Action and Event Recognition in Videos by Learning From Heterogeneous Web Sources

Li Niu, Xinxing Xu, Lin Chen, Lixin Duan, and Dong Xu, *Senior Member, IEEE*

***Abstract*— In this paper, we propose new approaches for action and event recognition by leveraging a large number of freely available Web videos (e.g., from Flickr video search engine) and Web images (e.g., from Bing and Google image search engines). We address this problem by formulating it as a new multi-domain adaptation problem, in which heterogeneous Web sources are provided. Specifically, we are given different types of visual features (e.g., the DeCAF features from Bing/Google images and the trajectory-based features from Flickr videos) from heterogeneous source domains and all types of visual features from the target domain. Considering the target domain is more relevant to some source domains, we propose a new approach named multi-domain adaptation with heterogeneous sources (MDA-HS) to effectively make use of the heterogeneous sources. In MDA-HS, we simultaneously seek for the optimal weights of multiple source domains, infer the labels of target domain samples, and learn an optimal target classifier. Moreover, as textual descriptions are often available for both Web videos and images, we propose a novel approach called MDA-HS using privileged information (MDA-HS+) to effectively incorporate the valuable textual information into our MDA-HS method, based on the recent learning using privileged information paradigm. MDA-HS+ can be further extended by using a new elastic-net-like regularization. We solve our MDA-HS and MDA-HS+ methods by using the cutting-plane algorithm, in which a multiple kernel learning problem is derived and solved. Extensive experiments on three benchmark data sets demonstrate that our proposed approaches are effective for action and event recognition without requiring any labeled samples from the target domain.**

***Index Terms*— Domain adaptation, learning using privileged information, multiple kernel learning.**

I. Introduction

RECENTLY, action and event recognition have attracted growing attention for real-world applications, such as video search and video surveillance. A large number of approaches have been proposed for action recognition [1]–[3] and event recognition [4]. In [1], the static and motion features were integrated for action recognition. To improve the action recognition performance, Wang and Schmid [2] developed the improved dense trajectory features. In [3], a two-stream deep convolutional network was proposed for action recognition by integrating both spatial and temporal information. For event recognition, Xu *et al.* [4] extracted the features from each video keyframe by using prelearned convolutional neural networks models, which are further integrated over the whole video for event detection. For the recent advances in event recognition, interested readers can refer to [5] for more details. Nevertheless, all the above works require sufficient labeled training samples in order to achieve reasonable action and event recognition performance.

However, it is often time-consuming and labor-intensive to collect labeled training videos based on human annotation. Meanwhile, we observe that abundant Web videos or images can be freely collected by using tag-based search [6]. Recently, researchers also developed new action and event recognition methods by employing Web data. Specifically, Duan *et al.* [7] developed a domain adaptation approach by learning from Web videos. In [6], a multi-domain adaptation scheme is also proposed for event recognition by using Web images from different sources. In [8], Web images that are incrementally collected were used for action recognition. However, simple actions like standing up and sitting down cannot be distinguished based on the works in [6] and [8] due to the lack of temporal information from the training Web images [7].

In this paper, we propose new approaches for action and event recognition without requiring any labeled videos in the target domain. In this paper, abundant Web images and videos are used as the loosely labeled training data. In addition to Web videos, Web images are also used for action and event recognition, because more Web images are available in the Internet and Web images are often accompanied with more accurate tags. Therefore, Web images can additionally be used to learn robust classifiers for improving action and event recognition performance. Motivated by [6] and [7], this task is formulated as a new multi-domain adaptation problem, in which heterogeneous sources are provided. Specifically, we are given different types of visual features (e.g., the DeCAF features from Web images and trajectory-based features from Web videos) from heterogeneous source domains and all types of visual features from the target domain. It is worth mentioning that the samples from each source domain are assumed to be associated with only one type of visual feature for ease of representation. If multiple types of visual features are extracted from the training samples in one source domain,

Manuscript received January 18, 2015; revised August 26, 2015 and December 24, 2015; accepted December 26, 2015. Date of publication March 11, 2016; date of current version May 15, 2017. This work was supported by the Faculty of Engineering and Information Technologies, The University of Sydney, through the Faculty Research Cluster Program.

L. Niu is with the School of Computer Engineering, Nanyang Technological University, Singapore 639798 (e-mail: lniu002@ntu.edu.sg).

X. Xu is with the Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore 138632 (e-mail: xuxinx@ihpc.a-star.edu.sg).

L. Chen and L. Duan are with Amazon, Seattle, WA 98109 USA (e-mail: gggchenlin@gmail.com; lxduan@gmail.com).

D. Xu is with the School of Electrical and Information Engineering, The University of Sydney, Sydney, NSW 2006, Australia (e-mail: dong.xu@sydney.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/TNNLS.2016.2518700

we can readily treat this source domain as multiple source domains with one type of visual features extracted from the same set of training samples in each source domain.

Based on our observation that the data distributions of some source domains are closer to that of the target domain, a new approach called multi-domain adaptation with heterogeneous sources (MDA-HS) is developed in Section III. To effectively cope with heterogeneous sources with different types of visual features, we seek for the optimal weights of different source domains and, at the same time, infer the labels of unlabeled samples in the target domain. In order to reduce the data distribution mismatch between each source domain and the target domain, we propose to learn an adapted classifier for each source domain by using the source classifier pretrained based on the loosely labeled training Web images/videos, in which the distance between the adapted classifier and the prelearned source classifier is measured based on their weight vectors. We propose a novel regularizer that adds up the weighted distances from all the source domains. We also propose a new target classifier by combining all the adapted classifiers with different weights. The new regularizer and target classifier are further incorporated into a new ρ-SVM-based objective function for domain adaptation. We also employ the cutting-plane method to solve the optimization problem in an iterative fashion, and a group-based multiple kernel learning (MKL) problem is also optimized at each iteration.

In addition to those visual features extracted from videos and images, we propose to utilize additional textual features extracted from surrounding textual descriptions (e.g., captions, titles, and tags) of training Web images and videos. With semantic meanings, those additional textual features are usually more discriminative than visual features, so a more robust target classifier is expected to be learned by effectively using those textual features. On the other hand, the videos in the target domain are generally not associated with any textual descriptions. As a result, we are facing the situation that each source sample (i.e., one Web video/image) is represented by one type of visual feature and the textual feature, while each target sample is represented by all types of visual features, as shown in Fig. 1. To handle this new setting, in Section IV, we propose to effectively utilize the additional textual features of source samples as privileged information, motivated by the learning using privileged information (LUPI) paradigm [9]. Specifically, we develop a new method called MDA-HS using privileged information (MDA-HS+). Moreover, a novel elastic-net-like regularization is further introduced for this newly proposed method, which leads to better results and more efficient optimization.

In Section V, extensive experiments are performed on three benchmark data sets. The results clearly show that our proposed methods are better than the related approaches for action and event recognition without requiring any labeled videos from the target domain.

The preliminary version of this paper was published in [10]. In this paper, we expand the work in [10] by proposing MDA-HS+ and MDA-HS+(ENR). This paper also provides more experiments for our new methods MDA-HS+ and MDA-HS+(ENR) and evaluate all methods using the Hollywood2 data set as well as conduct in-depth investigation of various aspects of the proposed approaches, such as robustness to the parameters and comparison of training time.

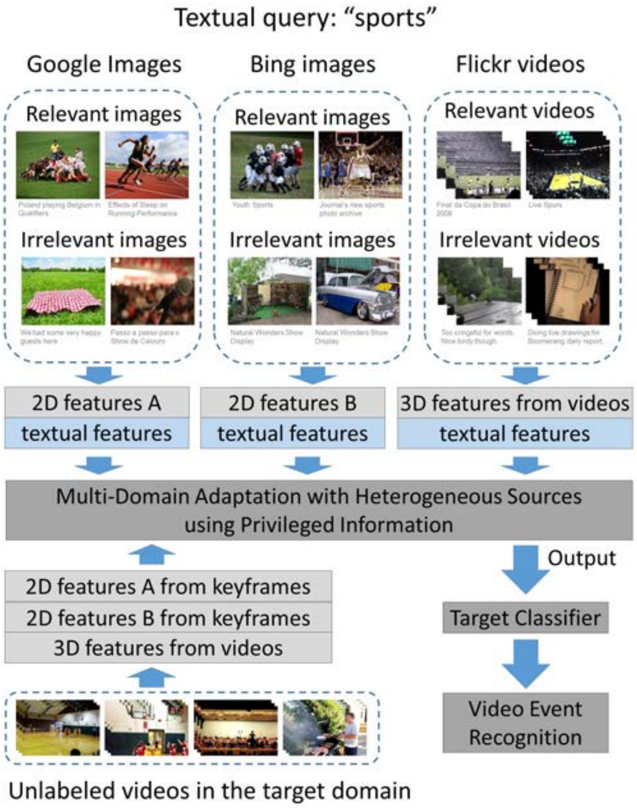

Textual query: "sports"
Google Images
Bing images
Flickr videos
Relevant images
Poland playing Belgium in Qualifiers
Effects of Sleep on Running Performance
Irrelevant images
We had some very happy guests here
Passo a passo para o Show de Calouro
Relevant images
Youth Sports
Journal's new sports photo archive
Irrelevant images
Natural Wonders Show Display
Natural Wonders Show Display
Relevant videos
Final da Copa do Brasil 2008
Live Spurs
Irrelevant videos
Too cringeful for words. Nice birdy though.
Doing live drawings for Boomerang daily report.
2D features A
textual features
2D features B
textual features
3D features from videos
textual features
Multi-Domain Adaptation with Heterogeneous Sources using Privileged Information
Output
2D features A from keyframes
2D features B from keyframes
3D features from videos
Target Classifier
Video Event Recognition
Unlabeled videos in the target domain

Fig. 1. Overview of our proposed multi-domain adaptation approach for action and event recognition by learning from heterogeneous Web sources. The source data contain Web images (resp., videos) and their surrounding textual descriptions from Google/Bing image search (resp., Flickr video search). The target domain contains unlabeled videos.

II. Related Work

Recently, domain adaptation approaches have been successfully used for different computer vision tasks, including object recognition [11]–[13] and event recognition [6], [7]. Most works focus on the single-source domain adaptation setting. For example, a few SVM-based approaches [7], [14] and distance metric learning approaches [13] were developed for domain adaptation. New domain adaptation methods were also developed in [11] and [12] by interpolating new subspaces to reduce the domain distribution mismatch between the two domains. A recent work in [15] proposed to learn a domain invariant subspace, while another approach in [16] learned the transform matrix to align the two subspaces from both domains.

Multi-domain adaptation methods were also developed [6], [17]–[20] when there are multiple source domains (i.e., the multi-domain adaptation setting). For example, the domain selection method was developed in [6] to select the most relevant source domains. Hoffman *et al.* [17] first discovered multiple latent source domains and then developed a multi-domain adaptation method by learning multiple transformations. A two-step approach was developed in [19], in which the weight for each source domain is first learned before learning the target classifier with the learned weights. However, in the existing approaches for multi-domain adaptation, the common assumption is that the training samples from all source and target domains are associated with the same type of feature. However, our new setting does not satisfy this assumption, because the samples from heterogeneous source domains are associated with only one type of visual feature, while the target domain data are associated with multiple types of visual features. As a result, these existing multi-domain adaptation methods [18]–[20] can only adopt the late-fusion strategy to fuse the prediction scores from multiple models, in which each model is trained by using one type of visual feature, or alternatively adopt the early fusion strategy to form a lengthy feature vector by concatenating multiple visual features as the feature representation of target data [6], [18]–[20]. In contrast, our work MDA-HS can seek for the optimal weights of different source domains and learn the optimal target classifier at the same time, while the samples in these source domains are represented by different types of visual features.

Our work is also different from heterogeneous domain adaption (HDA) [13], [21]. In HDA, different types of features are used to represent the samples from the source and the target domains. On the contrary, we assume the samples in the target domain have all types of visual features in our MDA-HS, so the samples from each pair of source and target domains are represented by the same type of visual feature. Labeled samples in the target domain are often required in the existing HDA methods [13], [21], while we do not require them in MDA-HS. Moreover, it is worth mentioning that our MDA-HS+ can take advantage of the additional textual features in the source domains as privileged information. How to utilize privileged information is not discussed in the above-existing domain adaptation methods.

Our work is different from zero-shot learning [22]–[24], which aims to transfer the knowledge from existing classes to unseen classes. In [22], multiple types of features are mapped to the high-dimensional concept space based on a large set of learned concept detectors. In [23], similarities among the concepts are utilized to fuse existing classifiers for recognizing the testing samples from an unseen class. In [24], each test sample is first classified as unseen classes or the existing classes by using a novelty detection method, and then, the test samples from unseen classes are classified as a specific class. In contrast to zero-shot learning, we aim to reduce the data distribution mismatch between the training and testing samples rather than transferring the knowledge from existing classes to unseen classes.

Our MDA-HS also differs from the existing multi-view domain adaptation methods [25], [26], in which multiple types of features are required for all the samples in the source and target domains. Besides, these works only focused on the single-source domain adaptation setting without learning the optimal weights of different source domains. How to learn the optimal weights is the key challenging issue in this paper.

Moreover, our work is related to the LUPI [9], in which training data are associated with additional features (i.e., privileged information) that are not available for test data. Some recent works utilized privileged information for different learning scenarios, such as learning to rank [27] and distance metric learning [28]. However, these works assume that the training data and test data are from the same data distribution. In [29], a new method was proposed to simultaneously take advantage of privileged information, handle label noise, and reduce domain distribution mismatch. However, this paper is not specifically designed for our new multi-domain adaptation setting, as shown in Fig. 2.

Recently, some works on action and event recognition [30], [31] have achieved the state-of-the-art results on several benchmark data sets. In [30], a rank SVM is trained for each video to learn a feature vector by exploiting temporal information. In [31], a set of decision values are first obtained by using prelearned classifiers from different classes on the subvolumes in one video, which are further aggregated as the input feature for this video. It is worth mentioning that the two approaches focus on learning feature representations from videos, and their features can be readily combined with our classification approaches in order to achieve better results. In contrast with these methods [30], [31], our methods are inherently not limited by any predefined lexicon, because we can readily collect a training data set with a large amount of freely available Web images/videos for any action/event class without additional human annotation efforts.

