Cross-Domain Collaborative Learning via Discriminative Nonparametric Bayesian Model

Shengsheng Qian, Tianzhu Zhang, *Member, IEEE*, and Changsheng Xu, *Fellow, IEEE*

***Abstract*—Cross-domain data analysis has been becoming more and more important, and can be effectively adopted for many applications. However, it is difficult to propose a unified cross-domain collaborative learning framework for cross-domain analysis in social multimedia, because cross-domain data have multidomain, multimodal, sparse, and supervised properties. In this paper, we propose a generic cross-domain collaborative learning (CDCL) framework via a discriminative nonparametric Bayesian dictionary learning model for cross-domain data analysis. Compared with existing cross-domain learning methods, our proposed model mainly has four advantages: First, to address the domain discrepancy, we utilize the shared domain priors among multiple domains to make them share a common feature space. Second, to exploit the multimodal property, we use the shared modality priors to model the relationship between different modalities. Third, to deal with the sparse property of media data in one domain, our goal is to learn a shared dictionary to bridge different domains and complement each other. Finally, to make use of the supervised property, we exploit class label information to learn the shared discriminative dictionary, and utilize a latent probability vector to select different dictionary elements for representation of each class. Therefore, the proposed model can investigate the superiorities of different sources to supplement and improve each other effectively. In experiments, we have evaluated our model for two important applications including cross-platform event recognition and cross-network video recommendation. The experimental results have showed the effectiveness of our CDCL model for cross-domain analysis.**

***Index Terms*—Social media, discriminative non-parametric Bayesian model, multi-modality.**

I. Introduction

NOWADAYS, more and more social media sites are popping up, like Facebook and Flickr, and make it workable for users to create and share rich online social media content.

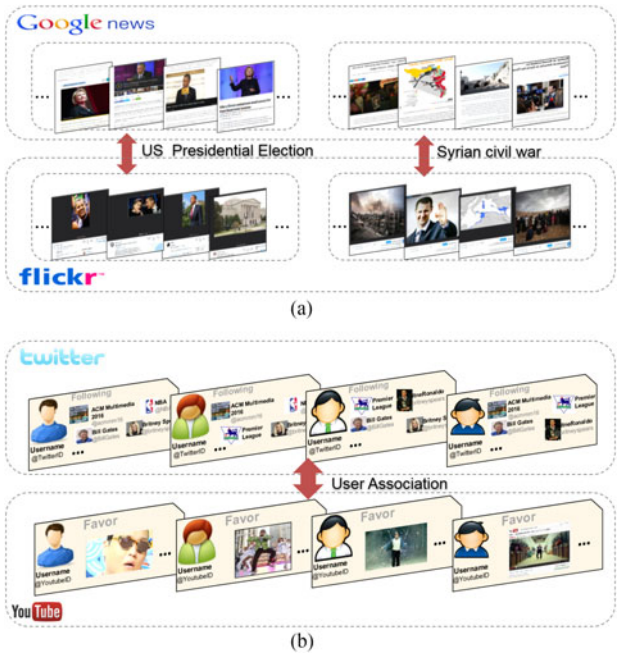


Fig. 1. There are two different scenarios in cross-domain analysis: (a) cross-platform data association (b) cross-network user association.

According to the media report[1], Flickr[2] has 87 million users and 8 billion photos, Instagram[3] has 100 million users producing 4 billion photos, Facebook[4] has about 1.11 billion users, and Twitter[5] has around 500 million enrolled users with over 200 million active users. As a result, social multimedia data increase enormously in cross-domain scenarios (different social media sites), such as cross-platform data association, and cross-network user association.

For cross-platform data association as shown in Fig. 1(a), when a popular event, such as United States Presidential Election or Syrian civil war, is going on around us, it can spread quickly and will have large amounts of event content information with multi-modality in Internet, such as texts, images, and videos. Here, we consider the Flickr and Google News as two different domains in Fig. 1(a), and exploit the shared social event, such as "United States Presidential Election", to asso-

Manuscript received February 13, 2017; revised June 30, 2017 and October 15, 2017; accepted November 20, 2017. Date of publication December 19, 2017; date of current version July 17, 2018. This work was supported in part by the National Natural Science Foundation of China under Grants 61720106006, 61432019, 61572498, 61532009, and 61572296; in part by the Key Research Program of Frontier Sciences, CAS, under Grant QYZDJ-SSW-JSC039; and in part by Beijing Natural Science Foundation (4172062). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Marco Bertini. *(Corresponding author: Changsheng Xu.)*

The authors are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with University of Chinese Academy of Sciences (e-mail: shengsheng.qian@nlpr.ia.ac.cn; tzzhang@nlpr.ia.ac.cn; csxu@nlpr.ia.ac.cn).

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/TMM.2017.2785227

[1]https://www.centillien.com/news/view/415595/the-history-and-evolution-of-social-media-an-infographic
[2]http://www.flickr.com/
[3]http://instagram.com/
[4]https://www.facebook.com/
[5]https://twitter.com/

ciate with their documents from different domains. Here, for the same event, there might be few authority reports on Google News, which can be complemented by many intriguing comments and pictures on Flickr. Besides, the related pictures of official reports caught by the professional writers can concentrate on a particular event on Google News, but most of the uploaded pictures by users are unprofessional on Flickr. As a result, if the related data of the same event can be aggregated across different platforms, we can make the strengths of one domain supplement the shortcomings of the others.

For cross-network user association as shown in Fig. 1(b), with the expansion of online networking, it has changed the way of people to share personal information, and users usually visit different social networks simultaneously for different needs. Here, there are many user accounts in the largest video-sharing YouTube network and the largest tweet-sharing Twitter network. For the same user, he/she may have tweeting practices on Twitter and video-related practices on YouTube, and the behavior of the same user in one network will be inevitably affected by the track of another social network. If we aggregate rich cross-network behavior information of the users across different social networks, they can supplement and improve each other, particularly for a new user with only little information in social network.

Based on the above discussions in two different cross-domain scenarios, it is clear that, to better know what happens among multiple domains, it is important to utilize the superiorities of multiple sources via collaborative learning methods, such as Google news, Filckr, Twitter, and YouTube. The cross-domain collaborative learning is useful in many cross-domain applications, like cross-domain event analysis [1]–[3], cross-domain collaboration recommendation [4], and cross-domain multiple event tracking [5], [6]. However, cross-domain data are innately heterogeneous and noisy, it is difficult to explore the helpful information across multiple domains.

Most existing cross domain learning methods could be categorized into two categories including (1) making multiple domains share a common feature space by introducing a cross-domain constraint for cross-domain feature learning [2], [7], [8]. (2) utilizing some cross-domain methods to model cross-network user behaviors with social links to conduct personalization recommendation tasks [9], [10]. However, few efforts have been made to propose a unified cross-domain collaborative learning framework because the media data have **multi-domain**, **multi-modal**, **sparse** and **supervised** properties. Specifically, the media data on different domains (e.g., Flickr, Google News, YouTube, and Twitter) are heterogenous, and they can supplement each other, but also have domain discrepancy. Each data instance can be represented with images and texts. Moreover, these texts and images can supplement each other, as shown in Fig. 1(a). In reality, the behavior information of users might be sparse in the media site, because only small part of photos and videos can be browsed by users. If a new user enrolls on YouTube, the system does not know anything about his/her associations on Youtube and cannot make effective video recommendations. In Fig. 1(a), data instances are about two types of events "United States Presidential Election" and "Syrian civil war" from different social media sites. The class label information can be exploited in the cross-domain learning to obtain discriminative feature representation.

To overcome the above issues, we design a novel unified Cross-Domain Collaborative Learning (CDCL) framework via the proposed discriminative non-parametric Bayesian dictionary learning model for cross-domain multi-modal data analysis. The proposed model can jointly utilize the multi-domain, multi-modality, sparse, and supervised properties. Here, we just show one example of cross-platform data association for social event. In the left panel of Fig. 2, we show two different social events "United States Presidential Election" and "Syrian civil war" frow two different domains (Flickr and Google News) with two modalities (texts and images). In the middle panel of Fig. 2, it shows that our model can learn the shared discriminative feature representation by using the supervised information, together with the domain and modality priors from cross-domain data. In the right panel of Fig. 2, we apply the proposed model for two different applications including cross-platform event recognition and cross-network video recommendation. Here, the cross-platform event recognition is to utilize multi-modal information among multiple domains to recognize the social event. The cross-network video recommendation is to exploit rich cross-network behavior information of users to gauge their preferences on other social networks. For example, by using the overlapped user account linkage between Twitter and YouTube, and considering both the Twitter tweeting activities and historical interactions with YouTube videos, we can design a cold-start recommendation task for the new YouTube user by the proposed CDCL method. We evaluate our model on two applications and the experimental results demonstrate its effectiveness for cross-domain data analysis. The contributions are as follows.

- We propose a generic cross-domain collaborative learning framework for cross-domain data analysis such as cross-platform data association and cross-network user association, which can effectively utilize the superiorities of multiple resources to supplement and improve each other.
- The proposed discriminative non-parametric Bayesian dictionary learning model is able to not only use the shared modality and domain priors to consider the multi-modal property and overcome the domain discrepancy, respectively, but also exploit supervised information of media data to obtain the discriminative dictionary.
- We evaluate the proposed model on two different applications, and the extensive experimental results show that our model can perform the best with comparison the existing methods. Besides, we collect a large-scale social event dataset for cross-domain analysis, and will release it for academic use.

