Multi-Group Adaptation for Event Recognition from Videos

Yang Feng, Xinxiao Wu, Han Wang and Jing Liu
Beijing Laboratory of Intelligent Information Technology, School of Computer Science
Beijing Institute of Technology, Beijing 100081, P.R. China
Email:{fengyangbit,wuxinxiao,wanghan,liu_jing}@bit.edu.cn

***Abstract*—Recognizing events in consumer videos is becoming increasingly important because of the enormous growth of consumer videos in recent years. Current researches mainly focus on learning from numerous labeled videos, which is time consuming and labor expensive due to labeling the consumer videos. To alleviate the labeling process, we utilize a large number of loosely labeled Web videos (e.g., from YouTube) for visual event recognition in consumer videos. Web videos are noisy and diverse, so brute force transfer of Web videos to consumer videos may hurt the performance. To address such a negative transfer problem, we propose a novel *Multi-Group Adaptation (MGA)* framework to divide the training Web videos into several semantic groups and seek the optimal weight of each group. Each weight represents how relative the corresponding group is to the consumer domain. The final classifier for event recognition is learned using the weighted combination of classifiers learned from Web videos and enforced to be smooth on the consumer domain. Comprehensive experiments on three real-world consumer video datasets demonstrate the effectiveness of *MGA* for event recognition in consumer videos.**

***Keywords*-video event recognition; transfer learning.**

I. Introduction

With the widespread of digital cameras and mobile phone cameras, ordinary users are capturing much more personal videos in daily life. Automatically recognizing events in consumer videos has become an important research topic due to its usefulness in video management and retrieval. However, consumer videos are generally captured by amateurs using hand-held cameras and they contain considerable camera motions and occlusions. There are also large intra-class variations within the same type of event videos, making it very challenging to recognize events in consumer videos.

In most pervious work [1, 2, 3, 4], a large number of training videos are collected and labeled first. Then, a robust model for event recognition in consumer videos is learned from these data. However, collecting enough labeled videos is labor expensive and time consuming. Several examples have shown the effectiveness of domain adaptation which enables target classifier to be trained using auxiliary data [5]. As search engines have become increasingly mature and they can offer abundant data with loose labels, researchers have begun collecting training data from searching on the Web instead of manual labeling. Duan et al. [6] proposed Adaptive Multiple Kernel Learning (A-MKL) method which transfers knowledge from YouTube videos to consumer domain by minimizing the structural risk function and the mismatch between the data distributions. In [7], Ikizler-Cinbis et al. collected images from the Web to learn representation of actions and used this knowledge to automatically annotate actions in videos. Recently, Chen et al. [8] proposed a new method called Multi-domain Adaptation with Heterogeneous Sources (MDA-HS) to learn an optimal target classifier, in which different source domains with different types of features can be integrated.

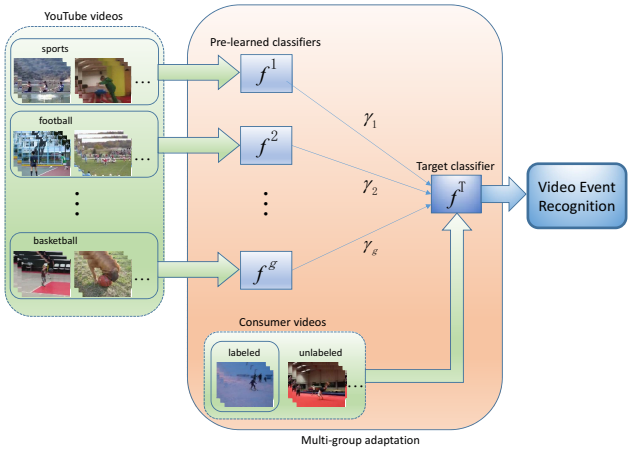

Figure 1. Illustration of our framework. Several classifier are trained from YouTube videos independently first. Then, the weights of pre-learned classifiers are optimized in a joint manner. Finally, robust classifier is learned for video event recognition.