III. Action and Event Recognition Using Heterogeneous Data Sources

Given abundant loosely labeled Web images and videos, we address the problem of recognizing actions and events in videos. To be exact, we use a 2-D visual descriptor (such as DeCAF features [32]) to represent each Web image, and a 3-D visual descriptor (such as trajectory-based features [2]) for each Web video. Each video in the target domain is represented using both 2-D and 3-D visual features. Since the domains have their own data distributions, and the visual feature spaces for the samples from different source domains are different, we need to address unsupervised domain adaptation problem with heterogeneous sources.

In this paper, we adopt the commonly used terminology. To be exact, we use target domain to denote the testing video domain, and use heterogeneous source domains to denote the Web video and image domains. It is worth mentioning that although multiple views of visual features are available for the target domain, only a single view of visual features is available for data in each source domain. Therefore, the task is to learn discriminative classifiers for classifying the videos in the target domain, by leveraging the unlabeled multi-view visual data from the target domain as well as the loosely labeled single-view visual data from the heterogeneous source domains.

In this paper, we only consider the binary classification problem. Let S denote the number of heterogeneous source domains. For the s-th source domain, there are n_s labeled single-view visual data denoted by $\{\mathbf{x}_i^s|_{i=1}^{n_s}\}$ and the corresponding labels $\{y_i^s|_{i=1}^{n_s}\}$ for each class, $\forall s = 1, \dots, S$.

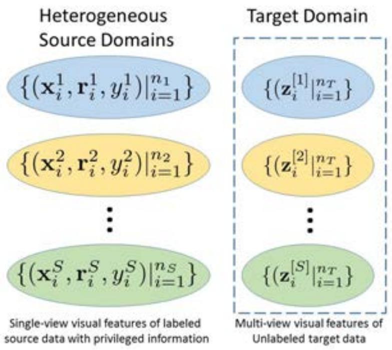

Heterogeneous Source Domains
Target Domain
$\{(\mathbf{x}_i^1, \mathbf{r}_i^1, y_i^1)|_{i=1}^{n_1}\}$
$\{(\mathbf{x}_i^2, \mathbf{r}_i^2, y_i^2)|_{i=1}^{n_2}\}$
$\{(\mathbf{x}_i^S, \mathbf{r}_i^S, y_i^S)|_{i=1}^{n_S}\}$
$\{(\mathbf{z}_i^{[1]}|_{i=1}^{n_T}\}$
$\{(\mathbf{z}_i^{[2]}|_{i=1}^{n_T}\}$
$\{(\mathbf{z}_i^{[S]}|_{i=1}^{n_T}\}$
Single-view visual features of labeled source data with privileged information
Multi-view visual features of Unlabeled target data

Fig. 2. Illustration of our setting (i.e., multi-domain adaptation with heterogeneous sources), which contains single-view visual features from labeled source data with privileged information and multi-view visual features of unlabelled target data.

In particular, each sample $\mathbf{x}_i^s$ comes from a data distribution $\mathcal{P}_s$ that is assumed to be fixed but unknown.

In Section IV, we assume that additional textual descriptions are available for the Web images or Web videos in each source domain. Let us use $\mathbf{r}_i^s$ to denote the textual feature of the i-th sample in the s-th source domain, so that there are labeled training samples $\{(\mathbf{x}_i^s, \mathbf{r}_i^s, y_i^s)|_{i=1}^{n_s}\}$ in the s-th source domain, whereas, in this section, we only discuss the case without considering such additional textual features $\mathbf{r}_i^s$'s.

For the target domain, there are n_T unlabeled multi-view visual samples denoted by $\{\mathbf{z}_i|_{i=1}^{n_T}\}$, in which each target domain sample $\mathbf{z}$ is represented by S visual views (i.e., $\mathbf{z} = (\mathbf{z}^{[1]}, \ldots, \mathbf{z}^{[S]})$) with $\mathbf{z}^{[s]}$ (drawn from $\mathcal{P}_T^{[s]}$) being the same view as $\mathbf{x}^s$. In other words, $\mathbf{z}^{[s]}$ and $\mathbf{x}^s$ share the same type of visual feature. Regarding the heterogeneous sources, we have $\mathcal{P}_i \neq \mathcal{P}_j, \mathcal{P}_T^{[i]} \neq \mathcal{P}_T^{[j]}$ and $\mathcal{P}_i \neq \mathcal{P}_T^{[i]}$ ($\forall i, j = 1, \ldots, S$ and $i \neq j$). We attempt to encourage the weighs for more relevant source domains to be larger and, simultaneously, reduce the data distribution mismatch between each pair of source domain and target domain.

In the rest of this paper, let $\mathbf{0}_n$ (resp., $\mathbf{1}_n$) denote $n \times 1$ vectors of all zeros (resp., ones). The superscript $'$ is used to indicate the transpose of a vector or matrix, while $\odot$ is used to denote the elementwise product between two vectors or matrices. Moreover, $\mathbf{a} \leq \mathbf{b}$ indicates that $a_i \leq b_i, \forall i$. Last but not least, we define a $n \times n$ identity matrix as $\mathbf{I}_n$, and a $n \times m$ matrix of all zeros as $\mathbf{O}_{n \times m}$.

A. Proposed Formulation

Inspired by MKL [33], we propose to learn the target classifier f^T for predicting each target domain sample $\mathbf{z}$, which takes the decision values from multiple views of visual data into consideration

$$f^T(\mathbf{z}) = \sum_{s=1}^{S} d_s \mathbf{w}_s' \phi_s(\mathbf{z}^{[s]}) \tag{1}$$

in which $\mathbf{w}_s$ is the learned weight vector with regards to the s-th visual view of target domain data, ϕ_s is the feature mapping function for the s-th visual view of target domain data (i.e., $\mathbf{z}^{[s]}$), and $d_s \geq 0$ is the weight.

Recall that the target domain data are not labeled. Since the existing domain adaptation approaches [11], [19] are not suitable for our setting with multiple heterogeneous source domains, they may not perform well under this setting. In the recent single-source domain adaptation work [34], the prelearned source classifier $\mathbf{u}'\phi(\mathbf{x})$ is utilized to learn the target classifier by using the regularizer $\|\mathbf{w}_T - \gamma \mathbf{u}\|_2^2$, in which $\mathbf{w}_T$ is the weight vector of the target classifier and γ is a tradeoff parameter to control how much knowledge in the source domain should be transferred to the target domain. Motivated by [34], we utilize a set of prelearned source classifiers $f^s(\mathbf{x}^s) = \mathbf{u}_s'\phi_s(\mathbf{x}^s)$'s obtained by utilizing the training samples from each source domain, and minimize the following newly proposed regularizer for multiple heterogeneous source domains:

$$\Omega_A = \frac{1}{2}\left(\sum_{s=1}^{S} d_s \left(\|\mathbf{w}_s - \gamma_s \mathbf{u}_s\|_2^2 + \theta {\gamma_s}^2\right)\right). \tag{2}$$

Specifically, the above regularizer linearly combines the distances between the weight vectors from the target classifiers and the weight vectors from the prelearned source classifiers from all views. Note that in the above regularizer, the same d_s in (1) is also used as the weight and the reason can be explained as follows. We conjecture that d_s should be larger when the data distribution of the s-th source domain is closer to that of the target domain based on the same view of visual feature. In this situation, the classifier from the s-th visual view is expected to contribute more to the target classifier in (1).

Note that either L_1 or L_2 norm [33] is usually employed to constrain $\mathbf{d} = [d_1, \ldots, d_S]'$. In this paper, we make the assumption that $\|\mathbf{d}\|_2^2 = 1$. In order to infer the labels y_i^T's for the unlabeled target domain data and, simultaneously, learn the target classifier in (1), we propose the following ρ-SVM-based objective function by using our regularizer in (2) as well as our target classifier in (1):

$$\min_{\mathbf{d}\in\mathcal{D},\mathbf{y}_T} \min_{\substack{\mathbf{w}_s,\gamma_s,\\ \rho,\xi_i^s,\xi_i^T}} \Omega_A - \rho + \frac{1}{2}\left(C_S \sum_{s=1}^{S}\sum_{i=1}^{n_s}(\xi_i^s)^2 + C_T \sum_{i=1}^{n_T}(\xi_i^T)^2\right) \tag{3}$$

$$\text{s.t. } y_i^T \in \{\pm 1\}, \quad y_i^T \sum_{s=1}^{S} d_s \mathbf{w}_s' \phi_s\left(\mathbf{z}_i^{[s]}\right) \geq \rho - \xi_i^T \quad \forall i \tag{4}$$

$$y_i^s d_s \mathbf{w}_s' \phi_s\left(\mathbf{x}_i^s\right) \geq \rho - \xi_i^s, \quad \forall s, \quad \forall i, \tag{5}$$

where $\theta, C_S, C_T > 0$ are the regularization parameters, $\mathcal{D} = \{\mathbf{d} | \|\mathbf{d}\|_2^2 = 1, \mathbf{d} \geq \mathbf{0}\}$ is the feasible set of $\mathbf{d}$, $\mathbf{y}_T = [y_1^T, \ldots, y_{n_T}^T]'$ is the label vector of target training samples, and ξ_i^T's (resp., ξ_i^s's) are the slack variables of training samples in the target domain (resp., the s-th source domain). Note that the target model based on the s-th view of visual features is enforced to achieve satisfactory performance on the corresponding labeled samples from the s-th source domain. Such supervision is assumed to be very crucial for the multi-domain adaptation problem, due to the following reasons.

1) The s-th source domain and the target domain have certain overlap when using the s-th view of visual features, and thus, a good model trained by using the

labeled source data are expected to also perform well on the target domain.

2) Since no labeled target domain data are available, the performance of our model will be degraded significantly if the constraints in (5) are removed (see Section V). We would also like to highlight that we need to solve a nontrivial optimization problem, which is a mixed-integer programming problem.

B. Dual Perspective

For ease of discussing the optimization problem in (3), we first make the following definitions. We define $\Phi_s = [\phi_s(\mathbf{x}_1^s), \ldots, \phi_s(\mathbf{x}_{n_s}^s)]$ (resp., $\Phi_T^{[s]} = [\phi_s(\mathbf{z}_1^{[s]}), \ldots, \phi_s(\mathbf{z}_{n_T}^{[s]})]$) as the data matrix after the nonlinear mapping in the s-th source domain (resp., the target domain in the s-th view), respectively. Moreover, $\mathbf{f}_s = [f^s(\mathbf{x}_1^s), \ldots, f^s(\mathbf{x}_{n_s}^s)]'$ and $\mathbf{f}_T^{[s]} = [f^s(\mathbf{z}_1^{[s]}), \ldots, f^s(\mathbf{z}_{n_T}^{[s]})]'$ are used to denote the decision values obtained from $f^s(\mathbf{x})$, $s = 1, \ldots, S$. In addition, h_s denotes the dimension of $\phi_s(\mathbf{x}^s)$, and $N(p, q) = \sum_{s=p}^{q} n_s$ denotes the number of training samples in the range from the p-th source domain to the q-th source domain ($q \geq p$). Based on Φ_s and $\Phi_T^{[s]}$, we further define $\Phi^{[s]}$ as follows with the columns for the samples from the target domain and the s-th source domain set as their corresponding values and the remaining columns set as zeros:

$$\Phi^{[s]} = \big[\mathbf{O}_{h_s \times N(1,s-1)}, \Phi_s, \mathbf{O}_{h_s \times N(s+1,S)}, \Phi_T^{[s]}\big]. \tag{6}$$

Based on $\mathbf{f}_s$ and $\mathbf{f}_T^{[s]}$, $\mathbf{f}^{[s]}$ can be similarly defined as

$$\mathbf{f}^{[s]} = \big[\mathbf{0}'_{N(1,s-1)}, \mathbf{f}'_s, \mathbf{0}'_{N(s+1,S)}, \mathbf{f}_T^{[s]\prime}\big]'. \tag{7}$$

In particular, when $s = 1$ (resp., $s = S$), $\mathbf{O}_{h_s \times N(1,s-1)}$ and $\mathbf{0}_{N(1,s-1)}$ (resp., $\mathbf{O}_{h_s \times N(s+1,S)}$ and $\mathbf{0}_{N(s+1,S)}$) in (6) and (7) will be degenerated as an empty matrix or vector.

To solve (3), we first derive the dual form of the inner optimization problem with respect to the primal variables $\rho, \mathbf{w}_s, \gamma_s, \xi_i^s$ and ξ_i^T, where $s = 1, \ldots, S$. Specifically, by introducing Lagrange multipliers α_i^T's (resp., α_i^s's) corresponding to the constraints in (4) [resp., (5)], the Lagrangian form of inner optimization problem in (3) can be written as follows:

$$\begin{aligned}\mathcal{L} = \frac{1}{2}\Bigg(&\sum_{s=1}^{S} d_s\big(\|\mathbf{w}_s - \gamma_s \mathbf{u}_s\|^2 + \theta\gamma_s^2\big) + C_S \sum_{s=1}^{S}\sum_{i=1}^{n_s} \xi_i^{s2} \\ &+ C_T \sum_{i=1}^{n_T} \xi_i^{T2}\Bigg) - \rho + \rho\boldsymbol{\alpha}'\mathbf{1}_n - \sum_{s=1}^{S}\sum_{i=1}^{n_s} \alpha_i^s \xi_i^s \\ &- \sum_{i=1}^{n_T} \alpha_i^T \xi_i^T - \sum_{s=1}^{S} d_s \mathbf{w}'_s \Phi^{[s]}(\boldsymbol{\alpha} \odot \mathbf{y})\end{aligned}$$

in which $\boldsymbol{\alpha} = [\boldsymbol{\alpha}^{1\prime}, \ldots, \boldsymbol{\alpha}^{S\prime}, \boldsymbol{\alpha}^{T\prime}]'$ is a vector containing dual variables with $\boldsymbol{\alpha}^s = [\alpha_1^s, \ldots, \alpha_{n_s}^s]'$ and $\boldsymbol{\alpha}^T = [\alpha_1^T, \ldots, \alpha_{n_T}^T]'$, and $\mathbf{y}$ is the label vector of all training samples. The feasible set of $\mathbf{y}$ is denoted by $\mathcal{Y} = \{\mathbf{y} | \mathbf{y} = [\mathbf{y}'_1, \ldots, \mathbf{y}'_S, \mathbf{y}'_T]', \mathbf{y}_T \in \{-1, 1\}^{n_T}\}$. Note that $\mathbf{y}_s = [y_1^s, \ldots, y_{n_s}^s]'$ is the label vector for the s-th source data.