A preliminary version of this paper was published in [11]. The extension includes the following four aspects: (1) In [11], an unsupervised non-parametric Bayesian dictionary method is proposed for cross-domain collaborative learning. Different from [11], we propose a novel discriminative model for multi-modal cross-domain data analysis by considering supervised information. The model in [11] is only a special case of

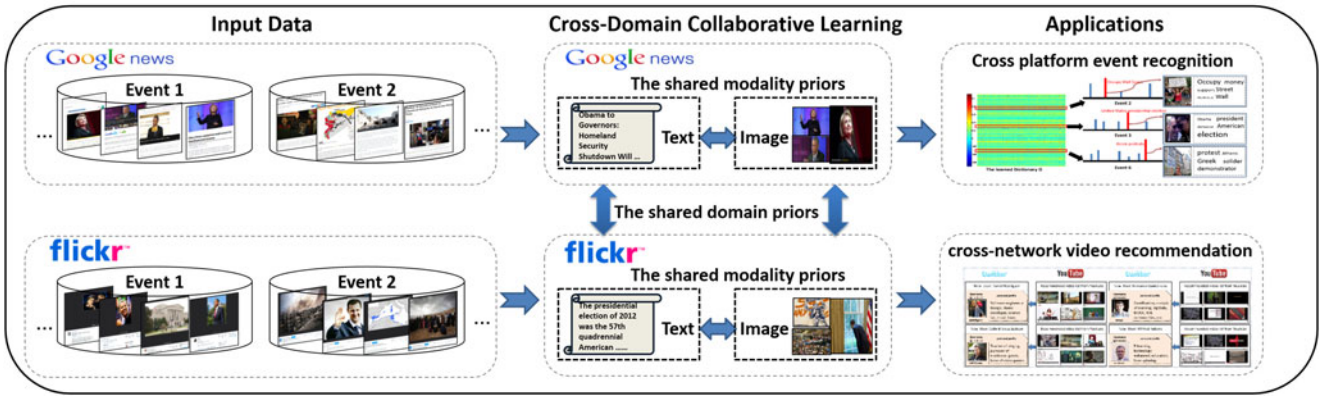


Fig. 2. Description of our cross-domain collaborative learning. For simplicity, we only show one example of cross-platform data association for social event. Here, the events have 2 classes (Event 1 and Event 2), and are from two domains (Flickr and Google News) with two different modalities (texts and images).

our model without considering supervised information. (2) In Section III, we present the details of the supervised cross-domain collaborative learning algorithm. (3) We introduce a more comprehensive survey of the related work about the discriminative dictionary learning method in Section II. (4) More quantitative results are shown to verify the effectiveness of our model, including: (a) The experimental results are added in Section V-A3. (b) A more focused investigation on discriminative level of the proposed model with different π_c is added in Figs. 6 and 7(c) We have added different feature representations of users with the users' social network and content information on Twitter in Section IV-B, and shown the detailed analysis in Section V-B and Fig. 11.

II. Related Work

Over the past few years, researchers have proposed various methods to cope with the problem of cross-domain collaborative learning. These studies can be generally classified into two categories as follows.

Cross-domain Feature Learning: In cross-domain feature learning, many algorithms use the prior knowledge of the auxiliary domain to help improve the task on target domain, and make multiple domains share the latent space by introducing some cross-domain constraints [2], [7], [8], [12]–[17]. In [7], a novel structural correspondence learning method is proposed to induce correspondences among features from different domains. Pan *et al.* [8] introduce a latent feature space to measure the domain similarity and reduce the distance among multiples domains. With the rapid development of multi-media information in Internet, more and more researchers have paid attention to the cross-media leaning in the past few years [18]–[26]. In [21], the CM-LDA method is proposed to adopt a shared latent variable to learn the relations of different media data. Chang *et al.* [14] propose a novel event search system by using concept classifiers collected from other sources to deal with the semantic event search and few-exemplar event detection. In [15], a bi-level semantic representation analyzing framework is proposed to deal with multiple semantic representations of MED videos learned from different sources. In [25], a novel deep feature learning paradigm is proposed to conduct cross-model feature learning by using social images and tags based on social collective intelligence. Liu *et al.* [26] propose a novel hierarchical clustering multi-task learning (HC-MTL) method for joint human action grouping and recognition, and this work assumes that all actions are either independent for individual learning or correlated for joint modeling.

Different from the above methods, the proposed CDCL method can associate with multiple domains by utilizing a shared dictionary learning strategy. Many existing dictionary learning approaches have been proposed [27]–[31]. Yang *et al.* [27] adopt the coupled dictionary model to associate the low resolution with high resolution patches for image super resolution. Here, the sparse coefficient can be learned by using the reconstructed residual error for the above dictionary learning methods. However, we cannot know the real value of the reconstructed residual error, and the performance might essentially degrade if the setting value is inconsistent with the ground truth. Instead, a new non-parametric Bayesian model is proposed to deal with the above problems [29], [30]. Zhou *et al.* [29] adopt a dictionary learning method by considering the beta process prior, where the appropriate dictionary size can be inferred naturally. The non-parametric Bayesian learning methods have a excellent performance in compressive sensing, image denoising, and human action recognition. However, the above methods only focus on unsupervised dictionary learning, and do not consider the supervised information to learn discriminative dictionary for feature representation. Wang *et al.* [32] introduce a novel supervised class-specific dictionary learning model to conduct action recognition, and the model can make dictionaries connected with various classes be independent via a dictionary disjointedness strategy for better targets classification. In [33], a novel K-SVD method based on the label consistent strategy is proposed to obtain the discriminative dictionary. Akhtar *et al.* [34] extend the unsupervised non-parametric Bayesian model to the supervised model, where the model deduces probability distributions of the dictionary elements to obtain the discriminative Bayesian dictionary by adopting the Beta Process prior. Inspired by their methods, we propose a novel cross-domain collaborative learning method via a discriminative non-parametric Bayesian

dictionary learning model for cross domain feature learning by use of the shared domain, modality and supervised information.

Cross-Network Collaborative Learning: Cross-network collaborative learning has attracted wide attentions in the past few years. Most of the existing methods mainly adopt multiple social networks' information for collaborative applications. In [35], a novel transfer learning framework via a real-time strategy is proposed to help solve some multimedia problems. In [36], the authors explore tag profiles of multiple social networks to help users find some consistent characteristics. However, how to utilize the shared content to bridge different domains to help complement each other is still a challenging problem. In many cross-network collaborative learning methods, our work is relevant to [10] and [37] in social media. Yan *et al.* [10] adopt cross-domain social relation data to conduct the friend recommendation task for the new users. Our work is different from [10], and mainly focuses on utilizing cross-network collaborative learning method to address the cold-start video recommendation problem. In [37], the authors propose a novel YouTube video promotion idea by using the dictionary learning strategy. In this work, we adopt a novel non-parametric Bayesian dictionary learning model to learn the shared dictionary collaboratively rather than utilize the coupled dictionary learning method. Furthermore, our model is also a generic framework for cross-domain analysis, and can be applied for many different applications in social multimedia.

III. Our Approach

In this section, we first show the details of our CDCL algorithm, and then introduce the model inference.

A. Cross-Domain Collaborative Learning

The cross-domain collaborative learning is to investigate the superiorities of multiple resources to supplement and improve each other. To achieve this goal, we propose a generic collaborative learning framework via a discriminative non-parametric Bayesian dictionary learning model. For multi-modal cross-domain data, we assume that data have J domains with M modalities from C classes. Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_J]$ represent all data instances of the J domains. Specifically, $\mathbf{x}_j = [\{\mathbf{x}_{j,1}^1, \cdots, \mathbf{x}_{j,1}^M\}, ..., \{\mathbf{x}_{j,c}^1, \cdots, \mathbf{x}_{j,c}^M\}, ..., \{\mathbf{x}_{j,C}^1, \cdots, \mathbf{x}_{j,C}^M\}]$ denote M modalities of data instances from C classes in the j-th domain. $\mathbf{x}_{j,c}^m \in \mathbf{R}^{n_j^m}$ denotes the m-th modality sample of data instance from the c-th class in the j-th domain, and the n_j^m denotes the feature dimensionality of the m-th modality sample in the j-th domain. Here, data instances $\mathbf{X}$ could be either social event data or user information, such as social events described by images and texts or the user account linkages between Twitter and YouTube.

We use the traditional sparse representation which has shown encouraging performance [38] to model these data instances. When given an data instance $\mathbf{x}_j^m$, we use the dictionary $\mathbf{D}_j^m$ and the reconstructed error term ε_j^m to represent this instance, as shown in (1).

$$\mathbf{x}_j^m = \mathbf{D}_j^m \mathbf{w}_j^m + \varepsilon_j^m \tag{1}$$

where $\mathbf{x}_j^m$ denotes the m-th modality sample of data instance in the j-th domain, the columns of the matrix $\mathbf{D}_j^m \in \mathbf{R}^{n_j^m \times K}$ denote the K dictionary elements, $\mathbf{w}_j^m$ denotes the sparse coefficient of feature, and ε_j^m denotes the reconstructed noise. Equation (1) uses the sparse representation to only consider the single modality of the instance in one domain, and cannot utilize the multi-domain, multi-modality, sparse, and supervised properties jointly.

In [11], we have proposed a unsupervised non-parametric Bayesian dictionary learning model in [11] to investigate the multi-domain and the multi-modality property of cross-domain data. However, this model does not consider the supervised information of data instance to learn discriminative feature representation. In this work, based on the method [11], we propose a novel generic cross-domain collaborative learning method via a discriminative non-parametric Bayesian dictionary learning model by considering the supervised information of data instances to learn the discriminative dictionary for feature representation, which can effectively boost the classification performance, as shown in Fig. 3.

In the traditional non-parametric Bayesian dictionary learning model, they usually consider the beta process prior to obtain the dictionary $\mathbf{D}_j^m$, which can non-parametrically infer the number of dictionary elements among multiple domains and obtain their relevant importance levels. Therefore, we can introduce the beta process priors to make the obtained results sparse instead of traditional ℓ_1 norm constraint. In [39], the authors develop the Beta process (BP) in details, and the BP with parameters $a_0 > 0$, $b_0 > 0$, and the base measure H_0 is denoted as $BP(a_0, b_0, H_0)$. The stick-breaking construction process $H \sim BP(a_0, b_{0,} H_0)$ is denoted as:

$$H(\psi) = \sum_{k=1}^{K} \pi_k \delta_{\psi_k}(\psi) \tag{2}$$

where $\pi_k \sim Beta(a_0/K, b_0(K-1)/K)$ and $\psi_k \sim H_0$. The $H(\psi)$ denotes the vector form of K probability values, where each value is related with a individual element ψ_k, and ψ_k denotes the element distribution according to the H_0. When the value of K is close to infinite, $H(\psi)$ can have an infinite dimensional vector representation of probability values, and each probability value will have a related element ψ_k drawn i.i.d. from H_0.

