Multiview Matrix Completion for Multilabel Image Classification

Yong Luo, Tongliang Liu, Dacheng Tao, *Fellow, IEEE*, and Chao Xu, *Member, IEEE*

***Abstract*— There is growing interest in multilabel image classification due to its critical role in web-based image analytics-based applications, such as large-scale image retrieval and browsing. Matrix completion (MC) has recently been introduced as a method for transductive (semisupervised) multilabel classification, and has several distinct advantages, including robustness to missing data and background noise in both feature and label space. However, it is limited by only considering data represented by a single-view feature, which cannot precisely characterize images containing several semantic concepts. To utilize multiple features taken from different views, we have to concatenate the different features as a long vector. However, this concatenation is prone to over-fitting and often leads to very high time complexity in MC-based image classification. Therefore, we propose to weightedly combine the MC outputs of different views, and present the multiview MC (MVMC) framework for transductive multilabel image classification. To learn the view combination weights effectively, we apply a cross-validation strategy on the labeled set. In particular, MVMC splits the labeled set into two parts, and predicts the labels of one part using the known labels of the other part. The predicted labels are then used to learn the view combination coefficients. In the learning process, we adopt the average precision (AP) loss, which is particular suitable for multilabel image classification, since the ranking-based criteria are critical for evaluating a multilabel classification system. A least squares loss formulation is also presented for the sake of efficiency, and the robustness of the algorithm based on the AP loss compared with the other losses is investigated. Experimental evaluation on two real-world data sets (PASCAL VOC' 07 and MIR Flickr) demonstrate the effectiveness of MVMC for transductive (semisupervised) multilabel image classification, and show that MVMC can exploit complementary properties of different features and output-consistent labels for improved multilabel image classification.**

***Index Terms*— Image classification, transductive, multi-label, multi-view, matrix completion, average precision.**

Manuscript received May 17, 2014; revised October 15, 2014; accepted March 31, 2015. Date of publication April 9, 2015; date of current version April 29, 2015. This work was supported in part by the National Basic Research Program of China under Grant 2011CB302400, in part by the National Natural Science Foundation of China under Grant 61375026, Grant 2015BAF15B00, and Grant JCYJ20120614152136201, and in part by the Australian Research Council under Project FT-130101457, Project DP-140102164, and Project LP-140100569. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Richard J. Radke.

Y. Luo and C. Xu are with the Key Laboratory of Machine Perception, School of Electronics Engineering and Computer Science, Ministry of Education, Peking University, Beijing 100871, China (e-mail: yluo180@gmail.com; xuchao@cis.pku.edu.cn).

T. Liu and D. Tao are with the Centre for Quantum Computation & Intelligent Systems and the Faculty of Engineering and Information Technology, University of Technology, Sydney, 81 Broadway Street, Ultimo,NSW 2007, Australia (email: tliang.liu@gmail.com, dacheng.tao@uts.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2015.2421309

I. Introduction

MULTI-LABEL image classification, where multiple labels are assigned to a given image, is useful in many web-based image analytic-based applications. For example, keywords can be automatically assigned to an uploaded web image so that annotated images may be searched directly using text-based image retrieval systems.

Dozens of multi-label algorithms have been proposed in the past decade [1]–[7]. However, none of these methods are able to handle missing features, or cases where parts of the training data labels are unknown. Many of the algorithms lack robustness to outliers and background noise. In order to overcome these limitations, matrix completion (MC) has recently been introduced as an alternative methodology for transductive (semi-supervised) multi-label classification [8], [9]. In particular, the MC-based multi-label classification concatenates the feature and label matrices, and then completes the unknown entries (either features or labels) in the concatenated matrix by the use of the rank minimization criterion. In this way, the MC-based methods can be used not only to infer labels of the unlabeled data, but also to estimate values of the missing features, and denoise the observed features and labels.

Although MC-based algorithms are robust for general transductive multi-label classification tasks, they cannot directly handle those image classification problems [10]–[12] that include images represented by multi-view features. A popular solution has been to concatenate all the features into a long vector, but this strategy not only ignores the physical interpretations of different features, but also encounters an over-fitting problem given frequently limited labeled training samples and high dimensional image features. Besides, the feature concatenation often leads to a very large matrix to be completed. Thus the time cost is very high and sometimes intolerable [13].

To avoid these drawbacks, we propose to weightedly combine the MC based classification outputs of different views, and develop a new framework, namely, multi-view matrix completion (MVMC) for handling multi-view features in semi-supervised multi-label image classification. To learn the view combination coefficients, we firstly perform a two-fold cross validation procedure on the labeled set for each view. That is, we divide the labeled training samples into two (usually equal) sets. Labels in one set are assumed to be unknown and are completed using the label information in the other set. In fact the labeled training samples have been annotated, and thus we propose to linearly combine the predicted labels of all the views to approximate

1057-7149 © 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

the ground-truth labels. In this way, we learn the combination coefficients of different views. In the learning of the coefficients, we propose to directly optimize average precision (AP), which is a critical criterion in evaluating a multi-label classification algorithm. We also present a formulation that adopts the least squares (LS) loss, which is quite efficient although not so proper as the AP loss for multi-label classification. Finally, for each view, the labels of the unlabeled and test data are predicted using matrix completion, and the obtained predictions of all the views are combined using the learned coefficients. The proposed algorithms tend to assign higher weight to the view carrying more discriminative information, and thus explore the complementary nature of different views.

In order to evaluate our MVMC algorithms (MVMC-LS and MVMC-AP), we have used two challenging datasets, PASCAL VOC' 07 [14] and MIR Flickr [15]. To the best of our knowledge, there are no other algorithms which employ multi-view matrix completion. Therefore, in order to assess performance, we first compared our algorithms with the best single view (BMC), concatenation of all the views (CMC), and average the outputs of different views (AMC) in terms of mean average precision (mAP), mean area under the ROC curve (mAUC) and hamming loss (HL). To further verify the effectiveness of MVMC, we compared MVMC-AP with some popular and competitive feature-level [16], [17] and classifier-level [18] multi-view approaches, as well as some competitive multi-label classification methods [4], [5]. The experimental results show that our novel approach outperforms the current state-of-the-art.

The main contributions of this paper are: 1) the cross validation strategy that learns the view combination coefficients in the proposed multi-view matrix completion framework for multi-label image classification; 2) the developed solutions for the constrained optimization problems with the AP loss, as well as the LS loss. The former is particular suitable for multi-label classification, and the latter is for the sake of efficiency; 3) the robustness analysis of the AP loss compared with the LS and hinge loss in MVMC.

The rest of the paper is organized as follows. We firstly review some related work on matrix completion and multi-view learning in Section II. Section III summarizes the recent work on utilizing matrix completion for transduction and multi-label classification. In Section IV, we present the proposed MVMC framework, as well as the MVMC-LS and MVMC-AP algorithms by choosing different loss functions. Moreover, the robustness of the different algorithms is analyzed. The experimental results are presented in Section V. Finally, we conclude this paper in Section VI and prove the main theorem of this paper in Section VII.

II. Related Work

A. Matrix Completion

Matrix completion (MC), as the name suggested, is to fill in the unknown entries of an uncompleted matrix M. Without any assumption about the nature of the matrix, the completion is impossible. We usually assume the matrix we want to recover is of low-rank [19]. Then our goal is to find a matrix X so that the errors between X and M on the known entries are as small as possible, and the rank of X is minimized. The rank minimization [20] problem is NP-hard and thus is of little practical use [21]. Fortunately, $\text{rank}(X)$ can be replaced by its convex envelope, the nuclear norm $\|X\|_*$, which is the sum of X's singular values [22]. On the basis of this relaxation, dozens of methods have been proposed for matrix completion [13], [21], [23], [24]. A representative work was done by Candès and Recht [21]. They showed that the minimization of $\|X\|_*$ and $\text{rank}(X)$ has the same solution under broad assumptions of incoherence, and proved that a limited number of samples are needed to recover a low-rank matrix. Besides, a semi-definite programming (SDP) algorithm was proposed to efficiently solve the nuclear norm minimization problem. However, this algorithm cannot handle large size matrix, and thus a singular value thresholding (SVT) algorithm was developed in [13]. SVT can be expressed exactly as a linearized Bregman iteration, and the key step is the derived matrix shrinkage operation. Although SVT is shown to be efficient when completing large size matrices, it may fail (diverges, or does not solve the problems in a tolerate time, or yields a very inaccurate solution) for matrices which are not of very low-rank (e.g., an 1000×1000 matrix of rank 50). Therefore, a more robust algorithm, fixed point continuation (FPC) [23], was proposed. Similar to SVT, the matrix shrinkage is also utilized in FPC. The difference is that FPC employed an operator splitting technique. To reduce the complexity and improve recoverability (the ability to recover matrices of either small or moderate rank with tolerate error and speed), the authors further presented an approximate SVD based FPC (FPCA), which was demonstrated empirically to have much better recoverability than SDP (for MC) and SVT, etc. For example, the authors show that "for matrices of size 1000×1000 and rank 50, FPCA can recover them with a relative error of 10^{-5}, in about 3 minutes by sampling only 20 percent of the matrix elements". But SDP and SVT, etc. do not have such a good recoverability property (since in these methods, the algorithm diverge, or the time cost is intolerant, or the solution is inaccurate).

To recovery a low-rank matrix corrupted with arbitrary large errors, a combination of the nuclear norm and the l_1-norm should be minimized. Lin et al. [25] extended the classical augmented Lagrange multipliers (ALM) for solving this minimization problem efficiently. In particular, the exact ALM (EALM) method proposed in [25] was proved to have a pleasing convergence speed. The improved version, inexact ALM (IALM), was shown to be more precise and much faster than the state-of-the-art solvers, such as the accelerated proximal gradient (APG) algorithm [26]. Recently, matrix completion was introduced for transductive (semi-supervised) multi-label learning and we will depict it in section III.

B. Multi-View Learning

Multi-view learning is an active research topic in recent years [27]. The multiple views can be the different viewpoints of an object in the camera, or the various descriptions of a given sample. We focus on the latter in this paper, and the goal is to learn to fuse the different descriptions.

Lots of methods have been proposed in the recent decades for multi-view classification [28], retrieval [18], clustering [29], etc. In this section, we mainly review the classification methods [30]–[35], although most of them are also amenable for other applications. According to the level of the fusion being carried out, the multi-view classification methods can be grouped into two major categories: feature-level fusion and classifier-level fusion. We further divide them into four sub-categories: similarity fusion and unified subspace learning for the feature level, output/decision fusion and interactive fusion for the classifier level.

1) Feature-Level Fusion: A direct strategy for feature-level fusion is to concatenate the different kinds of features into a long vector. This often leads to the curse of dimensionality problem and thus it is not practical. To this end, many sophisticated techniques are developed, which include those similarity space fusion (mostly kernel fusion by now) and multi-view subspace learning approaches.

Similarity Space Fusion: As far as we know, most of the current works on similarity space fusion are implemented in the form of kernel fusion. Multiple kernel learning (MKL) [36] is one of the most representative framework for kernel fusion. For example, in [37], a combination of different kernels built on different features sets were utilized for protein prediction. MKL was also used for dimensionality reduction of the multi-view data based on graph embedding [38]. McFee and Lanckriet [39] presented a method to combine multiple kernels in a proposed partial order embedding algorithm. The method learns a set of kernel mappings to induce a unified multi-modal similarity space, where the human perceptual information expressed by relative comparisons is incorporated. Kloft et al. [17] extended the traditional l_1-norm MKL to arbitrary norms, and showed that the non-sparse MKL was superior to the state-of-the-art in combining different feature sets for biometrics recognition.

