Video Annotation via Image Groups from the Web

Han Wang, Xinxiao Wu, and Yunde Jia

***Abstract*—Searching desirable events in uncontrolled videos is a challenging task. Current researches mainly focus on obtaining concepts from numerous labeled videos. But it is time consuming and labor expensive to collect a large amount of required labeled videos for training event models under various circumstances. To alleviate this problem, we propose to leverage abundant Web images for videos since Web images contain a rich source of information with many events roughly annotated and taken under various conditions. However, knowledge from the Web is noisy and diverse, brute force knowledge transfer of images may hurt the video annotation performance. Therefore, we propose a novel Group-based Domain Adaptation (GDA) learning framework to leverage different groups of knowledge (source domain) queried from the Web image search engine to consumer videos (target domain). Different from traditional methods using multiple source domains of images, our method organizes the Web images according to their intrinsic semantic relationships instead of their sources. Specifically, two different types of groups (*i.e.*, event-specific groups and concept-specific groups) are exploited to respectively describe the event-level and concept-level semantic meanings of target-domain videos. Under this framework, we assign different weights to different image groups according to the relevances between the source groups and the target domain, and each group weight represents how contributive the corresponding source image group is to the knowledge transferred to the target video. In order to make the group weights and group classifiers mutually beneficial and reciprocal, a joint optimization algorithm is presented for simultaneously learning the weights and classifiers, using two novel data-dependent regularizers. Experimental results on three challenging video datasets (*i.e.*, CCV, Kodak, and YouTube) demonstrate the effectiveness of leveraging grouped knowledge gained from Web images for video annotation.**

***Index Terms*—Concept-specific group, domain adaptation, event-specific group, video annotation.**

I. Introduction

DIGITAL cameras and mobile phone cameras have become popular in our daily life. The ever expanding video collections have motivated a real necessity to provide effective tools to support video annotation and retrieval. However, video annotation still remains a challenging problem due to the highly cluttered background, large intra-class variations and significant camera motions [1], [2], [3], [4]. In this paper, we focus on the event annotation of real-world unconstraint consumer videos, which have long-term spatially and temporally dynamic object interactions that happen under certain scene settings [5]. Recently, a number of previous methods have been proposed to effectively analyze events in videos [6], [7], [8]. These works require labeled training videos to learn robust classifiers and can achieve promising results with sufficient labeled training data. However, the labeling process is time consuming and labor expensive that users are generally reluctant to annotate abundant videos.

Since it is difficult to acquire enough knowledge from labeled videos, many researchers have tried to seek another source of labeled data and transfer the related knowledge from these data to videos. Fortunately, Web image searching engines have become increasingly mature and can offer abundant and easily accessible knowledge. Moreover, the image datasets from the Web are more diverse and less biased than home-grown datasets, which makes them more realistic for real-world tasks. Recently, several methods [9], [10], [11] are proposed to address the problem of knowledge transformation across the image domain and the video domain. In [9], Web images are incrementally collected to learn classifiers for action recognition in videos. Wang *et al.* [10] proposed to obtain knowledge for consumer videos from both labeled Web images and a small amount of labeled videos. Duan *et al.* [11] developed a multi-domain adaptation scheme by leveraging Web images from different sources. The main motivation behind their methods is that the keyword based search can be readily used to collect a large number of relevant Web images without human annotation.

Though it is beneficial to learn from Web knowledge, noisy images of little relevance with consumer videos still exist due to random noting and subjective understanding. Under this circumstance, brute force transferring may degrade the performance of classifiers for videos, which is known as negative transfer. Therefore, it is necessary to effectively summarize Web knowledge and transfer the most relevant pieces. One strategy to decrease the risk of negative transfer is assigning different weights to different source domains based on their relevances to the target domain. Recently, several domain adaptation methods were proposed to learn robust classifiers with diverse training data from multiple source domains. Luo *et al.* [12] proposed to maximize the consensus of predictions from multiple sources. Duan *et al.* [11] developed a multi-domain adaptation scheme by leveraging web images from different source domains. In their work, weights are assigned to the images according to their sources, ignoring the intrinsic semantic meaning among the source-domain data.

Manuscript received July 06, 2013; revised November 28, 2013 and February 09, 2014; accepted February 27, 2014. Date of publication March 20, 2014; date of current version July 15, 2014. This work was supported in part by the Natural Science Foundation of China (NSFC) under Grant No. 61203274, the Specialized Research Fund for the Doctoral Program of Higher Education of China (20121101120029), the Specialized Fund for Joint Building Program of Beijing Municipal Education Commission, and the Excellent Young Scholars Research Fund of BIT (2013). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Vasileios Mezaris.

The authors are with Beijing Lab of Intelligent Information Technology and the School of Computer Science, Beijing Institute of Technology, Beijing 100081, China (e-mail: wanghan@bit.edu.cn; wuxinxiao@bit.edu.cn; jiayunde@bit.edu.cn). X. Wu is the corresponding author.

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

Digital Object Identifier 10.1109/TMM.2014.2312251

We observe that it is more beneficial to measure the relevances between Web images and consumer videos according to their semantic meanings instead of their sources. In this paper, we propose to leverage Web images organized by groups, and each group of images stands for one event-related concept. Specifically, we manually define several concept-level query keywords to construct multiple groups in which the images of the same group have similar concepts. We refer this kind of group as concept-specific group. In addition, we propose another kind of groups called event-specific groups to represent events with more descriptive and discriminative abilities. The event class name is utilized as the query keyword to collect event-level Web images, and the returned images based on the same keyword construct several event-specific groups.

To use the image groups, one may consider a typical approach [13] which involves training source-domain group classifiers and using the outputs of group classifiers for the video annotation in the target domain. This approach fuses the decisions from multiple models in a late-fusion fashion without considering the mutual influence between the group classifiers and group weights. In this paper, we propose an approach which is able to automatically learn group classifiers together with group weights, in which the group weights and group classifiers are tightly correlated. The group classifiers accurately reflect the event semantic meaning and are more suitable for specific events. In the domain adaptation methods [12], [11], numerous unlabeled instances in the target domain are often ignored. As shown in [13], [14], the constraints on unlabeled instances can provide additional information to improve generalization performance. Based on this observation, the unlabeled data in our work is explored by two data-dependent regularizers, namely pseudo-loss regularizer and label-independent regularizer, which help incorporate extra informative cues into the target classifier and further enhance the generalization ability of the target classifier.