Let $\mathbf{w}_{j,c}^m$ denotes the sparse feature coefficient of the m-th modality sample of data instance of the c-th class in the j-th domain. Mathematically, $\mathbf{x}_{j,c}^m = \mathbf{D}_j^m \mathbf{w}_{j,c}^m + \varepsilon_{j,c}^m$, where $\mathbf{x}_{j,c}^m \in \mathbf{R}^{n_j^m}$ denotes an data instance of the c-th class with the m-th modality, $\mathbf{D}_j^m \in \mathbf{R}^{n_j^m \times K}$ denote the K dictionary elements shared by all the classes, $\mathbf{w}_{j,c}^m$ denotes the sparse feature coefficient, and $\varepsilon_{j,c}^m$ denotes measurement noise. We can set the sparse coefficient as $\mathbf{w}_{j,c}^m = \mathbf{z}_{j,c}^m \odot \mathbf{s}_{j,c}^m$, where $\odot$ denotes the Hadamard multiplication (element-wise) form of these two vectors. The $\mathbf{z}_{j,c}^m \in \{0,1\}^K$ represents a binary vector, and can show which element of the dictionary D_j^m is utilized to conduct the representation of the instance. The $\mathbf{s}_{j,c}^m \sim N(0, \gamma_{c,s}^{-1}\mathbf{I}_K)$ represents the weight values, and is used to ensure that the obtained sparse reconstruction values are not always binary. The $\gamma_{c,s}$

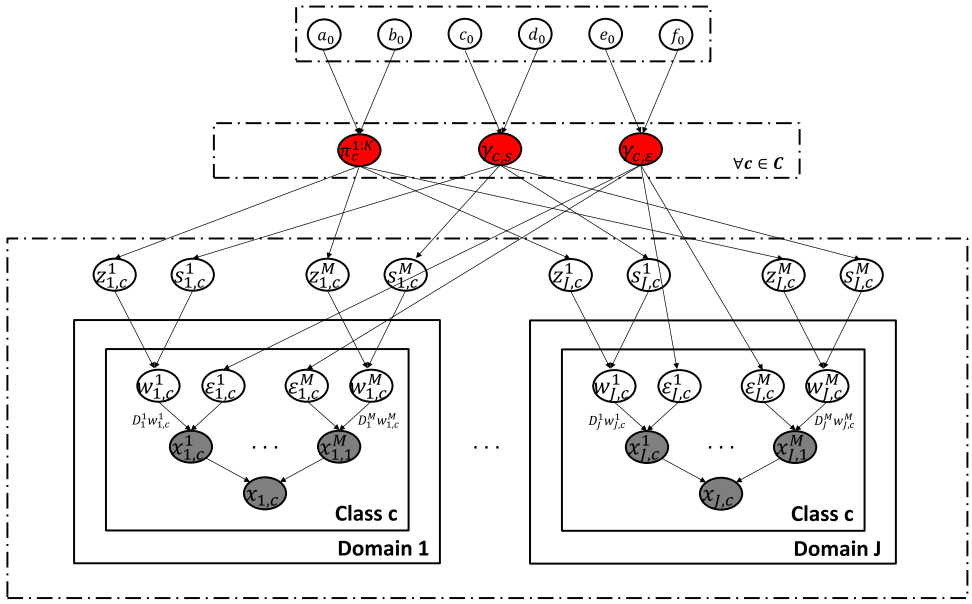


Fig. 3. The graphical representation of the proposed model is shown. Here, the red circles and the gray circles are represented as the shared priors and the observation data, respectively.

represents the inverse variance or the precision. Specifically, let the elements ψ_k be connected with the candidate members of the dictionary D_j^m, and the k-th element of the binary vector $\mathbf{z}_{j,c}^m$ can be drawn by $\mathbf{z}_{j,c,k}^m \sim \text{Bernoulli}(\pi_{c,k})$. The shared priors $\pi_c, \gamma_{c,s}, \gamma_{c,\varepsilon}$ can be used to obtain the shared dictionary for feature representation among multiple domains in the c-th class, as shown in Fig. 3.

The hierarchical form of our method could be denoted as:

$$\begin{aligned}
\mathbf{x}_{j,c}^m &= \mathbf{D}_j^m \mathbf{w}_{j,c}^m + \varepsilon_{j,c}^m \\
\mathbf{D}_j^m &= [d_{j,1}^m, \cdots, d_{j,K}^m] \\
\mathbf{w}_{j,c}^m &= \mathbf{z}_{j,c}^m \odot \mathbf{s}_{j,c}^m \\
\varepsilon_{j,c}^m &\sim N(0, \gamma_{c,\varepsilon}^{-1} \mathbf{I}_{n_j^m}),
\end{aligned} \tag{3}$$

where $m = 1, \ldots, M, j = 1, \ldots, J, c = 1, \ldots, C$, $d_{j,k}^m \sim N(0, n_j^{m(-1)} \mathbf{I}_{n_j^m})$ is shared by all classes, $\mathbf{z}_{j,c}^m \sim \prod_{k=1}^K \text{Bernoulli}(\pi_{c,k})$, $\pi_{c,k} \sim \text{Beta}(a_0/K, b_0(K-1)/K)$, $\mathbf{s}_{j,c}^m \sim N(0, \gamma_{c,s}^{-1} \mathbf{I}_K)$, $\gamma_{c,s} \sim \Gamma(c_0, d_0)$, and $\gamma_{c,\varepsilon} \sim \Gamma(e_0, f_0)$. The gamma hyper-priors over $\gamma_{c,s}$, $\gamma_{c,\varepsilon}$ are non-informative. Here, independent conjugacy Gaussian priors for $d_{j,k}^m$, $\mathbf{s}_{j,c}^m$, and $\varepsilon_{j,c}^m$ are adopted for simplicity. As a result, a latent probability vector π_c with $\pi_{c,k}$ as its elements are connected with the dictionary elements in the feature representation of data instances from the c-th class. Therefore, our method is able to not only utilize the knowledge of each domain, but also combine superiorities of other domains to supplement and improve each other by using the shared domain and modality priors. Moreover, the model can utilize the learned dictionary which is shared by all classes and the learned latent probability vector π_c of each class to obtain the discriminative dictionary for feature representation.

B. Model Inference

The full likelihood probability of the proposed CDCL method is factorized as:

$$\begin{aligned}
&P(X, D, Z, S, \pi, \gamma_s, \gamma_\varepsilon) = \\
&\prod_{j=1}^J \prod_{m=1}^M \prod_{c=1}^C N\left(x_{j,c}^m; D_j^m (s_{j,c}^m \odot z_{j,c}^m), \gamma_{c,\varepsilon}^{-1} I_{n_j^m}\right) \\
&\quad N(s_{j,c}^m; 0, \gamma_{c,s}^{-1} I_K) \\
&\prod_{j=1}^J \prod_{m=1}^M \prod_{k=1}^K N\left(d_{j,k}^m; 0, n_j^{m(-1)} I_{n_j^m}\right) \\
&\prod_{j=1}^J \prod_{m=1}^M \prod_{c=1}^C \prod_{k=1}^K \text{Bernoulli}\left(z_{j,c,k}^m; \pi_{c,k}\right) \\
&\prod_{c=1}^C \prod_{k=1}^K \text{Beta}(\pi_{c,k}; a_0, \mathrm{b}_0) \\
&\prod_{c=1}^C \Gamma(\gamma_{c,\varepsilon}; e_0 \mathrm{f}_0) \Gamma(\gamma_{c,s}; c_0, \mathrm{d}_0).
\end{aligned} \tag{4}$$

There are many latent variables, namely $\mathbf{D}$, $\mathbf{Z}$, $\mathbf{S}$, π, γ_s, γ_ε, to be estimated, and we utilize popular Gibbs sampling [40] strategy to estimate latent parameters. In a traditional Gibbs sampler, new assignments of latent variables will be iteratively sampled based on the distributions conditioned on the previous state of the model. This process is analogous to the atom-by-atom dictionary update step of K-SVD [41]. The derivation rules are listed for the latent variables $\mathbf{D}$, $\mathbf{Z}$, $\mathbf{S}$, π, γ_s, γ_ε as follows:

Sampling $\mathbf{D}_j^m$*:* Firstly, we sample the dictionary variable $\mathbf{D}_j^m = [d_{j,1}^m, \ldots, d_{j,K}^m]$ based on the posterior probability in (5).

$$P(d_{j,k}^m|-) \propto \prod_{c=1}^{C} N\left(x_{j,c}^m; D_j^m(s_{j,c}^m \odot z_{j,c}^m), \gamma_{c,\varepsilon}^{-1} I_{n_j^m}\right) N\left(d_{j,k}^m; 0, n_j^{m(-1)} I_{n_j^m}\right) \tag{5}$$

Here, the $d_{j,k}^m$ is shared by all classes and is updated using the complete training data, we can draw the $d_{j,k}^m$ from a normal distribution $p(d_{j,k}^m|-) \sim N(u_{d_{j,k}^m}, \Sigma_{d_{j,k}^m})$.

Sampling $\mathbf{z}_{j,c}^m$*:* We sample the binary vector $\mathbf{z}_{j,c}^m = [z_{j,c,1}^m, \ldots, z_{j,c,K}^m]$ based on the following posterior probability, as shown in (6).

$$P(z_{j,c,k}^m|-) \propto N\left(x_{j,c}^m; D_j^m(s_{j,c}^m \odot z_{j,c}^m), \gamma_{c,\varepsilon}^{-1} I_{n_j^m}\right) \text{Bernoulli}(z_{j,c,k}^m; \pi_{c,k}) \tag{6}$$

Here, when $z_{j,c,k}^m = 1$, we can write its posterior probability:

$$P_1 \propto N(x_{j,c}^m; D_j^m(s_{j,c}^m \odot z_{j,c}^m), \gamma_{c,\varepsilon}^{-1} I_{n_j^m}) \cdot \pi_{c,k} \tag{7}$$

when $z_{j,c,k}^m = 0$, the posterior probability can be written as:

$$P_0 \propto N\left(x_{j,c}^m; D_j^m(s_{j,c}^m \odot z_{j,c}^m), \gamma_{c,\varepsilon}^{-1} I_{n_j^m}\right) \cdot (1 - \pi_{c,k}) \tag{8}$$

We can draw $z_{j,c,k}^m$ according to the Bernoulli distribution $z_{j,c,k}^m \sim \text{Bernoulli}\ (\frac{\mathrm{P}_1}{\mathrm{P}_1 \mathrm{P}_0})$.