Unified Subspace Learning: In the feature-level fusion, another set of algorithms is on multi-view subspace learning [40], [41]. Canonical correlation analysis (CCA) [42] is one of the most popular methods for two-view learning, and seeks a subspace where the given two views are maximally correlated. SVM-2K [43] combines Kernel CCA (KCCA) and support vector machine (SVM) in a single optimization problem. Recently, White et al. [44] proposed a convex formulation for learning a shared subspace of multiple sources. In the learned subspace, conditional independence constraints are enforced.

2) Classifier-Level Fusion: Schemes in this category either learn classifiers of different views independently or interactively.

Output or Decision Fusion: Individual classifiers are created for different views and then the outputs or decisions are fused. In [45], the SVM outputs are firstly converted to probabilistic scores, and then concatenated as the input of an SVM for final classification. This method was shown to outperform the simple feature concatenation, followed by an SVM. Such an approach was called hierarchical SVM in [18], and is compared with several other popular classifier fusion strategies, such as weighted sum of outputs and majority voting. Fumera and Roli [46] gave a theoretical analysis of the linear combination of multiple classifiers, and the effectiveness of their analytical model was confirmed by the experimental results. A thorough study on the weighted voting methods for classifier fusion was presented in [47], where the neural network was adopted to estimate the combination weights.

Interactive Fusion: Methods in this sub-category communicate information with other views when learning classifier of the current view. Lots of these methods are semi-supervised and naturally two-view, such as co-training [48], co-regularization [49], etc. In the co-training framework, unlabeled samples classified by one view with high confidence were put in the labeled pool of the other view. Such a process was repeated until the classification performance on the validation dataset decreased. Although succeed empirically, co-training is based on the compatibility and class conditional independence assumptions, which are usually too restrictive to be satisfied [50]. The co-EM algorithm [51] bootstrapped samples in a similar way like co-training, but the unlabeled samples were labeled probabilistically in a batch mode using expectation-maximization (EM). By formulating the linear classifier in a probabilistic framework, SVM was introduced as the base classifier in co-EM [52]. In [49], classifiers of different views are enforced to be agreed on unlabeled data by the use of a regularization term, in a graph-based semi-supervised framework. This is called co-regularization and a similar idea was utilized in [53], under the theme of manifold regularization [54].

It was shown in [18] and [45] that classifier-level fusion outperforms simple feature concatenation, while sophisticated feature-level fusion can usually be better than classifier-level fusion since the raw information is preserved [55]. The proposed MVMC method belongs to the classifier-level fusion, and is particular suitable for transductive (semi-supervised) multi-label classification. It was demonstrated empirically in this paper that MVMC can outperform some competitive classifier-level and feature-level multi-view learning approaches when limited labeled data are available. Experiments also show that MVMC is superior to competitive multi-label [4] and the recently proposed semi-supervised multi-label classification method [5].

III. Transduction With Matrix Completion

This section briefly introduces the framework of utilizing matrix completion for transductive (semi-supervised) multi-label learning [8]. Given a set of n samples $D_n = (\mathrm{x}_j^0, \mathrm{y}_j^0)_{j=1}^n$ with each $\mathrm{x}_j^0 \in \mathbb{R}^d$, and the corresponding label vector $\mathrm{y}_j^0 \in \{-1, 1\}^m$. We construct the feature matrix $X^0 = [\mathrm{x}_1^0, \ldots, \mathrm{x}_n^0] \in \mathbb{R}^{d\times n}$ and the label matrix $Y^0 = [\mathrm{y}_1^0, \ldots, \mathrm{y}_n^0] \in \mathbb{R}^{m\times n}$. Some entries of X^0 may be missing, and our aim is to predict the unknown entries in Y^0. To solve this problem, we consider the linear classification model

$$\mathrm{y}_j = W\mathrm{x}_j + \mathrm{b} = [W\ \mathrm{b}]\begin{bmatrix}\mathrm{x}_j\\ 1\end{bmatrix}, \tag{1}$$

where $W \in \mathbb{R}^{m\times d}$ is the weight matrix, $\mathrm{b} \in \mathbb{R}^m$ is the bias vector, $\mathrm{y}_j \in \mathbb{R}^m$ contains the soft labels of the j'th sample

and the hard label y_{ij}^0 can be obtained by $\text{sign}(y_{ij})$. Then we can construct the underlying feature matrix $X = [\mathrm{x}_1, \ldots, \mathrm{x}_n]$, the soft label matrix $Y = [y_1, \ldots, y_n]$, and concatenate them into an $(m + d + 1) \times n$ matrix $Z = [Y; X; 1^T]$. The stacked matrix is of low rank since each row of Y can be represented by a linear combination of the rows in $[X; 1^T]$, i.e., $\text{rank}([Y; X; 1^T]) = \text{rank}([X; 1^T])$. Actually, the feature matrix X can also be assumed to be of low rank, i.e., $\text{rank}(X) \ll min(d, n)$. This assumption is reasonable since high dimensional data typically lies on a manifold, and the assumption succeed in many component analysis techniques, e.g., principal component analysis (PCA) and Fisher's linear discriminant analysis (FLDA).

Let $Z^0 = [Y^0; X^0; 1^T]$ and Ω_X, Ω_Y be the set of known entries in X^0 and Y^0 respectively. To recover the underlying feature matrix and predict the soft labels, Goldberg et al. [8] proposed to minimize the rank of Z, as well as the error between Z and Z^0 for the entries in Ω_X, Ω_Y. Thus the optimization problem, MC-1, is

$$\begin{aligned} \underset{Z}{\text{argmin}} \quad & \mu \|Z\|_* + \frac{1}{|\Omega_X|} \sum_{i,j \in \Omega_X} c_x(z_{ij}, z_{ij}^0) \\ & + \frac{\lambda}{|\Omega_Y|} \sum_{i,j \in \Omega_Y} c_y(z_{ij}, z_{ij}^0), \\ \text{s.t.} \quad & \mathrm{z}_{(m+d+1)} = \mathbf{1}^T. \end{aligned} \tag{2}$$

where $\|Z\|_*$ is the nuclear norm (sum of singular values) of Z, c_x and c_y are the loss function for the features and labels respectively. Both μ and λ are the trade-off parameters. In [8], c_x is chosen to be the least squares loss and c_y is the log loss.

The problem (2) can be solved by a modified fixed point continuation (FPC) algorithm, which is to alternate between the gradient descent, $A^k = Z^k - \tau g(Z^k)$, and the shrinkage $Z^{k+1} = S_{\tau\mu}(A^k)$. The step size τ can be easily computed according to [8] and [9]. Here, k is the iteration step and $g(Z^k)$ is the matrix gradient given by

$$g(z_{ij}) = \begin{cases} \dfrac{\lambda}{|\Omega_Y|} \dfrac{-z_{ij}^0}{1 + \exp(z_{ij}^0 z_{ij})}, & \text{if} z_{ij} \in \Omega_Y \\ \dfrac{1}{|\Omega_X|}(z_{ij} - z_{ij}^0), & \text{if} z_{ij} \in \Omega_X \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

However, images are usually represented by feature histograms, which consist of positive values naturally. The formulation (2) will introduce negative values to the features. Therefore, in [9], all feature vectors are constrained to be positive or belong to a simplex, and the corresponding MC-Pos and MC-Simplex algorithms are induced. Besides, it is suggested in [9] to use the Pearson's χ^2 distance for c_x to take the asymmetry of the feature histogram into consideration. For the labels, the authors used a generalized version of the log loss, i.e., $c_y(u, v) = \frac{1}{\gamma}\log(1 + \exp(-\gamma uv))$, which is approximate to the hinge loss, and more suitable for classification than the original log loss.

IV. Multi-View Matrix Completion

In this section, we first present our multi-view matrix completion (MVMC) framework, and then develop an algorithm that directly optimizes average precision for transductive (semi-supervised) multi-label image classification. The MVMC framework is depicted in Fig. 1. For all the labeled, unlabeled and test images, we extract different kinds of features, such as SIFT [56] and GIST [57]. Then we construct the stacked matrix $Z^{(v)}$ for the v'th view. To combine the multiple views for matrix completion, a natural idea is to weightedly sum the different feature matrices $X^{(v)}$, $v = 1, \ldots, V$, where V is the number of views. However, the dimensionality of different views varies. Although we can utilize some dimensionality reduction algorithm, such as kernel PCA (KPCA), to preprocess the features, the summation of different matrices lacks physical interpretation. Therefore, we propose to combine the output label matrices of different views. KPCA is still employed to preprocess the features to significantly reduce the time complexity of the MC algorithm. The MC-1 algorithm is adopted to complete the matrix for each view since the processed features can be either positive or negative. We use the least squares loss for c_x and the generalized log loss (a smooth approximation of the hinge loss) for c_y. Subsequently, we proposed a two-step algorithm to learn the output combination coefficients of the different views:

- *Generate Training Data for Each View:* The training data we referred here is not the features to complete the matrix $Z^{(v)}$, but the output labels to learn the combination θ. We generate the data by assuming parts (e.g. half) of the labeled data as unlabeled, and predict their labels using the other parts. Such a process is then conducted conversely. In this way, we obtain the predicted labels $Y_l^{(v)}, v = 1, \ldots, V$ of the labeled data, whose labels Y_l^0 are actually known.
- *Learn the Combination Coefficients:* In our formulation, the final output is a linear combination of the multiple outputs obtained from different views. Thus, we use the weighted summation $Y_l = \sum_{v=1}^V \theta_v Y_l^{(v)}$ to approximate the ground-truth Y_l^0. By minimizing the approximation error, we can learn the weight θ.

Finally, for each view, we predict the labels of the unlabeled and test data by utilizing all the labeled data. The multiple predictions are combined with the learned θ.