By setting the derivatives of $\mathcal{L}$ with respect to the primal variables ρ, $\mathbf{w}_s$'s, γ_s's, ξ_i^s's, and ξ_i^T's to zeros, we have

$$\partial_{\xi_i^s}\mathcal{L} = C\xi_i^s - \alpha_i^s = 0 \tag{8}$$

$$\partial_{\xi_i^T}\mathcal{L} = C_T\xi_i^T - \alpha_i^T = 0 \tag{9}$$

$$\partial_{\rho}\mathcal{L} = -1 + \boldsymbol{\alpha}'\mathbf{1}_n = 0 \tag{10}$$

$$\partial_{\mathbf{w}_s}\mathcal{L} = d_s(\mathbf{w}_s - \gamma_s\mathbf{u}_s) - d_s\Phi^{[s]}(\boldsymbol{\alpha} \odot \mathbf{y}) = 0 \tag{11}$$

$$\partial_{\gamma_s}\mathcal{L} = d_s\mathbf{u}'_s\mathbf{u}_s\gamma_s - d_s\mathbf{u}'_s\mathbf{w}_s + \theta d_s\gamma_s = 0. \tag{12}$$

The equality in (12) can be rewritten as

$$\gamma_s = (\mathbf{u}'_s\mathbf{u}_s + \theta)^{-1}\mathbf{u}'_s\mathbf{w}_s. \tag{13}$$

By substituting (13) into (11), we have the following equality based on the Woodbury formula[1]:

$$\mathbf{w}_s = \left(\mathbf{I} + \frac{1}{\theta}\mathbf{u}_s\mathbf{u}'_s\right)\Phi^{[s]}(\boldsymbol{\alpha} \odot \mathbf{y}). \tag{14}$$

Furthermore, γ_s can be simplified as follows by substituting (14) into (13):

$$\gamma_s = \frac{1}{\theta}\mathbf{u}'_s\Phi^{[s]}(\boldsymbol{\alpha} \odot \mathbf{y}). \tag{15}$$

With (8)–(10), (14), and (15), the Lagrangian $\mathcal{L}$ can rewritten as

$$\mathcal{L} = -\frac{1}{2}\boldsymbol{\alpha}'\left(\sum_{s=1}^{S} d_s\tilde{\mathbf{K}}^{[s]} \odot \mathbf{y}\mathbf{y}' + \tilde{\mathbf{I}}\right)\boldsymbol{\alpha} \tag{16}$$

in which

$$\tilde{\mathbf{K}}^{[s]} = \Phi^{[s]\prime}\Phi^{[s]} + \frac{1}{\theta}\Phi^{[s]\prime}\mathbf{u}_s\mathbf{u}'_s\Phi^{[s]} = \mathbf{K}^{[s]} + \frac{1}{\theta}\mathbf{f}^{[s]}\mathbf{f}^{[s]\prime}$$

$$\tilde{\mathbf{I}} = \operatorname{diag}\{[\mathbf{1}'_{n_1}/C_S, \ldots, \mathbf{1}'_{n_S}/C_S, \mathbf{1}'_{n_T}/C_T]'\}.$$

Based on (8)–(10), we can obtain the feasible set of $\boldsymbol{\alpha}$ as $\mathcal{A} = \{\boldsymbol{\alpha} | \boldsymbol{\alpha}'\mathbf{1}_n = 1, \boldsymbol{\alpha} \geq \mathbf{0}_n\}$. Then, with the inner problem replaced with its dual form, the optimization problem in (3) can be reformulated as follows:

$$\min_{\mathbf{d}\in\mathcal{D}, \mathbf{y}\in\mathcal{Y}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} -\frac{1}{2}\boldsymbol{\alpha}'\left(\sum_{s=1}^{S} d_s\tilde{\mathbf{K}}^{[s]} \odot \mathbf{y}\mathbf{y}' + \tilde{\mathbf{I}}\right)\boldsymbol{\alpha}. \tag{17}$$

Convex Relaxation: Since the problem in (17) is NP-hard, we relax it to the following group-based MKL problem, which is a convex optimization problem:

$$\begin{aligned}&\min_{\mathbf{D}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} -\frac{1}{2}\boldsymbol{\alpha}'\left(\sum_{s=1}^{S}\sum_{o:\mathbf{y}^o\in\mathcal{Y}} d_{so}\mathbf{G}^{so} + \tilde{\mathbf{I}}\right)\boldsymbol{\alpha} \\ &\text{s.t. } \|\mathbf{D}\|_{2,1} = 1, \quad d_{so} \geq 0, \quad \forall s, \quad \forall o,\end{aligned} \tag{18}$$

in which $\mathbf{G}^{so}$ is a base label-kernel defined as $\mathbf{G}^{so} = \tilde{\mathbf{K}}^{[s]} \odot (\mathbf{y}^o\mathbf{y}^{o\prime})$ with $\mathbf{y}^o$ denoting the o-th feasible labeling candidate for $\mathbf{y}$, $\mathbf{D} = [d_{so}] \in \mathbb{R}^{S\times|\mathcal{Y}|}$ is the kernel coefficient matrix (note $|\mathcal{Y}|$ is the cardinality of $\mathcal{Y}$), and $\|\mathbf{D}\|_{2,1} = \sum_{o=1}^{|\mathcal{Y}|}(\sum_{s=1}^{S} d_{so}^2)^{1/2}$ is the mixed $L_{2,1}$ norm.

Theoretically, we have the following proposition regarding the relaxation.

[1] $(\mathbf{A} - \mathbf{U}\mathbf{C}^{-1}\mathbf{V})^{-1} = \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{U}(\mathbf{C} - \mathbf{V}\mathbf{A}^{-1}\mathbf{U})^{-1}\mathbf{V}\mathbf{A}^{-1}$.

Proposition 1: The objective value of the optimization problem in (17) is lower bounded by the optimal value of the problem in (18).

Proof: Based on the theoretical results in [35], the objective value of (17) is lower bounded by the optimal value of the following optimization problem:

$$\min_{\mathbf{d},\boldsymbol{\mu}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} -\frac{1}{2}\boldsymbol{\alpha}' \left(\sum_{s=1}^{S} \sum_{o:\mathbf{y}^o\in\mathcal{Y}} d_s \mu_o \tilde{\mathbf{K}}^{[s]} \odot \mathbf{y}^o \mathbf{y}^{o\prime} + \tilde{\mathbf{I}} \right) \boldsymbol{\alpha} \quad (19)$$

$$\text{s.t. } \|\mathbf{d}\|_2^2 = 1,\ \ \mathbf{d} \geq \mathbf{0}, \quad \|\boldsymbol{\mu}\|_1 = 1, \quad \boldsymbol{\mu} \geq \mathbf{0} \quad (20)$$

where $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_{|\mathcal{Y}|}]'$. Intuitively, rather than directly solving the mixed-integer problem in (18), we optimize the linear combination of $\mathbf{y}^o\mathbf{y}^{o\prime}$'s in (19) (see [35] for the detailed proof).

By setting $d_{\text{so}} = d_s \mu_o$, we have $\|\mathbf{D}\|_{2,1} = 1$. Then, we show that the objective value of (19) is no less than the optimal objective value of (18). In order to verify this, we denote $\mathbf{d}^*$, $\boldsymbol{\mu}^*$, and $\boldsymbol{\alpha}^*$ as the optimal solution to (19) and $\text{obj2}(\mathbf{d}^*, \boldsymbol{\mu}^*; \boldsymbol{\alpha}^*)$ as the optimal objective value of (19). Therefore, we have $\|\mathbf{d}^*\|_2^2 = 1$, $\|\boldsymbol{\mu}^*\|_1 = 1$ and $\boldsymbol{\alpha}^* \in \mathcal{A}$. Then, we define $\tilde{\mathbf{D}} = [\tilde{D}_{\text{so}}] \in \mathcal{R}^{S\times|\mathcal{Y}|}$ with $\tilde{D}_{\text{so}} = d_s^* \mu_o^*$, so that $\|\tilde{\mathbf{D}}\|_{2,1} = \sum_{o=1}^{|\mathcal{Y}|}(\sum_{s=1}^{S} \tilde{D}_{\text{so}}^2)^{1/2} = \sum_{o=1}^{|\mathcal{Y}|}(\sum_{s=1}^{S} (d_s^*\mu_o^*)^2)^{1/2} = \sum_{o=1}^{|\mathcal{Y}|} \mu_o^*(\sum_{s=1}^{S} d_s^{*2})^{1/2} = \sum_{o=1}^{|\mathcal{Y}|} \mu_o^* = 1$. Therefore, $\tilde{\mathbf{D}}$ also falls into the feasible set of (18). By denoting the objective value of (18) when $\mathbf{D} = \tilde{\mathbf{D}}$ and $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$ as $\text{obj1}(\tilde{\mathbf{D}}; \boldsymbol{\alpha}^*)$, we arrive at $\text{obj1}(\tilde{\mathbf{D}}; \boldsymbol{\alpha}^*) = \text{obj2}(\mathbf{d}^*, \boldsymbol{\mu}^*; \boldsymbol{\alpha}^*)$. Furthermore, let $\mathbf{D}^*$ denote the optimal solution to (18). Considering $(\tilde{\mathbf{D}};\ \boldsymbol{\alpha}^*)$ may not be the optimal solution to (18), we have $\text{obj1}(\mathbf{D}^*; \boldsymbol{\alpha}^*) \leq obj1(\tilde{\mathbf{D}}; \boldsymbol{\alpha}^*) = \text{obj2}(\mathbf{d}^*, \boldsymbol{\mu}^*; \boldsymbol{\alpha}^*)$. As a result, the objective value of (19) is no less than the optimal objective value of (18). ■

C. Detailed Algorithm

It is worth mentioning that, the size of $\mathcal{Y}$ increases exponentially with the number of unlabeled target data, which makes that the optimization of the problem in (18) computationally expensive if abundant unlabeled target domain data are provided. Fortunately, one may adopt the cutting-plane [35] method by iteratively choosing a small number of most violated labeling candidates (i.e., $\mathbf{y}^o$'s) to obtain an approximated but reasonably good solution. We present the detailed algorithm in Algorithm 1. In particular, for Step 3 in Algorithm 1, we provide the optimization details in Algorithm 2.

Algorithm 1 A Cutting-Plane Algorithm for Solving (18)

```
1: Initialize $\mathbf{y}^1$ by using the outputs from the source classifiers
   and set $o = 1$, $\mathcal{Y}^o = \{\mathbf{y}^1\}$
2: repeat
3:     Update $\boldsymbol{\alpha}$ and $\mathbf{D}$ in the optimization problem (18) with
       $\mathcal{Y} = \mathcal{Y}^o$ by using Algorithm 2
4:     Obtain the most violated labeling candidate $y^{o+1}$ by
       solving the problem in (21)
5:     Set $\mathcal{Y}^{o+1} = \mathcal{Y}^o \cup \{\mathbf{y}^{o+1}\}$
6:     $o \leftarrow o + 1$
7: until The objective of (18) converges
```

Finding the Most Violated $\mathbf{y}^o$*:* At each iteration of Algorithm 1, after obtaining $\mathbf{D}$ and $\boldsymbol{\alpha}$ in Step 3, we fix them and solve the following problem with respect to each s to find the most violated $\mathbf{y}^o$:

$$\max_{\mathbf{y}^o\in\mathcal{Y}} \boldsymbol{\alpha}'(\tilde{\mathbf{K}}^{[s]} \odot \mathbf{y}^o\mathbf{y}^{o\prime})\boldsymbol{\alpha} = \max_{\mathbf{y}^o\in\mathcal{Y}} \|\mathbf{U}^{[s]\prime}\boldsymbol{\alpha} \odot \mathbf{y}^o\|_2^2 \quad (21)$$

in which $\tilde{\mathbf{K}}^{[s]} = \mathbf{U}^{[s]}\mathbf{U}^{[s]\prime}$ is the eigenvalue decomposition of $\tilde{\mathbf{K}}^{[s]}$. Note the problem in (21) is an integer programming problem. Inspired by [35] and [36], we propose an efficient algorithm to solve it by using the L_∞ norm to relax the L_2 norm

$$\max_{\mathbf{y}^o\in\mathcal{Y}} \|\mathbf{U}^{[s]\prime}\boldsymbol{\alpha} \odot \mathbf{y}^o\|_\infty = \max_{j=1,\ldots,n} \left(\max_{\mathbf{y}^o\in\mathcal{Y}} \left| \sum_{i=1}^{n} \alpha_i y_i^o U_{ij}^{[s]} \right| \right) \quad (22)$$

in which $U_{ij}^{[s]}$ denotes the entry in the ith row and jth column of $\mathbf{U}^{[s]}$. The corresponding solution can be efficiently obtained by simply sorting the coefficients $\alpha_i U_{ij}^{[s]}$'s for each j. Note that, since the source label vectors $\mathbf{y}_s$'s are available, we just need to infer the labels of unlabelled target data, namely, $\mathbf{y}_T \in \{-1, 1\}^{n_T}$.

