least one sample per image is not associated with the null class. To this end, we modify the PPM definition in (1) as follows:

$$\mathcal{P}_n = \left\{ \mathbf{P}_n \in \{0,1\}^{K_n\times\bar{K}} \left| \begin{array}{c} \mathbf{1}_{K_n}^T\mathbf{P}_n\mathbf{t}_n = \mathbf{1}_{K_n}^T\mathbf{P}_n\begin{bmatrix}\mathbf{1}_{\bar{K}-1}\\0\end{bmatrix}, \\ 1 \le \mathbf{1}_{K_n}^T\mathbf{P}_n\mathbf{t}_n, \\ \mathbf{1}_{K_n}^T\mathbf{P}_n\begin{bmatrix}\mathbf{I}_{\bar{K}-1}\\\mathbf{0}_{\bar{K}-1}^T\end{bmatrix} \le \mathbf{1}_{\bar{K}-1}^T, \\ \mathbf{P}_n\mathbf{1}_{\bar{K}} = \mathbf{1}_{K_n}, \end{array} \right. \right\} \quad (2)$$

where $\mathbf{I}_c$ is a $c \times c$ identity matrix. Similar to the PPM definition in (1), the first row in (2) prohibits our method from choosing a label not appeared in the image caption except the null class. The second row enforces that at least one label from the caption must be chosen to avoid the trivial solution that assigning all the samples to the null class. The third and forth rows in (2) enforce the uniqueness and non-redundancy constraints respectively. Based on the new PPM definition (2), we arrive at the following optimization problem:

$$\min_{\substack{\{\mathbf{L}_i,\mathbf{E}_i\}_{i=1}^{\bar{K}-1},\\ \{\mathbf{P}_n\in\mathcal{P}_n\}_{n=1}^{N}}} \sum_{i=1}^{\bar{K}-1} \|\mathbf{L}_i\|_* + \lambda\|\mathbf{E}_i\|_1,$$
$$\text{s.t. } \mathcal{F}(\{\mathbf{P}_n\}_{n=1}^{N}) = [(\mathbf{L}_1+\mathbf{E}_1)^T, ..., (\mathbf{L}_{\bar{K}}+\mathbf{E}_{\bar{K}})^T]^T.$$

To simplify the subsequent notation, we perform the change of variables, and rewrite the formulation as follows:

$$\min_{\substack{\{\mathbf{L}_i,\mathbf{E}_i\}_{i=1}^{\bar{K}-1},\\ \{\boldsymbol{\theta}_n\}_{n=1}^{N}}} \sum_{i=1}^{\bar{K}-1} \|\mathbf{L}_i\|_* + \lambda\|\mathbf{E}_i\|_1, \quad (3)$$
$$\text{s.t. } \mathbf{S}_n\boldsymbol{\theta}_n \le \mathbf{Q}_n,\ \mathbf{X}_n\boldsymbol{\theta}_n = \mathbf{R}_n,$$
$$\boldsymbol{\theta}_n \in \{0,1\}^{\bar{K}K_n},\ \forall n \in \{1,\cdots,N\},$$
$$[\mathbf{Z}_1\boldsymbol{\theta}_1,\cdots,\mathbf{Z}_N\boldsymbol{\theta}_N] = \mathbf{K},$$

where

$$\boldsymbol{\theta}_n = \text{vec}(\mathbf{P}_n),\ \mathbf{Z}_n = \mathbf{I}_{\bar{K}}\otimes\mathbf{F}_n,\ \mathbf{R}_n = [0, \mathbf{1}_{K_n}^T]^T,\ \mathbf{Q}_n = [\mathbf{1}_{\bar{K}-1}^T, -1]^T,$$
$$\mathbf{S}_n = \begin{bmatrix}[\mathbf{I}_{\bar{K}-1}, \mathbf{0}_{\bar{K}-1}]\\ -\mathbf{t}_n^T\end{bmatrix}\otimes\mathbf{1}_{K_n}^T,\ \mathbf{X}_n = \begin{bmatrix}([\mathbf{1}_{\bar{K}-1}^T, 0] - \mathbf{t}_n^T)\otimes\mathbf{1}_{K_n}^T\\ \mathbf{1}_{\bar{K}}^T\otimes\mathbf{I}_{K_n}\end{bmatrix},$$
$$\mathbf{K} = [(\mathbf{L}_1+\mathbf{E}_1)^T, ..., (\mathbf{L}_{\bar{K}}+\mathbf{E}_{\bar{K}})^T]^T,$$