Sampling $\mathbf{s}_{j,c}^m$*:* We sample the weight variable $\mathbf{s}_{j,c}^m = [s_{j,c,1}^m, \ldots, s_{j,c,K}^m]$, as shown in (9).

$$P(s_{j,c,k}^m|-) \propto N\left(x_{j,c}^m; D_j^m(s_{j,c}^m \odot z_{j,c}^m), \gamma_{c,\varepsilon}^{-1} I_{n_j^m}\right) \times N(s_{j,c}^m; 0, \gamma_{c,s}^{-1} I_K) \tag{9}$$

Here, like $d_{j,k}^m$, we can draw the $s_{j,c,k}^m$ from a normal distribution $p(s_{j,c,k}^m|-) \sim N(u_{s_{j,c,k}^m}, \Sigma_{s_{j,c,k}^m})$.

Sampling π_c*:* Based on the discriminative model, we can obtain the posterior probability distribution over $\pi_c = [\pi_{c,1}, \ldots, \pi_{c,K}]$ as:

$$P(\pi_{c,k}|-) \propto \text{Beta}(\pi_{c,k}; a_0, b_0) \prod_{j=1}^{J} \prod_{m=1}^{M} \text{Bernoulli}(z_{j,c,k}^m; \pi_{c,k}) \tag{10}$$

due to the conjugacy between the two distributions, it can be easily shown that $\pi_{c,k}$ can be drawn from a Beta distribution as:

$$P(\pi_{c,k}|-) \propto \text{Beta}\left(\frac{a_0}{K} + \sum_{j=1}^{J}\sum_{m=1}^{M} z_{j,c,k}^m, \frac{b_0(K-1)}{K} + |\mathrm{I}|_c - \sum_{j=1}^{J}\sum_{m=1}^{M} z_{j,c,k}^m\right) \tag{11}$$

Sampling $\gamma_{c,s}$*:* The posterior probability distribution over $\gamma_{c,s}$ is:

$$P(\gamma_{c,s}|-) \propto \Gamma(\gamma_{c,s}; c_0, d_0) \prod_{j=1}^{J}\prod_{m=1}^{M} N(s_{j,c}^m; 0, \gamma_{c,s}^{-1} I_K) \tag{12}$$

Sampling $\gamma_{c,\varepsilon}$*:* Like $\gamma_{c,s}$, the posterior probability distribution over $\gamma_{c,\varepsilon}$ is written as:

$$P(\gamma_{c,\varepsilon}|-) \propto \Gamma(\gamma_{c,\varepsilon}; e_0, f_0) \prod_{j=1}^{J}\prod_{m=1}^{M} N\left(x_{j,c}^m; D_j^m(s_{j,c}^m \odot z_{j,c}^m), \gamma_{c,\varepsilon}^{-1} I_{n_j^m}\right) \tag{13}$$

The j, m, c denote the corresponding j-th domain, m-th modality, and c-th class, respectively. The priors $\pi_c, \gamma_{c,s}, \gamma_{c,\varepsilon}$ are shared among multiple domains and modalities.

IV. Applications

In this section, we introduce the details of our model for two different cross-domain applications including *(A) cross-platform event recognition* and *(B) cross-network video recommendation*.

A. Cross-Platform Event Recognition

The cross-platform event recognition is to utilize multi-modal information from multiple domains for event recognition. When a popular event is going on around us, it can spread quickly and will generate large amounts of event content information with multi-modality in Internet, such as texts, images, and videos. As a result, it is important and necessary to automatically detect and recognize the interesting and popular social events from large amounts of media data. How to effectively utilize the cross-domain multi-modality data is a key challenge. In the traditional event recognition task, textual features are usually adopted. However, event data also contain rich visual information. For the same event from different social media sources, it might have different types of textual information because of different users, such as tags and comments. However, its visual information may be similar, such as similar images or videos. For instance, for the event "Syrian civil war", the images of the Asad in different sites are extremely related. Aa a result, it is helpful to use a multi-modality combination strategy for event recognition. Moreover, different platforms also supplement and improve each other. Specifically, media data of most event are usually official on Google News, while they are casual on Flickr like some personal comments or interesting pictures. Therefore, if we can aggregate the related data of an event among multiple domains, we can make the strengths of one domain supplement the shortcomings of the other and improve each other.

Next, we will introduce how to take two different domains into consideration for cross-platform event recognition by the proposed CDCL method. Here, we consider the Flickr and Google News as two different domains. For each social event, it has many documents and we consider each document including texts and images as an event instance. Our target is to use the data information from different domains to classify each document. Specifically, we can view these two domains as the auxiliary domain and target domain, and consider the prior knowledge of the auxiliary domain to help improve the task of the target domain. When there are not large enough training samples, this will be

Algorithm 1: The proposed CDCL method for cross-platform event recognition

Input: The data of the auxiliary domain D_a; The training and testing data of the target domain D_t; The iteration number of Gibbs sampling method T_{Gibbs};
Output: Conduct the prediction of the class labels for the testing documents in the target domain D_t.
// Learn the shared domain priors in the auxiliary domain D_a.
1: Initialize the dictionary variable via the K-SVD, and latent variables $\mathbf{z}, \mathbf{s}, \pi, \gamma_s, \gamma_\varepsilon$
2: **for** $t := 1 \rightarrow T_{Gibbs}$ **do**
3: Run the Gibbs sampling method for all instances in the auxiliary domain D_a based on the Eq. (5) $\sim$ Eq. (13)
4: **end for**
// Predict the class labels for the testing documents in the target domain D_t
5: Initialize the domain priors with the learned values of the auxiliary domain D_a
6: Learn the sparse representation $\mathbf{w}$ of all instances in the target domain D_t
7: Conduct the prediction of the class labels for the testing data using Linear SVM.

useful to boost the performance of target domain. We show the overall process in Algorithm 1 for cross-domain event recognition. Firstly, we use the proposed CDCL model to alternately sample data of the auxiliary domain, which can learn shared domain priors. Specifically, we utilize the data information of the auxiliary domain to deduce the shared domain and modality priors $\pi, \gamma_s, \gamma_\varepsilon$. Then, the learned priors can be utilized to learn the CDCL model, and obtain the feature representation $\mathbf{w}$ of all data instances in the target domain D_t. Next, we train the Linear SVM [42] to conduct the prediction of the testing data.

B. Cross-Network Video Recommendation

The cross-network video recommendation is to use rich cross-network behavior information of the users across different social networks to help conduct users' preference estimation, particularly for the new user with only few records in the new social network. In this paper, the social activity behaviors of users are obtained from the auxiliary social network. This information can be utilized to assist another network to deal with the cold-start video recommendation by the proposed CDCL. Here, we consider the YouTube and Twitter as the two different network platforms, and these two platforms can be associated by the overlapped user account link information between YouTube and Twitter. When a user enrolls on YouTube, we do not know anything about his/her association on Youtube, and cannot conduct video recommendation. In this work, we use the proposed CDCL model to conduct a YouTube video recommendation application by leveraging both the tweeting behaviors of Twitter and historical video records of YouTube, which can cope with the cold-start video recommendation task.

In cross-network video recommendation, there are a set of the overlapped users U, and we represent each user $u \in U$ as a 2-dimensional tuple $\langle \mathbf{u}^T, \mathbf{u}^Y \rangle$. Here, we consider the user u as an instance, and the variables $\mathbf{u}^T, \mathbf{u}^Y$ are denoted as the user's feature representation on Twitter and on YouTube, respectively. For Twitter users, we conduct user topic modeling with tweeting formation and the friend-following behaviors to obtain user's feature representation, respectively. **Content-Based Topic modeling**: Since our task is to recommend YouTube videos to users on cross-domain semantic level, we describe user's feature in the semantic topic space [43]. Specifically, tweet behaviors of each user can be considered as the document. Then, we use the standard Latent Dirichlet Allocation to learn the users' topic distribution from the dataset composed by all Twitter users. The learned topic space of Twitter users may contain some co-occurred semantic concepts that are also learned on YouTube user semantic space. **Network-Based Topic modeling**: On Twitter, users follow each other and have their social link networks. As in [44], each user's friends (words) can be considered as one document, and we can also use the standard Latent Dirichlet Allocation to obtain the feature representation of each user. Because the topic modeling method utilizes co-occurrence associations, the learned Twitter topic space can capture some user interest shared by a subset of the Twitter friends. After topic modeling, the feature description of the Twitter user $\mathbf{u}^T$ can be denoted as (1) Content-based tweet topic distribution, $\mathbf{u}^{T_t} = \{u_1^{T_t}, \cdots, u_{K_T}^{T_t}\}$, as well as (2) Network-based friend topic distribution, $\mathbf{u}^{T_f} = \{u_1^{T_f}, \cdots, u_{K_T}^{T_f}\}$. For YouTube users, we first obtain the feature representation of each video interested by users. Here, each video $v \in \mathcal{V}$ can be described as $\mathbf{v} = \{v_1, \cdots, v_{K^Y}\}$ by using the modified iCorr-LDA model to model the visual and tag information of videos as in [43]. Then, the feature representation of the YouTube user $\mathbf{u}^Y$ is learned by his/her interested video set. Here, for each user $\mathbf{u}^Y$ of the YouTube, we can obtain the interested video set $\mathcal{V}_u \subset \mathcal{V}$ from his/her uploaded videos, favorite videos, and videos of the playlists. When given the YouTube video $v \in \mathcal{V}_u$ and its feature representation $\mathbf{v}$, we use the commonly max-pooling to obtain the feature representation of the YouTube user. Then, we represent $\mathbf{u}^Y$ as $\mathbf{u}^Y = \{u_1^Y, \cdots, u_{K_Y}^Y\}$, where K_Y is the topic number of the YouTube video. For a new user $u \in U$ on YouTube, the goal of cross-network video recommendation is to recommend a ranking list of videos V_u based on the user's behaviors by considering the user's tweet activities $\mathbf{u}^T$.