Solving $\boldsymbol{\alpha}$ *and* $\mathbf{D}$*:* After obtaining $\mathbf{y}^o$, we fix $\mathcal{Y} = \mathcal{Y}^o$ and solve the group-based MKL problem in (18) by alternatingly updating $\boldsymbol{\alpha}$ and $\mathbf{D}$. To be exact, with $\mathbf{D}$ fixed, the optimization problem in (18) becomes a standard SVM problem, so that $\boldsymbol{\alpha}$ can be updated by using off-the-shelf optimization tools, e.g., LIBSVM [37]. Then, with $\boldsymbol{\alpha}$ fixed, after reformulating (18) in its primal form and removing the terms irrelevant to $\mathbf{D}$, we can update $\mathbf{D}$ by addressing the following problem:

$$\min_{\mathbf{D}\in\mathcal{M}} \frac{1}{2} \sum_{s=1}^{S} \sum_{o=1}^{|\mathcal{Y}|} \frac{\|\mathbf{v}_{\text{so}}\|_2^2}{d_{\text{so}}} \quad (23)$$

where $\mathcal{M} = \{\mathbf{D} \big| \|\mathbf{D}\|_{2,1} = 1, d_{\text{so}} \geq 0 \forall s, \forall o\}$ is the feasible set of $\mathbf{D}$, $\|\mathbf{v}_{\text{so}}\|_2 = d_{\text{so}}(\boldsymbol{\alpha}'\mathbf{G}^{\text{so}}\boldsymbol{\alpha})^{1/2}, \forall s, \forall o$. Fortunately, the problem in (23) can be solved in closed form as follows:

$$d_{\text{so}} = \frac{\|\mathbf{v}_{\text{so}}\|_2^{2/3} \left(\sum_{l=1}^{S} \|\mathbf{v}_{\text{lo}}\|_2^{4/3} \right)^{1/4}}{\sum_{o=1}^{|\mathcal{Y}|} \left(\sum_{l=1}^{S} \|\mathbf{v}_{\text{lo}}\|_2^{4/3} \right)^{3/4}}. \quad (24)$$

The derivation details can be found in [38].

Target Classifier: With the optimal $\boldsymbol{\alpha}, \mathbf{D}$, and $\mathbf{y}^o$'s, we can rewrite the target classifier in (1) as

$$f^T(\mathbf{z}) = \sum_{s=1}^{S} \boldsymbol{\beta}_s' \left(\Phi^{[s]\prime} \phi_s(\mathbf{z}^{[s]}) + \frac{1}{\theta} \mathbf{f}^{[s]} f^s(\mathbf{z}^{[s]}) \right)$$

in which $\boldsymbol{\beta}_s = \boldsymbol{\alpha} \odot (\sum_{o=1}^{|\mathcal{Y}|} d_{\text{so}} \mathbf{y}^o)$.

IV. MDA-HS Using Privileged Information

Note that textual descriptions are generally available for Web videos and images, and such textual information is normally more discriminative than visual information extracted from videos and images. In this paper, we, therefore, propose

Algorithm 2 Group-Based MKL for Solving (18)

```
Initialize $\mathbf{D}^1$ with each entry as the same value such that $\|\mathbf{D}^1\|_{2,1} = 1$ and set $\tau = 1$
repeat
    With fixed $\mathcal{Y}$, update $\boldsymbol{\alpha}$ by solving the standard SVM problem with $\mathbf{D}^\tau$ in (18)
    Update $\mathbf{D}^{\tau+1}$ according to (24)
    $\tau \leftarrow \tau + 1$
until The objective of (18) converges with fixed $\mathcal{Y}$
```

to make use of the textual features of Web data for obtaining a more robust classifier for our action/event recognition task. However, it is worth noting that raw videos in the target domain are not associated with such textual descriptions. In this section, we develop a new method to deal with a new learning scenario, in which textual features are only available for source data, but not for target data.

Specifically, we have a set of labeled training samples $\{(\mathbf{x}_i^s, \mathbf{r}_i^s, y_i^s)|_{i=1}^{n_s}\}$ from the s-th source domain, where $\mathbf{x}_i^s$ represents the visual feature of the i-th sample, $\mathbf{r}_i^s$ denotes the additional textual feature (referred to as privileged information in [9]), $y_i^s \in \{-1, 1\}$ is the label of $\mathbf{x}_i^s$, and $s = 1, \ldots, S$. Moreover, as the textual features are not available for the target samples, we still use the same set of unlabeled multiple visual view samples $\{\mathbf{z}_i|_{i=1}^{n_T}\}$ as in MDA-HS (see Fig. 2 for the feature correspondences). Specifically, each sample $\mathbf{z} = (\mathbf{z}^{[1]}, \ldots, \mathbf{z}^{[S]})$ has S visual views, and the s-th visual view $\mathbf{z}^{[s]}$ is the same type of feature with the same dimension as $\mathbf{x}^s$. The goal of our work is to utilize additional privileged information (i.e., the textual descriptions associated with Web videos and images) to help learn more robust target classifiers. Due to the utilization of privileged information, we name our method as MDA-HS+. MDA-HS+ can be further extended as MDA-HS+(ENR) by using a novel elastic-net-like regularization.

A. Formulation of MDA-HS+

Let us define $\Psi_s = [\psi_s(\mathbf{r}_1^s), \ldots, \psi_s(\mathbf{r}_{n_s}^s)]$ as the data matrix after using a nonlinear mapping function on the textual features (i.e., privileged information) of all the labeled training samples in the s-th domain. Motivated by the recent LUPI paradigm [9], we model the slack function in the source domain as a function of privileged information, i.e., $\mathbf{p}_s'\psi_s(\mathbf{r}_i^s)$. Therefore, in order to learn the target classifier in (1) and, meanwhile, infer the labels y_i^T for the target domain samples, we introduce our optimization problem as follows:

$$\begin{aligned}
\min_{\mathbf{d}\in\mathcal{D},\mathbf{y}_T} \min_{\substack{\mathbf{w}_s,\mathbf{p}_s,\gamma_s,\\ \rho,\xi_i^T}} \; & \Omega_A - \rho + \frac{\boldsymbol{\lambda}}{2}\sum_{s=1}^{S}\|\mathbf{p}_s\|^2 \\
& + \frac{1}{2}\left(C_S\sum_{s=1}^{S}\sum_{i=1}^{n_s}\left(\mathbf{p}_s'\psi_s(\mathbf{r}_i^s)\right)^2 + C_T\sum_{i=1}^{n_T}\xi_i^{T2}\right) \qquad (25)\\
\text{s.t. } & y_i^T \in \{\pm 1\}, \quad y_i^T\sum_{s=1}^{S} d_s\mathbf{w}_s'\phi_s\left(\mathbf{z}_i^{[s]}\right) \geq \rho - \xi_i^T, \quad \forall i,\\
& y_i^s d_s\mathbf{w}_s'\phi_s\left(\mathbf{x}_i^s\right) \geq \rho - \mathbf{p}_s'\psi_s(\mathbf{r}_i^s), \quad \forall s, \quad \forall i,
\end{aligned}$$

where $\theta, \boldsymbol{\lambda}, C_S,$ and $C_T > 0$ are the tradeoff parameters, $\mathcal{D} = \{\mathbf{d}|\|\mathbf{d}\|_2^2 = 1, \mathbf{d} \geq \mathbf{0}\}$ is the feasible set of $\mathbf{d}$, $\mathbf{y}_T = [y_1^T, \ldots, y_{n_T}^T]'$ is the label vector of the target samples, and ξ_i^T's are the slack variables of training samples in the target domain.

B. Dual Form of MDA-HS+

We can similarly derive its dual problem, as described in Section III-B. The dual problem of (25) can be solved approximately by using the following proposition.

Proposition 2: The objective value of (25) is lower bounded by the optimal value of the following optimization problem:

$$\begin{aligned}
\min_{\mathbf{D}}\max_{\boldsymbol{\alpha}\in\mathcal{A}} & -\frac{1}{2}\boldsymbol{\alpha}'\left(\sum_{s=1}^{S}\sum_{o=1}^{|\mathcal{Y}|} d_{so}\mathbf{G}^{so} + \tilde{\mathbf{Q}}\right)\boldsymbol{\alpha} \qquad (26)\\
\text{s.t. } & \|\mathbf{D}\|_{2,1} = 1, \quad d_{so} \geq 0 \quad \forall s \quad \forall o
\end{aligned}$$

where $\mathbf{G}^{so}$ is a base label-kernel defined as $\mathbf{G}^{so} = \tilde{\mathbf{K}}^{[s]} \odot (\mathbf{y}^o\mathbf{y}^{o\prime})$ with $\mathbf{y}^o$ denoting the o-th feasible labeling candidate for $\mathbf{y}$, $\mathbf{D} = [d_{so}] \in \mathbb{R}^{S\times|\mathcal{Y}|}$ is the kernel combination coefficient matrix, and $\|\mathbf{D}\|_{2,1} = \sum_{o=1}^{|\mathcal{Y}|}(\sum_{s=1}^{S} d_{so}^2)^{1/2}$ is the mixed $L_{2,1}$ norm. Note that $\tilde{\mathbf{Q}}$ in (26) is defined as

$$\tilde{\mathbf{Q}} = \begin{bmatrix} \tilde{\mathbf{Q}}_1 & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & & \\ \mathbf{0} & & \tilde{\mathbf{Q}}_S & \mathbf{0} \\ \mathbf{0} & & \mathbf{0} & \frac{1}{C_T}\mathbf{I} \end{bmatrix} \qquad (27)$$

where $\tilde{\mathbf{Q}}_s = (1/\boldsymbol{\lambda})(\mathbf{Q}_s - \mathbf{Q}_s((\boldsymbol{\lambda}/C_s)\mathbf{I}_{n_s} + \mathbf{Q}_s)^{-1}\mathbf{Q}_s)$ and $\mathbf{Q}_s = \Psi_s'\Psi_s$.

The solution to (26) is similar to the solution to (18) (see Algorithm 1) except that $\tilde{\mathbf{I}}$ is replaced by $\tilde{\mathbf{Q}}$.

C. Margin Regularization for MDA-HS+

In order to investigate the properties of the formulation in (26), we derive its dual form as follows:

$$\begin{aligned}
\max_{\boldsymbol{\alpha}\in\mathcal{A},\eta_{so},\eta} & -\frac{1}{2}\boldsymbol{\alpha}'\tilde{\mathbf{Q}}\boldsymbol{\alpha} - \eta \qquad (28)\\
\text{s.t. } & \frac{1}{2}\boldsymbol{\alpha}'\left(\mathbf{G}^{so}\right)\boldsymbol{\alpha} \leq \eta_{so}, \quad \forall s, \quad \forall o,\\
& \sqrt{\sum_{s=1}^{S}\eta_{so}^2} \leq \eta, \quad \forall o,
\end{aligned}$$

where η_{so} and η are the newly introduced dual variables.

The problem in (28) is a quadratic constraint quadratic programming (QCQP) problem. Note that privileged information is encoded in $\tilde{\mathbf{Q}}$, and $\boldsymbol{\alpha}'\tilde{\mathbf{Q}}\boldsymbol{\alpha}$ is a quadratic regularization term with respect to $\boldsymbol{\alpha}$. Moreover, the group structure has been encoded in the last inequality constraint (i.e., $(\sum_{s=1}^{S}\eta_{so}^2)^{1/2} \leq \eta$). However, as the last constraint should be satisfied inside each group, it is possible that some of η_{so}'s inside one group will become large, leading to a large value of $(1/2)\boldsymbol{\alpha}'(\mathbf{G}^{\text{so}})\boldsymbol{\alpha}$ accordingly. In order to prevent the problem due to one large value of $(1/2)\boldsymbol{\alpha}'(\mathbf{G}^{\text{so}})\boldsymbol{\alpha}$ inside each group, we further enforce each term $(1/2)\boldsymbol{\alpha}'(\mathbf{G}^{\text{so}})\boldsymbol{\alpha}$ should be upper bounded by a global margin ζ similarly as in the standard ℓ_1-MKL [33] method because of its good control for all terms [33]. Specifically, we propose the following optimization problem to improve the regularization of our MDA-HS+:

$$\min_{\boldsymbol{\alpha}\in\mathcal{A},\zeta,\eta,\eta_{\text{so}}} \frac{1}{2}\boldsymbol{\alpha}'\tilde{\mathbf{Q}}\boldsymbol{\alpha} + \eta + \tilde{\boldsymbol{\lambda}}\zeta \tag{29}$$

$$\text{s.t. } \frac{1}{2}\boldsymbol{\alpha}'(\mathbf{G}^{\text{so}})\boldsymbol{\alpha} \le \eta_{\text{so}}, \quad \forall s, \quad \forall o,$$

$$\frac{1}{2}\boldsymbol{\alpha}'(\mathbf{G}^{\text{so}})\boldsymbol{\alpha} \le \zeta, \quad \forall s, \quad \forall o,$$

$$\sqrt{\sum_{s=1}^{S}\eta_{\text{so}}^2} \le \eta, \quad \forall o, \tag{30}$$

where $\tilde{\boldsymbol{\lambda}} > 0$ is a tradeoff parameter. The group structure is enforced in the constraint (30), and privileged information is encoded in $\tilde{\mathbf{Q}}$ in the objective function.

Proposition 3: The minimization problem in (29) can be equivalently written as the following min-max optimization problem:

$$\min_{\mathbf{D},\mu_{\text{so}}}\max_{\boldsymbol{\alpha}\in\mathcal{A}} -\frac{1}{2}\boldsymbol{\alpha}'\left(\sum_{s=1}^{S}\sum_{o=1}^{|\mathcal{Y}|}(d_{\text{so}}+\mu_{\text{so}})\,\mathbf{G}^{\text{so}}+\tilde{\mathbf{Q}}\right)\boldsymbol{\alpha}$$

$$\text{s.t. } \|\mathbf{D}\|_{2,1} = 1, \quad d_{\text{so}} \ge 0, \quad \forall s, \quad \forall o, \tag{31}$$

$$\sum_{s=1}^{S}\sum_{o=1}^{|\mathcal{Y}|}\mu_{\text{so}} = \tilde{\boldsymbol{\lambda}}, \quad \mu_{\text{so}} \ge 0, \quad \forall s, \quad \forall o,$$

where d_{so} and μ_{so} are the newly introduced dual variables.

In (31), two sets of kernel combination coefficients (i.e., d_{so} and μ_{so}) are introduced for the optimization problem, which is different from the existing standard MKL problems [33] with only one set of kernel combination coefficients. The two sets of kernel combination coefficients have different types of regularization terms. If $\tilde{\boldsymbol{\lambda}} = 0$, the constraints on μ_{so}'s enforce μ_{so}'s to be zeros. Thus, the optimization problem in (26) can be deemed as a special case of the optimization problem in (31) when setting $\tilde{\boldsymbol{\lambda}} = 0$.