and $\otimes$ denotes the Kronecker product.

3.4. Optimization using ADMM

The number of variables in the optimization problem (3) is usually large, thus a scalable algorithm is essential for its practical use. In this work, we consider using the fast first-order method Alternating Direction Method of Multipliers (ADMM) [5] to solve (3). To facilitate the algorithmic development, we define the following notations:

$$\mathbf{W} = [\mathbf{Z}_1\boldsymbol{\theta}_1,\cdots,\mathbf{Z}_N\boldsymbol{\theta}_N] - \mathbf{K}, \quad (4)$$
$$I_C(\boldsymbol{\theta}_n) = \begin{cases} 0 & \text{if } \mathbf{S}_n\boldsymbol{\theta}_n \le \mathbf{Q}_n,\ \mathbf{X}_n\boldsymbol{\theta}_n = \mathbf{R}_n,\ \boldsymbol{\theta}_n \in \{0,1\}^{\bar{K}K_n} \\ \infty & \text{otherwise} \end{cases},$$

where $I_C(\boldsymbol{\theta}_n)$ is the indicator function of the inequality associated with $\boldsymbol{\theta}_n$. The augmented Lagrangian function of the optimization problem (3) can be written as:

$$\mathcal{L}_\mu(\{\mathbf{L}_i,\mathbf{E}_i\}_{i=1}^{\bar{K}-1}, \{\boldsymbol{\theta}_n\}_{n=1}^{N}, \mathbf{Y}) \quad (5)$$
$$= \sum_{i=1}^{\bar{K}-1}(\|\mathbf{L}_i\|_* + \lambda\|\mathbf{E}_i\|_1) + <\mathbf{Y},\mathbf{W}> + \sum_{n=1}^{N} I_C(\boldsymbol{\theta}_n) + \frac{\mu}{2}\|\mathbf{W}\|_F^2,$$

where $\mathbf{Y} \in \mathbb{R}^{d\bar{K}\times N}$ is the Lagrange multiplier matrix, μ is a positive scalar, $<.,.>$ denotes the matrix inner product and $\|\cdot\|_F$ denotes the Frobenius norm. An iteration for solving (5) is given as:

$$\{\{\mathbf{L}_i,\mathbf{E}_i\}_{i=1}^{\bar{K}-1}, \{\boldsymbol{\theta}_n\}_{n=1}^{N}\}^{t+1} = \underset{\{\mathbf{L}_i,\mathbf{E}_i\}_{i=1}^{\bar{K}-1}, \{\boldsymbol{\theta}_n\}_{n=1}^{N}}{\arg\min} \mathcal{L}_\mu, \quad (6)$$
$$\mathbf{Y}^{t+1} = \mathbf{Y}^t + \mu\mathbf{W}^{t+1}, \quad (7)$$

where t is the current iteration number, μ follows the updating rule $\mu^{t+1} = \rho\mu^t$ for some $\rho > 1$ as in [27], $\mathcal{L}_\mu$ is defined in (5) and $\mathbf{W}^{t+1}$ is computed as in (4). We apply the coordinate descent method to minimize $\mathcal{L}_\mu$ with respect to $\{\mathbf{L}_i,\mathbf{E}_i\}, \forall i \in [i,\cdots,\bar{K}-1]$ and $\boldsymbol{\theta}_n, \forall n \in \{1,\cdots,N\}$ alternatingly, and both of them are relatively easy to solve.