Our proposed cross-network video recommendation solution has two stages, as shown in Algorithm 2. Firstly, we investigate the proposed CDCL model to discover the shared latent information from different networks by considering the obtained Twitter and YouTube user feature representation $\mathbf{u}^T, \mathbf{u}^Y$. Here, when given the user feature representation $\mathbf{u}^T = \{u_1^T, \cdots, u_{K_T}^T\}$ and $\mathbf{u}^Y = \{u_1^Y, \cdots, u_{K_Y}^Y\}$, the shared dictionary variables, $\mathbf{D}^T = \{d_1^T, \cdots, d_K^T\}$ and $\mathbf{D}^Y = \{d_1^Y, \cdots, d_K^Y\}$, can be learned on Twitter and YouTube, respectively. By the obtained shared dictionary $\mathbf{D}^T$ and $\mathbf{D}^Y$, we can conduct the transform of feature representation of different social networks for users. As a result, we can deal with the video recommendation task for the new

Algorithm 2: The proposed CDCL method for cross-network video recommendation

Input: The feature representation of the user $\mathbf{u}^T \in U$ on Twitter; The feature representation of the user $\mathbf{u}^Y \in U$ on YouTube; The candidate YouTube video set $v_t \in V$; The feature representation of the test user $u_t \in U_t$ on YouTube.

Output: A ranked list of videos V_u for the test user u_t.

// Conduct the cross-network dictionary learning

1: Learn the shared dictionary $\mathbf{D}^T = \{d_1^T, \cdots, d_K^T\}$ on Twitter based on the Eq. (5) $\sim$ Eq. (13)

2: Learn the shared dictionary $\mathbf{D}^Y = \{d_1^Y, \cdots, d_K^Y\}$ on YouTube based on the Eq. (5) $\sim$ Eq. (13)

//Conduct the video recommendation for the new YouTube user u_t

3: Learn the feature representation of the user $\mathbf{u}^T \in R^{K_T \times 1}$ by the user's tweet history information on Twitter

4: Estimate the corresponding sparse coefficient $\mathbf{w}$ by Eq. (14)

5: Learn the feature representation of the user $\mathbf{u}^Y$ on YouTube by Eq. (15)

6: Recommend a ranked list of videos V_u for u_t by Eq. (16)

user of the YouTube by adopting his/her tweet history behaviors. Secondly, given new user on YouTube, based on the obtained $\mathbf{D}^T$ and $\mathbf{D}^Y$, we need to recommend a ranking list of videos V_u. Here, we can learn the feature representation of each user $\mathbf{u}^T \in R^{K_T \times 1}$ on Twitter by using his/her tweet behavior information. Then, we can obtain the sparse feature coefficient of the user $\mathbf{w}$ via (14) for $\mathbf{u}^T$.

$$\mathbf{u}^T = \mathbf{D}^T \mathbf{w} + \varepsilon_j \tag{14}$$

The common user among multiple domains has the shared dictionary ($\mathbf{D}^T$, $\mathbf{D}^Y$) and the corresponding sparse feature coefficient ($\mathbf{w}$), and these obtained parameters can be exploited to conduct the transform of feature distribution of different social networks for users. As a result, we can represent the feature representation of the new user on YouTube as:

$$\mathbf{u}^Y = \mathbf{D}^Y \mathbf{w} \tag{15}$$

Therefore, when given the new user $\mathbf{u}^Y$ and candidate YouTube videos $\mathbf{v}_t \in V$ which are represented in the same feature space, we can rank the recommended videos from YouTube by (16).

$$sim(\mathbf{u}^Y, \mathbf{v}_t) = \langle \mathbf{u}^Y, \mathbf{v}_t \rangle = \sum_{k=1}^{K_T} u_k^T \cdot v_{k,t} \tag{16}$$

V. Experiments

In this section, we conduct experimental evaluations of the proposed CDCL algorithm on two different applications: cross-platform event recognition and cross-network video recommendation. The experimental results show the effectiveness of our CDCL method for cross-domain collaborative learning in social multimedia.

A. Cross-Platform Event Recognition

1) Dataset Collection: In this paper, we construct the evaluation dataset from online social platforms for social event recognition. There is one publicly available event dataset called MediaEval social event detection (SED) [45]. However, event data in the SED dataset do not have multi-modality cross-domain information. Besides, this dataset does not contain social events happening currently. To better analyze event data with the multi-domain and the multi-modal properties, we concentrate on 8 complex social events which happened in the past couple of years, and gather a large amounts of event documents from Flickr and Google News. The collected 8 social events cover a wide range of topics including politics, economics, military, society, and so on. Therefore, the collected dataset has very rich types including multiple social events from a wide range of topics. For these 8 events, we manually create the introduction page of each event or download it from the Wikipedia page[6], which contains the whole stories of each event. Then, based on the whole timeline of each social event, we seek and download related content information including text and images from Flickr and Google News using these keywords. The detail of our collected dataset is given in Table I. Here, each social event has around 2000 to 8000 documents which contain texts and the corresponding images. We adopt simple rules to delete the unnecessary documents without including the queried keywords of the event, and ensure the reliability of the most of documents. In the collected dataset, some events are very similar, such as "North Korea nuclear program" and "Senkaku Islands dispute", "War in Afghanistan" and "Syrian civil war". Due to the similar topics in those events, it brings great challenges for social event analysis.

2) Feature Extraction: For textual representation, we adopt the stemming method and stop words elimination strategy to obtain clean data, and save words with the word frequency not less than 15 in the whole dataset. Then, textual information can be represented by the traditional vector space model. For visual representation, we use the popular sparse coding method [46], [47]. Specifically, we first conduct SIFT points sampling for images. Then, a simple K-means method is used to obtain a codebook. Next, we adopt Localized Soft-assigment Coding (LSC) method to acquire their descriptors, and the final image representation can be obtained by adopting the Spatial Pyramid Matching (SPM) and max pooling methods.

3) Results and Analysis: For our experiment setting, we set the hyperparameter values as $c_0 = d_0 = e_0 = f_0 = 10^{-6}$, and the beta-distribution parameter values as $a_0 = K, b_0 = 1$, as in [29]. We initially set the topic numbers to be $K = 100$. Here, we do not use all K dictionary elements, and decide the final number of the shared dictionary elements by the shared priors. We set the iteration number to be $T_{gibbs} = 100$. In our

[6]http://www.wikipedia.org

TABLE I
DESCRIPTION OF THE EVENT NAME, DURATION TIME, AND NUMBER OF DOCUMENTS FOR EACH SOCIAL EVENT IN OUR COLLECTED DATASET

Event ID	Event Name	Start Time	End Time	Google News		Flickr	
				#Images	#Text	#Images	#Text
1	Senkaku Islands dispute	2008.06	2012.12	3743	2495	6617	6617
2	Occupy Wall Street	2011.09	2012.09	5601	3108	7151	7151
3	United States Presidential Election	2009.10	2013.01	5169	3446	7352	7352
4	War in Afghanistan	2001.10	2012.08	5373	2915	7172	7172
5	North Korea nuclear program	2000.01	2012.04	3969	2640	8635	8635
6	Greek protests	2011.05	2012.04	3900	2630	7385	7385
7	Mars Reconnaissance Orbiter	2005.04	2012.08	3901	2600	7188	7188
8	Syrian civil war	2011.01	2013.01	4899	3266	7426	7426

experiment, we select half of the dataset as the training data and the left as the testing data.

We compare the proposed model with four baseline methods including BOW, CCA, SRC-L1, SRC-L1-DL, and LC-KSVD1.

- BOW: The text and image features are concatenated to conduct the representation of each document, as introduced in Section V-A2.
- Canonical Correlation Analysis (CCA) [48]: The representation of each document is obtained by using maximally correlated subspace constraint to learn a latent feature space in the multi-modal data.
- SRC-L1: The representation of each document is learned by using the traditional sparse representation method, and we use the traditional K-SVD method to learn the dictionary.
- SRC-L1-DL: The representation of each document is learned by using the traditional sparse representation method, and we use the proposed non-parametric model to learn the dictionary, as shown in Algorithm 1.
- LC-KSVD1 [33]: The representation of each document is learned by a label consistent K-SVD (LC-KSVD) algorithm with a discriminative dictionary for sparse coding.

In this experiment, our target is to classify the data instances on Google News/Flickr domain with the assistance of the auxiliary domain Flickr/Google News, respectively. There are 4 methods CDCL*, CDCL-c, CDCL-s, and CDCL in different experimental settings. The CDCL* is a special case of the CDCL without considering the supervised information in [11]. The CDCL-c is learned with the assistance of the auxiliary domain, which only concatenates the text and image features rather than adopts the fusion of multi-modal information. The CDCL-s is learned without the assistance of the auxiliary domain, but adopts the fusion of multi-modal information in the single domain. The CDCL is learned with the assistance of the auxiliary domain, and adopts the multi-modal property and supervised property. Once data feature representation is obtained, we utilize the Linear SVM to learn the classifier.

We show all classification results and give the performance comparison for each class, as show in Table II, Figs. 4 and 5. From the results, we can have the following conclusions. (1) The BOW model obtains worse performance than other methods. This shows that the BOW cannot effectively distinguish the relationships of the multi-modal data in modeling textual and visual words. (2) The CCA and the proposed CDCL methods obtain better performance, this shows that the fusion of the text and image information is helpful for social event analysis. (3) The SRC-L1-DL obtains much better results than the SRC-L1 and LC-KSVD1. This shows that the better feature representation can be obtained in the dictionary learning process with the

TABLE II
THE EVENT CLASSIFICATION RESULTS WITH DIFFERENT METHODS

Methods	Accuracy	
	#Google News	#Fickr
BOW	0.803	0.853
CCA	0.764	0.859
SRC-L1	0.825	0.858
SRC-L1-DL	0.839	0.865
LC-KSVD1	0.857	0.869
CDCL*	**0.862**	**0.881**
CDCL-s	**0.879**	**0.883**
CDCL-c	**0.888**	**0.884**
CDCL	**0.915**	**0.899**

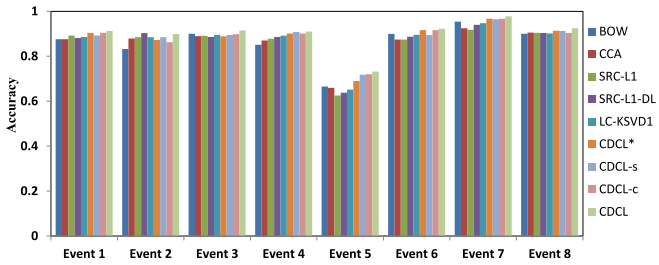


Fig. 4. The classification results for each event with different methods on Flickr.