D. Solution to MDA-HS+ With Margin Regularization

We propose an optimization algorithm to solve (31). Similarly as in MDA-HS, the size of $\mathcal{Y}$ increases exponentially with the number of unlabeled target training samples. For MDA-HS+ with margin regularization, we also iteratively select a small number of most violated labeling candidates (i.e., $\mathbf{y}^o$'s) by employing the cutting-plane algorithm. Therefore, at each iteration, we infer the labeling candidates and solve the optimization problem (31) in the same manner as in MDA-HS. In order to solve (31), we first introduce the following proposition to convert it into an equivalent problem.

Proposition 4: The optimization problem in (31) is equivalent to its primal form as follows:

$$\min_{\substack{d_{\text{so}},\mu_{\text{so}},\rho\\ \tilde{\mathbf{p}},\mathbf{v}_{\text{so}},\tilde{\mathbf{v}}_{\text{so}}}} \frac{1}{2}\sum_{s=1}^{S}\sum_{o=1}^{|\mathcal{Y}|}\left(\frac{\|\mathbf{v}_{\text{so}}\|_2^2}{d_{\text{so}}}+\frac{\|\tilde{\mathbf{v}}_{\text{so}}\|_2^2}{\mu_{\text{so}}}\right) - \rho$$

$$+\frac{1}{2}\sum_{s=1}^{S}\|\tilde{\mathbf{p}}_s\|^2 + \frac{1}{2}C_T\sum_{i=1}^{n_T}{\xi_i^T}^2 \tag{32}$$

$$\text{s.t. } \sum_{\tilde{s}=1}^{S}\sum_{o=1}^{|\mathcal{Y}|}\left(\mathbf{v}_{\text{so}}'\tilde{\phi}_{\text{so}}\left(\mathbf{x}_i^{\tilde{s}}\right)+\tilde{\mathbf{v}}_{\text{so}}'\tilde{\phi}_{\text{so}}\left(\mathbf{x}_i^{\tilde{s}}\right)\right) \ge \rho - \tilde{\mathbf{p}}_{\tilde{s}}'\tilde{\psi}_{\tilde{s}}(\mathbf{r}_i^{\tilde{s}}) \ \forall\tilde{s}, \ \forall i,$$

$$\sum_{s=1}^{S}\sum_{o=1}^{|\mathcal{Y}|}\left(\mathbf{v}_{\text{so}}'\tilde{\phi}_{\text{so}}(\mathbf{z}_i^{[s]}) + \tilde{\mathbf{v}}_{\text{so}}'\tilde{\phi}_{\text{so}}(\mathbf{z}_i^{[s]})\right) \ge \rho - \xi_i^T \quad \forall i,$$

$$\sum_{o=1}^{|\mathcal{Y}|}\sqrt{\sum_{s=1}^{S}d_{\text{so}}^2} = 1, \ d_{\text{so}} \ge 0 \quad \forall s, \quad \forall o,$$

$$\sum_{o=1}^{|\mathcal{Y}|}\sum_{s=1}^{S}\mu_{\text{so}} = \tilde{\boldsymbol{\lambda}}, \ \mu_{\text{so}} \ge 0 \quad \forall s, \quad \forall o,$$

where $\mathbf{v}_{\text{so}}$, $\tilde{\mathbf{v}}_{\text{so}}$ and $\tilde{\mathbf{p}}$ are the newly introduced primal variables, $\tilde{\phi}_{\text{so}}(\mathbf{x}_i)$ is a mapping function induced from the kernel matrix $\mathbf{G}^{\text{so}} = [\tilde{\phi}_{\text{so}}(\mathbf{x}_i)'\tilde{\phi}_{\text{so}}(\mathbf{x}_j)] \in \mathcal{R}^{n\times n}$, and $\tilde{\psi}_s(\mathbf{r}_i^s)$ is a mapping function induced from the kernel matrix $\tilde{\mathbf{Q}}_s = [\tilde{\psi}_s(\mathbf{r}_i^s)'\tilde{\psi}_s(\mathbf{r}_j^s)] \in \mathcal{R}^{n_s\times n_s}$. Since $((\|\mathbf{v}_{\text{so}}\|_2^2/d_{\text{so}}) + (\|\tilde{\mathbf{v}}_{\text{so}}\|_2^2/\mu_{\text{so}}))$ is an elastic-net-like regularization [33], we name our MDA-HS+ with margin regularization as MDA-HS+(ENR). Note that as discussed in Section IV-C, MDA-HS+ is a special case of MDA-HS+(ENR) when setting $\tilde{\boldsymbol{\lambda}}$ as 0.

Finally, we employ the coordinate descent method to solve (31) as follows.

1) Update $\boldsymbol{\alpha}$*:* With fixed μ_{so} and d_{so}, we have the following optimization problem:

$$\max_{\boldsymbol{\alpha}\in\mathcal{A}} \ -\frac{1}{2}\boldsymbol{\alpha}'\left(\sum_{s=1}^{S}\sum_{o=1}^{|\mathcal{Y}|}(d_{\text{so}}+\mu_{\text{so}})\,\mathbf{G}^{\text{so}}+\tilde{\mathbf{Q}}\right)\boldsymbol{\alpha} \tag{33}$$

which is a standard QP problem and can be solved by any existing QP solver.

2) Update μ_{so} *and* d_{so}*:* After obtaining the optimal $\boldsymbol{\alpha}$, we can then recover $\|\mathbf{v}_{\text{so}}\|_2^2$ (resp., $\|\tilde{\mathbf{v}}_{\text{so}}\|_2^2$) by using (34) [resp., (35)], which can be easily derived from the equations $\mathbf{v}_{\text{so}} = d_{\text{so}}\tilde{\Phi}_{\text{so}}\boldsymbol{\alpha}$ and $\tilde{\mathbf{v}}_{\text{so}} = \mu_{\text{so}}\tilde{\Phi}_{\text{so}}\boldsymbol{\alpha}$ in the proof of Proposition 4

$$\|\mathbf{v}_{\text{so}}\|_2^2 = d_{\text{so}}^2\boldsymbol{\alpha}'\mathbf{G}^{\text{so}}\boldsymbol{\alpha} \tag{34}$$

$$\|\tilde{\mathbf{v}}_{\text{so}}\|_2^2 = \mu_{\text{so}}^2\boldsymbol{\alpha}'\mathbf{G}^{\text{so}}\boldsymbol{\alpha}. \tag{35}$$

Algorithm 3 An Alternate Updating Algorithm for Solving (31)

```
Initialize $d_{so}$ (resp., $\mu_{so}$) by using the same value such that $\sum_{o=1}^{|\mathcal{Y}|}\sqrt{\sum_{s=1}^{S} d_{so}^2} = 1$ (resp., $\sum_{o=1}^{|\mathcal{Y}|}\sum_{s=1}^{S}\mu_{so} = \tilde{\boldsymbol{\lambda}}$)
repeat
    With fixed $\mu_{so}$ and $d_{so}$, update $\boldsymbol{\alpha}$ by solving the quadratic programming problem in (33) with the existing QP solver
    Update $\|\mathbf{v}_{so}\|_2^2$ (resp., $\|\tilde{\mathbf{v}}_{so}\|_2^2$) by using (34) (resp., (35))
    With fixed $\|\mathbf{v}_{so}\|_2^2$ and $\|\tilde{\mathbf{v}}_{so}\|_2^2$, update $d_{so}$ by using the closed-form solution as in (24) and update $\mu_{so} = \tilde{\boldsymbol{\lambda}}\frac{\|\tilde{\mathbf{v}}_{so}\|}{\sum_{o=1}^{|\mathcal{Y}|}\sum_{s=1}^{S}\|\tilde{\mathbf{v}}_{so}\|}$
until The objective of (31) converges with fixed $\mathcal{Y}$
```

With fixed $\|\mathbf{v}_{so}\|_2^2$ and $\|\tilde{\mathbf{v}}_{so}\|_2^2$, the optimization problem becomes

$$\begin{aligned}\min_{d_{so},\mu_{so}} \quad & \frac{1}{2}\sum_{s=1}^{S}\sum_{o=1}^{|\mathcal{Y}|}\left(\frac{\|\mathbf{v}_{so}\|_2^2}{d_{so}} + \frac{\|\tilde{\mathbf{v}}_{so}\|_2^2}{\mu_{so}}\right)\\ \text{s.t.} \quad & \sum_{o=1}^{|\mathcal{Y}|}\sqrt{\sum_{s=1}^{S} d_{so}^2} = 1, \quad d_{so} \geq 0, \quad \forall s, \quad \forall o,\\ & \sum_{o=1}^{|\mathcal{Y}|}\sum_{s=1}^{S}\mu_{so} = \tilde{\boldsymbol{\lambda}}, \quad \mu_{so} \geq 0, \quad \forall s, \quad \forall o.\end{aligned}$$

We update d_{so} by using the closed-form solution as in (24) and update $\mu_{so} = \tilde{\boldsymbol{\lambda}}(\|\tilde{\mathbf{v}}_{so}\|/\sum_{o=1}^{|\mathcal{Y}|}\sum_{s=1}^{S}\|\tilde{\mathbf{v}}_{so}\|)$ according to [33].

We solve (31) by iteratively updating $\boldsymbol{\alpha}$, μ_{so}, and d_{so} until the objective value converges. The algorithm to solve (31) is summarized in Algorithm 3.

Time Complexity Analysis: Our MDA-HS+(ENR) method employs the cutting-plane algorithm, in which we iteratively add the most violated label candidates and solve the subproblem, i.e., the group-based MKL problem with elastic-net-like regularization (see Algorithm 3). Assume that our whole algorithm runs T iterations and the training time of MKL is $O(\text{MKL})$, then the total time complexity of our MDA-HS+(ENR) method is $T \cdot O(\text{MKL})$.

However, for each subproblem, the time complexity of MKL [i.e., $O(\text{MKL})$] has not been theoretically studied. In general, Algorithm 3 converges after several iterations, in which the most time-consuming step is to solve the QP problem (33) by using the existing QP solver. Since the time complexity for solving the QP problem is $O(n^{2.3})$ with n being the number of training samples, the time complexity of MKL can be roughly estimated as $t \cdot O(n^{2.3})$ with t being the number of iterations in MKL. Since our MDA-HS and MDA-HS+ methods also employ the cutting-plane algorithm and solve an MKL subproblem at each iteration, their time complexities can be analyzed in a similar way.

Convergence Analysis: When using the cutting-plane algorithm to solve our optimization problems, the objectives will monotonically decrease as the number of iterations increases. Let us take the objective function of MDA-HS+(ENR) in (31) as an example for the detailed explanation. At each iteration, we add the most violated label candidates and solve an MKL problem by minimizing the objective function in (31) with respect to $\boldsymbol{\alpha}$, d_{so}'s, and μ_{so}'s (see Algorithm 3). In the worst case, the optimal solution of MKL at the current iteration should be at least the same as that at the previous iteration by simply setting the entries of d_{so}'s and μ_{so}'s corresponding to the newly added label kernels as zeros. Therefore, the objective value of (31) decreases monotonically when the number of iterations increases.

V. Experiments

In the experiments, our MDA-HS is compared with SVM, and the existing single-source domain adaptation algorithms geodesic flow kernel (GFK) [11], sampling geodesic flow (SGF) [12], subspace alignment (SA) [16], domain invariant projection (DIP) [15], transfer component analysis (TCA) [39], kernel mean matching (KMM) [40], and domain adaptation SVM (DASVM) [14], as well as the existing multisource domain adaptation approaches domain adaptation machine (DAM) [18], conditional probability based multisource domain adaptation (CPMDA) [19], maximal margin target label learning (MMTLL) [20], and domain selection machine (DSM) [6].

In order to demonstrate the effectiveness of our new regularizer in (2) and the constraint in (5), we report the results of two simplified versions of our method MDA-HS, which are named MDA-HS_sim1 and MDA-HS_sim2, respectively. Specifically, we set the parameter $\theta = \infty$ in MDA-HS_sim1. In this case, we have $\gamma_s = 0$ in (3), and our regularizer in (2) becomes $(1/2)\sum_{s=1}^{S} d_s \|\mathbf{w}_s\|_2^2$, so the prelearned source/auxiliary classifiers will not be used for calculating the kernel [see (17)]. In MDA-HS_sim2, we exclude the constraints in (5), so that the source data will not be employed.

To demonstrate the effectiveness after using privileged information (i.e., the additional textual features), we report the results of our method MDA-HS+(ENR). We additionally compare our MDA-HS+(ENR) with SVM+ [9] and sMIL-PI-DA [29] as well as the existing multi-view learning methods KCCA [41] and SVM-2K [42]. In order to show it is beneficial to utilize the elastic-net-like regularization term, we also report the results of MDA-HS+. Note that MDA-HS+(ENR) reduces to MDA-HS+ when setting the parameter $\tilde{\boldsymbol{\lambda}}$ as 0.

A. Action and Event Recognition Using Heterogeneous Web Sources

1) Data Sets and Features: All methods are evaluated on three benchmark data sets (i.e., Kodak [7], CCV [43], and Hollywood2 [44]). We collect three training data sets as the heterogeneous sources by crawling Web videos from Flickr and Web images from Bing/Google. We do not take any extra efforts to manually annotate the three training data sets, so the labels of training data in the three source domains are noisy. Below we introduce the details of the six data sets.

Kodak Data Set: The Kodak data set consists of 195 videos and their ground-truth annotations for six event classes (i.e., show, sports, wedding, birthday, picnic, and parade). This data set was used in [6] and [7].

CCV Data Set: The CCV data set [43] consists of the videos from 20 semantic categories, including 4659 videos in the training set and 4658 videos in the test set. According to [6], only the videos from the event related categories are used and similar categories are further merged. Finally, we have 2440 videos from five event classes (i.e., show, sports, wedding, birthday, and parade).