3.4.1 Update $\mathbf{L}_i$ and $\mathbf{E}_i, \forall i \in \{1, \cdots, \bar{K}-1\}$

The minimization problem (6) with respect to $\{\mathbf{L}_i, \mathbf{E}_i\}_{i=1}^{\bar{K}-1}$ can be decomposed into $\bar{K}-1$ independent subproblems. The i^{th} subproblem updating $\mathbf{L}_i$ and $\mathbf{E}_i$ can be equivalently written as follows:

$$\begin{aligned}\{\mathbf{L}_i^{t+1}, \mathbf{E}_i^{t+1}\} = \arg\min_{\mathbf{L}_i, \mathbf{E}_i} \quad & \|\mathbf{L}_i\|_* + \lambda\|\mathbf{E}_i\|_1 \\ & + <\mathbf{Y}_i^t, \mathbf{D}_i^t - \mathbf{L}_i - \mathbf{E}_i> + \frac{\mu}{2}\|\mathbf{D}_i^t - \mathbf{L}_i - \mathbf{E}_i\|_F^2,\end{aligned}$$

where $\mathbf{D}_i^t \in \mathbb{R}^{d\times N}$ (*resp.* $\mathbf{Y}_i^t \in \mathbb{R}^{d\times N}$) is a matrix containing the i^{th} sub-matrix of $[\mathbf{Z}_1\boldsymbol{\theta}_1^t, \cdots, \mathbf{Z}_N\boldsymbol{\theta}_N^t]$ (*resp.* $\mathbf{Y}^t$). This optimization problem could be solved by Singular Value Thresholding (SVT) [8]:

$$\begin{aligned}[\mathbf{U}, \mathbf{S}, \mathbf{V}] &= \text{svd}(\mathbf{D}_i^t - \mathbf{E}_i^t + \mu^{-1}\mathbf{Y}_i^t), \\ \mathbf{L}_i^{t+1} &= \mathbf{U}S_{\mu^{-1}}[\mathbf{S}]\mathbf{V}^T, \\ \mathbf{E}_i^{t+1} &= S_{\lambda\mu^{-1}}[\mathbf{D}_i^t - \mathbf{L}_i^{t+1} + \mu^{-1}\mathbf{Y}_i^t], \end{aligned} \tag{8}$$

where $S_\tau[\mathbf{X}]$ is the shrinkage operator for the matrix $\mathbf{X}$ that applies $S_\tau[x] = \text{sign}(x)\cdot\max\{|x|-\tau, 0\}$ to all the elements of $\mathbf{X}$.

Applying SVT on large matrices (*e.g.,* $1.5K \times 10K$ dimension) is computationally expensive. To overcome this bottleneck, we apply the algorithm in [9]. Instead of performing SVT, this algorithm solves the dual problem of the original low-rank optimization, which only involves polar decomposition and projection. Specifically, the low-rank matrix $\mathbf{L}_i$ can be computed in the following steps:

1. Compute the polar decomposition as follows:

$$\mathbf{D}_i^t - \mathbf{E}_i^t + \mu^{-1}\mathbf{Y}_i^t = \mathbf{U}\mathbf{V}, \tag{9}$$

   where $\mathbf{U}$ is a unitary matrix and $\mathbf{V}$ is a symmetric non-negative definite matrix.

2. Solve the following optimization problem:

$$\mathcal{P}_\mu(\mathbf{V}) = \arg\min_{\|\mathbf{L}_i\|_2 \le \mu^{-1}} \|\mathbf{L}_i - \mathbf{V}\|_F. \tag{10}$$

3. Update $\mathbf{L}_i$ as follows:

$$\mathbf{L}_i^{t+1} = \mathbf{D}_i^t - \mathbf{E}_i^t + \mu^{-1}\mathbf{Y}_i^t - \mathbf{U}\mathcal{P}_\mu(\mathbf{V}). \tag{11}$$