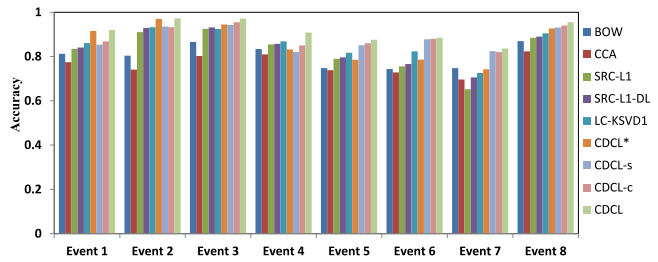


Fig. 5. The classification results for each event with different methods on Google News.

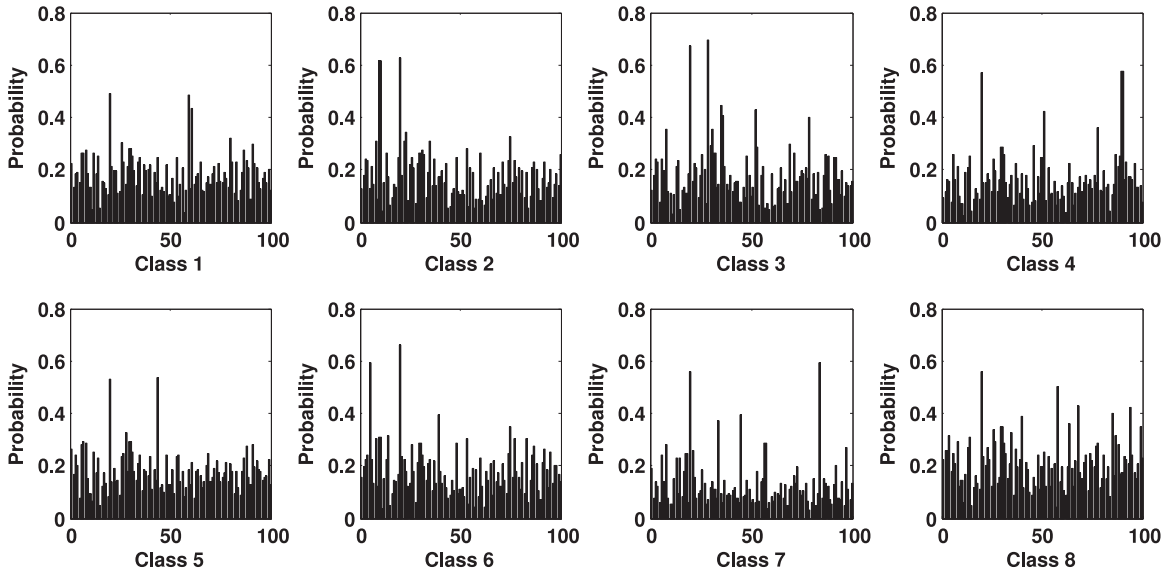


Fig. 6. The Dictionary weights π_c for each event on Google News.

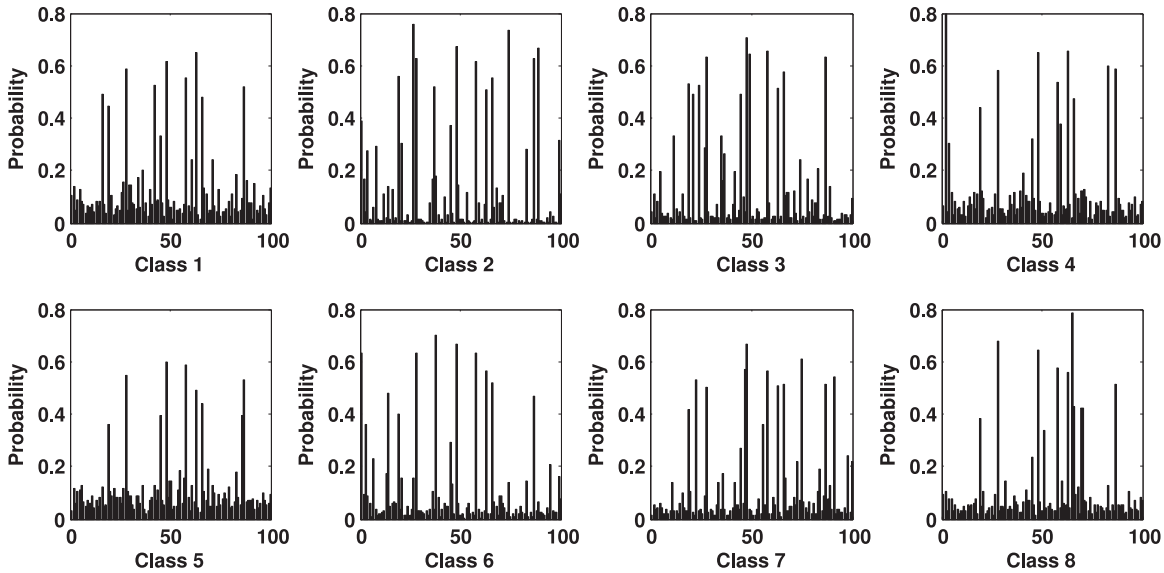
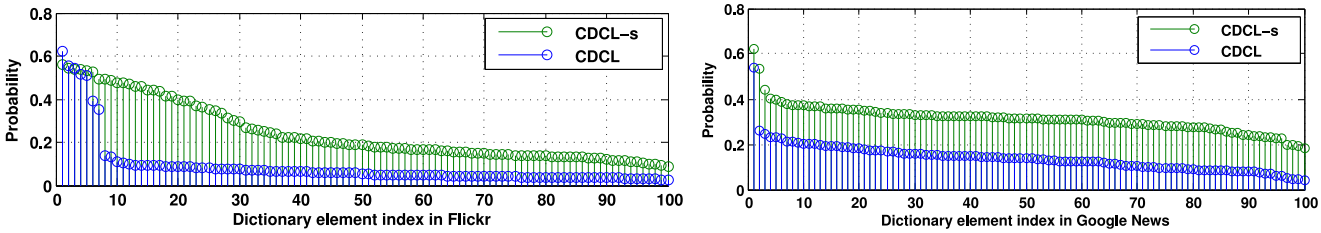


Fig. 7. The Dictionary weights π_c for each event on Flickr.

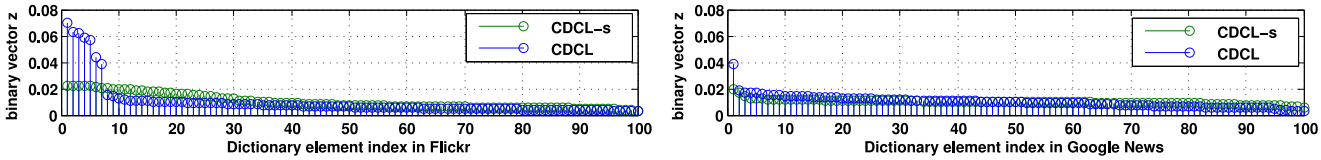
assistance of the auxiliary domain. (4) The CDCL has better performance than the CDCL*. This shows that the discriminative Bayesian dictionary learning can use the cross-domain data information to learn discriminative representation, which can boost the classification performance. (5) Overall, our CDCL method consistently outperforms other existing methods. The major reason is that the proposed model can expliot the shared domain, modality and supervised properties to jointly learn the feature representation. Therefore, the proposed model can investigate the superiorities of different sources to supplement and improve each other effectively.

Then, we conduct the parameter analysis of the proposed model in details. We visualize the dictionary weight $\pi_c, c \in C$ of each class, as shown in Figs. 6 and 7. Here, $\pi_{c,k}$ is followed by all the k-th components of the feature representation jointly for the c-th class. Because the learned dictionary is shared by all c classes, if the training data are commonly represented by the k-th dictionary element, we should expect a high value of $\pi_{c,k}$. From Figs. 6 and 7, we can observe that the high values appear at different locations in the dictionary weight $\pi_c, c \in C$ for different classes, which shows the learned dictionary is discriminative. For example, for the dictionary weights of two different classes (event 6 and event 7), the maximum top-5 element indexes of the two classes on the Google News are 20, 5, 39, 75, 14 and 84, 20, 45, 34, 57 respectively, and most of them are different. Similarly, for two similar events, such as event 2 and 3, the maximum top-5 element indexes of the two classes on the Google News are 20, 10, 23, 75, 22 and 29, 20, 35, 52, 36 respectively. We can observe that most of them are also different, which demonstrates the discriminative character of the learned dictionary. We also see that some non-zero values appear with similar locations in different classes. These results



(a) (b)

Fig. 8. The comparison of dictionary weights π_k by CDCL-s and CDCL on Flickr and Google News, respectively. (a) Dictionary weights π_k (Flickr). (b) Dictionary weights π_k (Google News).

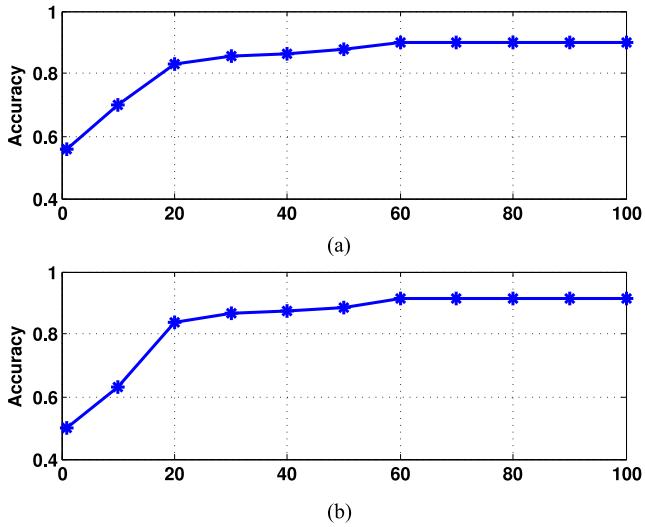


(a) (b)

Fig. 9. The statistics results of binary vector z by CDCL-s and CDCL on Flickr and Google News, respectively. (a) Statistics of binary vector z (Flickr). (b) Statistics of binary vector z (Google News).

show these dictionary elements are shared among the feature representation of different classes. Therefore, it is useful and effective to apply our discriminative dictionary learning method for the event classification task.