In summary, we propose a novel event annotation framework called Group-based Domain Adaptation (GDA) for consumer videos by leveraging a large amount of loosely labeled Web images. Our work is based on the observation that a large amount of loosely labeled Web images can be readily obtained by keywords based search. Using keywords as queries, we can easily collect labeled source data with semantic content at both concept level and event level. We treat an image set returned by a concept-level keyword as a concept-specific group. Besides concept-specific groups, we also learn several event-specific groups from an image set returned by an event-level keyword. Unlike the traditional multi-domain adaptation methods which leverage image data according to their sources, we propose to weight different source-domain groups according to their relevances to the target domain. Moreover, the group weights and group classifiers are simultaneously learned by a joint optimization problem to make them mutually beneficial and reciprocal. We also exploit the unlabeled consumer videos in the target domain to optimize the group weights and classifiers for building the target classifier. Specifically, two new data-dependent regularizers are introduced to enhance the generalization ability and adaptiveness of the target classifier. The experimental results on three real-world consumer video datasets (*i.e.*, CCV, Kodak and YouTube) demonstrate the effectiveness of our method for video annotation.

II. Related Work

A. Video Annotation

In recent decades, event annotation in consumer videos has become a challenging problems due to multiple concepts and their complex interactions underlying videos. Several approaches have been proposed to deal with the problem of detecting multiple concepts and modeling the relations between concepts, such as human-object interaction [15], [16], visual context for object and scene recognition [17], scene and action combination [18], and object, person, and activity relations [19]. These methods followed the conventional learning framework by assuming that the training and testing samples have the same feature distribution from the same domain. In contrast, our work focuses on annotating consumer videos by leveraging a large amount of loosely labeled Web images, in which the training and testing data come from different domains with different data distributions.

B. Domain Adaptation for Video Annotation

Domain adaptation [20] (cross-domain learning or transfer learning) methods have been employed over a wide variety of applications, such as sign language recognition [21], text classification [22], and WiFi localization [23]. Roughly speaking, there are two settings of domain adaptation: unsupervised domain adaptation where the target domain is completely unlabeled, and semi-supervised domain adaptation where the target domain contains a small amount of labeled data [24]. Since the labeled data alone is insufficient to construct well generalized target classifier, a very fruitful line of work has been focusing on effectively using unlabeled target-domain data. In [14], Bruzzone proposed a Domain Adaptation Support Vector Machine (DASVM) to iteratively learn the target classifier by labeling the unlabeled samples in the source domain. Gopalan [25] and Gong [26] used both labeled source-domain data and unlabeled target-domain data to infer new subspaces for domain adaptation. Saenko [27] proposed a metric learning method to make the intra-class samples from two domains become closer to each other. Our method belongs to the unsupervised domain adaptation methods, in which the training data consists of a large number of labeled Web images and a few of unlabeled consumer videos.

Recently, applying domain adaptation to multimedia content analysis has attracted more and more attentions of researchers [28], [25], [29], [11]. Yang *et al.* [30] proposed an Adaptive Support Vector Machine (A-SVM) method to learn a new SVM classifier for the target domain, which is adapted from a pre-trained classifier from a source domain. Duan *et al.* [31] proposed to simultaneously learn the optimal linear combination of base kernels and the target classifier by minimizing a regularized structural risk function. And then, they proposed A-MKL [6] to add the pre-learned classifiers as the prior. Their methods mainly focus on the single source domain setting. To utilize numerous labeled image data in the Web, multiple source domains adaptation methods [32], [11], [33] are proposed to leverage different pre-computed classifiers learned from multiple source domains. In these methods, different weights are assigned to different source domains without taking account of intrinsic semantic relations between source domains. In this paper, we leverage

different groups of images queried by different associational keywords to the Web. We insure that the samples in each group are of the same concept, and also ensure that different groups within the same event class are correlated to each other.

Several recent methods have been proposed to investigate the knowledge transform from Web images to consumer videos. In [9], Web images are incrementally collected to learn classifiers for action video recognition. Tang *et al.* [34] introduced a novel self-paced domain adaptation algorithm to iteratively adapt the detector from source images to target videos. Recently, Duan *et al.*[11] developed a domain selection method to select the most relevant source domains. In these existing works, the pre-learned classifiers are primarily learned using training data from different source domains and then the target classifiers are learned from pre-learned classifiers in a late-fusion fashion. In contrast, our work can simultaneously learn the optimal classifiers and weights of different source-domain groups to construct the target classifier.

III. The Proposed Framework

A. Traditional Approach

Given an abundant number of loosely labeled Web images as source domain, and unlabeled consumer videos as target domain, the aim is to learn a target predictive function $f_t(\cdot)$ by leveraging the knowledge from both source and target domains. Under the setting of *transductive learning* [20], some unlabeled data in the target domain can be seen at the training stage. Let S source domains be $\mathcal{D}^s = \{(x_i^s, y_i^s)|_{i=1}^{n_s}\}, (s = 1, \ldots, S)$, where x_i^s is the i-th image from the s-th source domain with its label y_i^s. And $\mathcal{D}^t$ is the target domain consisting of unlabeled videos x_i^t, $(1 \leq i \leq N_t)$, where N_t is the total number of the target-domain videos. The annotation of an input video x_i^t can be solved by

$$f_t(x_i^t) = \sum_{s=1}^{S} \alpha_s f^s(x_i^t), \tag{1}$$

where $f_t(x_i^t)$ is the event classifier of the target domain, $f^s(x_i^t)$ is the decision value of x_i^t from the s-th source classifier, and α_s is the weight of the s-th source domain. Then, a general approach to train the target classifier can be formulated by minimizing the following objective function:

$$\min_{f_t} \sum_{s=1}^{S} \sum_{i=1}^{n_s} \ell(f_t, y_i^s) + \lambda \Omega(f_t), \tag{2}$$

where $\ell(\cdot, \cdot)$ is a loss function of the target classifier f_t on the labeled instances of the target domain, $\Omega(f_t)$ is a regularization function on f_t, and λ is a regularization parameter. Once f_t is learned, we can use it for event annotation. Clearly, in this objective function, three main components need to be properly designed: the source-domain weights α_s, the loss function $\ell(\cdot, \cdot)$, and the regularization function $\Omega(\cdot)$.

Traditional multi-source adaptation methods typically adopt a simple two-step process for video annotation: (1) pre-learn event classifiers for each source domain to predict event labels for all the training videos; (2) learn the source-domain weights to fuse the predictions from multiple source domains for building the target classifier. In the first step, S source-domain classifiers $\{f^1, \ldots, f^S\}$ are pre-learned by the following optimization problem:

$$\min_{f^s} \sum_{i=1}^{n_s} \widetilde{\ell}(f^s(x_i^s), y_i^s) + \lambda \widetilde{\Omega}(f^s), \tag{3}$$

where $\widetilde{\ell}(\cdot, \cdot)$ and $\widetilde{\Omega}(\cdot)$ are the loss function and the regularization function of the pre-learned classifiers f^s on the labeled instances of the source domain, respectively. Once the S classifiers $\{f^1, \ldots, f^S\}$ are obtained, we convert the original feature representation x_i^s to the domain-based representation $f^s(x_i^s)$. In the second step, the event classifier function f_t can be trained based on the new domain representation in the same way of Eq. (2), *i.e.*,

$$\begin{aligned} &\min_{f_t} \ell(f_t, y_i) + \lambda \Omega(f_t) \\ &\Rightarrow \min_{f_t} \sum_{s=1}^{S} \sum_{i=1}^{n_s} \ell(\alpha_s f^s(x_i^s), y_i) + \lambda \Omega(f_t). \end{aligned} \tag{4}$$