3.4.2 Update $\boldsymbol{\theta}_n, \forall n \in \{1, \cdots, N\}$

The minimization problem (6) with respect to $\{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_N\}$ can be decoupled into N independent subproblems. The n^{th} subproblem updating $\boldsymbol{\theta}_n$ can be equivalently formulated as the following integer constrained quadratic programming (QP) problem:

$$\begin{aligned}\boldsymbol{\theta}_n^{t+1} = \arg\min_{\boldsymbol{\theta}_n} \ & \frac{1}{2}\boldsymbol{\theta}_n^T\mathbf{Z}_n^T\mathbf{Z}_n\boldsymbol{\theta}_n + \mathbf{e}_n^T(\frac{1}{\mu}\mathbf{Y}^t - \mathbf{K})^T\mathbf{Z}_n\boldsymbol{\theta}_n \\ \text{s.t. } & \mathbf{S}_n\boldsymbol{\theta}_n \le \mathbf{Q}_n,\ \mathbf{X}_n\boldsymbol{\theta}_n = \mathbf{R}_n,\ \boldsymbol{\theta}_n \in \{0,1\}^{\bar{K}K_n},\end{aligned} \tag{12}$$

where $\mathbf{e}_n$ is a unit column vector with all the entries set to 0 except the n^{th} entry set to 1. If each feature is l_2 normalized, (12) can be approximated by the following linear programming (LP) similar to [21]:

$$\begin{aligned}\boldsymbol{\theta}_n^{t+1} = \arg\min_{\boldsymbol{\theta}_n} \ & \mathbf{e}_n^T(\frac{1}{\mu}\mathbf{Y}^t - \mathbf{K})^T\mathbf{Z}_n\boldsymbol{\theta}_n, \\ \text{s.t. } & \mathbf{S}_n\boldsymbol{\theta}_n \le \mathbf{Q}_n,\ \mathbf{X}_n\boldsymbol{\theta}_n = \mathbf{R}_n,\ \mathbf{0}_{\bar{K}K_n} \le \boldsymbol{\theta}_n \le \mathbf{1}_{\bar{K}K_n},\end{aligned} \tag{13}$$

whose solution can be efficiently solved by the standard LP solver. The whole algorithm is summarized in Algorithm 1.

Algorithm 1: The optimization algorithm

Data: $[\mathbf{F}_1, \cdots, \mathbf{F}_N], \lambda, \rho, \mathbf{Y}^0, \{\mathbf{E}_i^0, \mathbf{L}_i^0\}_{i=1}^{\bar{K}}, \{\boldsymbol{\theta}_n^0\}_{n=1}^N$
Result: $\boldsymbol{\theta}_n, \forall n \in \{1, \cdots, N\}$
$t = 0$;
while *not converge* **do**
  for $i = 1$ **to** $\bar{K}-1$ **do**
    Update $\mathbf{L}_i^{t+1}$ as in (9) - (11);
    Update $\mathbf{E}_i^{t+1}$ as in (8);
  end
  for $n = 1$ **to** N **do**
    Update $\boldsymbol{\theta}_n^{t+1}$ as in (13);
  end
  Update $\mathbf{Y}^{t+1}$ as in (7);
  $t = t + 1$;
end

3.5. Classification for unseen data

To classify samples in unseen images, we train a SVM with the Gaussian kernel (*i.e.,* $K(i,j) = \exp(-\gamma D^2(\mathbf{f}_i, \mathbf{f}_j))$), where $D(\mathbf{f}_i, \mathbf{f}_j)$ is the distance between $\mathbf{f}_i$ and $\mathbf{f}_j$, and the kernel parameter $\gamma = \frac{1}{A}$ with A being the mean value of the square distances between all the training samples as in [14]. Specifically, we apply the proposed method on the training set to obtain the partial permutation matrices, which can be readily used to obtain the labels for the training samples. We then use the obtained labels to train a one-vs-all SVM classifier. Of course, we can train a more sophisticated supervised learning model to consider all the label constraints within one image during this step, similarly as did in the work [12, 24]. However, we use the off-the-shelf SVM model to demonstrate the effectiveness of the proposed method. We therefore refer to the proposed method as LR-SVM.

