Finding Event Videos via Image Search Engine

Han Wang
School of Information Science and Technology
Beijing Forestry University
Beijing,China
wanghan@bjfu.edu.cn

Xinxiao Wu
School of Computer Science
Beijing Institute of Technology
Beijing,China
wuxinxiao@bit.edu.cn

***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 to model events under various circumstances. To alleviate the labeling process, we propose to learn models for videos by leveraging abundant Web images which contains a rich source of information with many events taken under various conditions and roughly annotated. However, knowledge from the Web is noisy and diverse, brute force knowledge transfer may hurt the retrieval performance. To address such negative transfer problem, we propose a novel Joint Group Weighting Learning (JGWL) framework to leverage different but related groups of knowledge (source domain) queried from the Web image searching engine to real-world videos (target domain). Under this framework, weights of different groups are learned in a joint optimization framework, and each weight represents how contributive the corresponding image group is to the knowledge transferred to the videos. Moreover, to deal with the feature distribution mismatching between video feature space and image feature space, we build a common feature subspace to bridge these two heterogeneous feature spaces in an unsupervised manner. Experimental results on two challenging video datasets demonstrate that it is effective to use grouped knowledge gained from Web images for video retrieval.**

***Index Terms*—video annotation; transfer learning; heterogeneous domain adaptation;**

I. Introduction

With ever expanding multimedia collections, video retrieval has found many applications ranging from Web searching to multimedia information delivery. Traditional simple text processing for video retrieval is generally insufficient, due to the lack of detailed textual annotation. In addition, the video in real world is highly unconstrained with significant camera motion and large intra-class variations, which makes finding desired events an extremely challenging task. However, it is known that collecting enough labeled videos covering a diverse set of conditions is time consuming and labor expensive. Since finding enough labeled videos is impossible, we try to seek another way to find labeled data and transfer the related knowledge from these data to videos. Fortunately, we find that the Web image searching engines, on the other hand, become increasingly mature and can offer abundant easily accessible knowledge. Moreover, the data collected from the Web are more diverse and less biased than home-grown datasets, which makes it more realistic for real-world tasks. This motivates us to acquire knowledge of videos by using labeled image data from the public available image searching engines (i.e. Google and Bing).

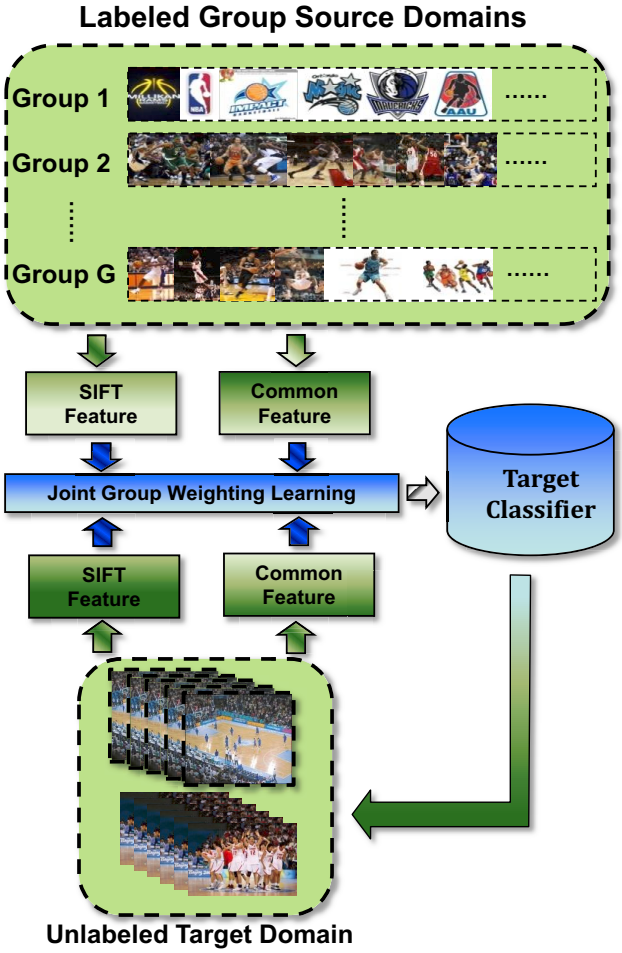

Fig. 1. Illustration of the framework.

To collect knowledge from the Web, we decide to use keywords related to the event to query from the searching engine. Since the events in real world are complex and vary widely, single query word is not sufficient to describe the complicated situations. Taking the event "basketball" for example, we may associate it with the keywords of "basketball match", "NBA", "Kobe", and "basketball dancer", etc. For

978-1-4673-8493-3/15 $31.00 © 2015 IEEE
DOI 10.1109/ICDMW.2015.78

each keyword, we collect a group of related Web images. For each event with multiple keywords, multiple groups of images are collected by querying the Web image searching engines.