Although traditional multi-domain adaptation methods are expected to decrease the risk of negative transfer by importing knowledge from multiple source domains rather than one source domain, we observe that it is more beneficial to assign different instances of the same source domain with different weights according to their relevances to the target domain, instead of treating the instances from the same source equally. The more relevant the source-domain instance is to the target domain, the higher its weight becomes, which benefits transferring more relevant instances and further alleviating the negative transfer problem. Thus, we apply multiple groups of the source domain to leverage Web knowledge and assign different groups with different weights based on the relevance between the corresponding group and the target domain. Different from traditional multiple source domains, the proposed multiple groups are able to describe different properties of events using different concepts as well as capture the semantic relationship between the source-domain data. For each event, according to the learned weights, we only choose several related groups of images for knowledge transfer. In order to improve the discriminative ability of multiple groups, we propose two types of groups, namely, event-specific group and concept-specific group, where event-specific groups are automatically learned from a set of event-level images while concept-specific groups are manually defined by concept-level keywords (Section III-B). Since unlabeled target-domain data can provide useful constraints to enhance the generalization of the target classifier, we employ unlabeled target-domain data as additional information to optimize the group weights and classifiers in a joint learning manner. (Section III-C).

B. Concept-Specific and Event-Specific Groups of the Source Domain

To generate the concept-specific groups of the source domain, we first manually define the event concept collection as $\mathcal{C} = \{C_1, C_2, \ldots, C_G\}$, where C_i represents the i-th concept. In this paper, we use 43 (*i.e.*, $G = 43$) concept-level keywords, including action related concepts, object related concepts, as well as scene related concepts. For each concept, we collect a group

of images by querying a concept-level keyword to the Web image search engine. We denote a set of images returned by one concept-level keyword as a *concept-specific group*, which represents one semantic concept of events. Using multiple groups, we can capture different event knowledge that relates to multiple semantic concepts. The labeled data of the s-th group is defined as $X^s = \{x_i^s\}_{i=1}^{N_s}$, $s \in \{1, \ldots, G\}$, where $x_i^s \in \mathbb{R}^{d_s}$ represents the i-th image in the s-th group with d_s the dimensionality of the source-domain image feature, and N_s represents the number of images in the s-th group. For each concept-specific group, we pre-learn a SVM classifier $g_s(\cdot)$ using the corresponding data X^s, and then G group classifiers are obtained from G concept-specific groups. In addition, we define N_t unlabeled videos in the target domain as $X^t = \{x_i^t\}_{i=1}^{N_t}$, where $x_i^t \in \mathbb{R}^{d_t}$ is the i-th video in the target domain, with d_t the dimensionality of target-domain video feature.

Although concept-specific groups provide high-level semantic information for improving the characterization of events, they still suffer from two practical problems in implementation: 1) the manually specified concepts are sometimes subjective without considering the potentially discriminative concepts; 2) the number of concepts is manually fixed and it remains unclear how many groups (*i.e.*, how many concepts) will be sufficiently reliable to describe the source domain. Therefore, we additionally propose the event-specific groups which are automatically learned from source-domain images. These images are collected by querying a event-level keyword to the Web image search engine. The event-level keyword is actually the name of event class such as "wedding" and "swimming", so each event-level keyword corresponds to one event class. Here we denote the image set returned from an event-level keyword as $X^e = \{x_i^e\}_{i=1}^{N_e}$, where N_e is the number of images and $x_i^e \in \mathbb{R}^{d_s}$ is the i-th image in this image set.

Combining concept-specific groups and event-specific groups constructs the whole group set of the source-domain images in our method. For each event class, we automatically learn the total S group classifiers using the event-level images. Among the S group classifiers, G group classifiers are learned from concept-level images by using standard SVM, and the rest $E = S - G$ group classifiers are randomly initialized to zero. Then, the event-specific images are applied to automatically learn the total S group classifiers as well as the group weights using the joint group weighting scheme described in Section III-C.

C. Regularizers

Given the pre-learned group classifiers $g_s(\cdot), s \in \{1, \ldots, G\}$ from the source-domain images, we aim to learn the target-domain classifiers by combing event-specific group classifiers $g_s(\cdot), s \in \{G+1, \ldots, S\}$ and concept-specific group classifiers $g_s(\cdot), s \in \{1, \ldots, G\}$ in a joint learning framework. For a target-domain video x_i^t, the target classifier $f_t(x_i^t)$ is formulated by

$$f_t(x_i^t) = \sum_{s=1}^{S} \alpha_s g_s(x_i^t), \tag{5}$$

where $g_s(x_i^t) = w_s^T x_i^t$ indicates the s-th group classifier from the source domain, and α_s represents the group weight.

Since we do not have any labeled data in the target domain, we propose to simultaneously minimize the loss of labeled training data from the source domain as well as different regulariers defined on the unlabeled data from the target domain. The proposed framework is then formulated as follows:

$$\min_{f_t} \Omega_C(f_t) + \lambda_L \sum_{i=1}^{N_e} \Omega_L(f_t(x_i^e)) + \lambda_D \sum_{i=1}^{N_t} \Omega_D(f_t(x_i^t)) + \lambda_P \sum_{i=1}^{N_t} \Omega_P(f_t(x_i^t)), \tag{6}$$

where $\lambda_L, \lambda_D > 0$ and $\lambda_P < 0$ are tradeoff parameters. The details of each term in Eq. (6) are described in the following.

$\Omega_C(f_t)$ is a data-independent regularizer lead to a sparse representation of the target classifier, which reduces the complexity of the target classifier f_t, defined as

$$\Omega_C(f_t) = \sum_{s=1}^{S} \|\alpha_s\|^2. \tag{7}$$

$\Omega_L(f_t)$ is a loss function of the target classifier f_t on the labeled instances of the source domain, defined as

$$\Omega_L(f_t) = \sum_{i=1}^{N_e} \Omega_L(f_t(x_i^e)) = \sum_{i=1}^{N_e} \sum_{s=1}^{S} \|\alpha_s(w_s^T x_i^e) - y_i^e\|^2. \tag{8}$$

Since there is no labeled training data in the target domain, we formulate the loss function on the event-level images from the source domain.