4. Experiment

In this section, we compare our newly proposed method LR-SVM with the SVM based methods for the face recognition problem on two ambiguously labeled datasets, one is

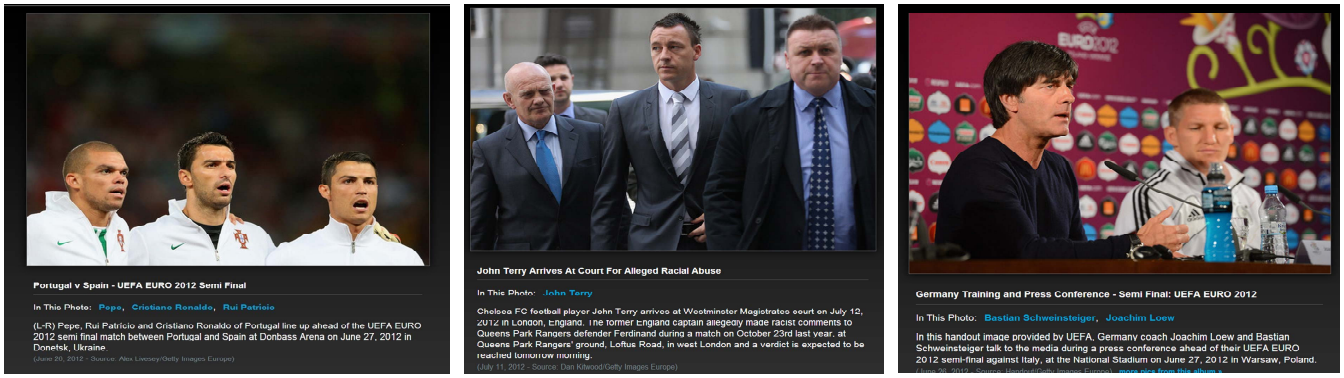


Figure 2. Sample images with the captions from the Soccer player dataset.

Table 1. Face Recognition Rate(%) comparison on the Soccer player dataset.

Algorithm	SIL-SVM [6]	mi-SVM [1]	sMIL [6]	consGMM [18]	PLL [12]	MMS [24]	LR-SVM
Accuracy	37.87	37.40	27.38	53.39	3.93	54.43	58.51

the Soccer player dataset and the other is the Labeled Yahoo! News dataset. We compare the baseline method SIL-SVM [6]. To train a classifier for each class, all the faces within an image are treated as positive samples if the corresponding label appears in the caption, otherwise they are negative samples. We also compare the result with Multiple Instance Learning methods such as mi-SVM [1], sMIL [6] and the works in [12] and [24] with ambiguous loss. In addition, we report the result from [18] (referred to as ConsGMM), in which it uses the caption for the prediction when constructing the bipartite graph. We use the recognition accuracy for performance evaluation, which is defined as the percentage of correctly classified faces (including the background ones) over all faces in the dataset. Note that the caption is not used during the final prediction except for ConsGMM. We fix the parameters $\lambda = 0.3$, $\rho = 1.05$ for the proposed LR-SVM on the two datasets, and tune $C = \{0.1, 1, 10\}$ for the SVM based methods. We first describe the two datasets used for our experiments, followed by a discussion on the results.

4.1. Soccer player dataset

We perform the experiment on the Soccer player dataset, which is collected from from www.zimbio.com website by querying with names of soccer players from famous European football clubs. This dataset contains 8934 images and 17878 faces detected by using [29] with 1579 names. We retain 170 names that occur at least 20 times and remove the images whose captions do not contain any of the 170 names, resulting in 8640 images and 17472 faces. We treat the names that exist in the remaining images but not in the 170 names as null class. Some of the representative images are shown in Fig. 2. As can be seen in the figure, these faces have significant appearance and pose variation. The number of ground-truth faces of each class is shown in Fig. 3. For each detected face, 13 interest points are detected by using [16] and each point is described by taking the vector of pixels in the elliptical region, which is further normalized for local photometric invariance as in [15]. Each descriptor is with 149-dimension. We concatenate all the descriptors within one face to form a 1937-dimension feature vector, and we further perform PCA to retain 90% energy, resulting in 279-dimension feature vector. The recognition accuracies on all faces in the dataset for all the methods are reported in Table 1.