Though learning from a large range of Web knowledge is valuable, noise images of little relevance with the events still exist due to random notes and subjective understanding. Under this circumstance, brute force transferring in case of weak relationships may lead to performance deterioration of the resulting classifiers, which is so-called negative transfer. Therefore, it is necessary to effectively summarize knowledge and pick the most relevant pieces to transfer. One strategy to decrease the risk of negative transfer is assigning different weights to different groups of knowledge based on their relevance to the target events. In this paper, we compute weights for all groups of each event in a joint manner to increase the chance of borrowing target domain related knowledge.

Since the Web image data and video data are represented by heterogenous features with different dimensions, it is essential to design a translator for bridging these two heterogeneous feature spaces. In this work, we introduce Canonical Correlation Analysis (CCA) to transfer knowledge from Web image data (source domain) to video data (target domain). By applying CCA, the image features from source domain and the video features from target domain are projected into a common intermediate subspace via their corresponding projection matrices. Thanks to the common feature subspace, the classifiers learned on the source domain can be adapted to the target domain.

Fig.1 illustrates our Joint Group Weighting Learning (JGWL) framework. The main contributions of our work are as follows:

(1) We propose to collect labeled image groups by querying different associational keywords from the Web image searching engines and analyze videos by leveraging different aspects of knowledge transferred from these images.

(2) We develop a novel Joint Group Weighting Learning (JGWL) method to enforce the relationship among different groups, as well as the relationship between groups and the target.

(3) We propose to use a new common feature subspace to incorporate heterogeneous feature spaces in transfer learning procedure.

II. Related work

Transfer learning has been deployed over a wide variety of applications, such as sign language recognition [1], text classification [2], and WiFi localization [3]. Recently, knowledge transfer in multimedia content analysis has attracted more and more attentions of researchers. Yang et al. [4] proposed Adaptive SVM (A-SVM) to enhance the prediction performance of video concept detection , in which the new SVM classifier is adapted from an existing classifier trained from the source domain. Duan et al. [5] proposed to simultaneously learn the optimal linear combination of base kernels and the target classifier by minimizing a regularized structural risk functional. And then, they proposed A-MKL [6] to add the prelearned 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 domain adaptation methods [7], [8], [9] are proposed,which leverage different precomputed classifiers learned from multiple source domains. However, these methods take no account on intrinsic connections among the data within and between groups. And weights are assigned to each independent source according to the difference between the target and particular source only. In this paper, we propose to leverage different groups of images queried by different associational keywords to the Web. By this means, we insure that the data in each group are of the same concept, and also insure that different groups within the same event are correlated to each other.

A few works have been done on investigation of the knowledge transform from image to video. In [10], transfer models are learned from loosely labeled web images. This work cannot distinguish actions like “standing-up” and “sitting-down” because it does not utilize temporal information of actions in the image-based model. Recently, Duan et al.[8] developed a new event recognition approach for consumer videos by using web images. In their work motion features and image features are integrated in a target decision function to jointly determine events in videos. Among these methods, features in different spaces are used separately, ignoring the potential connections among different feature spaces.

To adapt classifiers in different feature spaces, one simple way is to translate all the training data into a target feature space. The idea has already been demonstrated successfulness in several applications such as cross-lingual text classification [11]. Recently, to solve more general translated learning problem such as the translation between documents and images, Canonical Correlation Analysis (CCA) [12], [13] is introduced to capture the relationship between heterogeneous features. In this paper, we apply CCA to translate image features and video features to a common feature subspace. Thus, classifiers learned on this common subspace can be adapted in both domains. To our best knowledge, it is the first time for CCA to be applied in connecting the video features and the image features.

III. Problem statement

A. Motivation

Our aim is to improve the learning of the target predictive function $f_t(\cdot)$ using the knowledge in both source and target domain. In our learning scheme, we assume that some unlabeled data in target-domain can be seen at the training stage, which is so-called *transductive learning*. Under such setting, one can adapt the predictive function learned in the source domain for use in the target domain through some unlabeled target-domain data.

Our method leverages different groups of source data and gradually adapts them to the target classifier by building a common feature subspace for two heterogeneous features. As a result we can take advantage of useful groups of knowledge

from various available sources which are found to be the most closely related to the target. Our work not only provides a framework for automatically discovering which part of knowledge is helpful for the target domain or tasks, but also provides an effective way to transfer this knowledge from the source to the target.

B. Problem description

We apply different associational keywords for each event to multiple image sets from source domains. Here we refer to an image set returned by one keyword as a *group*, and each group represents one concept of the event. Resort to multiple groups, we have knowledge of the events covering different concepts. We define the g-th group data as $X^g = \{\mathbf{x}_i^g\}_{i=1}^{N_g}$, where $g \in \{1, ..., G\}$ and $\mathbf{x}_i^g \in \mathbb{R}^{d_s}$ is the i-th image in the g-th group. d_s represents the dimensionality of the source domain features and N_g represents the number of images in the g-th group. In addition, we define m unlabeled videos in target domain as $X^t = \{\mathbf{x}_i^t\}_{i=1}^{N_t}$, where $\mathbf{x}_i^t \in \mathbb{R}^{d_t}$ is the i-th video in the target domain and d_t represents the dimensionality of target domain features.