Recall that the learned classifier by using labeled instances from the source domain may not perform well due to the feature distribution mismatch between the source and target domains. Thus, if we only adopt the loss function on labeled training source data, the learned classification hyperplane may overfit and the generalization ability of the target classifier may be degraded. In this paper, we use the unlabeled instances in the target domain to improve the generalization ability of the learned classifier by controlling the decision value of the target classifier. $\Omega_D(f_t)$ is a label-independent regularizer to control the complexity of the target classifier f_t, defined as

$$\Omega_D(f_t) = \sum_{i=1}^{N_t} \Omega_D(f_t(x_i^t)) = \sum_{i=1}^{N_t} \|f_t(x_i^t)\|^2. \tag{9}$$

We use the f_t to predict the labels of the unlabeled target-domain data, called "pseudo labels". As mentioned before, though we do not have any labeled data in the target domain, we still want to maximize the margin between the target-domain data whose pseudo labels are different. Consequently, for the target classifier f_t, a pseudo-loss regularizer $\Omega_P(f_t)$ is defined as

$$\Omega_P(f_t) = \sum_{i=1}^{N_t} \Omega_P(f_t(x_i^t)) = \sum_{i=1}^{N_t} \tilde{y}_i f_t(x_i^t), \tag{10}$$

where $\tilde{y}_i$ is the pseudo label of the i-th unlabeled target-domain data x_i^t.

Putting everything together, we have the following optimization problem:

$$\min_{\alpha_1,\ldots,\alpha_S,w_1,\ldots,w_S} \sum_{s=1}^{S} \|\alpha_s\|^2 + \lambda_L \sum_{i=1}^{N_e} \sum_{s=1}^{S} \|\alpha_s(w_s^T x_i^e) - y_i^e\|^2 + \lambda_D \sum_{i=1}^{N_t} \|f_t(x_i^t)\|^2 + \lambda_P \sum_{i=1}^{N_t} \tilde{y}_i f_t(x_i^t),$$
$$s.t. \quad \sum_{s}^{S} \alpha_s = 1. \tag{11}$$

By denoting $A = [\alpha_1, \ldots, \alpha_S] \in R^{1*S}$, $W = [w_1, \ldots, w_S] \in R^{d_t * S}$, $X^e = [x_1^e, \ldots, x_{N_e}^e] \in R^{d_t * N_e}$, $Y^e = [y_1^e, \ldots, y_{N_e}^e] \in R^{1*N_e}$, $X^t = [x_1^t, \ldots, x_{N_t}^t] \in R^{d_t * N_t}$ and $\tilde{Y} = [\tilde{y}_1, \ldots, \tilde{y}_{N_t}] \in R^{1*N_t}$, we rewrite Eq. (11) as

$$\begin{aligned} &\min_{A,W} \|A\|^2 + \lambda_L \|AW^T X^e - Y^e\|^2 + \lambda_D \|AW^T X^t\|^2 \\ &\quad + \lambda_P \tilde{Y}(AW^T X^t)^T. \\ &\quad s.t. \quad \|A\|_1 = 1. \end{aligned} \tag{12}$$

We propose an iterative approach to solve the optimization problem in Eq. (12). We define $f_t^{(m)}$ as the target function learned in the m-th iteration. The pseudo label used in the m-th iteration is obtained by calculating the target function $f_t^{(m-1)}$ learned from the previous step.

Our GDA algorithm is made up of two main phases. In the first phase, G concept-specific groups are learned by standard SVM and E event-specific groups are initialized randomly. The second phase comes to iterative adapt all these pre-learned group classifiers to the target classifier. By iteratively holding group classifiers W and group weights A fixed, the optimization of Eq. (12) can be directly solved by a Quadratic Problem. This iterative approach is summarized in Algorithm 1.

Algorithm 1: Group-based Domain Adaptation.

Input:

- $\{X^s\}_{s=1}^S$ Z: S image groups;
- X^t Z: unlabeled target videos.

Output:

- $\{w_s\}_{s=1}^S$ Z: group classifiers;
- $\{\alpha_s\}_{s=1}^S$ Z: group weights.

1: Initialize G concept-specific group classifiers $\{w_s^{(0)}\}_{s=1}^G$ using standard SVM;

2: Initialize E event-specific group classifiers $\{w_s^{(0)}\}_{s=G+1}^S$ randomly;

3: Compute $\alpha_s^{(0)}$ by solving the standard Quadratic Programming problem in Eq. (12) with fixed $w_s^{(0)}$;

4: Set $m = 0$;

5: **repeat**

Compute $\tilde{y}_i^{(m)}$ using $w_s^{(m)}$ and $\alpha_s^{(m)}$ according to the target classifier in Eq. (5);

With fixed $\alpha_s^{(m)}$, compute $w_s^{(m+1)}$ in the Eq. (12) by using standard Quadratic Programming;

With fixed $w_s^{(m+1)}$, compute $\alpha_s^{(m+1)}$ in the Eq. (12) by using standard Quadratic Programming;

until *Convergence*;

6: Return α_s and w_s.

IV. Experiments

A. Datasets

We evaluate our method on three benchmark video datasets (*i.e.*, Kodak [3], YouTube [6] and CCV [1]). To collect Web images, we use Google image search engine with pre-defined keywords.

1) Video Datasets: **CCV dataset** [1] contains a training set of 4,659 videos and a testing set of 4,658 videos which are annotated to 20 semantic categories. Since our work focuses on event annotation, we do not consider the non-event categories (*i.e.*, "playground", "bird", "beach", "cat" and "dog"). In order to facilitate the keyword based image collection using the Web search engine, the events of "wedding ceremony", "wedding reception" and "wedding dance" are merged into one event as "wedding". The events of "non-music performance" and "music performance" are merged into "performance". Finally, there are twelve event categories: "basketball", "baseball", "soccer", "iceskating", "biking", "swimming", "skinning", "graduation", "birthday", "wedding", "show", and "parade".

Kodak dataset is collected by Kodak [3] from about 100 real users over one year, consisting of 195 consumer videos with their ground truth labels of six event classes (*i.e.*, "wedding", "birthday", "picnic", "parade", "show" and "sports").

YouTube dataset [6] contains 906 consumer videos from YouTube with labels of the same six event classes as in the Kodak dataset (*i.e.*, "wedding", "birthday", "picnic", "parade", "show" and "sports").

According to the setting of transductive learning [20], we assume that some unlabeled data can be seen during both training and testing stages. For CCV dataset, we use the 4,659 unlabeled videos for training and evaluate the performance on all 9,317 videos. And for Kodak and Youtube datasets, we use all the unlabeled videos at both training and testing stages. For each video, one frame is randomly sampled as the keyframe and then 128-dimensional SIFT descriptors [35] are extracted to represent the keyframe. In our experiments, we use the existing tool[1] from Professor David Lowe in a sparse pattern.