A. The General Formulation for Learning θ

To simplify the derivation, we convert the original 2D index (i, j) of a sample to 1D k by the use of vectorization. Let us define Ω_{Y_l} as the set of all the known labels' entries, and $\{p_k^{(v)}\}, k \in \Omega_{Y_l}$ as the set of generated data (predicted values of the labeled data) from the v'th view. Thus our prediction function can be written as $f(\mathrm{p}_k) = \sum_{v=1}^V \theta_v p_k^{(v)} = \theta^T \mathrm{p}_k$, where $\mathrm{p}_k = [p_k^{(1)}, \ldots, p_k^{(V)}]^T$ is the predictions of all the views for the k'th sample. Then we have the following optimization problem,

$$\begin{aligned} \underset{\theta}{\text{argmin}} \quad & \frac{1}{2N} \sum_{k=1}^N L\left(f(\mathrm{p}_k), y_k^0\right) + \frac{\eta}{2}\|\theta\|_2^2, \\ \text{s.t.} \quad & \sum_{v=1}^V \theta_v = 1, \quad \theta_v \geq 0, \end{aligned} \tag{4}$$

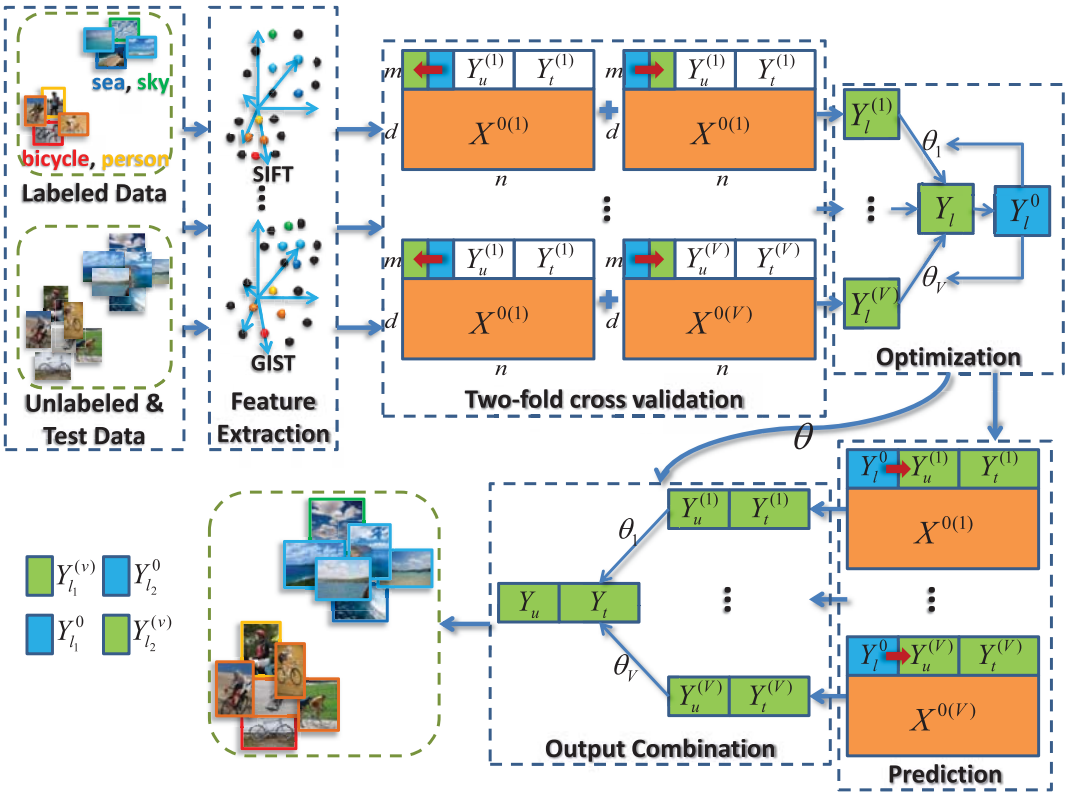


Fig. 1. System diagram of the multi-view matrix completion framework for transductive (semi-supervised) multi-label image classification. For all the data, features from different views (e.g. SIFT [56], GIST [57], etc.) are extracted. Then a two-step algorithm is performed: 1) generate data for learning θ by performing a two-fold cross validation procedure on the labeled data set for each view; 2) learn θ with the constructed optimization problem and the generated data. Finally, for all the views, we predict the labels of the unlabeled and test data, and combine the outputs using the learned θ.

where $N = |\Omega_{Y_l}|$ is the number of entries in Ω_{Y_l}, which is equal to $n_l \times m$. Here, n_l is the number of labeled samples. The regularization term $\|\theta\|_2^2$ is used to control the model complexity, and η is the trade-off parameter. Here, L is some pre-defined convex loss. It is easy to verify that the problem (4) is convex, and thus we can obtain the global solution. In this paper, we first choose L to be the least squares (LS) loss, which will lead to a quite efficient solution.

B. A Least Squares Formulation of MVMC (MVMC-LS)

If we choose L to be the least squares loss, i.e. $L(f(x), y) = (f(x) - y)^2$, then the optimization problem becomes

$$\begin{aligned} \underset{\theta}{\operatorname{argmin}} \quad & \frac{1}{2N}\sum_{k=1}^{N}\left((\theta^T p_k) - y_k^0\right)^2 + \frac{\eta}{2}\|\theta\|_2^2, \\ \text{s.t.} \quad & \sum_{\upsilon=1}^{V}\theta_\upsilon = 1, \quad \theta_\upsilon \geq 0, \end{aligned} \tag{5}$$

Let $P = [p_1, \ldots, p_N] \in \mathbb{R}^{V\times N}$ and $\mathrm{y}^0 = [y_1^0, \ldots, y_N^0]^T$. The objective of the above problem can be written in a compact form as $\frac{1}{2N}\|P^T\theta - \mathrm{y}^0\|_F^2 + \frac{\eta}{2}\|\theta\|_2^2$, and then the problem (5) can be reformulated as

$$\begin{aligned} \underset{\theta}{\operatorname{argmin}} \quad & \frac{1}{2}(\theta^T H\theta - 2\theta^T \mathrm{h}) + \frac{\eta}{2}\|\theta\|_2^2, \\ \text{s.t.} \quad & \sum_{\upsilon=1}^{V}\theta_\upsilon = 1, \quad \theta_\upsilon \geq 0, \end{aligned} \tag{6}$$

where the constant term $\frac{1}{N}(\mathrm{y}^0)^T\mathrm{y}^0$ has been omitted, $H = \frac{1}{N}PP^T$ is a $V \times V$ semi-definite and usually full-rank matrix, and $h = [h_1, \ldots, h_V]^T$ with each $h_\upsilon = \frac{1}{N}P_{(\upsilon)}^T\mathrm{y}^0$. Here, $P_{(\upsilon)}$ denotes the υ'th row of the matrix P, that is the prediction from the υ'th view. To solve this problem, the coordinate descent algorithm is adopted. In each round of iteration, only two variables θ_i and θ_j are selected to update, while the others are fixed. We firstly only consider the constraint $\sum_{\upsilon=1}^{V}\theta_\upsilon = 1$. By using the Lagrange method and note that $\theta_i + \theta_j$ do not change in the iteration, we have the following solution for updating θ_i and θ_j:

$$\begin{cases} \theta_i^* = \dfrac{\eta(\theta_i + \theta_j) + (h_i - h_j) + \varepsilon_{ij}}{(H_{ii} - H_{ij} - H_{ji} + H_{jj}) + 2\eta}, \\ \theta_j^* = \theta_i + \theta_j - \theta_i^*, \end{cases} \tag{7}$$

where $\varepsilon_{ij} = (H_{ii} - H_{ij} - H_{ji} + H_{jj})\theta_i - \sum_k (H_{ik} - H_{jk})\theta_k$. By further taking the constraint $\theta_\upsilon \geq 0$ into consideration, we have

$$\begin{cases} \theta_i^* = 0, \theta_j^* = \theta_i + \theta_j, \text{if } \eta(\theta_i + \theta_j) + (h_i - h_j) + \varepsilon_{ij} \leq 0, \\ \theta_j^* = 0, \theta_i^* = \theta_i + \theta_j, \text{if } \eta(\theta_i + \theta_j) + (h_j - h_i) + \varepsilon_{ji} \leq 0. \end{cases}$$

From solution (7), we can see that larger h_i implies larger θ_i. This is because $h_i = \frac{1}{N}P_{(i)}^T\mathrm{y}^0$ denotes the similarity between the prediction of the i'th view and the ground-truth.

In spite of the efficiency of the LS formulation, the LS loss is designed to optimize the accuracy performance, which is not appropriate for multi-label classification [58], [59].

TABLE I
A TOY EXAMPLE OF THE PREDICTIONS AND RANKINGS

$r(o)$	1	0	0	0	1	0	0	0
$\hat{o}_1$	0.4	0.3	0.2	0.1	-0.1	-0.2	-0.3	-0.4
$r(\hat{o}_1)$	8	7	6	5	4	3	2	1
$\hat{o}_2$	-0.4	-0.3	-0.2	-0.1	0.1	0.2	0.3	0.4
$r(\hat{o}_2)$	1	2	3	4	5	6	7	8

Thus the obtained solution may be unsatisfactory. The hinge loss used in support vector machine (SVM) [60] is not adopted for the same reason. Besides, the least squares and hinge loss are not robust loss functions [61], [62]. Therefore, we propose to directly optimize the average precision (AP), which is a critical criterion for evaluating the multi-label classification performance, and we can prove that the algorithm that utilizes the AP loss is more robust than that adopts the least squares or hinge loss. To this end, better view combination coefficients $\{\theta_v\}$ can be found hopefully. In the following, we first present the AP formulation of MVMC, and then give some theoretical analysis of the proposed algorithm, i.e., the robustness of the algorithm that adopts the AP loss compared with the least squares and hinge loss.

C. Optimizing Average Precision in MVMC (MVMC-AP)

Different from the least squares formulation, where the loss is point-wise and calculated on two scalar elements, the average precision (AP) loss is computed over two vectors for each label (category) and thus list-wise. Before presenting the AP loss, we first introduce the AP score. Suppose the input space is $\mathcal{C}$ and the output space is $\mathcal{O}$ (rankings[1] over a corpus $\mathcal{S} = \{d_1, \ldots, d_{|\mathcal{S}|}\}$, each d_i is a sample). In this paper (transductive multi-label classification), $\mathcal{C}$ consists of the different labels (categories), which correspond to the possible queries in information retrieval [63]. For a certain label, if $\hat{o}$ is a prediction vector for the samples in $\mathcal{S}$ and o is the corresponding ground-truth, then the AP score can be defined as

$$\mathrm{AP}(r(o), r(\hat{o})) = \frac{1}{N_{pos}} \sum_k Prec@k, \tag{8}$$

where $r(\cdot)$ is a vector of rank values. The ground-truth ranking $r(o)$ has only two rank values, i.e., 1 for the positive samples and 0 otherwise. The prediction ranking $r(\hat{o})$ is the sorting result of the predictions in $\hat{o}$, a larger prediction value corresponding to a higher rank value. Here, $N_{pos} = |\{i : r_i(o) = 1\}|$ is the total number of positive samples, and $Prec@k$ is the percentage of positive samples in the top k samples, where the samples in the corpus are assumed to have been sorted according to the prediction $\hat{o}$.

Table I is a toy example of two predictions. The AP scores are calculated as $\mathrm{AP}(r(o), r(\hat{o}_1)) = \frac{1}{2}(1/1 + 2/5) = 0.7$ and $\mathrm{AP}(r(o), r(\hat{o}_2)) = \frac{1}{2}(1/4 + 2/8) = 0.25$ respectively. However, the accuracy of the two predictions are both 0.5 if we choose the threshold to be zero, and classify the samples predicted to be higher than the threshold as positive. We find that the accuracy is not appropriate in this case since the first classifier assigns much higher rank values to the positive samples than the second classifier, and thus performs much better. Such cases are common in multi-label classification since there are usually much more negative samples than the positive ones for each label. This is an intuition of why adopting the AP loss here, which is defined as

$$\Delta_{ap}(o, \hat{o}) = 1 - \mathrm{AP}(r(o), r(\hat{o})), \tag{9}$$

Usually, the mean of the AP scores of all labels, i.e., mAP is adopted for evaluation. Therefore, the loss calculation will be preformed over all labels. In the following, we show how to incorporate the AP loss into an optimization problem for learning θ.

The central idea of optimizing AP in MVMC is to transform the multi-label classification into a retrieval problem, and regard each label as a query. Given a certain label t, the aim is to find a ranking o that maximizes the discriminant function:

$$F(t, o, \theta) = \theta^T \Psi(t, o), \tag{10}$$

which is assumed to be linear (parameterized by θ) in some combined feature representation given by:

$$\Psi(t, o) = \frac{1}{|\mathcal{S}^t| \cdot |\bar{\mathcal{S}}^t|} \sum_{i:d_i \in \mathcal{S}^t} \sum_{j:d_j \in \bar{\mathcal{S}}^t} \left[o_{ij} (\phi(t, d_i) - \phi(t, d_j)) \right], \tag{11}$$

where d_i and d_j are the samples, $\mathcal{S}^t$ and $\bar{\mathcal{S}}^t$ denote the set of positive and negative samples of S for label t. Here, the pairwise orderings is utilized, i.e., $\mathcal{O} \subset \{-1, 0, +1\}^{|\mathcal{S}| \times |\mathcal{S}|}$. For each $o \in \mathcal{O}$, $o_{ij} = +1$ if d_i is ranked ahead of d_j, $o_{ij} = 0$ if d_i and d_j have equal rank, and $o_{ij} = -1$ otherwise. We assume that the rankings is complete, i.e., o_{ij} is either $+1$ or -1 (never 0).