IV. Transferring knowledge from Web images to videos

In this section we first introduce how to build a common feature subspace for image and video domains, and then describe the Joint Group Weighting Learning method for leveraging different groups of knowledge.

A. Building common feature subspace

As image features and video features are from two heterogeneous feature spaces, classifiers learned on images cannot be directly applied to videos. To overcome this inapplicability, we introduce a common space for the source and the target data, in which heterogeneous features from these two domains can be compared. Any source sample (Web images) and any target sample (videos) can be projected onto this common space by using two projection matrices $w_s \in \mathbb{R}^{d_c \times d_s}$ and $w_t \in \mathbb{R}^{d_c \times d_t}$, respectively. Here d_c is the dimension of common feature. Specifically, we define a general function by

$$\psi(x) = \begin{cases} \phi(x) = w_s x & \text{if } x \in \mathbb{R}^{d_s} \\ \varphi(x) = w_t x & \text{if } x \in \mathbb{R}^{d_t} \end{cases}, \tag{1}$$

where x is wither the source sample or the target sample. Canonical Correlation Analysis (CCA) is a technique for joint dimensionality reduction across two (or more) feature spaces, providing heterogeneous representation of the same instance. The assumption is that the representations in these two spaces contain some joint information that is reflected in correlations between them. Note that CCA is a supervised method, but videos in the target domain are unlabeled. So we cannot directly use CCA to connect features from these two domains. Fortunately, there is a correspondence between video and its keyframe, which provides a natural correlation between image/keyframe feature space and video feature space. Formally, given N samples of paired data $\{(I_1, V_1), ..., (I_N, V_N)\}$, where each $I_i \in \mathbb{R}^m$ and $V_i \in \mathbb{R}^n$ denote the image feature space and video feature space,respectively. The goal is to learn two projection matrices $w_s \in \mathbb{R}^m$ and $w_t \in \mathbb{R}^n$ by maximizing the canonical correlation:

$$\begin{aligned} &\max_{w_s, w_t} \frac{\widehat{E}[\langle I, w_s\rangle \langle V, w_t\rangle]}{\sqrt{\widehat{E}[\langle I, w_s\rangle^2]\widehat{E}[\langle V, w_t\rangle^2]}} \\ &= \max_{w_s, w_t} \frac{{w_s}^T C_{st} w_t}{\sqrt{{w_s}^T C_{II} w_s {w_t}^T C_{VV} w_t}}, \end{aligned} \tag{2}$$

where $\widehat{E}$ denotes the empirical expectation. C_{IV} denotes the between-sets covariance matrix. C_{II} and C_{VV} denote the auto-covariance matrices for image domain I and video domain V, respectively. It is worth to mention that the proposed common subspace feature for the source and target samples can be readily incorporated into different methods, making these methods applicable for the transductive learning problem.

B. Training pre-learned classifiers

In this section, we discuss the pre-learned classifiers on the G groups of query images. Formally, we define the g-th classifier $f_s^g(x^{s,g})$ in source domain as

$$f_s^g(x^{s,g}) = w_1 \psi(x^{s,g}) + w_2 \upsilon(x^{s,g}), \tag{3}$$

where $\mathbf{w} = [w_1; w_2]$ is the weighting parameter. $x^{s,g}$ is the image of the g-th group from the source domain. $\psi(x^{s,g})$ and $\upsilon(x^{s,g})$ are the common feature and the SIFT feature for $x^{s,g}$, respectively.

Note that the distribution of the image/keyframe features in target-domain videos are different from those in source-domain images. Therefore, we employ DASVM [14] for training pre-learned group classifiers to make them more adaptive to the target domain. In DASVM [14], source-domain samples are only used for initializing the pre-learned classifiers of the target-domain problem. After initialization, the source domain samples are gradually replaced by the target domain samples which are used to learn the final separation hyperplane.

C. Jointly learning group weights

We propose a novel Joint Group Weighting Learning scheme to integrate the pre-learned classifiers of all the source groups into the target classifier. The target classifier of our Joint Group Weighting Learning method is defined as

$$f_t(x) = \sum_{g=1}^{G} \alpha_g f_s^g(x), \tag{4}$$

where $\alpha_g > 0$ is the weight for the g-th group. We assume that the weights are normalized, that is, $\sum_{g=1}^{G} \alpha_g = 1$.

Based on the smoothness assumption for different groups, we minimize both the loss of the labeled source data and the difference between different group classifiers on the unlabeled target data. The proposed framework is given by

$$\min_{f_t} \Omega(f_t) + \lambda_L \Omega_L(f_t) + \lambda_T \Omega_T(f_t) + \lambda_G \Omega_G(f_t), \tag{5}$$

Algorithm 1 Joint Group Weighting Learning.

Input:

$\{X^g\}_{g=1}^G$: the set of image groups;

X^t : unlabeled target videos;

Output:

$\{f^g\}_{g=1}^G$: pre-learned classifiers;

$\{\alpha_g\}_{g=1}^G$: group weights.