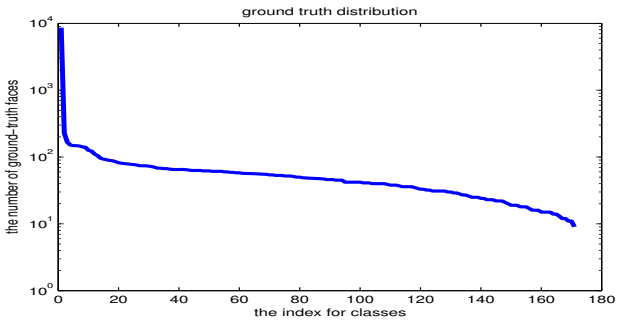


Figure 3. The number of ground-truth faces of each class in the Soccer player dataset. The average number of faces for each class is 52.3 with the standard deviation as 33.4. Note that classes are sorted in descending order of the number of ground-truth faces.

4.2. Labeled Yahoo! News dataset

The Labeled Yahoo! News dataset contains news pictures and captions from Yahoo News. It was collected by Berg [3] and was further processed by Guillaumin *et al.* [19] by extracting the 128-dimension SIFT [23] from 3 scales at 13 landmark points detected by [16]. This dataset contains 20071 images and 31147 detected faces. Following the experimental protocol in [24], we retain the 214 names oc-

Table 2. Face recognition accuracy (%) on the testing set of the Labeled Yahoo! News dataset.

Algorithm	SIL-SVM [6]	mi-SVM [1]	sMIL [6]	consGMM [18]	PLL [12]	MMS [24]	LR-SVM
Split 1	31.01	28.18	44.51	66.00	22.53	82.54	81.53
Split 2	31.32	28.09	45.34	66.34	22.62	83.33	80.30
Split 3	31.59	28.49	43.94	65.45	22.75	84.59	80.47
Split 4	30.92	28.62	45.11	64.74	21.97	84.63	80.76
Split 5	31.71	28.71	43.72	64.74	22.89	83.16	80.96
Avg Accuracy	31.31	28.41	44.52	65.45	22.55	83.65	80.80

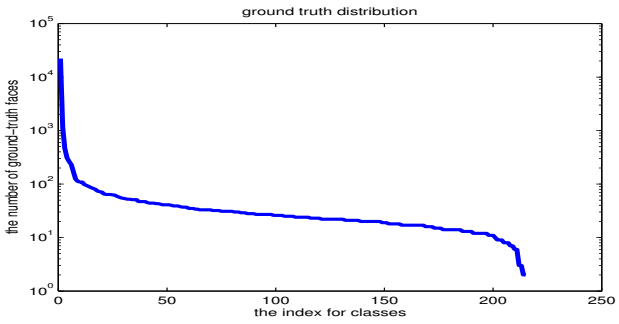


Figure 4. The number of ground-truth faces of each class in the Labeled Yahoo! News Dataset. The average number of faces for each class is 41.5 with the standard deviation as 90.5. Note that classes are sorted in descending order of the number of ground-truth faces.

curred at least 20 times in the captions and treat the other names as the null class. Fig. 4 shows the number of ground-truth faces of each class. The experiments are performed over 5 different random training/testing splits, by randomly sampling 80% of the images as the training set the remaining samples for testing. During the split, we also try to maintain this ratio for each class. We observe that PLL can not handle such high dimension data, following [12], we apply PCA and project the data onto 50 dimensions. The result is summarized in Table 2.