2) Web Image Dataset: According to the consumer video datasets, we collect Web images covering thirteen events: "basketball", "baseball", "soccer", "iceskating", "biking", "swimming", "graduation", "birthday", "wedding", "skinning", "show", "parade" and "picnic". For each input keyword, the top ranked 200 images are downloaded and the corrupted images with invalid URLs are discarded. Finally, 5,942 images and 1,647 images are collected for concept-specific groups and event-specific groups, respectively. Table I shows the keywords used in our experiment. We want to mention that the concept-specific keywords are selected according to the events instead of the datasets. All the concepts are shared among all the three datasets. In Table I, the left-most column lists the keywords for concept-specific groups where each keyword corresponds to one concept-specific group. The top row lists the keywords used for event-specific groups. For event-specific groups, we directly use the event class names as keywords with one keyword for one event class. In Table I, the element in the i-th row and the j-th column indicates whether the i-th concept-level keyword occurs in the j-th event

[1]http://www.cs.ubc.ca/ lowe/keypoints/

TABLE I
THE KEYWORD USED FOR QUERY WEB IMAGES. THE LEFT-MOST COLUMN LISTS THE KEYWORDS FOR CONCEPT-SPECIFIC GROUPS WHERE EACH KEYWORD CORRESPONDS TO ONE CONCEPT-SPECIFIC GROUP. THE TOP ROW LISTS THE KEYWORDS USED FOR EVENT-SPECIFIC GROUPS. FOR EVENT-SPECIFIC GROUP, WE DIRECTLY USE THE EVENT CLASS NAMES AS KEYWORDS WITH ONE KEYWORD FOR EACH EVENT CLASS

concept_specific \ event_specific	biking	birthday_party	graduation	iceskating	parade	picnic	play_baseball	play_basketball	play_soccer	show	skiing	swimming	wedding
academic_dress			×										
arm_pull_swimming												×	
ball							×	×	×				
ball_shot									×				
baseball_field							×						
baseball_pitcher							×						
basketball_court								×					
bike	×												
bike_riding	×												
blowing_candles		×											
bride													×
cheering		×	×										×
chorus										×			
clapping		×	×							×			×
college_cap			×										
dancing										×			×
gribble													×
eating		×				×							
football_field									×				
groom													×
hugging		×	×										×
ice_skate				×									
jumping								×					
kicking								×					
kissing													×
laughing		×	×										×
leg_bending	×												
marching					×								
music_instrument									×				
outside													×
pass_the_baseball								×					
picnic_food						×							
running					×	×	×	×					
singing		×								×			
skating_rink				×									
skis											×		
snow_field				×									
stage										×			
swimming_pool												×	
swimwear												×	
throw							×						
walking_down_the_aisle													×
waving								×					

(*i.e.*, "×" for occurrence and blank for nonoccurrence). For each image, the 128-dimensional SIFT descriptors are extracted in the same way used in video keyframes.

B. Experimental Setup

The bag-of-words representation is used for both image and video features. Specifically, we cluster the SIFT descriptors [35] extracted from all the training Web images and keyframes of the training consumer videos, into 2,000 words by using k-means clustering method. Each image (video keyframe) is then represented as a 2,000-dimensional token frequency (TF) feature by quantizing its SIFT descriptors with respect to the visual codebook. For consumer videos, we directly use the 5,000-dimensional features provided by [1], and 2,000-dimensional features provided by and [6] for the Kodak and YouTube dateset, respectively.

To pre-learn a classifier for each event class in one group, the positive samples are constructed by the images belonging to the corresponding event class in the corresponding group and the negative samples consist of randomly selected 300 images of other events classes from other groups. At the training stage, for the CCV dataset, the training set defined by [1] is used as the unlabeled target domain. For the Kodak and YouTube datasets, the target domains contain 195 and 906 videos, respectively. Consequently, the training data includes the labeled Web image groups from the source domain and unlabeled videos from the target domain.

We compare our method with several baseline methods, including the standard SVM (S_SVM) [36], the single domain adaptation methods of Geodesic Flow Kernel (GFK) [26] and Domain Adaptive SVM (DASVM) [14], the multi-domain adaptation methods of Domain Adaptation Machine (DAM) [13], Conditional Probability based Multi-source Domain Adaptation (CPMDA) [37] and Domain Selection Machine (DSM) [11]. To validate the effectiveness of event-specific source-domain groups, we also report the results of simplified

TABLE II
COMPARISON OF MAPS BETWEEN DIFFERENT METHODS ON THE CCV, KODAK AND YOUTUBE DATASETS

Method	S_SVM	CPMDA [37]	DASVM [14]	DAM [13]	DSM [11]	GFK [26]	GDA_sim	GDA
CCV	0.0977	0.0923	0.0973	0.1027	0.0974	0.0966	0.1051	**0.1205**
Kodak	0.2016	0.1966	0.1684	0.2574	0.1787	0.1742	0.2060	**0.2860**
YouTube	0.2006	0.1974	0.1708	0.2004	0.1824	0.1722	0.1985	**0.2130**

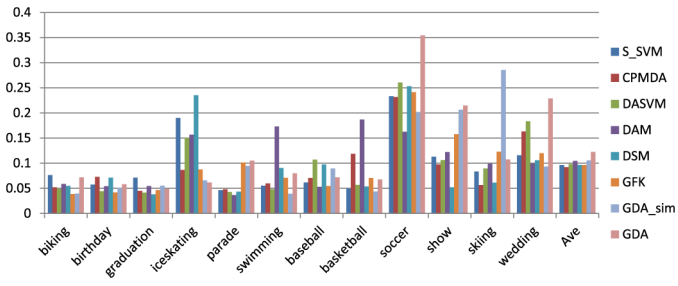

Fig. 1. Per-event Average Precisions (APs) of all methods on the CCV dataset.

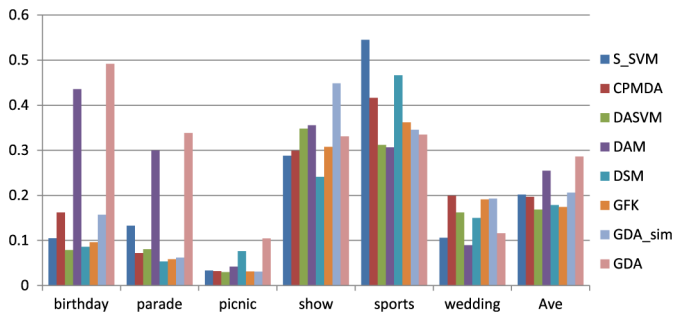

Fig. 2. Per-event Average Precisions (APs) of all methods on the Kodak dataset.

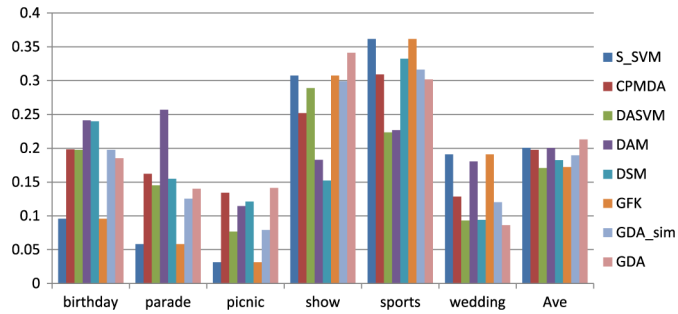

Fig. 3. Per-event Average Precisions (APs) of all methods on the YouTube dataset.