Hollywood2 Data Set: The Hollywood2 data set [44] contains 810 videos in the training set and 884 videos in the test set as well as their ground-truth annotations for 12 action classes. In our experiments, we use the test set as our target domain for performance evaluation.

Google/Bing Image Data Set: We use the related keywords as queries (e.g., wedding reception, wedding ceremony, and wedding dance are used for the event class wedding) to collect the top ranked 200 images for each event class. After removing invalid links, we collect 1049 (resp., 870, 2400) Google images for six (resp., 5, 12) event classes in the Kodak (resp., CCV, Hollywood2) data set. Similarly, we also collect 1134 (resp., 945, 2400) Bing images for the Kodak (resp., CCV, Hollywood2) data set.

Flickr Data Set: We use six (resp., 5, 12) event class names from the Kodak (resp., CCV, Hollywood2) data set as queries to crawl Web videos from Flickr. For each query, the top 200 relevant Web videos are downloaded. Similarly as the Google and Bing image data sets, we use 1200 (resp., 1000, 2400) videos as the training set when using Kodak (resp., CCV, Hollywood2) as the test set.

Features: We extract the DeCAF features [32] for each image in the Bing and Google data sets. Following [32], we use the outputs from the sixth layer (i.e., the 4096-dim DeCAF$_6$ features) as the visual features. For each video in the Kodak, CCV, and Hollywood2 data sets, we extract the DeCAF features from video keyframes, which are sampled from each video with one keyframe sampled per two seconds. To compare each image from the Google/Bing data set and each video from the Kodak/CCV/Hollywood2 data set when using the DeCAF features, we first calculate the similarity between each image and each video keyframe using the Gaussian kernel and, then, use the average similarity over all video keyframes of one video to form the kernel matrix for the SVM-based methods.

For each video in the Flickr, Kodak, CCV, and Hollywood2 data sets, improved dense trajectory (IDT) descriptors are also extracted, which include trajectory, histogram of oriented gradient, histogram of optical flow, and motion boundary histogram. The source code provided in [2] is used to extract the IDT descriptors by using 16 for the sampling stride and 50 for the trajectory length as well as setting the remaining parameters as their default values. Following the Fisher vector encoding method in [2], we then train 256 Gaussian mixture models by using the IDT descriptors from the videos in the Flickr training data set and generate the 128 000-dim Fisher vectors as 3-D visual features for each video in the training and test data sets.

2) Experimental Setups: In our experiments, we treat the Bing/Google image data set and the Flickr video data set as $S = 3$ heterogeneous source domains, while the Kodak/CCV/Hollywood2 data set is used as the target domain.

TABLE I

MAPs (%) of Different Methods on the Kodak, CCV, and Hollywood2 Data Sets. We Do Not Consider the Additional Textual Features of Source Domain Data in This Table

Method	Kodak	CCV	Hollywood2
SVM_A [37]	63.41	58.91	52.04
SS-SVM [45]	63.88	59.84	53.19
DAM [18]	65.74	61.90	54.05
DSM [6]	66.20	62.19	54.30
CPMDA [19]	65.48	61.84	54.13
MMTLL [20]	64.98	47.76	34.89
KMM [40]	64.26	61.84	54.18
DASVM [14]	65.68	60.95	53.67
GFK [11]	62.60	61.53	53.58
SGF [12]	66.27	60.48	53.71
SA [16]	64.98	60.37	53.54
DIP [15]	62.76	57.41	48.54
TCA [39]	64.77	61.04	55.83
MDA-HS_sim1	67.02	62.79	56.73
MDA-HS_sim2	33.84	25.70	21.24
MDA-HS	**68.12**	**63.49**	**57.33**

Labeled videos are not available in the target domain during the training process, so SVM is also referred to as SVM_A in this paper. Specifically, we first train S independent SVM classifiers (i.e., f^s's) based on each individual source/auxiliary domain and use each SVM classifier to predict the test data in the target domain based on the same type of visual feature. Finally, we average the prediction scores from all the classifiers to generate the final prediction score of each test sample. The same late-fusion strategy is used for the self-training semi-supervised SVM method [45] as well as the single-source domain adaptation methods, including GFK [11], SGF [12], SA [16], DIP [15], TCA [39], KMM [40], and DASVM [14]. The traditional multiple source domain adaptation methods CPMDA [19], DAM [18], DSM [6], and MMTLL [20] are not specifically designed for our setting illustrated in Fig. 2. For these methods, we train S prelearned classifiers based on the training samples in the source domains, and also calculate a new kernel matrix by averaging the S kernels with each kernel constructed based on one view of visual data in the target domain. Then, the S prelearned source classifiers and the average kernel for target domain data can be used as the input for CPMDA, DAM, DSM, and MMTLL.

In this paper, we train one-versus-rest SVMs by using the Gaussian kernel $k_s(\mathbf{x}_i, \mathbf{x}_j) = \phi_s(\mathbf{x}_i)'\phi_s(\mathbf{x}_j) = \exp(-(1/\nu)^{1/2}\|\mathbf{x}_i - \mathbf{x}_j\|_2^2)$, in which wet set the default bandwidth parameter according to [10]. We empirically fix the parameters $C_S = C_T = 10$, and set the parameter $\theta = 0.1$ for MDA-HS and MDA-HS_sim2. For the baseline methods, we tune their parameters based on the test data and report their best results by using the optimal parameters. As in [6] and [7], we choose average precision (AP) for performance evaluation and report mean AP (MAP) over all the action/event classes for each method.

3) Results: The MAPs of all methods are reported in Table I. Compared with the preliminary conference version of this paper [10], our experimental setting is different in two aspects. First, we use the Flickr video data set to replace the YouTube data set as one of the source domains. Observing the CCV data set is also collected from YouTube, we additionally use the videos from another Web site (i.e., Flickr) in order to have larger domain distribution mismatch between the training and testing videos. Second, we use more discriminant features in this paper. Specifically, we use the DeCAF features to replace the bag-of-word representation based on SIFT features for the Web images and video keyframes. While IDT descriptors are still used for the videos in the Flickr/CCV/Kodak data set, we use the Fisher vectors to replace the BOW features. After using more discriminant visual features, the performances for all methods reported in this paper are improved when compared with those in [10]. Based on the results in Table I, we have the following observations.

1) Most existing single-source domain adaptation methods (i.e., KMM, DASVM, GFK, SGF, SA, and TCA) generally outperform SVM_A by explicitly reducing the domain distribution mismatch between each source domain and the target domain. Moreover, the existing multisource domain adaptation algorithms (i.e., DAM, DSM, CPMDA, and MMTLL) are also generally better than SVM_A, which shows it is effective to adapt the prelearned classifiers from multiple source domains to the target domain. DSM achieves better results than DAM on all data sets, which indicates the benefits of selecting the relevant source domains.
2) MDA-HS_sim2 is worse than MDA-HS and MDA-HS_sim1, which demonstrates that it is important to learn the target classifier by using the labeled samples from the source domains. Besides, MDA-HS also outperforms MDA-HS_sim1, which shows the effectiveness of our new regularizer in (2) by utilizing the prelearned source/auxiliary classifiers.
3) Our method MDA-HS achieves the best results on all three data sets, which clearly demonstrates our MDA-HS method after using the new target classifier and new regularizer in (2) is effective for action and event recognition.

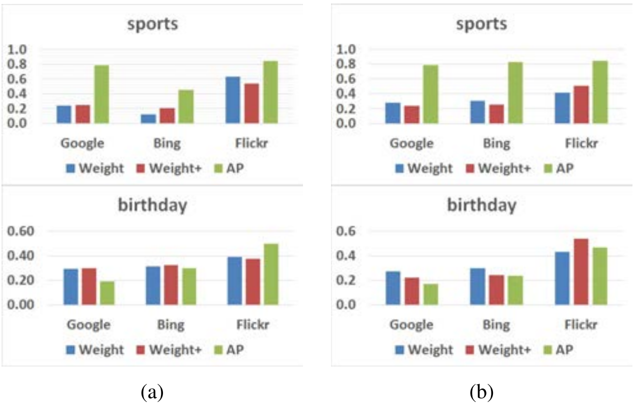
B. Action and Event Recognition Using Heterogeneous Web Sources With Privileged Information

1) Data Sets and Features: In the following experiments, we use the same data sets as in Section V-A and additionally utilize the textual descriptions of Google/Bing images and Flickr videos as privileged information. Specifically, we download the associated textual descriptions for each image from the Google/Bing data set and each video from the Flickr data set. Because the word distributions from three search engines are different, we extract the textual features independently. For each image or video, its textual feature is represented as a term-frequency feature. Eventually, we construct 4545 (resp., 2322, 2000)-dimensional term-frequency features for Google images (resp., Bing images, Flickr videos).

2) Experimental Setups: We evaluate our MDA-HS+ and MDA-HS+(ENR) on the Kodak, CCV, and Hollywood2 data set. We compare our methods with SVM+ [9] in which we employ the late-fusion strategy by fusing S SVM+ classifiers independently trained based on the training data from each source domain. Moreover, we additionally report the results of KCCA [41] and SVM-2K [42]. Note that they can also utilize both visual features and textual features of training samples from each individual source domain. Specifically, we employ KCCA based on the textual features and visual features of training samples and use the common representations of visual features to train the SVM classifier, and then use the projected visual features of test samples in the common subspace for prediction. For SVM-2K, we train the SVM-2K classifiers by using the visual features and textual features of training samples and, then, apply the visual feature-based classifier to predict the test samples. Finally, we use the late-fusion strategy to fuse the prediction scores from the classifiers of three source domains. Moreover, our methods are also compared with sMIL-PI-DA [29], in which the late-fusion strategy is used again to fuse the sMIL-PI-DA classifiers from three source domains. Note that label noise is not considered in this paper, so we use the fully supervised version of sMIL-PI-DA by setting the bag size and positive ratio as 1.

We empirically fix $\boldsymbol{\lambda} = 100$ for MDA-HS+ and MDA-HS+(ENR) and $\tilde{\boldsymbol{\lambda}} = 0.01$ for MDA-HS+(ENR) on all three data sets. The other experimental settings are the same as in Section V-A.

3) Results: Based on the results in Table II, we have the following observations.



(a) (b)

Fig. 3. Illustration of three learned domain weights and per-event APs for three source domains. We report the results for two events, i.e., sports and birthday, on the Kodak and CCV data sets. Specifically, we show the learned domain weights by using our MDA-HS (denoted by Weight) and MDA-HS+(ENR) (denoted by Weight+) as well as the per-event APs of three SVMs with each SVM trained by using the labeled training samples from one source domain. (a) Kodak data set. (b) CCV data set.

TABLE II

MAPs (%) OF DIFFERENT METHODS ON THE KODAK, CCV, AND HOLLYWOOD2 DATA SETS. IN THIS TABLE, THE ADDITIONAL TEXTUAL FEATURES ARE EMPLOYED IN ALL METHODS EXCEPT SVM_A AND MDA-HS

Method	Kodak	CCV	Hollywood2
SVM_A [37]	63.41	58.91	52.04
SVM+ [9]	65.04	61.55	55.23
KCCA [41]	64.91	60.71	54.31
SVM-2K [42]	64.94	59.89	54.95
sMIL-PI-DA [29]	66.71	63.45	57.82
MDA-HS	68.12	63.49	57.33
MDA-HS+	69.69	64.51	59.09
MDA-HS+(ENR)	**70.87**	**66.18**	**60.86**

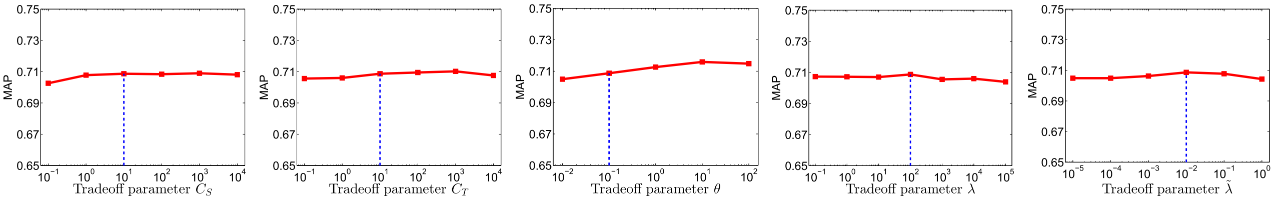


Fig. 4. MAPs of MDA-HS+(ENR) on the Kodak data set when using different tradeoff parameters. Vertical dash lines: default parameters.

TABLE III

TRAINING TIME OF ALL METHODS WITHOUT USING PRIVILEGED INFORMATION ON THE KODAK DATA SET. OUR METHOD IS DENOTED IN BOLDFACE

Methods	SVM_A	SS-SVM	DAM	DSM	CPMDA	MMTLL	KMM	DASVM	GFK	SGF	SA	DIP	TCA	MDA-HS
Time(s)	2.32	6.77	2.50	2.54	2.65	5.54	46.59	74.24	176.45	1105.87	82.26	3109.54	251.96	**351.54**

TABLE IV

TRAINING TIME OF ALL METHODS USING PRIVILEGED INFORMATION ON THE KODAK DATA SET. OUR METHODS ARE DENOTED IN BOLDFACE

Methods	SVM+	KCCA	SVM-2K	sMIL-PI-DA	MDA-HS+	MDA-HS+(ENR)
Time(s)	195.94	268.51	69.15	705.72	**349.65**	**370.79**

1) SVM+, KCCA, and SVM-2K outperform SVM_A on all data sets, which indicates that it is beneficial to utilize both visual features and textual features of training samples. Moreover, SVM+ achieves better results than KCCA and SVM-2K on all data sets by using the additional textual features as privileged information.
2) sMIL-PI-DA outperforms SVM+. A possible explanation is sMIL-PI-DA additionally handles the domain distribution mismatch. sMIL-PI-DA is also better than the existing single-source domain adaptation methods in Table I (i.e., GFK, SGF, SA, DIP, TCA, KMM, and DASVM), which shows that it is beneficial to leverage the additional textual features as privileged information.
3) Our newly proposed methods MDA-HS+ and MDA-HS+(ENR) outperform MDA-HS on all data sets, which shows that the additional textual descriptions of Web images and videos in the source domains are helpful for training a more robust model. MDA-HS+ and MDA-HS+(ENR) also outperform SVM+, KCCA, and SVM-2K on all data sets. The baseline methods SVM+, KCCA, and SVM-2K assume that the training and test samples are with the same data distribution. MDA-HS+ and MDA-HS+(ENR) are better than them, because they can additionally handle the domain distribution mismatch by leveraging the prelearned source/auxiliary classifiers in our new regularizer in (2).
4) Our new method MDA-HS+(ENR) achieves the best results, and it is also better than MDA-HS+ on all data sets, which clearly shows that it is beneficial to utilize the elastic-net-like regularization term in order to train a more robust classifier.