4.3. Discussion on the experiments

Based on the results in Table 1 and Table 2, we have the following observations:

1. The proposed LR-SVM outperforms the MIL-based methods on both datasets. An explanation is that these MIL methods treat each class independently, where the non-redundancy constraint is not explicitly taken into consideration during the training process. In contrast, the proposed method explicitly takes this constraint into account when seeking for the optimal partial permutation matrices (PPMs). Together with the rank constraint, it can recover more accurate labels for the subsequent supervised learning problems, resulting in better performance.

2. LR-SVM also outperforms ConsGMM on both datasets. Take note that for ConsGMM, the caption is used during the prediction stage to enforce the uniqueness constraint, while ours does not use any caption once the PPMs are obtained. This clearly shows the rank constraint is a more effective way to associate faces of the same classes when compared with the Constrained Gaussian Mixture Model.

3. LR-SVM performs better than PLL. Although it is proved in [12] that the ambiguous loss is a tight upper bound for the ground-truth 0/1 loss when there is only one face in an image, such loss is generally not tight for the cases where more than one face appear in the image without considering the uniqueness constraint. On the other hand, by providing less noisy labels for the standard SVM training, which uses much simpler form of hinge loss, we have shown the effectiveness of the proposed framework.

4. There is no consistent winner between LR-SVM and MMS. LR-SVM is better than MMS on the Soccer player dataset, while MMS outperforms LR-SVM on the Labeled Yahoo! News dataset. One possible explanation is that the low rank assumption is effective in discovering the low-dimensional subspace when the number of samples from each class is large enough. When the number of samples per class is small, it is generally difficult to use the rank to measure the sample similarity. As shown in Fig. 3, the average number of samples from each class on the Soccer players dataset is 52.3 with the standard deviation as 33.4, and almost all the classes of this dataset have more than 10 samples. In this case, rank constraint can be used effectively to discover the low-dimensional subspace, thus the learnt PPMs can lead to more accurate labels for the subsequent supervised learning, resulting in better performance over MMS. On the other hand, as shown in Fig. 4, on the Labeled Yahoo! News dataset, the average number of samples from each class is 41.5 with the standard deviation as 90.5, in which the number of samples from each class ranges from a few hundred to two. In this case, the low rank assumption is less effective in discovering the low-dimensional subspace with less training samples per class, hence our

method obtains more noisy labels for the subsequent supervised learning problems, resulting in worse result when compared with the MMS. On the other hand, if we are interested in automatically learning the classifiers for celebrities from the Internet images in which case the images for each person are sufficient and our low rank assumption can be well satisfied, the proposed method will lead to better results.

5. Conclusion

In this paper, we have proposed a novel framework to address the problem of learning from ambiguously labeled data. In contrast to the existing methods which directly formulate it as a classification problem, we provide a novel perspective by formulating it as a sample-label correspondence task with partial permutation matrix (PPM) optimization, where the intra-image and inter-image sample-label relations are used in a principled way. To efficiently solve the the proposed formulation, a scalable algorithm based on ADMM is proposed to cope with medium and even large scale data. Once the sample-label correspondences are obtained, we can adopt the standard supervised learning method like SVM for unseen data. Experiments on two datasets show that the proposed method outperforms most of the existing algorithms, which shows the efficacy of our method.

Acknowledgement: This study is supported by the research grant for the Human Sixth Sense Programme at the Advanced Digital Sciences Center from Singapore's Agency for Science, Technology and Research (A*STAR) and the Multi-plAtform Game Innovation Centre (MAGIC) in Nanyang Technological University, Singapore.