We can predict a ranking (of the samples) for label t with a learned θ. However, this is not the point of this paper and we only concentrate on learning the weight vector θ. Here, we define the feature mapping function ϕ as $\phi(t, d) = p$. Each element of p corresponds to the prediction of a certain view, and thus $\theta^T \phi(t, d)$ is a combined prediction of different views. For different labels, the feature mappings are the same, only the ground-truth orderings change. In this way, we can learn θ to combine different views by the use of the average precision (AP) loss.

Following [64], we use the structural SVM formulation to learn θ, where an additional simplex constraint is added:

$$\begin{aligned} \arg\min \quad & \frac{1}{2}\|\theta\|^2 + C \sum_{t=1}^{m} \xi_t, \\ \text{s.t.} \quad & \forall t, \forall o \in \mathcal{O} \backslash o_t : \theta^t \delta\Psi_t(o) \geq \Delta(o_t, o) \\ & - \xi_t, \quad \xi_t \geq 0, \sum_v \theta_v = 1, \quad \theta_v \geq 0, \end{aligned} \tag{12}$$

where o_t is the ground-truth ranking for the label t. $C = \frac{1}{2N\eta}$, and we have defined $\delta\Psi_t(o) = \Psi(t, o_t) - \Psi(t, o)$. We propose to solve the problem (12) using an alternating algorithm in

[1] It should be noted that the ranking we refer to here is an ordered sequence and the rank value is a number in the sequence, not the notation "rank" of a matrix we used in Section III.

the dual formulation. Note that for $\theta_v \geq 0, v = 1, \ldots, V$, the constraint $\sum_{v=1}^{V} \theta_v = 1$ can be satisfied by a simple normalization, so we left this sum-to-one constraint to be considered later. By introducing the Lagrangian, we obtain

$$\mathcal{L}(\theta, \xi, \alpha, \beta, \zeta) = \frac{1}{2}\|\theta\|^2 + C\sum_{t=1}^{m}\xi_t - \sum_t \beta_t\xi_t - \sum_v \zeta_v\theta_v - \sum_{t,o\neq o_t} \alpha_{to}\left(\theta^T\delta\Psi_t(o) - \Delta(o_t, o) + \xi_t\right), \tag{13}$$

Taking the partial derivatives of L w.r.t. θ, ξ_t and setting them to be zero,

$$\frac{\partial\mathcal{L}}{\partial\theta} = 0, \Rightarrow \theta = \sum_{t,o\neq o_t}\alpha_{to}\delta\Psi_t(o) + \zeta,$$
$$\frac{\partial\mathcal{L}}{\partial\xi_t} = 0, \Rightarrow C - \alpha_{to} - \beta_t = 0.$$

where $\zeta = [\zeta_1, \ldots, \zeta_V]^T$ and the second identity is equivalent to $0 \leq \alpha_{to} \leq C$ since the Lagrange multiplier $\beta_t \geq 0$. By substituting θ back into (12), we obtain the dual

$$\begin{aligned}\arg\max_{\alpha,\zeta} \quad & \sum_{t,o\neq o_t}\alpha_{to}\Delta(o_t, o) - \frac{1}{2}\sum_{t,o\neq o_t}\sum_{s,\bar{o}\neq o_s}\alpha_{to}\alpha_{s\bar{o}}K_{(to)(s\bar{o})} \\ & - \zeta^T\sum_{t,o\neq o_t}\alpha_{to}\delta\Psi_t(o) - \frac{1}{2}\zeta^2, \\ \text{s.t.} \quad & 0 \leq \alpha_{to} \leq C, \quad \zeta_v \geq 0.\end{aligned} \tag{14}$$

where $K_{(to)(s\bar{o})} = \langle\delta\Psi_t(o), \delta\Psi_s(\bar{o})\rangle$. Let $\alpha = [\alpha_{1,o\neq o_1}^T, \ldots, \alpha_{m,o\neq o_m}^T]^T \in \mathbb{R}^{m|\mathcal{S}|-m}$, $\Delta = [\Delta^T(o_1, o \neq o_1), \ldots, \Delta^T(o_m, o \neq o_m)]^T$ and $\delta\Psi = [\delta\Psi_1(o \neq o_1), \ldots, \delta\Psi_m(o \neq o_m)]$ with each $\delta\Psi_t(o \neq o_t) \in \mathbb{R}^{V\times(|\mathcal{S}|-1)}$, the problem (14) can be written in a compact form as

$$\begin{aligned}\arg\max_{\alpha,\zeta} \quad & \Delta^T\alpha - \frac{1}{2}\alpha^T K\alpha - \zeta^T\delta\Psi\alpha - \frac{1}{2}\zeta^T\zeta, \\ \text{s.t.} \quad & 0 \leq \alpha_{to} \leq C, \quad \zeta_v \geq 0.\end{aligned} \tag{15}$$

We propose to solve this problem using the alternating optimization strategy. For fixed ζ, the problem (15) becomes

$$\begin{aligned}\arg\max_{\alpha} \quad & (\Delta^T - \zeta^T\delta\Psi)\alpha - \frac{1}{2}\alpha^T K\alpha, \\ \text{s.t.} \quad & 0 \leq \alpha_{to} \leq C,\end{aligned} \tag{16}$$

which is a structural SVM [65] formulation with the linear part $(\Delta^T - \zeta^T\delta\Psi)\alpha$. A cutting-plane algorithm is introduced to solve this problem and the most violated constraint can be found using the algorithm presented in [64]. For fixed α, the problem (15) can be reformulated as

$$\begin{aligned}\max_{\zeta} \quad & (-\alpha^T\delta\Psi^T)\zeta - \frac{1}{2}\zeta^T\zeta, \\ \text{s.t.} \quad & \zeta_v \geq 0.\end{aligned} \tag{17}$$

This is a quadratic programming (QP) problem and can be solved quite efficiently using a standard SVM solver. It can be easily verified that the Hessian matrix of (15) $H_e(\alpha, \zeta) = -\begin{bmatrix} K & \delta\Psi^T \\ \delta\Psi & I_V \end{bmatrix}$ is negative semi-definite, where I_V is the $V \times V$ identity matrix, and thus the problem (15) is jointly concave w.r.t. α and ζ. Besides, the sub-problems (16) and (17) are concave w.r.t. α and ζ respectively. Therefore, by alternatively solving (16) and (17), the algorithm will converge to the global solution of (15).

D. Complexity Analysis

The complexity of MVMC has two parts: the first is determined by the MC-based classification of each view, and the second is determined by the learning of the view combination coefficients. In this paper, the MC-based classification is carried out by optimizing the MC-1 problem (2). We can reduce the complexity of the MC-based classification algorithm MC-1 [9] to $(st(m + \bar{d} + n))$ by exploiting the approximate SVD based fixed point continuation (FPCA) [23] algorithm, where m and n are the number of class labels and all samples respectively, $\bar{d}$ is the average feature dimension, t is the iteration numbers, and s is the number of elements in the μ sequence of the continuation step. In common, $s < 50$ and $t < 100$. Thus, the adopted MC-1 can solve large matrix rank minimization problems efficiently.

With regard to the learning of the combination coefficients ${\theta_v}_{v=1}^{V}$, we use the weighted summation $Y_l = \sum_{v=1}^{V}\theta_v Y_l^{(v)}$ to approximate the groundtruth Y_l^0. Weights $\{\theta_v\}_{v=1}^{V}$ are obtained by minimizing the approximation error. The size of each $Y_l^{(v)}$ is $n_l \times m$, where n_l is the number of labeled data. This means that the complexity of the proposed view combination procedure is not dependent on the amount of all samples, but only on the labeled sample size, which is usually small in transductive (semi-supervised) classification. For MVMC-LS, the time complexity is $(V^2(n_l \times m))$, where V is the number of views and is usually smaller than 10. For MVMC-AP, structure SVM is exploited for optimization. According to [66], the time complexity of the cutting-plane method used in structural SVMs is linear in the number of training samples. Thus the computational time cost of learning the view combination coefficients is $(T(n_l \times m))$, where T is the number of iterations and is independent on the number of samples [66]. To this end, the view combination procedure is also very efficient, and according to our experience, it is more efficient than the MC-based classification.

Therefore, the time complexity of MVMC-LS and MVMC-AP are $(Vst(m+\bar{d}+n)+V^2(n_l\times m))$ and $(Vst(m + \bar{d} + n) + T(n_l \times m))$, respectively. The former is often more efficient since V is usually small.

E. Robustness Analysis

In this section, we aim to prove that AP loss is more robust than the least squares and hinge loss when learning θ for MVMC. That is, AP is more tolerate of noise. We present the definition of the robustness here for completeness.

Definition 1 (Robustness [67]): Let $\mathcal{Z}$ be the set that each sample is drawn from, and $\mathbf{s}$ be a training sample set $\{s_1, \ldots, s_n\}$. A learning algorithm $\mathcal{A}$ is $(K, \epsilon(\cdot))$ robust, for $K \in \mathbb{N}$ and $\epsilon(\cdot) : \mathcal{Z}^n \mapsto \mathbb{R}$, if $\mathcal{Z}$ can be partitioned into K disjoint sets, denoted by $\{\mathcal{C}_i\}_{i=1}^{K}$, such that the following holds

for all $\mathbf{s} \in \mathcal{Z}^n$:

$$\forall s \in \mathbf{s}, \quad \forall z \in \mathcal{Z}, \quad \forall i = 1, \ldots, K: \\ \text{if} s, \quad z \in \mathcal{C}_i, \quad \text{then } |l(h_{\mathbf{s}}, s) - l(h_{\mathbf{s}}, z)| \leq \epsilon(\mathbf{s}), \quad (18)$$

where $h_{\mathbf{s}}$ represents the hypothesis learned using $\mathcal{A}$ on the training set $\mathbf{s}$. The sets $\{\mathcal{C}_i\}_{i=1}^K$ can be regarded as sets with elements having some similarity to each other, such as the distance.

Let $\mathcal{A}_{\text{ap}}$, $\mathcal{A}_{\text{ls}}$ and $\mathcal{A}_{\text{svm}}$ denote the learning algorithms which adopt the AP loss, least squares loss and hinge loss in (4) respectively. It is well known that the ranking based criteria are critical for multi-label classification [59], [68]–[70], and we can also prove that $\mathcal{A}_{\text{ap}}$ is more robust than $\mathcal{A}_{\text{ls}}$ and $\mathcal{A}_{\text{svm}}$ when learning θ for MVMC. This is easy to understand since the AP loss is list-wise, while the LS and hinge loss are both point-wise.

Theorem 1: We can partition the sample space $\mathcal{Z}$ into N disjoint sets $\{\mathcal{C}_i\}_{i=1}^N$ such that $\mathcal{A}_{\text{ap}}$ is $(N, 0)$ robust, $\mathcal{A}_{\text{ls}}$ is (N, a) robust and $\mathcal{A}_{\text{svm}}$ is (N, b) robust, where N is a positive integer and $a, b \neq 0$. That is, the $\mathcal{A}_{\text{ap}}$ will tolerate to the noise which do not change the partition of the disjoint sets $\{\mathcal{C}_i\}_{i=1}^N$, while the $\mathcal{A}_{\text{ls}}$ and $\mathcal{A}_{\text{svm}}$ will not.