The comparisons of the dictionary weights π_k are visualized in Fig. 8(a) and (b), where the results are ordered with the learned probability and can be obtained by CDCL-s and CDCL on Google News and Flickr, respectively. The results show that the CDCL model obtains the low probability values in most of the dictionary elements. But the probability values of the CDCL-s model are more than 0.2. Therefore, our CDCL can obtain more sparse feature representation. This further verifies that the proposed model can exploit the auxiliary domain as the prior knowledge to help improve the shared dictionary learning. We also visualize the statistics results of binary vector $\mathbf{z}$ in Fig. 9(a) and (b), where the results are obtained by figuring the expected value of binary components on Flickr and Google News, respectively. We observe that the obtained binary vector $\mathbf{z}$ is coherent with the obtained dictionary weights π_k, and this verifies that our model is sensible. In Fig. 10, we give the classification performances with different Gibbs sampling iterations on Google News and Flickr, respectively. We observe that accessible results can be obtained with 20 iterations. We run 100 Gibbs sampling iterations, and this process takes around 15 minutes on the Flickr, and around 22 minutes on the Google News on average. Results are produced on an Intel Core i7 CPU at 3.6 GHz with 16 GB RAM running Matlab.



(a)

(b)

Fig. 10. The classification results with the numbers of different iteration in the Gibbs sampling method on Flickr and Google News, respectively. (a) On Flickr. (b) On Google News.

B. Cross-Network Video Recommendation

1) Dataset Collection: In this paper, the cross-network user dataset is used as in [37]. This dataset contains 143,259 Google+ users with user account linkage between the YouTube and Twitter, where there are 38,540 users in the YouTube account, 39,400 users in the Twitter account, and 11,850 common users both YouTube and Twitter accounts. In this public dataset, Twitter users have no tweet activity information. Based on the users' account information provided by [37], we download the latest 1,000 tweets created by each user through the Twitter APIs. In the experiment, only users having both the Twitter and YouTube accounts are used, and we also only keep the users who are interacted with no less than 8 unique video recordings on YouTube. Finally, 1,655 cross-network users and 5,105 YouTube videos are obtained for the experimental evaluation.

In our experiment setting, we expect the proposed video recommendation method to help improve cold-start recommendation problem for the new user of the YouTube. We first randomly choose 900 active users to build the training dataset and learn the

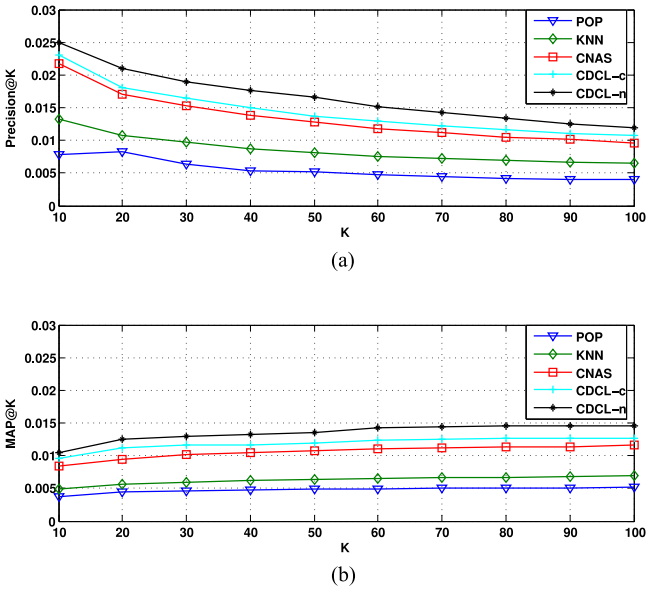


Fig. 11. The Precision and MAP results of cross-network video recommendation for the new YouTube users. (a) Precision@K. (b) MAP@K.

shared dictionary. Then, we consider the remaning 755 users as the cold-start test users, which are represented as U^{new}. For each testing user $u_t \in U^{new}$, we hide all the watched video-related associations in the testing step and take them as the ground truth for experimental evaluation.

2) Evaluation Metrics: The top-ranked recommendation results are concerned by users in the practical video recommendation task. The goal of the customized video recommendation is to give each user a video ranking list. In the experiment evaluation, we utilize Precision@K, and Mean Average Precision (MAP@K) to quantify the performance of recommended videos, which are similar to traditional information retrieval task. We represent Precision@K as $Precision@K = \sum_{k=1}^{K} r_k/K$. The MAP@K is the mean of average precision scores over test users U^{new}, and is denoted as:

$$MAP@K = \frac{1}{U^{new}} \sum_{u=1}^{U^{new}} \frac{\sum_{k=1}^{K} Precision@uk * r_{u,k}}{L_u}, \tag{17}$$

where r_k denotes the relevant level at the index k, if the value is zero, it represents "Not Relevant" otherwise represents "Relevant". And, $r_{u,k}$ denotes the relevant level of the user u at index k. $Precision@uk$ denotes the precision of the user u at index k. The result can be acquired by testing whether the recommended videos are in the interested video recordings of the user u. We set different truncation levels K, $K \in 10, 20, 30, 40, 50, 60, 70, 80, 90, 100$ in our experiment.

3) Results and Analysis: In the experiment, we compare our models (CDCL-c, CDCL-n) with three baseline methods:

- Popularity (POP): It is to adopt the video's popularity to conduct the same recommendation list of the YouTube videos for new users, and the view counts of the videos are considered as their popularity.

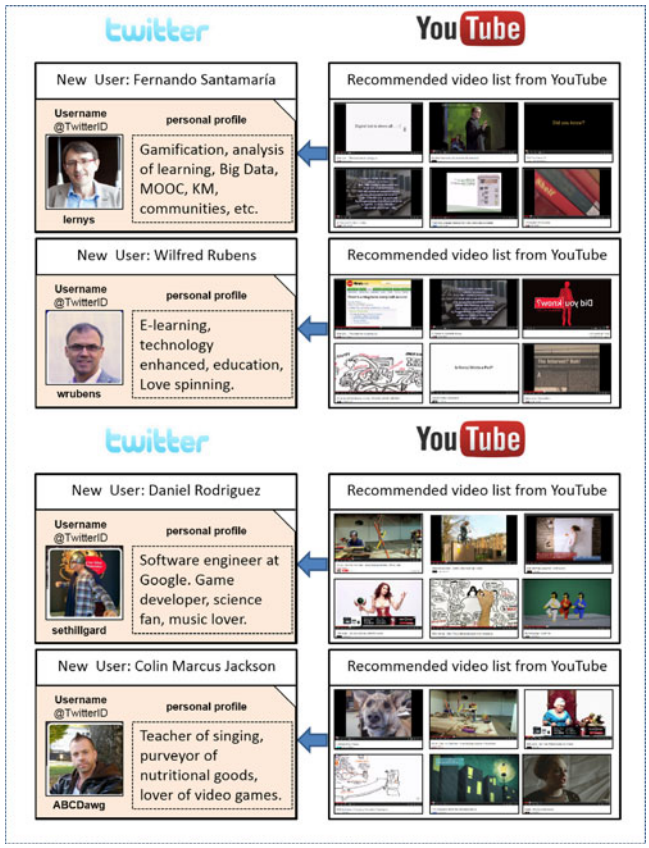


Fig. 12. Four examples are shown on cross-network video recommendation from Twitter to YouTube users.

- KNN: It is to adopt user's Twitter information to acquire most relevant Twitter users by the KNN method for a new YouTube user. Then, based on the most relevant users we can obtain the related videos.
- Cross-network Association (CNAS) [37]: It is to adopt a coupled dictionary sparse learning method to learn the common dictionary space based on the same users among multiple networks.

In our experiment setting, we have 2 variant methods including CDCL-c and CDCL-n with different feature representation on Twitter. The CDCL-c and CDCL-n methods are the same as our model, but are conducted to obtain Twitter users' feature representation by the content-based topic modeling and the network-based topic modeling, respectively. We visualize the experimental results with different methods in Fig. 11. From the results, we can have the following observations: (1) The POP model obtains worse performance, and it is because this method is lack of the ability to learn user's personalized demands, and does not consider cross-network user behavior information. (2) The KNN and CNAS methods obtain better results than the POP model, and this shows that it is helpful to utilize the cross-network user behaviors for the cold-start video recommendation. (3) The proposed CDCL-c and CDCL-n methods obtain the best recommendation performance. This shows the proposed model can effectively exploit cross-network activity behaviors of the users to collaboratively learn the shared

dictionary, and can deal with the user cold-start recommendation problem. (4) The proposed CDCL-n achieves much better performance than the CDCL-c, which shows the effectiveness of the network-based feature representation. And, a possible reason is that network-based topic modeling by user interest indicator is more stable than noisy tweet content information on Twitter, which can obtain better feature representation.

We show the results of four new YouTube users recommended by the proposed model in Fig. 12. Here, we show their tweet history data and the related recommended video list. We consider the user of the Fig. 12 for instance, and the user's name is Daniel "Rodriguez". We can see that the user is a software engineer, and likes science and music on Twitter. From the results, the related recommended video list from YouTube contains some prominent music, science innovation and game outline, which can make the new YouTube user obtain a good experience. Therefore, these results demonstrate the effectiveness of the proposed cross-network video recommendation solution.

VI. Conclusion

In this paper, we have proposed a generic cross-domain data analysis framework based on the discriminative non-parametric Bayesian dictionary model. The proposed discriminative learning model is able to not only introduce the shared domain and the modality priors to cope with the domain gap as well as consider the multi-modal property, but also exploit the class label information of data to obtain the discriminative dictionary for feature representation. We evaluate the proposed model on two different applications, and shows that it achieves the best performance. In the future, we will apply the proposed model to more different applications, like cross-domain event association and cross-domain user representation.