Phase-I

1: **Initialize:** $\alpha_g = 1/G$;

2: Get initial source classifier $f_s^{g(0)}(x)$ using standard SVMs

3: Set $i = 1$

4: **repeat**

5: Calculate $f_s^{g(i)}(x)$ for all target and source samples

6: Choose largest τ target samples falling in margin band $\mathcal{M} = \{x| - 1 \leq f_s^{g(i)}(x^t) \leq 1\}$

7: Remove largest τ source data according to the g-th group classifier $f_s^{g(i)}(x^s)$

8: Update $f_s^{g(i)}$

9: $i = i + 1$

10: **until** Convergence

Phase-II

11: Initialize target classifier $f_t = \sum_{g=1}^G \alpha_g f_s^g(x)$

12: Use Quadratic Programming to minimize (9) to obtain α_g

13: **return** f_s^g and α_g

where $\lambda_L, \lambda_G, \lambda_T > 0$ are tradeoff parameters. The details of each term in Eq. (5) are described in the following.

$\Omega(f_t) = \frac{1}{2} \parallel \alpha \parallel^2$ controls the complexity of the target classifier f_t, where $\alpha = [\alpha_1, \alpha_2, ..., \alpha_g]$ are the weights of all the groups.

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

$$\Omega_L(f_t) = \sum_{i=1}^{N_s} \parallel f_t(x_i^s) - y_i^s \parallel^2, \tag{6}$$

where x_i^s is the i-th Web image, y_i^s is the event label of x_i^s and N_s is the total number of the event. This regularizer enforces the decision value of the target classifier f_t on the source domain similar to the event label of the ground truth.

$\Omega_G(f_t)$ is a group loss function based on the smoothness assumption on target domain data, parameterized as

$$\Omega_G(f_t) = \sum_{i=1}^{N_t} \sum_{g=1}^{G} \alpha_g \sum_{k=1, k\neq g}^{G} \parallel f_s^k(x_i^t) - f_s^g(x_i^t) \parallel^2 . \tag{7}$$

This loss function expects different groups belonging to the same event to have similar decision values. For example, if two groups g and k are from the same event, we ensure that $f_s^k(x)$ is close to $f_s^g(x)$. Actually, we introduce $\Omega_G(f_t)$ to penalize those groups far from majority event-related groups.

So far, the decision function is learned only with the labeled source domain data and a limited number of pseudo labeled target data. The target function may overfitting on these data, and the generalization ability may be degraded. As shown in the traditional transductive learning methods [15], [16], unlabeled data can be employed to improve the classification performance. In our framework, we also use the instances in the target domain to enhance the generalization ability of the classification model. So we have the target data-driven regularizer formulated by

$$\Omega_T(f_t) = \sum_{i=1}^{N_t} \parallel f_t(x_i^t) \parallel^2, \tag{8}$$

where N_t is the number of unlabeled target samples.

Putting everything together, we have the following optimization problem:

$$\begin{aligned} \min_{\alpha} \frac{1}{2} \parallel \alpha \parallel^2 + \lambda_L \sum_{i=1}^{N_s} \parallel f_t(x_i^s) - y_i^s \parallel^2 + \lambda_T \sum_{i=1}^{N_t} \parallel f_t(x_i^t) \parallel^2 \\ + \lambda_G \sum_{i=1}^{N_t} \sum_{g=1}^{G} \alpha_g \sum_{k=1, k\neq g}^{G} \parallel f_s^k(x_i^t) - f_s^g(x_i^t) \parallel^2 \\ s.t. \sum_{g=1}^{G} \alpha_g = 1, \alpha_g > 0 \end{aligned} \tag{9}$$

The optimization problem of Eq. (9) can be solved by a standard Quadratic Programming.

The algorithm framework is summarized in Algorithm 1. In Phase-I of training stage, we train the pre-learned classifiers for each individual group in the source domain. In Phase-II of training stage, weights are assigned to different groups in a joint manner based on Eq. 9 to generate the target classifier f^t. In testing stage, for each testing video the static image features are extracted from the keyframes, and the motion feature is represented by the common features. We input these features to the target classifier f_t to obtain the final event label.

V. Experiments

A. Datasets

We use Google image searching engine to collect a large number of knowledge for all events. Each event is represented by five groups of knowledge. Specifically, each group is obtained by querying an associational keyword of the event from the image searching engine. Fig. 2 shows some examples of grouped images for the basketball event. The first column of the figure is the associational keywords we used for the basketball event and each row corresponds to an associational keyword. We collect Web images for thirteen events: basketball, baseball, soccer, iceskating, biking, swimming, graduation, birthday, wedding, show, parade, picnic. Table I lists the associational keywords for each event in our experiment. Each row of the table corresponds to the five groups of an event and each column is one group of associational keywords. For each image, we extract 128-dimensional SIFT features from salient regions detected by the Difference of Gaussians (DoG) detectors [17].

For target domain, we apply two real-world video datasets for performance evaluation:

Fig. 2. Example groups of basketball event.