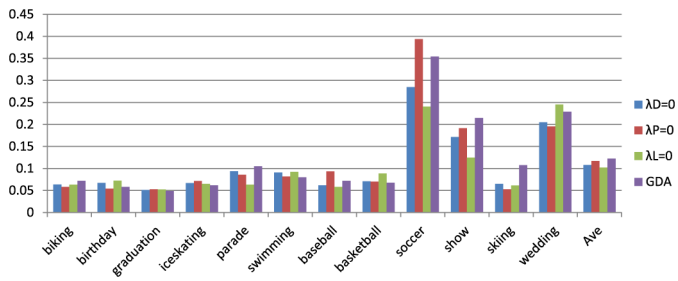

Fig. 4. Performances of different regularizers on the CCV dataset.

version (referred to as GDA_sim) of our method which only uses concept-specific source-domain groups. Since the S_SVM can only handle data from a single group, the Web images from all the event-specific groups are collected as a single group to train SVM classifiers in S_SVM. For DASVM and GFK, the target classifiers are trained using the labeled images from the Web and the keyframes of unlabeled videos from the target domain. In CPMDA, DAM and DSM, we treat the G concept-specific image groups as G source domains and define all the event-specific image groups as the $(G+1)$-th source domain. In GDA_sim, we only adopt the G concept-specific image groups. In our method GDA, the total group number is set by $S = 15$.

For all the methods, Average Precision (AP) is used for performance evaluation and mean Average Precision (mAP) is defined as the mean of APs over all event classes.

C. Performance Comparisons

1) Results of Different Methods: We report the per-event APs of all the methods on the CCV, Kodak and YouTube datasets in Fig. 1, 2 and 3, respectively. We also show the mAPs of all the methods on these datasets in Table II. It can be seen that our method is consistently competitive with other methods. Zooming into the details, we have the following observations:

- S_SVM outperforms the existing domain adaptation methods GFK, CPMDA, DSM and DASVM. This indicates that simply weighting instances according to their sources may bring negative information that hurts the transferring performance. Among multi-domain adaptation methods, DAM outperforms CPMDA and DSM. A possible explanation is that the manifold assumption employed in CPMDA may not hold well for real-word consumer videos, which degrades the annotation performance of CPMDA. In DSM, the most relevant group is selected for domain adaptation. However, the consumer video always contain complex semantic meanings, it is not reasonable to apply only one group to describe complex knowledge.
- DAM is better than S_SVM. The results demonstrate the effectiveness of applying multiple groups of knowledge in video annotation task. Moreover, the usage of target-domain instances in DAM verifies that the instances from both source domain and target domain are capable of improving generalization performance of target-domain classifiers.
- On all the datasets, the GDA_sim is consistently better than the other six methods (*i.e.*, S_SVM, CPMDA, DASVM, DAM, DSM, GFK). The results clearly demonstrate that it is beneficial to leverage multiple groups with semantic meanings by employing unlabeled target-domain instances. Our proposed method GDA outperforms GDA_sim, which further demonstrates the effectiveness of integrating the event-specific groups.

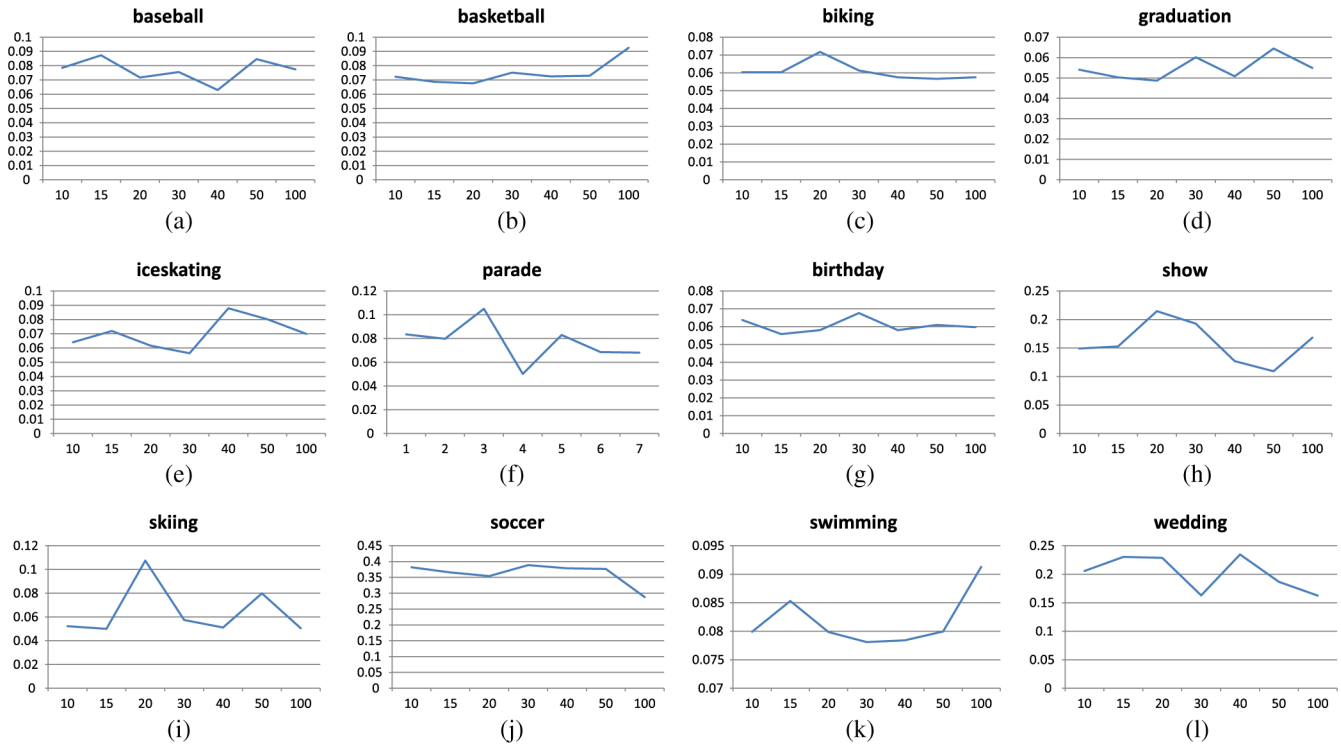

Fig. 5. Performance variations of different group numbers S on the CCV dataset. The horizontal axis indicates the group number S and the vertical axis indicates the mean Average Precision performance.

- The GDA achieves the best results in all three datasets, which shows that jointly learning different concept groups of knowledge is beneficial for positive transform. Compared with S_SVM, CPMDA, DASVM, DAM, DSM and GFK on the CCV dataset, the relative improvements of our method are 11.95%, 23.40%, 14.77%, 19.17% and 19.83%. On the Kodak dataset (*resp.*, the YouTube dataset), the relative improvement of our method is no less than 10%.