References

[1] X. Wu, C.-W. Ngo, and A. G. Hauptmann, "Multimodal news story clustering with pairwise visual near-duplicate constraint," *IEEE Trans. Multimedia*, vol. 10, no. 2, pp. 188–199, Feb. 2008.

[2] X. Yang, T. Zhang, and C. Xu, "Cross-domain feature learning in multimedia," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 64–78, Jan. 2015.

[3] S. Qian, T. Zhang, C. Xu, and M. S. Hossain, "Social event classification via boosted multimodal supervised latent dirichlet allocation," ACM Trans. Multimedia Comput., Commun., Appl., vol. 11, no. 2, pp. 27:1–27:22, 2014.

[4] J. Tang, S. Wu, J. Sun, and H. Su, "Cross-domain collaboration recommendation," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 1285–1293.

[5] T. Zhang and C. Xu, "Cross-domain multi-event tracking via CO-PMHT," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 10, no. 4, 2014, Art. no. 31.

[6] S. Qian, T. Zhang, C. Xu, and J. Shao, "Multi-modal event topic model for social event analysis," *IEEE Trans. Multimedia*, vol. 18, no. 2, pp. 233–246, Feb. 2016.

[7] J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2006, pp. 120–128.

[8] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," in *Proc. 21st Int. Joint Conf. Artif. Intell.*, 2009, pp. 1187–1192.

[9] F. Abel, S. Araújo, Q. Gao, and G. Houben, "Analyzing cross-system user modeling on the social web," in *Proc. Int. Conf. Web Eng.*, Paphos, Cyprus, 2011, pp. 28–43.

[10] M. Yan, J. Sang, T. Mei, and C. Xu, "Friend transfer: Cold-start friend recommendation with cross-platform transfer learning of social knowledge," in *Proc. IEEE Int. Conf. Multimedia Expo.*, San Jose, CA, USA, Jul. 2013, pp. 1–6.

[11] S. Qian, T. Zhang, R. Hong, and C. Xu, "Cross-domain collaborative learning in social multimedia," in *Proc. 23rd Annu. ACM Conf. Multimedia Conf.*, Brisbane, Qld., Australia, Oct. 2015, pp. 99–108.

[12] X. Yang, T. Zhang, C. Xu, and M. Yang, "Boosted multifeature learning for cross-domain transfer," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 11, no. 3, 2015, Art. no. 35.

[13] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan, "Learning consistent feature representation for cross-modal multimedia retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 370–381, Mar. 2015.

[14] X. Chang, Y. L. Yu, Y. Yang, and E. P. Xing, "They are not equally reliable: Semantic event search using differentiated concept classifiers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1884–1893.

[15] X. Chang, Z. Ma, Y. Yang, Z. Zeng, and A. G. Hauptmann, "Bi-level semantic representation analysis for multimedia event detection," *IEEE Trans. Cybern.*, vol. 47, no. 5, pp. 1180–1197, May 2017.

[16] B. K. Bao, G. Liu, C. Xu, and S. Yan, "Inductive robust principal component analysis," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3794–3800, Aug. 2012.

[17] B. K. Bao, G. Zhu, J. Shen, and S. Yan, "Robust image analysis with sparse representation on quantized visual features," *IEEE Trans. Image Process.*, vol. 22, no. 3, pp. 860–871, Mar. 2013.

[18] Y. Yang, Y. Yang, and H. T. Shen, "Effective transfer tagging from image to video," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 9, no. 2, 2013, Art. no. 14.

[19] Y. Yang, Z. Zha, Y. Gao, X. Zhu, and T. Chua, "Exploiting web images for semantic video indexing via robust sample-specific loss," *IEEE Trans. Multimedia*, vol. 16, no. 6, pp. 1677–1689, Oct. 2014.

[20] X. Yang, T. Zhang, C. Xu, and M. S. Hossain, "Automatic visual concept learning for social event understanding," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 346–358, Mar. 2015.

[21] J. Bian, Y. Yang, H. Zhang, and T. Chua, "Multimedia summarization for social events in microblog stream," *IEEE Trans. Multimedia*, vol. 17, no. 2, pp. 216–228, Feb. 2015.

[22] X. Chen, A. O. Hero III, and S. Savarese, "Multimodal video indexing and retrieval using directed information," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 3–16, Feb. 2012.

[23] H. Li, Y. Wang, T. Mei, J. Wang, and S. Li, "Interactive multimodal visual search on mobile device," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 594–607, Apr. 2013.

[24] W. Liu, T. Mei, and Y. Zhang, "Instant mobile video search with layered audio-video indexing and progressive transmission," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2242–2255, Dec. 2014.

[25] H. Zhang, X. Shang, H. Luan, M. Wang, and T.-S. Chua, "Learning from collective intelligence: Feature learning using social images and tags," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 13, no. 1, pp. 1:1–1:23, 2017.

[26] A. A. Liu, Y. T. Su, W. Z. Nie, and M. Kankanhalli, "Hierarchical clustering multi-task learning for joint human action grouping and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 102–114, Jan. 2017.

[27] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. S. Huang, "Coupled dictionary training for image super-resolution," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3467–3478, Aug. 2012.

[28] S. Wang, L. Zhang, Y. Liang, and Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2216–2223.

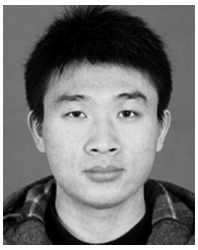
[29] M. Zhou *et al.*, "Non-parametric Bayesian dictionary learning for sparse image representations," in *Proc. 22nd Int. Conf. Neural Inf. Process. Syst.*, 2009, pp. 2295–2303.

[30] C. Yuan, W. Hu, G. Tian, S. Yang, and H. Wang, "Multi-task sparse learning with beta process prior for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 423–429.

[31] X. He, H. Zhang, M. Y. Kan, and T. S. Chua, "Fast matrix factorization for online recommendation with implicit feedback," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2016, pp. 549–558.

[32] H. Wang, C. Yuan, W. Hu, and C. Sun, "Supervised class-specific dictionary learning for sparse modeling in action recognition," *Pattern Recognit.*, vol. 45, no. 11, pp. 3902–3911, 2012.

[33] Z. Jiang, Z. Lin, and L. S. Davis, "Label consistent K-SVD: Learning a discriminative dictionary for recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2651–2664, Nov. 2013.
[34] N. Akhtar, F. Shafait, and A. S. Mian, "Discriminative Bayesian dictionary learning for classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 12, pp. 2374–2388, Dec. 2016.
[35] S. D. Roy, T. Mei, W. Zeng, and S. Li, "Socialtransfer: Cross-domain transfer learning from social streams for media applications," in *Proc. ACM Int. Conf. Multimedia*, 2012, pp. 649–658.
[36] F. Abel, E. Herder, G. Houben, N. Henze, and D. Krause, "Cross-system user modeling and personalization on the social web," *User Model. User-Adapt. Interact.*, vol. 23, no. 2/3, pp. 169–209, 2013.
[37] M. Yan, J. Sang, and C. Xu, "Mining cross-network association for youtube video promotion," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 557–566.
[38] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.
[39] J. Paisley and L. Carin, "Nonparametric factor analysis with beta process priors," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 777–784.
[40] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Nat. Acad. Sci.*, vol. 101, pp. 5228–5235, 2004.
[41] M. Aharon, M. Elad, and A. Bruckstein, "rmK-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.
[42] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, 2008.
[43] M. Yan, J. Sang, C. Xu, and M. S. Hossain, "YouTube video promotion by cross-network association: @Britney to advertise Gangnam style," *IEEE Trans. Multimedia*, vol. 17, no. 8, pp. 1248–1261, Aug. 2015.
[44] Y. Cha and J. Cho, "Social-network analysis using topic models," in *Proc. 35th Int.ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, Portland, OR, USA, Aug. 2012, pp. 565–574.
[45] T. Reuter *et al.*, "Social event detection at mediaeval 2013: Challenges, datasets, and evaluation," in *Proc. MediaEval Multimedia Benchmark Workshop*, 2013, pp. 18–19.
[46] T. Zhang, B. Ghanem, S. Liu, C. Xu, and N. Ahuja, "Low-rank sparse coding for image classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 281–288.
[47] T. Zhang, S. Liu, N. Ahuja, and M.-H. Yang, "Robust visual tracking via consistent low-rank sparse learning," *Int. J. Comput. Vis.*, vol. 111, no. 2, pp. 171–190, 2015.
[48] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 251–260.



Shengsheng Qian received the B.E. degree from Jilin University, Changchun, China, in 2012, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2017. He is currently an Assistant Professor with the Institute of Automation, Chinese Academy of Sciences. His current research interests include social media data mining and social event content analysis.

Tianzhu Zhang (S'09–M'11) received the B.S. degree in communications and information technology from the Beijing Institute of Technology, Beijing, China, in 2006, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2011. He is currently an Associate Professor with the Institute of Automation, Chinese Academy of Sciences. His current research interests include computer vision and multimedia, especially action recognition, object classification, and object tracking.

Changsheng Xu (M'97–SM'99–F'14) is a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, and the Executive Director for the China–Singapore Institute of Digital Media. He holds 30 granted/pending patents and has authored or coauthored more than 200 refereed research papers in these areas. He is an Associate Editor for the IEEE TRANSACTIONS ON MULTIMEDIA, *ACM Transactions on Multimedia Computing, Communications, and Applications*, and *ACM/Springer Multimedia Systems Journal*. His research interests include multimedia content analysis/indexing/retrieval, pattern recognition, and computer vision. Prof. Xu was the recipient of the Best Associate Editor Award of ACM Transactions on Multimedia Computing, Communications, and Applications in 2012 and the Best Editorial Member Award of ACM/Springer *Multimedia Systems Journal* in 2008. He was the Program Chair for ACM Multimedia 2009. He was an Associate Editor, Guest Editor, General Chair, Program Chair, Area/Track Chair, Special Session Organizer, Session Chair, and TPC Member for more than 20 IEEE and ACM prestigious multimedia journals, conferences, and workshops. He is a Fellow of IAPR and a Distinguished Scientist of ACM.