Remark 1: It can also be seen from a general perspective that $\mathcal{A}_{\text{ap}}$ is more robust than $\mathcal{A}_{\text{ls}}$ and $\mathcal{A}_{\text{svm}}$. That is, for any partitions, the least squares and hinge loss based algorithms cannot be $(K, 0)$ robust, where $K \in \mathbb{N}$, while we can find at least one set of partition in which the AP loss based algorithm is $(K, 0)$ robust.

Remark 2: In Theorem 1, the claim that $\mathcal{A}_{\text{ap}}$ is $(N, 0)$ robust does not mean that for all instances in the training set, the corresponding objective function losses are equal to each other. Instead, $\mathcal{A}_{\text{ap}}$ is $(N, 0)$ robust with respect to a particular partition $\{\mathcal{C}_i\}_{i=1}^N$, and such an assertion means that for all instances in a fixed set $\mathcal{C}_i, i \in \{1, \ldots, N\}$, the objective function losses are equal to each other. The central idea of Theorem 1 is that, as we stated in the proof, for every point P in a fixed prediction space $\mathcal{P}$, there has been a finite number of different rankings. Thus, we can partition the prediction space into a finite number of sets $\{\mathcal{C}_i\}_{i=1}^N$ such that all points in one set $\mathcal{C}_i$ share the same ranking. This is why the objective function losses for all instances in a fixed set are equal to each other and $\mathcal{A}_{\text{ap}}$ is $(N, 0)$ robust. However, $\mathcal{A}_{\text{ls}}$ and $\mathcal{A}_{\text{svm}}$ cannot be $(N', 0)$ robust, where N' is a finite counting number. This is because the corresponding losses have infinitely different values and we could never find a finite number of sets such that all points in one particular set share the same loss value.

The following two lemmas are useful for proving Theorem 1.

Lemma 1 [67]: Fix $\gamma > 0$ and metric ρ of $\mathcal{Z}$, If algorithm $\mathcal{A}$ satisfies

$$|l(\mathcal{A}_{\mathcal{Z}^n}, z_1) - l(\mathcal{A}_(\mathcal{Z}^n), z_2)| \leq \epsilon(\mathcal{Z}^n), \\ \forall z_1, \quad z_2 \in \mathcal{Z}^n, \quad \rho(z_1, z_2) \leq \gamma, \quad (19)$$

and $N(\gamma/2, \mathcal{Z}, \rho) \leq \infty$, then $\mathcal{A}$ is $(N(\gamma/2, \mathcal{Z}, \rho), \epsilon(\mathcal{Z}^n))$ robust.

Lemma 2 [71], [72]: Let B be a ball of radius r in an N-dimensional Banach space and $\epsilon > 0$. There exists a subset $B_\epsilon \subset B$ such that $|B_\epsilon| \leq (4r/\epsilon)^N$ and $\forall z \in B$, $\exists z' \in B_\epsilon$ with $\rho(z, z') \leq \epsilon$, where ρ is the metric of the Banach space.

We defer the detailed proof of Theorem 1 in the last section.

V. Experiments

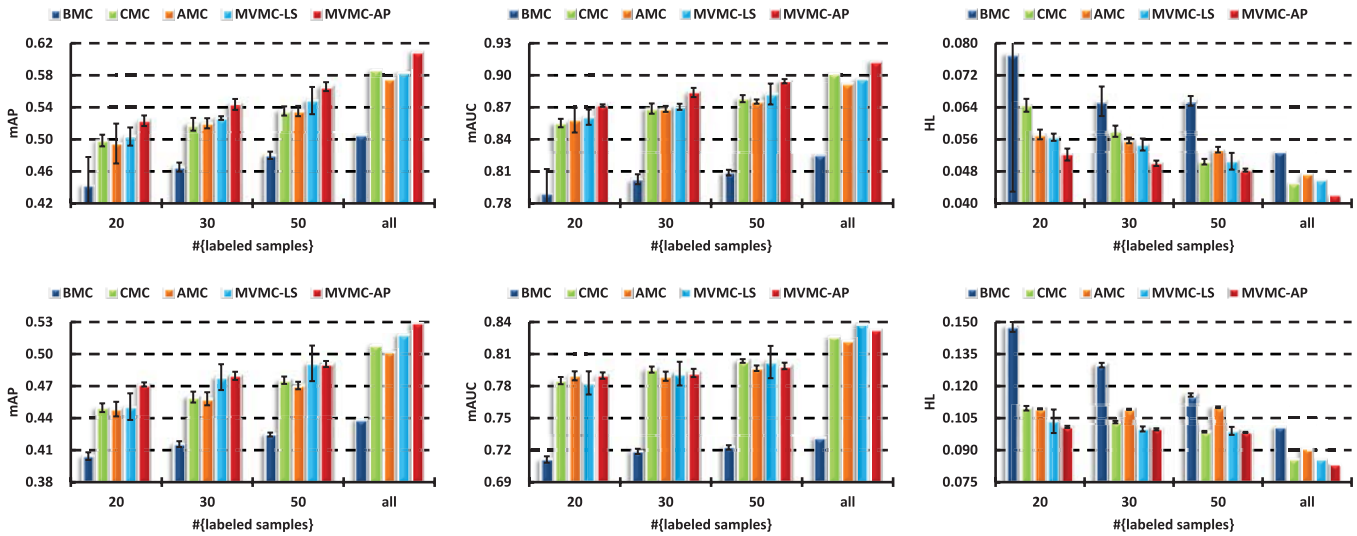
In the experimental evaluation, we firstly compare the proposed MVMC-LS and MVMC-AP algorithms with the following MC-based strategies: 1) BMC, which is to use a single view that performs the best in MC; 2) CMC, which is to simply concatenate features of the multiple views in MC; 3) AMC, which is to average the MC outputs of the multiple views. Then we show the view combination coefficients learned by our MVMC-AP method. Finally, we compare MVMC-AP with some popular and competitive feature-level fusion [16], [17], and classifier-level fusion [18] algorithms, as well as some competitive multi-label [4] and recently proposed semi-supervised multi-label classification approaches [5]. Before all of these evaluations, we present the datasets and features we used, as well as our experimental settings.

A. Datasets, Features and the Evaluation Criteria

Our experiments are conducted on two popular datasets, PASCAL VOC' 07 (VOC for short) [14] and MIR Flickr (MIR for short) [15]. There are around $10,000$ natural images ($5,011$ training, $4,952$ test) from 20 categories in the VOC dataset. The MIR dataset consists of $25,000$ images (half training, half test) and 38 categories.

In natural image classification and web image annotation, the taxonomy of the images is very large and the category defined for a certain image may vary in different fields. Therefore, it is impossible to manually label abundant training images for every category, and it is common that limited labeled samples are given in natural image classification or annotation [59], [73], especially for the new defined or uncommon categories. One of the advantages of the proposed transductive algorithms is the ability to handle small labeled sample size, since it can effectively utilize the information contained in the images from different views. To empirically demonstrate the effectiveness of MVMC, for both datasets, we only randomly select $n_l = \{20, 30, 50\}$ samples for each category in the training set as labeled, and all the others unlabeled. The labels of the unlabeled and test data are then inferred by our algorithms. Five random choices of the labeled data are used in our experiments. We also present the results of using all the labeled training samples to compare the performance of the algorithms in the fully supervised scenario. Besides, twenty percent of the test data are used for validation, which means that the parameters corresponding to the best performance on the validation set are used for unlabeled and test inference.

The features we used here are from [59], where 15 different image representations and tags are provided. Actually, the 15 representations are constructed from six kinds of visual features, which include two kinds of local features (Hue [74], SIFT [56]), a global representation (GIST [57]), and three global color histograms (Hsv, Lab, Rgb). In this paper, we



(a) (b) (c)

Fig. 2. A comparison of the MC-based approaches. Both the mean and standard deviation are presented. The arrow "↑" indicates the larger the better, while "↓" means a smaller value is better. (Top row: PASCAL VOC' 07; Bottom row: MIR Flickr.) (a) mAP ↑. (b) mAUC ↑. (c) HL ↓.

regard each kind of visual feature as a single view. Then we have seven different views in all, with a tag view included. The dimensionality of the SIFT features we used is $1,000$, and the color features (Hsv, Lab, Rgb) is around $4,000$. The existing MC algorithms either fail to recover such a large size matrix (25000×4000) or the time cost is intolerable. Thus we preprocess the features by KPCA to reduce the time complexity in the MC based image classification. For all the compared methods in this paper, the result dimension after KPCA is fixed to be 50, since we found the different algorithms perform well under this setting.

In the VOC and MIR datasets, the positive and negative samples are quite unbalanced. Thus the traditional accuracy criterion is not proper any more, and we introduce three popular criteria in multi-label classification for evaluation. They are average precision (AP) [75], area under ROC curve (AUC) [76] and hamming loss (HL) [77]. In this paper, AP and AUC are the ranking performance computed under each label. Usually, the mean value over all labels, i.e., mAP and mAUC are reported. HL is utilized to evaluate the label set prediction for each instance. It is not a ranking based criterion but widely used in multi-label classification [69]. A smaller value in HL indicates a better performance.

B. A Comparison With the MC-Based Strategies

There are several direct strategies to make use of the different views in MC: BMC, CMC and AMC. In this section, we demonstrate that learning the output combination coefficients is superior to all of these approaches. In particular, the algorithms compared are:

- *BMC [8], [9]:* using the best single view, i.e., one that achieves the best MC-based classification performance. The MC-based transductive multi-label classification is performed by optimizing (2), where c_x is chosen to be the least squares loss for the KPCA processed features, c_y is tuned with γ in $\{1, 3, 30\}$. The candidate set for choosing λ is $\{10^i \mid i = -4, \ldots, 2\}$. The parameter μ is initialized as $\mu_0 = 0.25\sigma_1$ (σ_1: the largest singular value of Z^0), and decreases with a factor 0.25 in the continuation steps until $\mu = 10^{-12}$.
- *CMC:* feature-level fusion. Concatenating the normalized features of all the views into a very long vector, and then performing MC-based classification.
- *AMC:* classifier-level fusion. Learning separate MC-based classifiers for different views, and then combining the outputs of all the classifiers uniformly. Before the combination, the outputs of each view are converted to the probability scores by the use of a sigmoid function $S(x) = 1/((1 + e^{-x}))$.
- *MVMC-LS:* the proposed multi-view MC algorithm, in which the least squares loss is utilized. Outputs are also converted to probability scores. The trade-off parameter η is tuned on the set $\{10^i \mid i = -2, \ldots, 5\}$.
- *MVMC-AP:* the proposed multi-view MC algorithm with the average precision (AP) loss.