2) Evaluation on Different Regularizers: We also investigate the effects of each term in our optimization function in Eq. (11). Fig. 4 shows the results when $\lambda_L = 0$, $\lambda_D = 0$ and $\lambda_P = 0$, respectively. From the result in Fig. 5, it is interesting to notice that the average precision of all the event classes degrades when any regularizers is removed from the optimization function. For some event classes such as "soccer" and "baseball", the performances increase when the term $\lambda_p \tilde{Y}(AW^T X^t)^T$ is removed. A possible explanation is that the prediction error exists in the initial pseudo-label, which degrades the performance of the term $\lambda_p \tilde{Y}(AW^T X^t)^T$. For the events of "basketball" and "wedding", the performances increase when the term $\lambda_L \|AW^T X^e - Y^e\|^2$ is removed from the objective function. The reason may be the existence of noisy Web images whose appearance or semantic meanings of these images are not consistent with those in consumer videos which degrades the performance of the term $\lambda_L \|AW^T X^e - Y^e\|^2$.

3) Overlap of Training Unlabeled Target Domain: Since some unlabeled target-domain data can be seen at training stage, we evaluate our GDA method on two different data settings: (1) all the testing data can be seen at the training stage (Table III); (2) all the testing data cannot be seen at the training stage (Table IV). The experiment is conducted on the CCV dataset which is separated into the training part and the testing part in [1] by the authors. Table III shows the performance of our method when the training part of the videos are also used as the testing data. Table IV shows the results when the training target domain videos are different from the testing videos (*i.e.*, the training part only appears at training stage and the testing part is only used for testing). From Table III and Table IV, it can be seen that better performance shows when the training and testing target domain data are the same. This demonstrate the effectiveness of using unlabeled target-domain data in training stage.

4) Group Number Sensitivity: Additional experiments are conducted on the CCV dataset to study how the group number (*i.e.*, S in Eq. (11)) affects the annotation performance. Fig. 5 demonstrates the Average Precision variations with regard to the increasing group number S, where the horizontal axis indicates the group number S and the vertical axis indicates the Average Precision performance. There is no consistent winner among the events. A possible explanation is that the inter-class variation exists in the consumer videos and the learned group weights cannot adapt to all the instances. Taking event "wedding" for example, there are many types of wedding, and the weights of groups for different wedding types may be different. The learned group weights cannot adapt to all the wedding events. For most event classes, the best performance is obtained when S ranges from 15 to 20. Either large or small values of S will degrade the performance. This may be explained that when the S is small, the algorithm is degraded into learning weights of manually defined concept-specific groups ignoring the event-specific groups. Meanwhile, when the S is large, the learnt target classifier will become overfit for too many meaningless groups are considered. In our experiment, we set $S = 15$ with good results for a wide range of datasets.

TABLE III
PER-EVENT APS ON THE CCV DATASETS WHEN THE TRAINING AND TESTING TARGET-DOMAIN SAMPLES ARE THE SAME

Method	S_SVM	CPMDA [37]	DASVM [14]	DAM [13]	DSM [11]	GFK [26]	GDA_sim	GDA
biking	**0.0762**	0.0518	0.0493	0.0587	0.0549	0.0387	0.0396	0.0718
birthday	0.0574	**0.0729**	0.0443	0.0542	0.0713	0.0418	0.0514	0.0580
graduation	**0.0713**	0.0449	0.0413	0.0547	0.0379	0.0465	0.0553	0.0487
iceskating	**0.1901**	0.0866	0.1506	0.1567	0.2353	0.0877	0.0656	0.0616
parade	0.0463	0.0479	0.0425	0.0366	0.0433	0.1015	0.0945	**0.1050**
picnic	0.0554	0.0596	0.0479	**0.1732**	0.0904	0.0708	0.0391	0.0799
baseball	0.0617	0.0704	**0.1071**	0.0528	0.0972	0.0544	0.0892	0.0718
basketball	0.0495	0.1186	0.0567	**0.1868**	0.0538	0.0707	0.0438	0.0677
soccer	0.2333	0.2318	0.2606	0.1625	0.2530	0.2414	0.2019	**0.3541**
show	0.1128	0.0974	0.1060	0.1219	0.0520	0.1577	0.2062	**0.2147**
skiing	0.0834	0.0565	0.0896	0.0989	0.0612	0.1226	**0.2854**	0.1075
wedding	0.1159	0.1632	0.1833	0.0998	0.1058	0.1199	0.0932	**0.2287**
mAP	0.0961	0.0918	0.0983	0.1047	0.0964	0.0961	0.1054	**0.1224**

TABLE IV
PER-EVENT APS ON THE CCV DATASETS WHEN THE TRAINING AND TESTING TARGET-DOMAIN SAMPLES ARE DIFFERENT

Method	S_SVM	CPMDA [37]	DASVM [14]	DAM [13]	DSM [11]	GFK [26]	GDA_sim	GDA
biking	0.0551	0.0533	0.0541	0.0601	0.0607	0.0407	0.0414	**0.0682**
birthday	0.0618	**0.0794**	0.0434	0.0601	0.0603	0.0443	0.0425	0.0701
graduation	**0.0796**	0.0479	0.0421	0.0647	0.0653	0.0570	0.0489	0.0513
iceskating	**0.1793**	0.1126	0.1463	0.1105	0.0502	0.0934	0.0926	0.0984
parade	0.0526	0.0458	0.0459	0.0496	**0.1657**	0.1097	0.0658	0.1013
swimming	0.0432	0.0656	0.0455	**0.1120**	0.0576	0.0839	0.0389	0.0772
baseball	0.0915	0.0643	0.0980	**0.1472**	0.0647	0.0592	0.0934	0.0756
basketball	0.0639	0.1044	0.0613	0.1034	0.0869	0.0623	0.0539	0.0742
soccer	0.1211	0.2516	0.2510	0.1843	0.2693	0.2356	**0.3202**	0.2455
show	0.0998	0.0759	0.1089	0.0945	0.1251	0.1465	**0.2533**	0.2126
skiing	0.0895	0.0521	0.0959	0.0905	0.0664	0.1277	**0.1180**	0.0930
wedding	0.1242	0.1599	0.1634	0.1329	0.1076	0.1044	0.0889	**0.2207**
mAP	0.0993	0.0928	0.0963	0.1008	0.0983	0.0970	0.1049	**0.1157**

V. CONCLUSION

We have presented a new framework, referred to GDA, for annotating consumer videos by leveraging a large amount of loosely labeled Web images. Specifically, we exploited concept-level and event-level images to learn concept-specific and event-specific group representation of source-domain Web images. The group classifiers and weights are jointly learned in a unified optimization algorithm to build the target-domain classifiers. In addition, we introduced two new data-dependent regularizers based on the unlabeled target-domain consumer videos for enhancing the generalization of the target classifier. Experimental results clearly demonstrate the effectiveness of our framework. To the best of our knowledge, this is the first attempt in transfer learning to weight data according to their semantic meaning instead of their sources. A possible future research direction is to develop a discriminative common feature space between Web images and consumer videos as well as investigate several criteria to deal with the data distribution mismatch between source and target domains. We are also going to apply our proposed method to other cross-domain applications, such as text-video domain adaptation.