Event	group 1	group 2	group 3	group 4	group 5
Baseball	baseball	baseballgames	foul line	softball	baseball American
Basketball	basketball	basket	NBA	rebound	basketball court
Biking	biking	bicycle	bike	cycle	biking clipart
Birthday	birthday	cake	candle	celebrate	birthday dinner
Graduation	graduation	finish school	scholar	graduation cap	graduate
Iceskating	iceskating	patinar	skate	skater	skating cartoon
Show	non-music performance	music performance	concert	show	acrobat performance
Parade	parade	demonstration	procession	protest	halloween parade
Skiing	skiing	skier	sled	slege	alpine skiing
Soccer	soccer	fifa	football	realmadrid	AC milan
Swimming	natation	swim	swimming	synchronised	aquatics
Wedding	wedding	wedding ceremony	wedding dance	wedding reception	wedding cake
Picnic	picnic	barbecue	cookout	food	dine together

TABLE I
EXAMPLES OF THE ASSOCIATIONAL QUERY KEYWORDS FOR EACH EVENT.

1) **CCV dataset:** This is a newly released consumer video dataset collected by Columbia University [18]. It contains a training set of $4,659$ videos and a test set of $4,658$ videos which are annotated to 20 semantic categories. Since our work focus on event analysis, 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 Web searching engine, we merge "wedding ceremony", "wedding reception" and "wedding dance" into one event as "wedding" and also merge "non-music performance" and "music performance" as "performance". Finally, there are twelve event types: "basketball", "baseball", "soccer", "iceskating", "biking", "swimming", "graduation", "birthday", "wedding", "show", "parade".

2) **Kodak dataset:** This dataset was collected by Kodak [19] from about 100 real users over the period of one year. There are $1,358$ consumer video clips in the Kodak dataset. Also, we only consider six event categories (i.e., "wedding", " birthday", "picnic", "parade", "show" and "sports") in our experiments.

For each video, we apply two types of features: motion feature and static feature. For motion features, we extract 144-dimensional 3D Space-Time Interest Point (STIP) [18] on the CCV dataset and use 96-dimensional Histograms of Oriented Gradients (HOG) and 108-dimensional Histograms of Optical Flow (HOF) [6] on the Kodak.

B. Experimental setup

In the experiments, we use the bag-of-words for both static and motion features. Specifically, we cluster the SIFT features, extracted from all the training Web images and keyframes of the videos, into $2,000$ words by using k-means clustering. Each image/video keyframe is then represented as a $2,000$-dimensional token frequency (TF) feature by quantizing its SIFT features with respect to the visual codebook. For the videos in both datasets, we directly use the 5000-dimensional and 2000-dimensional features provided by [18] and [6] for CCV dataset and Kodak dateset, respectively.

For a given event, we use five associational keywords as query to search relevant images, and collect the original top 300 images for each query keyword. To train an initial pre-learned classifier for each group, we use the queried 300 images in the corresponding group as positive samples and randomly select 300 images from other groups as negative samples. At the training stage, for CCV dataset we use the training set defined by work in [18] and the total 195 videos for Kodak dataset as our unlabeled target domain. Then we have labeled Web image groups in the source domain and unlabeled videos in the target domain to form our training set.

We compare our method with standard SVM, Domain Adaptation SVM (DASVM) [14], and Domain Selection Machine(DSM) [8], as these methods can work when there is no labeled training data in target domain. The standard SVM and DASVM can transfer knowledge only when the source

data and the target data are with the same type of features. Therefore, they can only use the static features to learn classifiers for the target domain. Specifically, since we do not have any labeled data in the target domain, we learn one classifier from each group for standard SVM. The SVM and DASVM could not handle multiple groups, and the final results are equally fusion all the source classifiers. For DASVM, we obtained the initial classifier by the labeled data in one group and gradually adapt it using the unlabeled videos from the target domain. Similar to the standard SVM, we equally fuse the DASVM classifiers from all five grouped source domains to obtain the final results. For the DSM, we use the non-linear χ^2 kernel and average the decision values of the video keyframes by using pre-learned SVM classifier to generate the prediction for each video.

For performance evaluation, we use the Average Precision (AP) as in [20] and define mean Average Precision (mAP) as the mean of APs over all events.

C. Results

We first compare our method with existing approaches on CCV dataset and Kodak dataset. We report the per-event APs of all methods on both datasets in Fig.3 and Fig.4, respectively. We also show the mAPs of all methods on these datasets in Table II.

Method	SVM	DASVM [14]	DSM [8]	Ours
CCV	8.52	10.90	12.63	13.52
Kodak	23.29	28.63	31.54	32.70

TABLE II
COMPARISON OF MAPS (%) BETWEEN OUR METHOD AND OTHER METHODS ON CCV AND KODAK DATASETS.

On the CCV datset, our method can achieve the relative improvements of 46.95%, 24.03% and 7% over the methods of standard SVM, DASVM, DSM, respectively. One the Kodak dataset, the relative improvements of our method are 40.40%, 14.21% and 3.7% over the methods of standard SVM, DASVM, DSM, respectively.