The results for all the compared methods are presented in Fig. 2. A self-test with different number of labeled samples is carried out to see the performance variation with respect to the labeled training size. We can see that the performance of all the presented methods improve when the number of labeled samples increase. By fusing all the views, either in the feature-level or classifier-level, can always be superior to the use of only the best single view. The concatenation (CMC) and average outputs (AMC) methods are comparable with each other. Although the proposed MVMC-LS algorithm is superior to CMC and AMC in most cases, the improvement is not significant, except for the mAP performance on the MIR dataset, while MVMC-AP consistently outperforms them in terms of all the evaluated criteria on the VOC dataset, and usually has small deviations. This demonstrates the effectiveness of the learned coefficients using the AP loss. Similar results can be seen on the MIR dataset, only the mAUC performance of the compared combination methods is comparable. In the

TABLE II

A COMPARISON OF THE MULTI-VIEW AND MULTI-LABEL ALGORITHMS ON THE VOC DATASET

	VOC mAP ↑ vs. #{labeled samples}				
Methods	20	30	50	all	Ranks
HierSVM [18]	0.459±0.032 (6)	0.488±0.026 (5)	0.499±0.020 (6)	0.552 (4)	5.25
SimpleMKL [16]	0.490±0.006 (3)	0.513±0.004 (3)	0.541±0.005 (3)	0.617 (2)	2.75
LpMKL [17]	0.498±0.008 (2)	0.515±0.006 (2)	0.553±0.005 (2)	**0.618 (1)**	1.75
KLS-CCA [4]	0.470±0.011 (5)	0.487±0.010 (6)	0.512±0.005 (5)	0.538 (5)	5.25
DLP [5]	0.481±0.006 (4)	0.499±0.004 (4)	0.513±0.004 (4)	0.533 (6)	4.5
MVMC-AP	**0.523±0.007 (1)**	**0.544±0.007 (1)**	**0.566±0.006 (1)**	0.610 (3)	**1.5**
	VOC mAUC ↑ vs. #{labeled samples}				
Methods	20	30	50	all	Ranks
HierSVM [18]	0.861±0.005 (2)	0.876±0.004 (2)	0.886±0.005 (3)	0.912 (3.5)	2.63
SimpleMKL [16]	0.848±0.005 (5)	0.866±0.004 (4)	0.882±0.003 (4)	0.912 (3.5)	4.13
LpMKL [17]	0.859±0.003 (3)	0.870±0.002 (3)	0.889±0.003 (2)	**0.913 (1.5)**	2.38
KLS-CCA [4]	0.831±0.012 (6)	0.843±0.010 (6)	0.854±0.008 (6)	0.861 (6)	6
DLP [5]	0.852±0.002 (4)	0.861±0.002 (5)	0.866±0.001 (5)	0.877 (5)	4.75
MVMC-AP	**0.872±0.001 (1)**	**0.884±0.004 (1)**	**0.894±0.002 (1)**	**0.913 (1.5)**	**1.13**
	VOC HL ↓ vs. #{labeled samples}				
Methods	20	30	50	all	Ranks
HierSVM [18]	0.059±0.003 (4)	0.061±0.002 (3)	0.061±0.014 (5)	0.058 (5)	4.25
SimpleMKL [16]	**0.052±0.002 (1.5)**	0.051±0.001 (2)	**0.048±0.001 (1.5)**	0.043 (2)	1.75
LpMKL [17]	0.074±0.004 (5)	0.077±0.005 (5)	0.050±0.001 (3.5)	0.050 (4)	4.38
KLS-CCA [4]	0.056±0.001 (3)	0.062±0.001 (4)	0.050±0.001 (3.5)	0.048 (3)	3.38
DLP [5]	0.229±0.000 (6)	0.229±0.000 (6)	0.229±0.000 (6)	0.229 (6)	6
MVMC-AP	**0.052±0.002 (1.5)**	**0.050±0.001 (1)**	**0.048±0.001 (1.5)**	**0.042 (1)**	**1.25**

(↑ indicates "the larger the better"; ↓ indicates "the smaller the better". Mean and std. are reported. The best result is highlighted in boldface. The values in the middle and last columns are average ranks.)

fully supervised case (the "all" column), we find that CMC is superior to AMC and comparable to MVMC sometimes. This is because when large amounts of labeled data are available, the curse-of-dimension problem caused by simple concatenation is alleviated.

C. Analysis of the View Combination Coefficients

In Fig. 3, we show the combination coefficients θ learned by MVMC-AP, together with the mAP by using MC for each view. From the results, we find that the tendency of the weights is consistent with the corresponding mAP in general, i.e., the views with a higher classification performance tend to be assigned larger weights, taking the DenseSIFT (the 2nd view) and the tags (the last view) for example. The three color histogram views (Hsv, Lab, Rgb) have similar discriminative power on the VOC dataset, and thus the weights for them are almost equivalent.

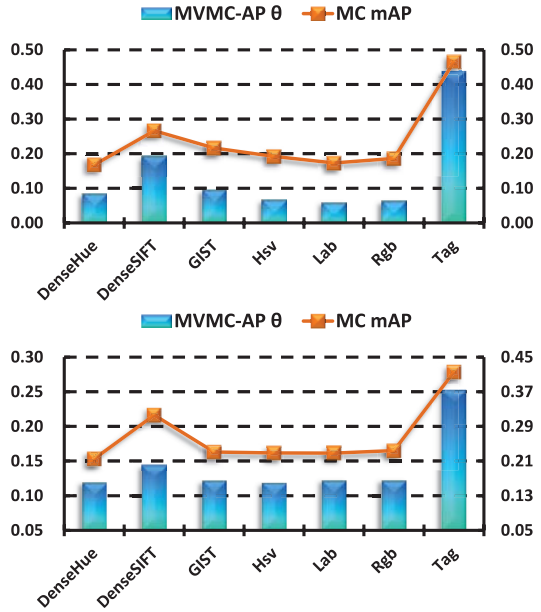


Fig. 3. The view combination coefficients learned by MVMC, together with the mAP of each MC classifier. The coefficients and the mAP scores agree in general. (Top: PASCAL VOC' 07; Bottom: MIR Flickr).

D. Compared With Other Multi-View and Multi-Label Algorithms

Our last set of experiments compare MVMC with some popular and competitive multi-view algorithms, where we learn a binary classifier for each label. Besides, extensive comparisons with the competitive and recently proposed multi-label classification algorithms are also performed. Specifically we compare MVMC-AP with the following methods:

- *HierSVM [18]:* learning separate SVM classifiers for each view, and then fusing the results by using an additional SVM classifier. This is called hierarchical SVM, and the trade-off parameter C for each SVM classifier is optimized over $\{10^i \mid i = -1, \ldots, 6\}$.
- *SimpleMKL [16]:* a very popular and competitive SVM-based multiple kernel learning algorithm. Constructing a kernel for each view, and then learning

TABLE III
A COMPARISON OF THE MULTI-VIEW AND MULTI-LABEL ALGORITHMS ON THE MIR DATASET

	MIR mAP ↑ vs. #{labeled samples}				
Methods	20	30	50	all	Ranks
HierSVM [18]	0.422±0.011 (6)	0.438±0.012 (6)	0.450±0.014 (6)	0.492 (6)	6
SimpleMKL [16]	0.449±0.003 (2)	0.456±0.005 (3)	0.470±0.003 (2)	**0.538 (1.5)**	2.13
LpMKL [17]	0.444±0.007 (4)	0.454±0.008 (4)	0.468±0.009 (3)	**0.538 (1.5)**	3.13
KLS-CCA [4]	0.436±0.007 (5)	0.448±0.008 (5)	0.457±0.008 (5)	0.502 (4)	4.75
DLP [5]	0.447±0.002 (3)	0.457±0.005 (2)	0.464±0.007 (4)	0.498 (5)	3.5
MVMC-AP	**0.472±0.002 (1)**	**0.480±0.004 (1)**	**0.491±0.003 (1)**	0.529 (3)	**1.5**
	MIR mAUC ↑ vs. #{labeled samples}				
Methods	20	30	50	all	Ranks
HierSVM [18]	0.792±0.008 (5)	0.799±0.006 (5)	0.802±0.005 (5)	0.836 (5)	5
SimpleMKL [16]	0.795±0.004 (4)	0.801±0.005 (4)	0.812±0.003 (4)	0.851 (2)	3.5
LpMKL [17]	**0.809±0.003 (1.5)**	0.810±0.003 (2.5)	0.818±0.003 (3)	**0.854 (1)**	**2**
KLS-CCA [4]	0.805±0.006 (3)	0.810±0.006 (2.5)	**0.821±0.003 (1)**	0.848 (3)	2.38
DLP [5]	**0.809±0.002 (1.5)**	**0.815±0.003 (1)**	0.819±0.001 (2)	0.837 (4)	2.13
MVMC-AP	0.790±0.003 (6)	0.792±0.004 (6)	0.799±0.003 (6)	0.833 (6)	6
	MIR HL ↓ vs. #{labeled samples}				
Methods	20	30	50	all	Ranks
HierSVM [18]	0.105±0.009 (2)	0.103±0.006 (2)	0.102±0.005 (2)	0.090 (2)	2
SimpleMKL [16]	0.111±0.001 (3)	0.119±0.001 (3)	0.114±0.001 (3)	0.091 (3)	3
LpMKL [17]	0.170±0.005 (5)	0.167±0.005 (5)	0.142±0.004 (5)	0.119 (4)	4.75
KLS-CCA [4]	0.123±0.000 (4)	0.123±0.000 (4)	0.123±0.000 (4)	0.123 (5)	4.25
DLP [5]	0.231±0.000 (6)	0.231±0.000 (6)	0.231±0.000 (6)	0.231 (6)	6
MVMC-AP	**0.101±0.001 (1)**	**0.100±0.001 (1)**	**0.098±0.001 (1)**	**0.083 (1)**	**1**

(↑ indicates "the larger the better"; ↓ indicates "the smaller the better". Mean and std. are reported. The best result is highlighted in boldface. The values in the middle and last columns are average ranks.)

a linear combination of the different kernels, as well as a classifier based on the combined kernel. The penalty factor C is tuned on the set $\{10^i \mid i = -1, \ldots, 6\}$.

- *LpMKL [17]:* a recent proposed MKL algorithm, which extend MKL to l_p-norm with $p \geq 1$. The penalty factor C is tuned on the set $\{10^i \mid i = -1, \ldots, 6\}$ and we choose the norm p from the set $\{1, 8/7, 4/3, 2, 4, 8, 16, \infty\}$.
- *KLS-CCA [4]:* a least-squares formulation of the kernelized CCA for multi-label classification. The ridge parameter is chosen from the candidate set $\{10^i | i = -3, \ldots, 3\}$. The different views are fused by combining the kernels (similarity matrices) with uniform weights.
- *DLP [5]:* an improved label propagation algorithm that is proposed recently for transductive multi-label learning by considering the label correlation. The parameters α and λ are optimized over the set $\{10^i | i = -5, \ldots, 1\}$ and $\{10^i | i = -4, \ldots, 2\}$ respectively. The parameter K is chosen from $\{10, 20, \ldots, 100\}$. The different views are fused by combining the similarity matrices with uniform weights.

The performance of the compared methods on the VOC and MIR datasets are reported in Table II and III, respectively. Both the mean and standard deviation of the three criteria are presented. From the experimental results, we observe that: 1) the performance improves with an increase of the labeled samples; 2) the multi-label classification methods (KLS-CCA and DLP) are better than HierSVM, but is inferior to other multi-view learning algorithms overall; 3) DLP outperforms KLS-CCA in most cases since the unlabeled information is utilized in transduction. But when the number of labeled data is increased, the improvement deceases and sometimes KLS-CCA is better (e.g., the mAUC performance in the fully supervised case (the "all" column) on the MIR dataset), since the significance of the unlabeled information decreases; 4) On the VOC dataset, the mAP and HL scores of the SimpleMKL method are larger than HierSVM, while the mAUC performance of the latter is better. In general, LpMKL is superior to SimpleMKL and HierSVM. The proposed algorithm consistently outperforms the other three methods; 5) On the MIR dataset, the other methods are comparable with or superior to our algorithm in terms of mAUC, while their mAP and HL performance are poor. Under the HL criterion, SimpleMKL and HierSVM perform well on only one of the two datasets (VOC and MIR respectively), while the proposed MVMC-AP achieves the best performance consistently on both datasets. In particular, we obtain a significant 6.3%, 5.7% and 4.9% improvement in terms of mAP compared with LpMKL, when 20, 30 and 50 labeled samples for each class are used, respectively; 6) In the fully supervised scenario (the "all" column), SimpleMKL and LpMKL are comparable to MVMC under the mAP and mAUC criteria, but the HL performance of our method is the best among all methods. Therefore, the proposed transductive classification approach is particular suitable for the small labeled sample

size problem, and is comparable to the state-of-the-art methods when large amount of labeled data are available.