References

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, 2003. 1, 2, 6, 7

[2] K. Barnard, P. Duygulu, D. Forsyth, N. Freitas, D. Blei, and M. Jordan. Matching words and pictures. *JMLR*, 3:1107–1135, 2003. 1, 2

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

[4] T. Berg, A. Berg, J. Edwards, M. Maire, R. White, Y.-W. Teh, E. Miller, and D. Forsyth. Faces and names in the news. In *CVPR*, 2004. 1

[5] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3:1–22, March 2010. 4

[6] R. C. Bunescu and R. J. Mooney. Multiple instance learning for sparse positive bags. In *ICML*, 2007. 1, 6, 7

[7] R. Cabral, F. Torre, J. Costeira, and A. Bernardino. Matrix completion for multi-label image classification. In *NIPS*, 2011. 3

[8] J. Cai, E. Candes, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM J. on Optimization*, 4:1956–1982, 2010. 5

[9] J. Cai and S. Osher. Fast singular value thesholding without singular value decomposition. In *UCLA CAM Report*, 2010. 5

[10] C. Chen, C. Wei, and Y. Wang. Low-rank matrix recovery with structural incoherence for robust face recognition. In *CVPR*, 2012. 1, 2, 3

[11] B. Cheng, G. Liu, J. Wang, Z. Huang, and S. Yan. Multi-task low-rank affinity pursuit for image segmentation. In *ICCV*, 2011. 3

[12] T. Cour, B. Sapp, and B. Taskar. Learning from partial labels. *Journal of Machine Learning Research*, 12:1501–1536, May 2011. 1, 2, 5, 6, 7

[13] W. Deng, J. Hu, and J. Guo. Extended SRC: Undersampled face recognition via intraclass variant dictionary. In *T-PAMI*, volume 34, pages 1864–1870, Sep 2012. 1

[14] L. Duan, D. Xu, I. W. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. In *CVPR*, 2010. 5

[15] M. Everingham, J. Sivic, and A. Zisserman. Hello! my name is... buffy–automatic naming of characters in TV video. In *BMVC*, 2006. 6

[16] M. Everingham, J. Sivic, and A. Zisserman. "who are you?" - learning person specific classifiers from video. In *BMVC*, 2006. 6

[17] M. Fazel. Matrix rank minimization with applications. *PhD Thesis*, March 2002. 4

[18] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Automatic face naming with caption-based supervision. In *CVPR*, 2008. 2, 4, 6, 7

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

[20] M. Jamieson, A. Fazly, S. Dickinson, S. Stevenson, and S. Wachsmuth. Learning structured appearance models from captioned images of cluttered scenes. In *ICCV*, 2007. 1, 2

[21] K. Jia, T.-H. Chan, Z. Zeng, and Y. Ma. ROML: A robust feature correspondence approach for matching objects in a set of images, submitted to *T-PAMI*. 2012. 5

[22] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *T-PAMI*, 35(1):171–184, Feb 2013. 3

[23] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 6

[24] J. Luo and F. Orabona. Learning from candidate labeling sets. In *NIPS*, 2010. 1, 2, 3, 4, 5, 6, 7

[25] O. Maron and A. Ratan. Multiple-instance learning for natural scene classification. In *ICML*, pages 341–349, 1998. 1

[26] R. Oliveira, J. Costeira, and J. Xavier. Optimal point correspondence throught the use of rank constraints. In *CVPR*, 2005. 3

[27] Y. Peng and A. Ganesh. RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images. In *CVPR*, 2010. 3, 4

[28] M. Turk and A. Pentland. Face recognition using eigenfaces. In *CVPR*, 1991. 1, 3

[29] P. Viola and M. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, May 2004. 6

[30] C. Wang, D. Blei, and F. Li. Simultaneous image classification and annotation. In *CVPR*, 2009. 1

[31] Y. Wang and G. Mori. A discriminative latent model of image region and object tag correspondence. In *NIPS*, 2010. 1, 2, 3

[32] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *T-PAMI*, 31(2):210–227, Feb 2009. 1

[33] O. Yakhnenko and V. Honavar. Multi-instance multi-label learning for image classification with large vocabularies. In *BMVC*, 2011. 1

[34] Z. Zeng, T.-H. Chan, K. Jia, and D. Xu. Finding correspondence from multiple images via sparse and low-rank decomposition. In *ECCV*, 2012. 3