REFERENCES

[1] Y. Jiang, G. Ye, S. Chang, D. Ellis, and A. Loui, "Consumer video understanding: A benchmark database and an evaluation of human and machine performance," in *Proc. ICMR*, 2011.

[2] M. R. Naphade and J. R. Smith, "On the detection of semantic concepts at TRECVID," in *Proc. 12th Annu. ACM Int. Conf. Multimedia*, 2004, pp. 660–667.

[3] A. Loui, J. Luo, S. Chang, D. Ellis, W. Jiang, L. Kennedy, K. Lee, and A. Yanagawa, "Kodak's consumer video benchmark data set: Concept definition and annotation," in *Proc. Workshop Multimedia Information Retrieval*, 2007, pp. 245–254.

[4] X. Wu, X. Dong, D. Lixin, L. Jiebo, and J. Yunde, "Action recognition using multi-level features and latent structural SVM," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 8, pp. 1422–1431, 2013.

[5] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah, "High-level event recognition in unconstrained videos," *Int. J. Multimedia Inf. Retriev.*, pp. 1–29, 2012.

[6] L. Duan, D. Xu, I. Tsang, and J. Luo, "Visual event recognition in videos by learning from web data," in *Proc. CVPR*, 2010, pp. 1959–1966.

[7] Z. Ma, A. G. Hauptmann, Y. Yang, and N. Sebe, "Classifier-specific intermediate representation for multimedia tasks," in *Proc. 2nd ACM Int. Conf. Multimedia Retrieval*, 2012, p. 50.

[8] N. Ikizler-Cinbis and S. Sclaroff, "Object, scene and actions: Combining multiple features for human action recognition," in *Proc. ECCV*, 2010, pp. 494–507.

[9] N. Ikizler-Cinbis, R. Cinbis, and S. Sclaroff, "Learning actions from the web," in *Proc. CVPR*, 2009, pp. 995–1002.

[10] X. W. Han Wang and Y. Jia, "Annotating video events from the web images," in *Proc. ICPR*, 2012.

[11] L. Duan, D. Xu, Tsang, and S.-F. Chang, "Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach," in *Proc. CVPR*, 2012, pp. 1959–1966.

[12] P. Luo, F. Zhuang, H. Xiong, Y. Xiong, and Q. He, "Transfer learning from multiple source domains via consensus regularization," in *Proc. 17th ACM Conf. Information and Knowledge Management*, 2008, pp. 103–112.

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

[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, 2010.

[15] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. M. Rehg, "A scalable approach to activity recognition based on object use," in *Proc. ICCV*, 2007, pp. 1–8.

[16] B. Yao and L. Fei-Fei, "Grouplet: A structured image representation for recognizing human and object interactions," in *Proc. CVPR*, 2010, pp. 9–16.

[17] A. Gupta and L. S. Davis, "Objects in action: An approach for combining action understanding and object perception," in *Proc. CVPR*, 2007, pp. 1–8.

[18] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Proc. CVPR*, 2009, pp. 2929–2936.
[19] M. S. Ryoo and J. Aggarwal, "Hierarchical recognition of human activities interacting with objects," in *Proc. CVPR*, 2007, pp. 1–8.
[20] S. Pan and Q. Yang, "A survey on transfer learning," *Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010.
[21] A. Farhadi, D. Forsyth, and R. White, "Transfer learning in sign language," in *Proc. CVPR*, 2007, pp. 1–8.
[22] P. Wang, C. Domeniconi, and J. Hu, "Using wikipedia for co-clustering based cross-domain text classification," *Data Mining*, pp. 1085–1090, 2008.
[23] Z. Sun, Y. Chen, J. Qi, and J. Liu, "Adaptive localization through transfer learning in indoor wi-fi environment," *Mach. Learn. Applicat.*, pp. 331–336, 2008.
[24] X. Wu, H. Wang, C. Liu, and Y. Jia, "Cross-view action recognition over heterogeneous feature spaces," in *Proc. ICCV*, 2013.
[25] R. Gopalan, R. Li, and R. Chellappa, "Domain adaptation for object recognition: An unsupervised approach," in *Proc. ICCV*, 2011, 2013, pp. 999–1006.
[26] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. CVPR*, 2012, pp. 2066–2073.
[27] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *Proc. ECCV*, 2010, pp. 213–226, Springer.
[28] A. Bergamo and L. Torresani, "Exploiting weakly-labeled web images to improve object classification: A domain adaptation approach," *Adv. Neural Inf. Process. Syst. (NIPS)*, 2010.
[29] B. Kulis, K. Saenko, and T. Darrell, "What you saw is not what you get: Domain adaptation using asymmetric kernel transforms," in *Proc. CVPR*, 2011, pp. 1785–1792.
[30] J. Yang, R. Yan, and A. Hauptmann, "Cross-domain video concept detection using adaptive SVMs," in *Proc. Int. Conf. Multimedia*, 2007, pp. 188–197.
[31] L. Duan, I. Tsang, D. Xu, and S. Maybank, "Domain transfer SVM for video concept detection," in *Proc. CVPR*, 2009, pp. 1375–1381.
[32] G. Doretto and Y. Yao, "Boosting for transfer learning with multiple auxiliary domains," in *Proc. CVPR*, 2010.
[33] G. Schweikert, C. Widmer, B. Schölkopf, and G. Rätsch, "An empirical analysis of domain adaptation algorithms for genomic sequence analysis," in *Proc. NIPS*, 2009.
[34] K. Tang, V. Ramanathan, L. Fei-Fei, and D. Koller, "Shifting weights: Adapting object detectors from image to video," *Adv. Neural Inf. Process. Syst.*, pp. 647–655, 2012.
[35] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.
[36] D. Meyer, F. Leisch, and K. Hornik, "The support vector machine under test," *Neurocomputing*, vol. 55, no. 1, pp. 169–186, 2003.
[37] 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. Discov. Data (TKDD)*, vol. 6, no. 4, p. 18, 2012.

Han Wang received the B.A. degree in computer science from National University of Defence Technology. She is currently a doctoral candidate in the School of Computer Science at the Beijing Institute of Technology. Her research interests include multimedia retrieval, machine learning and computer vision.

Xinxiao Wu received the Ph.D. degree in Computer Science from the Beijing Institute of Technology. She is currently a lecturer in the School of Computer Science at the Beijing Institute of Technology. Her research interests include machine learning, computer vision, video analysis and event recognition.

Yunde Jia received the Ph.D. degree in mechatronics from the Beijing Institute of Technology in 2000. He is currently a professor of computer science and also a director of the Lab of Media Computing and Intelligent Systems, School of Computer Science, Beijing Institute of Technology. His research interests include computer vision, media computing, human computer interaction and intelligent systems.