Besides, the average rank of the proposed alogrithm is smaller than all the other methods in terms of all the three criteria on the VOC dataset, as well as the mAP and HL criteria on the MIR dataset. According to the Friedman test [78], the statistics F_F of mAP, mAUC and HL on the two datasets are (16.09, 11.57), (31.49, 11.79) and (27.82, 137.00) respectively. We can see that all of them are larger than the critical value $F(5, 15) = 2.27$, so we reject the null-hypothesis (the compared algorithm perform equally well).

VI. Conclusion

Matrix completion (MC) has recently been used in transductive (semi-supervised) multi-label classification. It has the advantageous of being efficient, robust to noise, and being able to handle missing data. In existing algorithms, only features from a single view can be used, but there is not a single feature perfect for image classification. We therefore present a multi-view framework to fuse different kinds of features for MC-based multi-label classification. Our framework has the advantage of being able to explore the complementary properties of different views. We have designed two algorithms under the framework, MVMC-LS and MVMC-AP, which differ by the choice of different losses. The robustness of the two algorithms is analyzed.

From the experimental validation on the challenging PASCAL VOC' 07 and MIR Flickr datasets, we mainly conclude that: 1) The classifier-level fusion is better than simple feature concatenation, which is in line with [18], [45]; 2) The learning of the combination coefficients is critical in the classifier-level fusion. Although the least squares formulation is quite efficient for learning the coefficients, the performance is usually not satisfactory. Thus more sophisticated loss should be adopted, such as the average precision loss utilized in this paper. Future works may be to extend MVMC for optimizing other criteria, such as AUC [79], [80], by changing the loss in MVMC.

VII. Proofs of Main Results

In this section, we prove Theorem 1.

Proof: Let $\mathcal{P} \subseteq \mathbb{R}^{V\times N}$, $N = n_l \times m$ be the prediction space, which contains all the possible predictions. For a certain prediction $P \in \mathcal{P}$, each column is a vector $p \in \mathbb{R}^V$, which is the predictions of V views for a certain sample, and a given label. The formulation (12) is to find a θ to combine all the views such that the combined prediction will have the largest mean average precision over all labels.

Let $\{c_1, \ldots, c_{\mathcal{N}(\gamma/2,p,\rho)}\}$ be a $\gamma/2$ cover of p, where ρ is a 2-norm metric. Suppose that every prediction returned by matrix completion has the range $[-b, b]$. Then, according to Lemma 2, we have

$$\mathcal{N}(\gamma/2, p, \rho) \leq (\frac{4\sqrt{V}b}{\gamma/2})^V.$$

We can partition $\mathcal{P}$ into $(n_l!)^{mV}2^{n_l m}\mathcal{N}(\gamma/2, p, \rho)^{n_l m}$ disjoint sets such that if (P_1, Y_1^0) and (P_2, Y_2^0) belong to the same set, then $Y_1^0 = Y_2^0$, rankings of different views are the same ($o_{P_1}^v = o_{P_2}^v, v = 1, \ldots, V$) and $(p_1^k, p_2^k) \leq \frac{\sqrt{V}\gamma}{2}$, $k = 1, \ldots, N$. Here, $(n_l!)^{mV}$ is the number of all possible rankings, and $2^{n_l m}$ is the number of all possible labels (since the ground-truth label only takes two values). Note that if P_1 and P_2 have the same ranking for each view, the combination of all views will have the same ranking. Suppose that θ_{ap}^* is the solution returned by (12), then we have

$$|\text{AP}(\theta_{\text{ap}}^*, P_1, Y_1^0) - \text{AP}(\theta_{\text{ap}}^*, P_2, Y_2^0)| = 0. \tag{20}$$

That is, the algorithm $\mathcal{A}_{\text{ap}}$ is $(N, 0)$ robust, where $N = (n_l!)^{mV}2^{n_l m}\mathcal{N}(\gamma/2, p, \rho)^{n_l m}$.

In the next, we will prove that $\mathcal{A}_{\text{ls}}$ is (N, a) robust and $\mathcal{A}_{\text{svm}}$ is (N, b) robust, but $a, b \neq 0$. Suppose that θ_{ls}^* is the solution returned by (5), then we have

$$\frac{1}{2N}\sum_{k=1}^{N}\left((\theta_{\text{ap}}^*)^T p_k - y_k^0\right)^2 + \frac{\eta}{2}\|\theta_{\text{ap}}^*\|_2^2 \leq \frac{1}{2N}\sum_{k=1}^{N}\left(p_k^V - y_k^0\right)^2 + \frac{\eta}{2}\|1\|_2^2 = \frac{\eta + (b+1)^2}{2}.$$

The right side of the inequality is obtained by setting $\theta = (0, \ldots, 0, 1)$, and the inequality holds because θ_{ls}^* is the minimizer of (5). We thus further obtain $\|\theta_{\text{ls}}^*\|_2 \leq \sqrt{((\eta + (b+1)^2)/\eta)}$. For any (p_1, y_1^0) and (p_2, y_2^0) in the same set of a partition, we have

$$\begin{aligned}
&|l(\theta_{\text{ls}}^*, (p_1, y_1^0)) - l(\theta_{\text{ls}}^*, (p_2, y_2^0))| \\
&= |((\theta_{\text{ls}}^*)^T p_1 - y_1^0)^2 - ((\theta_{\text{ls}}^*)^T p_2 - y_2^0)^2| \\
&= |2y_2^0(\theta_{\text{ls}}^*)^T p_2 - 2y_1^0(\theta_{\text{ls}}^*)^T p_1 + ((\theta_{\text{ls}}^*)^T p_1)^2 - ((\theta_{\text{ls}}^*)^T p_2)^2| \\
&\leq |2(\theta_{\text{ls}}^*)^T(p_1 - p_2)| + |((\theta_{\text{ls}}^*)^T(p_1 + p_2))((\theta_{\text{ls}}^*)^T(p_1 - p_2))| \\
&\leq 2\|\theta_{\text{ls}}^*\|_2\|p_1 - p_2\|_2 + 2b\|\theta_{\text{ls}}^*\|_2\|p_1 - p_2\|_2 \\
&\leq \sqrt{V}\gamma(b+1)\sqrt{\frac{(b+1)^2 + \eta}{\eta}},
\end{aligned} \tag{21}$$

where we have utilized that $y_1^0 = y_2^0$ and $y_1^0, y_2^0 \in \{-1, 1\}$. The second inequality follows from the Cauthy-Schwarz inequality. The last inequality holds because $\|p_1 - p_2\|_2 \leq \sqrt{V}\gamma/2$ and $\|\theta_{\text{ls}}^*\|_2 \leq \sqrt{(\eta + (b+1)^2)/\eta}$. Therefore, $\mathcal{A}_{\text{ls}}$ is $(N, \gamma(b+1)\sqrt{V((b+1)^2 + \eta)/\eta})$ robust. We now prove that $\mathcal{A}_{\text{ls}}$ is not $(N, 0)$ robust. Because p is a vector of real values which are chosen from an infinite continuous space, there must exists a set C' that contains more than three different choices of p in any partition. Then for some $(p_0, y^0), (p_0 + \delta p, y^0) \in C'$, the following inequality holds

$$\begin{aligned}
&|l(\theta_{\text{ls}}^*, (p_0 + \delta p, y^0)) - l(\theta_{\text{ls}}^*, (p_0, y^0))| \\
&\quad = |2y^0(\theta_{\text{ls}}^*)^T\delta p + ((\theta_{\text{ls}}^*)^T p_0 + \delta p)^2 - ((\theta_{\text{ls}}^*)^T p_0)^2| > 0.
\end{aligned} \tag{22}$$

This is because there are at most two different δp that make $2y^0(\theta_{\text{ls}}^*)^T\delta p + ((\theta_{\text{ls}}^*)^T p_0 + \delta p)^2 - ((\theta_{\text{ls}}^*)^T p_0)^2 = 0$ while the choices of δp is more than two in the set C'.

According to (21) and (22), we conclude that $\mathcal{A}_{\text{ls}}$ is (N, a) robust and a cannot be 0. Similarly, we can prove that $\mathcal{A}_{\text{svm}}$ is (N, b) robust and $b \neq 0$. Therefore, the AP loss will tolerate to the small perturbation added to p, as long as the perturbation

does not change the partition of $\mathcal{P}$, while the least squares and hinge loss will not. This completes the proof. ■

ACKNOWLEDGMENT

The authors would like to thank the handling associate editor Prof. Dr. Richard Radke and all the anonymous reviewers for their constructive comments.

REFERENCES

[1] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognit.*, vol. 37, no. 9, pp. 1757–1771, Sep. 2004.

[2] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *Int. J. Data Warehousing Mining*, vol. 3, no. 3, pp. 1–13, 2007.

[3] B. Hariharan, L. Zelnik-Manor, S. Vishwanathan, and M. Varma, "Large scale max-margin multi-label classification with priors," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 423–430.

[4] L. Sun, S. Ji, and J. Ye, "Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 194–200, Jan. 2011.

[5] B. Wang, Z. Tu, and J. K. Tsotsos, "Dynamic label propagation for semi-supervised multi-class multi-label classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 425–432.

[6] Y. Luo, D. Tao, B. Geng, C. Xu, and S. J. Maybank, "Manifold regularized multitask learning for semi-supervised multilabel image classification," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 523–536, Feb. 2013.

[7] G. Sundaramoorthi and B.-W. Hong, "FAST LABEL: Easy and efficient solution of joint multi-label and estimation problems," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3126–3133.

[8] A. Goldberg, X. Zhu, B. Recht, J. Xu, and R. Nowak, "Transduction with matrix completion: Three birds with one stone," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 757–765.

[9] R. S. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino, "Matrix completion for multi-label image classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 190–198.

[10] Y. Luo, T. Liu, D. Tao, and C. Xu, "Decomposition-based transfer distance metric learning for image classification," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3789–3801, Sep. 2014.

[11] D. Tao, X. Lin, L. Jin, and X. Li, "Principal component 2D long short-term memory for font recognition on single Chinese characters," *IEEE Trans. Cybern.*, 2015, doi: 10.1109/TCYB.2015.2414920.

[12] D. Tao, J. Cheng, X. Lin, and J. Yu, "Local structure preserving discriminative projections for RGB-D sensor-based scene classification," *Inf. Sci.*, 2015, doi: 10.1016/j.ins.2015.03.031.

[13] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.

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

[15] M. J. Huiskes and M. S. Lew, "The MIR Flickr retrieval evaluation," in *Proc. Int. Conf. Multimedia Inf. Retr.*, 2008, pp. 39–43.

[16] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *J. Mach. Learn. Res.*, vol. 9, pp. 2491–2521, Nov. 2008.