Analysis on the Learned Domain Weights Using Our Methods MDA-HS and MDA-HS+(ENR): We analyze the domain weights of three source domains, which are learned by using our methods MDA-HS and MDA-HS+(ENR). The per-event APs of three SVMs are additionally reported, in which each SVM classifier is learned by using the training samples from one single-source domain (i.e., Flickr, Bing, or Google). If the data distribution of one source domain is closer to that of the target domain when using the same type of visual feature, the per-event AP from the corresponding SVM classifier is also expected to be higher, and we also expect to learn a larger domain weight for this source domain. As the objective of our MDA-HS is relaxed to the one in (18), we, therefore, analyze d_{so} in (18) instead of $\mathbf{d}$ in (17), and we also report the three coefficients of the column in $\mathbf{D}$, which has the largest L_1 norm. Similarly, for our MDA-HS+(ENR), we analyze $(d_{\mathrm{so}} + \mu_{\mathrm{so}})$ in (31) instead of $\mathbf{d}$ in (25). Specifically, we use $\bar{\mathbf{D}} \in \mathbb{R}^{S\times|\mathcal{Y}|}$ to denote the matrix with each entry being $(d_{\mathrm{so}} + \mu_{\mathrm{so}})$, and we also report the three coefficients of the column in $\bar{\mathbf{D}}$, which has the largest L_1 norm. In Fig. 3, we take the events, i.e., sports and birthday, as two examples to show the per-event APs of three SVMs as well as the domain weights for three source domains (i.e., the three learned coefficients) on the Kodak and CCV data sets. Note that we use Weight and Weight+ to indicate the domain weights learned by using our MDA-HS and MDA-HS+(ENR), respectively. Based on these results, we observe that the highest weight can be correctly assigned to the most relevant source domain (i.e., Flickr) by using our methods MDA-HS and MDA-HS+(ENR), in which the corresponding per-event AP is also the best. The results demonstrates that our methods MDA-HS and MDA-HS+(ENR) can effectively combine multiple heterogeneous source domains for domain adaptation. We have similar observations for other event classes on all data sets.

Robustness to the Parameters: Let us take the Kodak data set as an example to study the performance variation of our MDA-HS+(ENR) method by varying one parameter while fixing other parameters as their default values. The results in Fig. 4 show our methods are relatively robust when the tradeoff parameters are set in certain ranges. We have similar observations for all our methods on all data sets.

Comparison of Training Time: We take the Kodak data set as an example to report the training time of our MDA-HS, MDA-HS+, and MDA-HS+(ENR) methods as well as other baseline methods in Tables III and IV. We observe that our methods are reasonably efficient when compared with other baseline methods. Specifically, our method MDA-HS is as efficient as TCA and GFK, and it is faster than SGF and DIP. Our methods MDA-HS+ and MDA-HS+(ENR) are also as efficient as SVM+ and KCCA.

VI. Conclusion

We have proposed new domain adaptation approaches for action and event recognition by leveraging a large number of freely available Web videos and Web images. By formulating this task as a multi-domain adaptation problem with heterogeneous sources, we introduce a new regularizer and a new target classifier based on the optimal weights of different source domains. To seek the optimal weights of different source domains and learn the optimal target classifier, we also propose a new method called MDA-HS, which can additionally infer the labels of unlabeled target data. Moreover, we also develop a new method called MDA-HS+ by utilizing the additional textual descriptions of Web data as privileged information, which is further extended as MDA-HS+(ENR) by using the elastic-net-like regularization. By leveraging a large amount of freely available Web data as the training data, our methods are inherently not limited by any predefined lexicon. Extensive experiments on three benchmark data sets (i.e., Kodak, CCV, and Hollywood2) clearly demonstrate that our newly proposed methods MDA-HS, MDA-HS+, and MDA-HS+(ENR) are effective for action and event recognition. Moreover, our experiments also demonstrate that it is beneficial to utilize additional textual information as privileged information and validate the effectiveness of our elastic-net-like regularization.

In the future, we will study how to explicitly handle label noise of training Web images and videos for learning more robust target classifiers as well as investigate how to automatically decide the optimal parameters for our methods MDA-HS, MDA-HS+, and MDA-HS+(ENR). Moreover, we will also investigate how to combine our approaches with the more discriminant features proposed in [30] and [31] to further improve the recognition performance.

Our proposed approaches can be used for many other real-world applications. For example, our work can be used to recognize multimedia objects collected from multimedia cyclopedia, in which images, audio, and text are jointly used to describe the same semantic concepts [46]. We can collect images, audio clips, and text documents from a set of predefined semantic concepts to construct multiple source domains in order to classify each multimedia object in the target domain. In another application for action recognition, the training videos in each source domain may be captured by the cameras from one viewpoint, while the testing videos in the target domain are captured by the cameras from all viewpoints. How to use our proposed approaches for those interesting applications will also be investigated in the future.

Appendix A
Proof of Proposition 3

Proof: By introducing the Lagrangian multipliers $\bar{d}_{\text{so}} \geq 0$, $\bar{\mu}_{\text{so}} \geq 0$ and $\tilde{d}_o \geq 0$, we write the Lagrangian of (29) as

$$\mathcal{L} = \frac{1}{2}\boldsymbol{\alpha}'\tilde{\mathbf{Q}}\boldsymbol{\alpha} + \eta + \tilde{\boldsymbol{\lambda}}\zeta + \sum_{s=1}^{S}\sum_{o}\bar{d}_{\text{so}}\left(\frac{1}{2}\boldsymbol{\alpha}'(\mathbf{G}^{\text{so}})\boldsymbol{\alpha} - \eta_{\text{so}}\right)$$
$$+\sum_{s=1}^{S}\sum_{o}\bar{\mu}_{\text{so}}\left(\frac{1}{2}\boldsymbol{\alpha}'(\mathbf{G}^{\text{so}})(\mathbf{G}^{\text{so}})\boldsymbol{\alpha} - \zeta\right)$$
$$+\sum_{o=1}^{|\mathcal{Y}|}\tilde{d}_o\left(\sqrt{\sum_{s=1}^{S}\eta_{\text{so}}^2} - \eta\right).$$

By setting the derivatives of $\mathcal{L}$ with respect to the variables ζ, η, η_{so} to be zeros, we obtain

$$\partial_{\zeta}\mathcal{L} = \tilde{\boldsymbol{\lambda}} - \sum_{s=1}^{S}\sum_{o=1}^{|\mathcal{Y}|}\bar{\mu}_{\text{so}} = 0 \tag{36}$$

$$\partial_{\eta}\mathcal{L} = 1 - \sum_{o=1}^{|\mathcal{Y}|}\tilde{d}_o = 0 \tag{37}$$

$$\partial_{\eta_{\text{so}}}\mathcal{L} = -\bar{d}_{\text{so}} + \tilde{d}_o\frac{\eta_{\text{so}}}{\sqrt{\sum_{s=1}^{S}\eta_{\text{so}}^2}} = 0. \tag{38}$$

From (36) we have $\sum_{s=1}^{S}\sum_{o=1}^{|\mathcal{Y}|}\bar{\mu}_{\text{so}} = \tilde{\boldsymbol{\lambda}}$. According to (37), we have $\sum_o \tilde{d}_o = 1$. From (38), we have $\bar{d}_{\text{so}} = \tilde{d}_o(\eta_{\text{so}}/(\sum_{s=1}^{S}\eta_{\text{so}}^2)^{1/2})$, which further gives $(\sum_{s=1}^{S}\bar{d}_{\text{so}}^2)^{1/2} = (\tilde{d}_o/(\sum_{s=1}^{S}\eta_{\text{so}}^2)^{1/2})(\sum_{s=1}^{S}\eta_{\text{so}}^2)^{1/2} = \tilde{d}_o$. Together with $\sum_{o=1}^{|\mathcal{Y}|}\tilde{d}_o = 1$, we therefore arrive at $\sum_{o=1}^{|\mathcal{Y}|}(\sum_{s=1}^{S}\bar{d}_{\text{so}}^2)^{1/2} = 1$. By substituting the obtained conditions back into $\mathcal{L}$, we arrive at the following optimization problem:

$$\max_{\bar{d}_{\text{so}},\bar{\mu}_{\text{so}}}\min_{\boldsymbol{\alpha}\in\mathcal{A}}\frac{1}{2}\boldsymbol{\alpha}'\left(\sum_{s=1}^{S}\sum_{o=1}^{|\mathcal{Y}|}(\bar{d}_{\text{so}} + \bar{\mu}_{\text{so}})\mathbf{G}^{\text{so}} + \tilde{\mathbf{Q}}\right)\boldsymbol{\alpha}$$
$$\text{s.t.}\ \sum_{o=1}^{|\mathcal{Y}|}\sqrt{\sum_{s=1}^{S}\bar{d}_{\text{so}}^2} = 1,\quad \bar{d}_{\text{so}} \geq 0,\quad \forall s,\quad \forall o,$$
$$\sum_{s=1}^{S}\sum_{o=1}^{|\mathcal{Y}|}\bar{\mu}_{\text{so}} = \tilde{\boldsymbol{\lambda}},\quad \bar{\mu}_{\text{so}} \geq 0,\quad \forall s,\quad \forall o, \tag{39}$$

which leads to the optimization problem as in (31) by multiplying -1 to the objective and switching the max and min operations. Note that the newly introduced Lagrangian multipliers $\bar{d}_{\text{so}}$ and $\bar{\mu}_{\text{so}}$ correspond to d_{so} and μ_{so} in (31), respectively. We thus prove the proposition. ■

Appendix B
Proof of Proposition 4

Proof: By introducing the Lagrangian multipliers $\alpha_i^s \geq 0$ and $\alpha_i^T \geq 0$, we obtain the Lagrangian of (32) as

$$\begin{aligned}\mathcal{L} = {} & \frac{1}{2}\sum_{s=1}^{S}\sum_{o=1}^{|\mathcal{Y}|}\left(\frac{\|\mathbf{v}_{\mathrm{so}}\|_2^2}{d_{\mathrm{so}}} + \frac{\|\tilde{\mathbf{v}}_{\mathrm{so}}\|_2^2}{\mu_{\mathrm{so}}}\right) - \rho + \frac{1}{2}\sum_{s=1}^{S}\|\tilde{\mathbf{p}}_s\|^2 \\ & - \sum_{\tilde{s}=1}^{S}\sum_{i=1}^{n_{\tilde{s}}}\alpha_i^{\tilde{s}}\left(\sum_{s=1}^{S}\sum_{o=1}^{|\mathcal{Y}|}\left(\mathbf{v}_{\mathrm{so}}'\tilde{\phi}_{\mathrm{so}}\left(\mathbf{x}_i^{\tilde{s}}\right) + \tilde{\mathbf{v}}_{\mathrm{so}}'\tilde{\phi}_{\mathrm{so}}\left(\mathbf{x}_i^{\tilde{s}}\right)\right) - \rho\right) \\ & - \sum_{\tilde{s}=1}^{S}\sum_{i=1}^{n_{\tilde{s}}}\alpha_i^{\tilde{s}}\tilde{\mathbf{p}}_{\tilde{s}}'\tilde{\psi}_{\tilde{s}}\left(\mathbf{r}_i^{\tilde{s}}\right) - \sum_{i=1}^{n_T}\alpha_i^T\xi_i^T + \frac{1}{2}C_T\sum_{i=1}^{n_T}\xi_i^{T\,2} \\ & - \sum_{i=1}^{n_T}\alpha_i^T\left(\sum_{s=1}^{S}\sum_{o=1}^{|\mathcal{Y}|}\left(\mathbf{v}_{\mathrm{so}}'\tilde{\phi}_{\mathrm{so}}\left(\mathbf{z}_i^{[s]}\right) + \tilde{\mathbf{v}}_{\mathrm{so}}'\tilde{\phi}_{\mathrm{so}}\left(\mathbf{z}_i^{[s]}\right)\right) - \rho\right).\end{aligned}$$

By setting the derivatives of $\mathcal{L}$ with respect to the primal variables $\rho, \mathbf{v}_{\mathrm{so}}, \tilde{\mathbf{v}}_{\mathrm{so}}, \tilde{\mathbf{p}}, \xi_i^T$ to be zeros, respectively, we have $1 = \sum_{s=1}^{S}\sum_{i=1}^{n_s}\alpha_i^s + \sum_{i=1}^{n_T}\alpha_i^T$, $(\mathbf{v}_{\mathrm{so}}/d_{\mathrm{so}}) = \sum_{\tilde{s}=1}^{S}\sum_{i=1}^{n_{\tilde{s}}}\alpha_i^{\tilde{s}}\tilde{\phi}_{\mathrm{so}}(\mathbf{x}_i^{\tilde{s}}) + \sum_{i=1}^{n_T}\alpha_i^T\tilde{\phi}_{\mathrm{so}}(\mathbf{z}_i^{[s]})$, $(\tilde{\mathbf{v}}_{\mathrm{so}}/\mu_{\mathrm{so}}) = \sum_{\tilde{s}=1}^{S}\sum_{i=1}^{n_{\tilde{s}}}\alpha_i^{\tilde{s}}\tilde{\phi}_{\mathrm{so}}(\mathbf{x}_i^{\tilde{s}}) + \sum_{i=1}^{n_T}\alpha_i^T\tilde{\phi}_{\mathrm{so}}(\mathbf{z}_i^{[s]})$, and $\tilde{\mathbf{p}}_s = \sum_{i=1}^{n_s}\alpha_i^s\tilde{\psi}_s(\mathbf{r}_i^s)$, $\xi_i^T = (1/C_T)\alpha_i^T$.