From the results, we observe that:

1) Our method achieves the best results on both datasets. This leads us to believe that jointly learning different groups of knowledge is beneficial to performing positive transform.

2) DSM and our method outperform the other two methods, which clearly demonstrates that it is helpful to employ multiple groups for knowledge transfer. A possible explanation is that the concepts in real-world events vary dramatically, so single query word cannot stands for all cases.

3) Our method is better than DSM on mAPs, which illustrates the benefit of using grouped event related images returned from Web searching engine and associating different weights to different groups with continuous values rather than binary assignment. We can also deduce that the data from all groups querying by associational keywords can help the final recognition results. With the help of image searching engines, all the groups can be used to help understand video events.

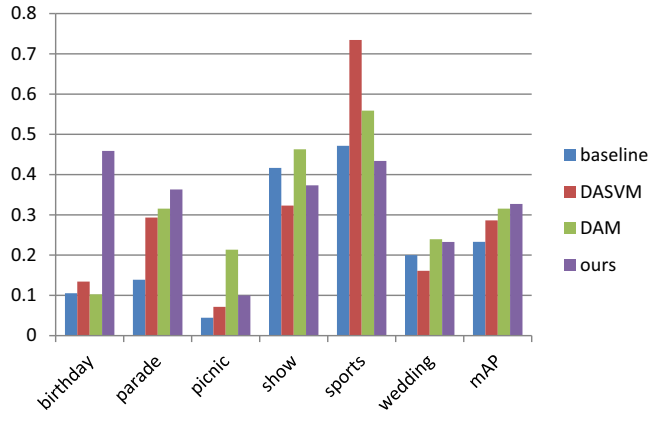

Fig. 4. Per-event Average Precisions (APs) of all methods on Kodak dataset.

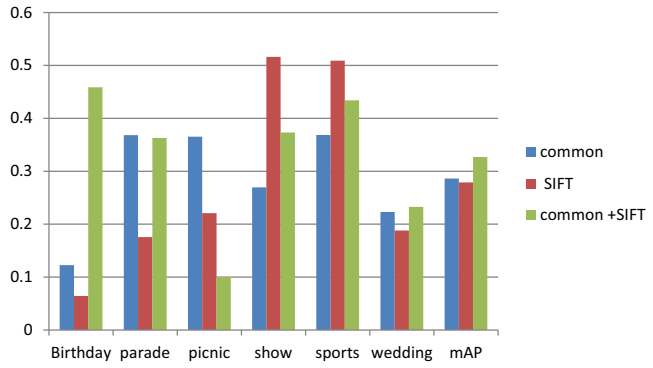

Fig. 6. Evaluation of different features on Kodak dataset.

4) In terms of per-event average precisions, there is no consistent winner among the four methods. This indicates that there exits some irrelevant data hinds these transfer learning methods to acquire good target classifier. Our method achieves more stable performance, which demonstrates that jointly weighting different groups can cope with noisy web images.

We then evaluate the proposed common feature subspace of the heterogeneous features of the source and target domains on both CCV and Kodak datasets, respectively. Fig.5 and Fig.6 shows per-event Average Precision results by using the SIFT feature and the common feature independently, as well as the combination of these two features. The last three bins show the mAP results on these these feature representation schemes. For average performance on most events, we observe that the proposed common feature achieves comparable and stable performance. Especially for those events closely related to motion such as “birthday”, “parade” and “picnic”, the common feature plays a more important role in significantly improving the performance. The best results appear when we integrate SIFT feature and common feature in one learning processing, which obviously demonstrates the beneficial of combining both static and motion features.