[17] M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien, "l_p-norm multiple kernel learning," *J. Mach. Learn. Res.*, vol. 12, pp. 953–997, Mar. 2011.

[18] J. Kludas, E. Bruno, and S. Marchand-Maillet, "Information fusion in multimedia information retrieval," in *Proc. 5th Int. Workshop Adapt. Multimedia Retr., Retr., User, Semantics*, 2008, pp. 147–159.

[19] S. Ono, T. Miyata, and I. Yamada, "Cartoon-texture image decomposition using blockwise low-rank texture characterization," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1128–1142, Mar. 2014.

[20] H. Arguello and G. R. Arce, "Rank minimization code aperture design for spectrally selective compressive imaging," *IEEE Trans. Image Process.*, vol. 22, no. 3, pp. 941–954, Mar. 2013.

[21] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, Dec. 2009.

[22] M. Fazel, "Matrix rank minimization with applications," Ph.D. dissertation, Dept. Elect. Eng., Stanford Univ., Stanford, CA, USA, 2002.

[23] S. Ma, D. Goldfarb, and L. Chen, "Fixed point and Bregman iterative methods for matrix rank minimization," *Math. Program.*, vol. 128, nos. 1–2, pp. 321–353, Jun. 2011.

[24] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2980–2998, Jun. 2010.

[25] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 612–620.

[26] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma, "Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix," Dept. Electron. Commun. Eng., Univ. Illinois, Urbana, IL, USA, Tech. Rep. UILU-ENG-09-2214, 2009.

[27] Y. Luo, J. Tang, J. Yan, C. Xu, and Z. Chen, "Pre-trained multi-view word embedding using two-side neural network," in *Proc. AAAI Conf. Artif. Intell.*, 2014, pp. 1982–1988.

[28] A. Zien and C. S. Ong, "Multiclass multiple kernel learning," in *Proc. Int. Conf. Mach. Learn.*, 2007, pp. 1191–1198.

[29] S. Bickel and T. Scheffer, "Multi-view clustering," in *Proc. Int. Conf. Data Mining*, 2004, pp. 19–26.

[30] A. Rodriguez, V. N. Boddeti, B. V. K. V. Kumar, and A. Mahalanobis, "Maximum margin correlation filter: A new approach for localization and classification," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 631–643, Feb. 2013.

[31] J. Lindblad and N. Sladoje, "Linear time distances between fuzzy sets with applications to pattern matching and classification," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 126–136, Jan. 2014.

[32] R. Ptucha and A. E. Savakis, "LGE-KSVD: Robust sparse representation classification," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1737–1750, Apr. 2014.

[33] Y. Luo, D. Tao, C. Xu, C. Xu, H. Liu, and Y. Wen, "Multiview vector-valued manifold regularization for multilabel image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 5, pp. 709–722, May 2013.

[34] Y. Luo, D. Tao, C. Xu, D. Li, and C. Xu, "Vector-valued multi-view semi-supervised learning for multi-label image classification," pp. 647–653, Jul. 2013.

[35] J. Gui, D. Tao, Z. Sun, Y. Luo, X. You, and Y. Tang, "Group sparse multiview patch alignment framework with view consistency for image classification," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 3126–3137, Jul. 2014.

[36] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," in *Proc. Int. Conf. Mach. Learn.*, 2002, pp. 323–330.

[37] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *J. Mach. Learn. Res.*, vol. 5, pp. 27–72, Jan. 2004.

[38] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh, "Dimensionality reduction for data in multiple feature representations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 961–968.

[39] B. McFee and G. Lanckriet, "Learning multi-modal similarity," *J. Mach. Learn. Res.*, vol. 12, pp. 491–523, Jan. 2011.

[40] C. Xu, D. Tao, and C. Xu, "Large-margin multi-view information bottleneck," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1559–1572, Aug. 2014.

[41] C. Xu, D. Tao, and C. Xu, "Multi-view intact space learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2015, doi: 10.1109/TPAMI.2015.2417578.

[42] D. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, Dec. 2004.

[43] J. D. R. Farquhar, D. R. Hardoon, H. Meng, J. Shawe-Taylor, and S. Szedmak, "Two view learning: SVM-2K, theory and practice," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 355–362.

[44] M. White, X. Zhang, D. Schuurmans, and Y.-L. Yu, "Convex multi-view subspace learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1682–1690.

[45] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proc. 13th Annu. ACM Int. Conf. Multimedia*, 2005, pp. 399–402.

[46] G. Fumera and F. Roli, "A theoretical and experimental analysis of linear combiners for multiple classifier systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 942–956, Jun. 2005.

[47] M. Wozniak and K. Jackowski, "Some remarks on chosen methods of classifier fusion based on weighted voting," in *Proc. 4th Int. Conf. Hybrid Artif. Intell. Syst.*, 2009, pp. 541–548.

[48] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. Annu. Conf. Comput. Learn. Theory*, 1998, pp. 92–100.

[49] B. Krishnapuram, D. Williams, Y. Xue, L. Carin, M. A. T. Figueiredo, and A. Hartemink, "On semi-supervised classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 721–728.

[50] C. M. Christoudias, R. Urtasun, A. Kapoorz, and T. Darrell, "Co-training with noisy perceptual observations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2844–2851.

[51] K. Nigam and R. Ghani, "Analyzing the effectiveness and applicability of co-training," in *Proc. Int. Conf. Inf. Knowl. Manage.*, 2000, pp. 86–93.

[52] U. Brefeld and T. Scheffer, "Co-EM support vector learning," in *Proc. Int. Conf. Mach. Learn.*, 2004, pp. 121–128.

[53] V. Sindhwani, P. Niyogi, and M. Belkin, "A co-regularization approach to semi-supervised learning with multiple views," in *Proc. ICML Workshop Learn. Multiple Views*, 2005, pp. 74–79.

[54] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, Nov. 2006.

[55] A. Klausner, A. Tengg, and B. Rinner, "Vehicle classification on multi-sensor smart cameras using feature- and decision-fusion," in *Proc. ACM/IEEE Int. Conf. Distrib. Smart Cameras*, Sep. 2007, pp. 67–74.

[56] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[57] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, May 2001.

[58] [Online]. Available: http://pascallin.ecs.soton.ac.uk/challenges/VOC/

[59] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 902–909.

[60] D. Tao, X. Tang, X. Li, and X. Wu, "Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 7, pp. 1088–1099, Jul. 2006.

[61] A. Christmann and I. Steinwart, "Consistency and robustness of kernel-based regression in convex risk minimization," *Bernoulli*, vol. 13, no. 3, pp. 799–819, 2007.

[62] Y. Wu and Y. Liu, "Robust truncated hinge loss support vector machines," *J. Amer. Statist. Assoc.*, vol. 102, no. 479, pp. 974–983, 2007.

[63] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[64] Y. Yue, T. Finley, F. Radlinski, and T. Joachims, "A support vector method for optimizing average precision," in *Proc. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2007, pp. 271–278.

[65] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *J. Mach. Learn. Res.*, vol. 6, pp. 1453–1484, Sep. 2005.

[66] T. Joachims, T. Finley, and C.-N. J. Yu, "Cutting-plane training of structural SVMs," *Mach. Learn.*, vol. 77, no. 1, pp. 27–59, Oct. 2009.

[67] H. Xu and S. Mannor, "Robustness and generalization," *Mach. Learn.*, vol. 86, no. 3, pp. 391–423, Mar. 2012.

[68] K. Aas and L. Eikvil, "Text categorization: A survey," Norwegian Computing Center, Oslo, Norway, Tech. Rep., Jun. 1999.

[69] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, Jul. 2007.

[70] H. Q. Minh and V. Sindhwani, "Vector-valued manifold regularization," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 57–64.

[71] T. Poggio and C. R. Shelton, "On the mathematical foundations of learning," *Amer. Math. Soc.*, vol. 39, no. 1, pp. 1–49, 2002.

[72] A. Maurer and M. Pontil, "K-dimensional coding schemes in Hilbert spaces," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5839–5846, Nov. 2010.

[73] Z.-Q. Zhao, H. Glotin, Z. Xie, J. Gao, and X. Wu, "Cooperative sparse representation in two opposite directions for semi-supervised image annotation," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 4218–4231, Sep. 2012.

[74] J. Van De Weijer and C. Schmid, "Coloring local feature extraction," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 334–348.

[75] M. Zhu, "Recall, precision and average precision," Dept. Statist. Actuarial Sci., Univ. Waterloo, Waterloo, ON, Canada, Tech. Rep. Working Paper 2004-09, 2004.

[76] T. Fawcett, "ROC graphs: Notes and practical considerations for researchers," *Mach. Learn.*, vol. 31, pp. 1–38, Mar. 2004.

[77] R. E. Schapire and Y. Singer, "BoosTexter: A boosting-based system for text categorization," *Mach. Learn.*, vol. 39, nos. 2–3, pp. 135–168, 2000.

[78] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Jan. 2006.

[79] A. Herschtal and B. Raskutti, "Optimising area under the ROC curve using gradient descent," in *Proc. Int. Conf. Mach. Learn.*, 2004, pp. 49–56.

[80] T. Joachims, "A support vector method for multivariate performance measures," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 377–384.

Yong Luo received the B.E. degree in Computer Science from the Northwestern Polytechnical University, Xi'an, China, in 2009, and the D.Sc. degree in the School of Electronics Engineering and Computer Science, Peking University, Beijing, China, in 2014. He was a visiting student in the School of Computer Engineering, Nanyang Technological University, and the Faculty of Engineering and Information Technology, University of Technology Sydney. His research interests are primarily on machine learning with applications on image classification and annotation.He has authored several scientific articles at top venues including IEEE T-IP, IEEE T-NNLS, and the AAAI conference on Artificial Intelligence.

Tongliang Liu received the B.E. degree in electronics engineering and information science from the University of Science and Technology of China, Hefei, China, in 2012. He is currently pursuing the Ph.D. degree in computer science with the University of Technology at Sydney, Sydney, NSW, Australia. He was a recipient of the best paper award in the IEEE International Conference on Information Science and Technology 2014.

His research interests include machine learning, computer vision, and optimization.

Dacheng Tao (F'15) is Professor of Computer Science with the Centre for Quantum Computation & Intelligent Systems, and the Faculty of Engineering and Information Technology in the University of Technology, Sydney. He mainly applies statistics and mathematics to data analytics and his research interests spread across computer vision, data science,image processing, machine learning, neural networks and video surveillance. His research results have expounded in one monograph and 100+ publications at prestigious journals and prominent conferences,such as IEEE T-PAMI, T-NNLS, T-IP, JMLR, IJCV, NIPS, ICML, CVPR,ICCV, ECCV, AISTATS, ICDM; and ACM SIGKDD, with several best paper awards, such as the best theory/algorithm paper runner up award in IEEEICDM'07, the best student paper award in IEEE ICDM'13, and the 2014ICDM 10 Year Highest-Impact Paper Award.

Chao Xu received the B.E. degree from Tsinghua University, in 1988, the M.S. degree from the University of Science and Technology of China, in 1991, and the Ph.D. degree from the Institute of Electronics, Chinese Academy of Sciences, in 1997.

He was an Assistant Researcher with the University of Science and Technology of China from 1991 to 1994. Since 1997, he has been with the Key Laboratory of Machine Perception (Ministry of Education), Peking University, where he has been a Professor since 2005. He has authored or co-authored over 100 publications and holds six patents in his research fields. His research interests are in image and video processing and multimedia technology.