Let us define $\boldsymbol{\alpha}^s = [\alpha_1^s, \ldots, \alpha_{n_s}^s]'$ and $\boldsymbol{\alpha}^T = [\alpha_1^T, \ldots, \alpha_{n_T}^T]'$, and then, we have $\boldsymbol{\alpha} = [\boldsymbol{\alpha}^{1'}, \ldots, \boldsymbol{\alpha}^{S'}, \boldsymbol{\alpha}^{T'}]' \in \mathcal{A}$. By defining $\tilde{\mathbf{\Phi}}_{\mathrm{so}} = [\tilde{\phi}_{\mathrm{so}}(\mathbf{x}_1^1), \ldots, \tilde{\phi}_{\mathrm{so}}(\mathbf{x}_{n_1}^1), \ldots, \tilde{\phi}_{\mathrm{so}}(\mathbf{x}_1^S), \ldots, \tilde{\phi}_{\mathrm{so}}(\mathbf{x}_{n_S}^S), \tilde{\phi}_{\mathrm{so}}(\mathbf{x}_1^T), \ldots, \tilde{\phi}_{\mathrm{so}}(\mathbf{x}_{n_T}^T)]$, and $\tilde{\Psi}_s = [\tilde{\psi}_s(\mathbf{r}_1^s), \ldots, \tilde{\psi}_s(\mathbf{r}_{n_s}^s)]$, we can simplify the above equations as follows, $\boldsymbol{\alpha}'\mathbf{1}_n = 1$, $\mathbf{v}_{\mathrm{so}} = d_{\mathrm{so}}\tilde{\Phi}_{\mathrm{so}}\boldsymbol{\alpha}$, $\tilde{\mathbf{v}}_{\mathrm{so}} = \mu_{\mathrm{so}}\tilde{\Phi}_{\mathrm{so}}\boldsymbol{\alpha}$, and $\tilde{\mathbf{p}}_s = \tilde{\Psi}_s\boldsymbol{\alpha}^s$.

By substituting the above equations back into the Lagrangian, we arrive at

$$\begin{aligned} & \min_{\mathbf{D},\mu_{\mathrm{so}}}\max_{\boldsymbol{\alpha}\in\mathcal{A}} -\frac{1}{2}\boldsymbol{\alpha}'\left(\sum_{s=1}^{S}\sum_{o=1}^{|\mathcal{Y}|}(d_{\mathrm{so}} + \mu_{\mathrm{so}})\,\tilde{\Phi}_{\mathrm{so}}'\tilde{\Phi}_{\mathrm{so}}\right)\boldsymbol{\alpha} \\ & \quad -\frac{1}{2}\sum_{s=1}^{S}\boldsymbol{\alpha}^{s\,\prime}\tilde{\Psi}_s'\tilde{\Psi}_s\boldsymbol{\alpha}^s - \frac{1}{2C_T}\boldsymbol{\alpha}^{T\,\prime}\boldsymbol{\alpha}^T \\ & \text{s.t. } \|\mathbf{D}\|_{2,1} = 1, \quad d_{\mathrm{so}} \geq 0 \ \ \forall s, \ \ \forall o, \\ & \quad \sum_{s=1}^{S}\sum_{o=1}^{|\mathcal{Y}|}\mu_{\mathrm{so}} = \tilde{\boldsymbol{\lambda}}, \quad \mu_{\mathrm{so}} \geq 0 \ \ \forall s, \ \ \forall o. \end{aligned} \tag{40}$$

Based on the definitions of $\tilde{\phi}_{\mathrm{so}}(\mathbf{x}_i)$, $\tilde{\psi}_s(\mathbf{r}_i^s)$, and with some simplifications, we can obtain the min–max problem, as shown in (31). Thus, we prove the proposition. ■

REFERENCES

[1] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos 'in the wild,"' in *Proc. 22nd IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami Beach, FL, USA, Jun. 2009, pp. 1996–2003.

[2] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 3551–3558.

[3] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. 28th Annu. Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2014, pp. 568–576.

[4] Z. Xu, Y. Yang, and A. G. Hauptmann, "A discriminative CNN video representation for event detection," in *Proc. 28th IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 1798–1807.

[5] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah, "High-level event recognition in unconstrained videos," *Int. J. Multimedia Inf. Retr.*, vol. 2, no. 2, pp. 73–101, Jun. 2013.

[6] L. Duan, D. Xu, and S.-F. Chang, "Exploiting Web images for event recognition in consumer videos: A multiple source domain adaptation approach," in *Proc. 25th IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 1338–1345.

[7] L. Duan, I. W.-H. Tsang, D. Xu, and J. Luo, "Visual event recognition in videos by learning from Web data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1667–1680, Sep. 2012.

[8] N. Ikizler-Cinbis and S. Sclaroff, "Web-based classifiers for human action recognition," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1031–1045, Aug. 2012.

[9] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural Netw.*, vol. 22, nos. 5–6, pp. 544–557, 2009.

[10] L. Chen, L. Duan, and D. Xu, "Event recognition in videos by learning from heterogeneous Web sources," in *Proc. 26th IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 2666–2673.

[11] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. 25th IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 2066–2073.

[12] R. Gopalan, R. Li, and R. Chellappa, "Domain adaptation for object recognition: An unsupervised approach," in *Proc. 13th Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 999–1006.

[13] B. Kulis, K. Saenko, and T. Darrell, "What you saw is not what you get: Domain adaptation using asymmetric kernel transforms," in *Proc. 24th IEEE Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, Jun. 2011, pp. 1785–1792.

[14] L. Bruzzone and M. Marconcini, "Domain adaptation problems: A DASVM classification technique and a circular validation strategy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 770–787, May 2010.

[15] M. Baktashmotlagh, M. T. Harandi, B. C. Lovell, and M. Salzmann, "Unsupervised domain adaptation by domain invariant projection," in *Proc. 14th Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 769–776.

[16] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *Proc. 14th Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 2960–2967.

[17] J. Hoffman, B. Kulis, T. Darrell, and K. Saenko, "Discovering latent domains for multisource domain adaptation," in *Proc. 12th Eur. Conf. Comput. Vis.*, Florence, Italy, Oct. 2012, pp. 702–715.

[18] L. Duan, D. Xu, and I. W. Tsang, "Domain adaptation from multiple sources: A domain-dependent regularization approach," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 3, pp. 504–518, Mar. 2012.

[19] R. Chattopadhyay, Q. Sun, W. Fan, I. Davidson, S. Panchanathan, and J. Ye, "Multisource domain adaptation and its application to early detection of fatigue," *ACM Trans. Knowl. Discovery Data*, vol. 6, no. 4, p. 18, Dec. 2012.

[20] C.-W. Seah, I. W. Tsang, and Y.-S. Ong, "Healing sample selection bias by source classifier selection," in *Proc. 12th Int. Conf. Data Mining*, Vancouver, BC, Canada, Dec. 2011, pp. 577–586.

[21] W. Li, L. Duan, D. Xu, and I. W. Tsang, "Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1134–1148, Jun. 2013.

[22] S. Wu, S. Bondugula, F. Luisier, X. Zhuang, and P. Natarajan, "Zero-shot event detection using multi-modal fusion of weakly supervised concepts," in *Proc. 27th IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 2665–2672.

[23] T. Mensink, E. Gavves, and C. G. M. Snoek, "COSTA: Co-occurrence statistics for zero-shot classification," in *Proc. 27th IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 2441–2448.

[24] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," in *Proc. 27th Annu. Conf. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, Dec. 2013, pp. 935–943.

[25] D. Zhang, J. He, Y. Liu, L. Si, and R. Lawrence, "Multi-view transfer learning with a large margin approach," in *Proc. 17th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, San Diego, CA, USA, Aug. 2011, pp. 1208–1216.

[26] M. Chen, K. Q. Weinberger, and J. C. Blitzer, "Co-training for domain adaptation," in *Proc. 25th Annu. Conf. Neural Inf. Process. Syst.*, Granada, Spain, Dec. 2011, pp. 2456–2464.

[27] V. Sharmanska, N. Quadrianto, and C. H. Lampert, "Learning to rank using privileged information," in *Proc. 14th Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 825–832.

[28] X. Xu, W. Li, and D. Xu, “Distance metric learning using privileged information for face verification and person re-identification,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 12, pp. 3150–3162, Dec. 2015.

[29] W. Li, L. Niu, and D. Xu, “Exploiting privileged information from Web data for image categorization,” in *Proc. 13th Eur. Conf. Comput. Vis.*, Zürich, Switzerland, Sep. 2014, pp. 437–452.

[30] B. Fernando, E. Gavves, M. J. Oramas, A. Ghodrati, and T. Tuytelaars, “Modeling video evolution for action recognition,” in *Proc. 28th IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 5378–5387.

[31] M. Hoai and A. Zisserman, “Improving human action recognition using score distribution and ranking,” in *Proc. 12th Asian Conf. Comput. Vis.*, Singapore, Nov. 2014, pp. 3–20.

[32] J. Donahue *et al.*, “DeCAF: A deep convolutional activation feature for generic visual recognition,” in *Proc. 31st IEEE Int. Conf. Mach. Learn.*, Beijing, China, Jun. 2014, pp. 647–655.

[33] X. Xu, I. W. Tsang, and D. Xu, “Soft margin multiple kernel learning,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 5, pp. 749–761, May 2013.

[34] Y. Aytar and A. Zisserman, “Tabula rasa: Model transfer for object category detection,” in *Proc. 13th Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 2252–2259.

[35] Y.-F. Li, I. W. Tsang, J. T. Kwok, and Z.-H. Zhou, “Tighter and convex maximum margin clustering,” in *Proc. 12th Int. Conf. Artif. Intell. Statist.*, Clearwater Beach, FL, USA, Apr. 2009, pp. 344–351.

[36] W. Li, L. Duan, D. Xu, and I. W. Tsang, “Text-based image retrieval using progressive multi-instance learning,” in *Proc. 13th Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 2049–2055.

[37] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.

[38] X. Xu, I. W. Tsang, and D. Xu, “Handling ambiguity via input-output kernel learning,” in *Proc. 12th Int. Conf. Data Mining*, Brussels, Belgium, Dec. 2012, pp. 725–734.

[39] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, “Domain adaptation via transfer component analysis,” *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.

[40] J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Schölkopf, “Correcting sample selection bias by unlabeled data,” in *Proc. 19th Annu. Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2006, pp. 601–608.

[41] D. R. 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, 2004.

[42] J. Farquhar, D. Hardoon, H. Meng, J. S. Shawe-Taylor, and S. Szedmak, “Two view learning: SVM-2K, theory and practice,” in *Proc. 18th Annu. Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2005, pp. 355–362.

[43] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. C. Loui, “Consumer video understanding: A benchmark database and an evaluation of human and machine performance,” in *Proc. 1st ACM Int. Conf. Multimedia Retr.*, Trento, Italy, Apr. 2011, p. 29.

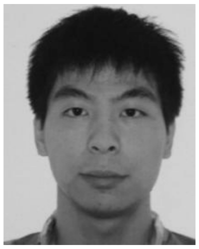
[44] M. Marszalek, I. Laptev, and C. Schmid, “Actions in context,” in *Proc. 22nd IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami Beach, FL, USA, Jun. 2009, pp. 2929–2936.

[45] Y. Li, C. Guan, H. Li, and Z. Chin, “A self-training semi-supervised SVM algorithm and its application in an EEG-based brain computer interface speller system,” *Pattern Recognit. Lett.*, vol. 29, no. 9, pp. 1285–1294, 2008.

[46] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan, “A multimedia retrieval framework based on semi-supervised ranking and relevance feedback,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 723–742, Apr. 2012.

Li Niu received the B.E. degree from the University of Science and Technology of China, Hefei, China, in 2011. He is currently pursuing the Ph.D. degree with the Interdisciplinary Graduate School, Nanyang Technological University, Singapore.

His current research interests include machine learning and computer vision.



Xinxing Xu received the B.E. degree from the University of Science and Technology of China, Hefei, China, in 2009, and the Ph.D. degree in computer engineering from the School of Computer Engineering, Nanyang Technological University, Singapore, in 2015.

He is currently a Scientist with the Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore. His current research interests include machine learning and its applications to computer vision.

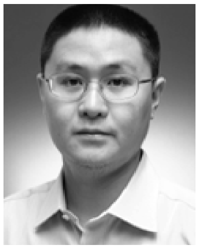
Lin Chen received the B.E. degree from the University of Science and Technology of China, Hefei, China, in 2009, and the Ph.D. degree from the School of Computer Engineering, Nanyang Technological University, Singapore, in 2014.

He is currently a Research Scientist with Amazon. His current research interests include computer vision and machine learning, in particular, deep learning with its application to computer vision tasks, such as object recognition, image/video retrieval, and classification.

Lixin Duan received the B.E. degree from the University of Science and Technology of China, Hefei, China, in 2008, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2012.

He is currently a Machine Learning Scientist with Amazon. His current research interests include transfer learning, multiple instance learning, and their applications in computer vision and data mining.

Dr. Duan was a recipient of the Microsoft Research Asia Fellowship in 2009, and the Best Student Paper Award at the IEEE Conference on Computer Vision and Pattern Recognition in 2010.



Dong Xu (M’07–SM’13) received the B.Eng. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2001 and 2005, respectively.

He was with Microsoft Research Asia, Beijing, China, and The Chinese University of Hong Kong, Honk Kong, for more than two years, during the Ph.D. study. He was also a Post-Doctoral Research Scientist with Columbia University, New York, NY, USA, from 2006 to 2007, and a Faculty Member with Nanyang Technological University, Singapore, from 2007 to 2015. He is currently a professor and Chair in Computer Engineering with the School of Electrical and Information Engineering, The University of Sydney, Sydney, NSW, Australia. His current research interests include computer vision, machine learning, and multimedia content analysis.

Dr. Xu co-authored a paper that received the Best Student Paper Award in the IEEE International Conference on Computer Vision and Pattern Recognition in 2010. Another his co-authored work also won the IEEE TRANSACTIONS ON MULTIMEDIA Prize Paper Award in 2014.