To verify the influence of the number of groups in our

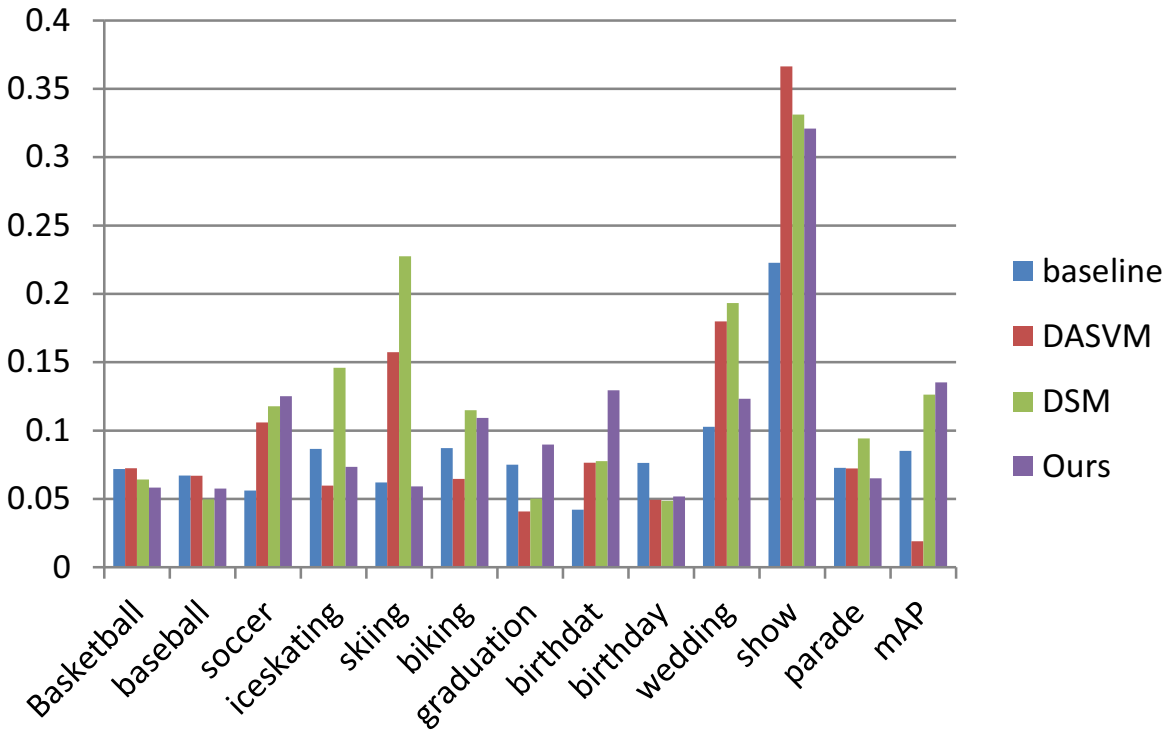

Fig. 3. Per-event Average Precisions (APs) of all methods on CCV dataset.

method, we vary the number of groups during training. As shown in Fig.7, the results of multiple groups consistently perform better than that on single group. This indicates that our group weighting scheme could gain useful information to guide video retrieval. We also notice that the mAPs do not monotonically increase with the increase group numbers. The possible explanation is that the ability to transfer knowledge from source to target depends on how they are related. The transferred knowledge becomes more useful as the relationships are stronger.

We also investigate the effects of each term in our optimization function in (9) for learning weights of groups. Table III shows the results when $\lambda_G = 0$ and $\lambda_T = 0$, respectively. From the results, we can observe that the mAP dramatically decreases when either $\Omega_G(f^t)$ or $\Omega_T(f^t)$ is removed from the optimization function. In Table III, we also show the result in case the weights of all the groups are the same, i.e. $\alpha_g = \frac{1}{G}$. The results firmly demonstrate the importance of leveraging different group with different weights.

	$\lambda_G = 0$	$\lambda_T = 0$	$\alpha_g = 1/G$	Our method
CCV	10.30	10.13	11.40	13.52
Kodak	24.90	20.90	19.79	32.70

TABLE III
EVALUATION ON DIFFERENT COMPONENTS OF THE OPTIMAL FUNCTION USING MAPS (%).

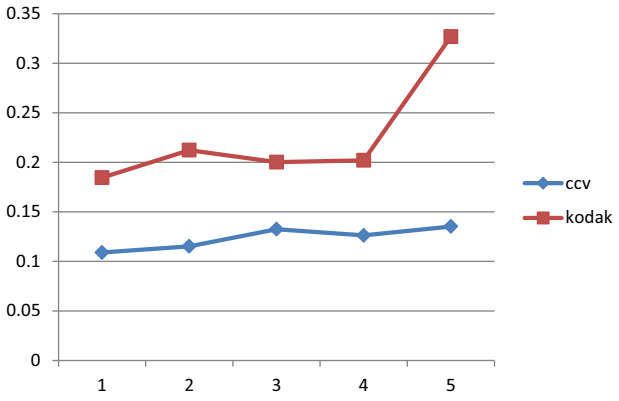

Fig. 7. Evaluation on different group numbers.

VI. CONCLUSION

In this paper, we have proposed a novel Jointly Group Weighting Learning (JGWL) method to utilize different groups of Web images to search unlabeled real-world video clips. In our framework, we divide the image data into different groups by querying different associational keywords to the Web image searching engine. We further assign weights to different groups in a joint manner which takes account of correlations among the source groups, as well as the correlations between the source and the target. From the experimental results, we

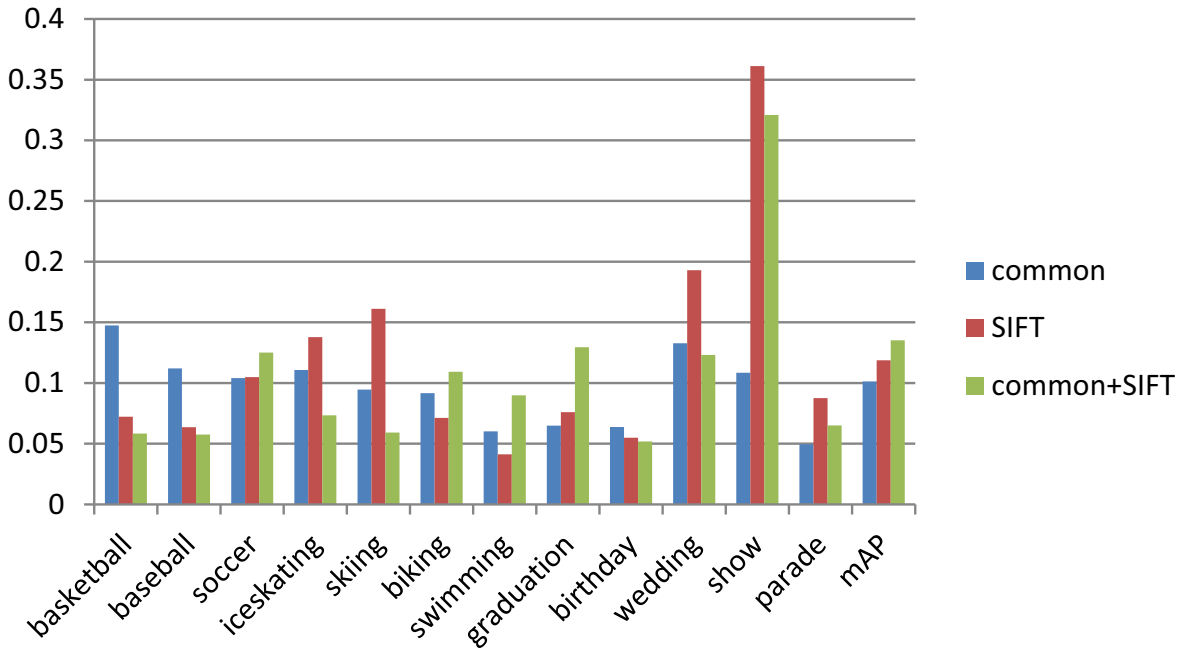

Fig. 5. Evaluation of different features on CCV dataset.

have the following observations: (1) Leveraging knowledge from various related aspects brings better results than that from only one-side. (2) When it comes to the problem of delivering knowledge from image to video, the connection between static and motion features should be considered to boost the performance. (3) Assigning different weights to different source groups is critical to the performance, and our assignment scheme has proved to be effective.

In our work, we assume the feature spaces is linear, and non linear kernel will be considered in generating intermediate feature space to get more robust representation. In future work, we also wish to develop a mechanism for automatically re-filtering images to obtain more clean source data.

ACKNOWLEDGMENT

This work was supported by the Fundamental Research Funds for the Central Universities with Grant No. BLX2014-28 and the Natural Science Foundation of China under Grant 61203274.

REFERENCES

[1] A. Farhadi, D. Forsyth, and R. White, "Transfer learning in sign language," in *CVPR*, 2007, pp. 1–8.
[2] P. Wang, C. Domeniconi, and J. Hu, "Using wikipedia for co-clustering based cross-domain text classification," in *Data Mining*, 2008, pp. 1085–1090.
[3] Z. Sun, Y. Chen, J. Qi, and J. Liu, "Adaptive localization through transfer learning in indoor wi-fi environment," in *Machine Learning and Applications*, 2008, pp. 331–336.
[4] J. Yang, R. Yan, and A. Hauptmann, "Cross-domain video concept detection using adaptive svms," in *International Conference on Multimedia*, 2007, pp. 188–197.
[5] L. Duan, I. Tsang, D. Xu, and S. Maybank, "Domain transfer svm for video concept detection," in *CVPR*, 2009, pp. 1375–1381.
[6] L. Duan, D. Xu, I. Tsang, and J. Luo, "Visual event recognition in videos by learning from web data," in *CVPR*, 2010, pp. 1959–1966.
[7] G. Doretto and Y. Yao, "Boosting for transfer learning with multipple auxiliary domains," in *CVPR*, 2010.
[8] 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 *CVPR*, 2012, pp. 1959–1966.
[9] G. Schweikert, C. Widmer, B. Schölkopf, and G. Rätsch, "An empirical analysis of domain adaptation algorithms for genomic sequence analysis," in *NIPS*, 2009.
[10] N. Ikizler-Cinbis, R. Cinbis, and S. Sclaroff, "Learning actions from the web," in *CVPR*, 2009, pp. 995–1002.
[11] N. Bel, C. Koster, and M. Villegas, "Cross-lingual text categorization," *Research and Advanced Technology for Digital Libraries*, pp. 126–139, 2003.
[12] D. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.
[13] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *ICMM*, 2010, pp. 251–260.
[14] L. Bruzzone and M. Marconcini, "Domain adaptation problems: A dasvm classification technique and a circular validation strategy," *PAMI*, vol. 32, no. 5, pp. 770–787, 2010.
[15] S. Pan and Q. Yang, "A survey on transfer learning," *Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.
[16] T. Joachims, "Transductive inference for text classification using support vector machines," in *Work shop on Machine Learning-International*, 1999, pp. 200–209.
[17] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.
[18] 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 *ICMR*, 2011, p. 29.
[19] 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 *Workshop on Multimedia Information Retrieval*, 2007, pp. 245–254.
[20] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *CVPR*, 2008, pp. 1–8.