In this paper, we aim to recognize visual events in consumer videos by utilizing a large number of loosely labeled Web videos and a few labeled consumer videos. Since the events in consumer videos are complex and various, videos returned by searching with single query word is not sufficient to describe the event. We use several event-related keywords to query videos from search engines. For example, we may associate the event "sports" with the keywords of "football", "baseball" and "basketball", etc. For each keyword, we collect a set of related Web videos regarded as a *group*. Several researchers also obtained training videos from YouTube, but they did not divide the videos into groups. Jiang et al. [9] released a database called CCV, containing 9,317 Web videos which were collected by searching a string "MVI" with each of the category names. In [10], Luo et al. also searched with several keywords for each concept, but the returned videos are mixed together to represent a concept.

Although it is easy to download a lot of Web videos by searching with different keywords, the obtained videos are always poor in quality. Either they are heavily compressed by

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.671

the Web server or they are not closely related to the keyword associated event, or even related to other events. Using the irrelevant videos to train the target classifier may be harmful to the classification performance, which is known as negative transfer. To decrease the risk of negative transfer, we propose a Multi-Group Adaptation (MGA) method which aims to make full use of Web videos by assigning different weights to different source video groups based on their relevances to target events. To optimize the weight of each group, a joint weighting scheme is introduced based on smoothness assumption, which enforces that similar consumer videos have similar decision values and positive labeled videos have higher decision values than negative labeled videos. Finally, the target classifier for event recognition is learned using the weighted combination of classifiers learned from the video groups and enforced to be smooth on the consumer domain. Fig. 1 illustrates our framework.

II. Related Work

Traditional machine learning algorithms are usually based on an assumption that the training data and test data must be drawn from the same distribution. These methods may fail when there are not enough training data in the domain we are interested in, but there are sufficient data in related domains. Several domain adaptation methods have been proposed and have succeeded in many computer vision tasks such as object recognition [11] and video concept detection [12]. In [12], Yang et al. proposed Adaptive SVM (A-SVM) to adapt concept classifiers across different video domains, in which the new classifier is the sum of an existing classifier trained from the source domain and a perturbation for the labeled target data. Wu et al. [13] proposed Heterogeneous Transfer Discriminat-analysis of Canonical Correlations (HTDCC) to transfer action models learned in source view to target view, where samples from these two views are represented by heterogeneous features. Bruzzone and Marconcini [14] proposed Domain Adaptation Support Vector Machine (DASVM) to iteratively label the unlabeled target domain data, removing some labeled source domain data at the same time. These methods are mainly designed to handle single source domain setting. When several source domains are available, researchers proposed multiple source domain adaptation methods. Duan et al. [15] proposed a new domain adaptation method called Domain Adaptation Machine (DAM), to learn target decision function by using a set of per-computed classifiers independently learnt from the labeled samples from multiple source domains. Wang et al. [16] proposed Group-based Domain Adaptation (GDA) in which multiple source domains of images are organized according to their intrinsic semantic relationships and weighted according to the relevances between the source and the target domain. In [17], Chattopadhyay et al. proposed a novel weighting scheme based on smoothness assumption on the probability distribution of the target domain data and the target classifier was learned using the weighted combination of source decision values.

Our framework is mainly related to two multi-source domain adaptation frameworks including Domain Adaptation Machine (DAM) [15] and Conditional Probability based Multi-Source Domain Adaptation (CP-MDA) [17]. In DAM, the weight assigned to each source domain is obtained by Maximum Mean Discrepancy (MMD) [18] which measures the marginal probability distribution difference between the target domain and source domain. The labels in source domains and connections among the source domains are not considered in the weighting scheme of DAM. In CP-MDA, the weights are optimized jointly by manifold based regularizer, and the precious labeled target data is not used. Our method computes wights for the source groups taking both the connections among the source groups and the labeled target data into account.

III. Proposed Framework

Following the terminology of domain adaptation, we refer to the Web video domain as the source domain, in which we have abundant loosely labeled videos, and the consumer video domain as the target domain, in which we have a few labeled and plenty of unlabeled videos. To obtain the training videos for one event, we search with several keywords associated with the event. We refer to the videos returned by one keyword search as a *group*. Consequently the videos retrieved from the search engine for one event are divided into different groups according to their corresponding search keywords. We define the s-th group of an event as $D^s = (\boldsymbol{x}_i^s, y_i^s)|_{i=1}^{n_s}, s \in \{1, ..., g\}$, where g is the total number of groups and n_s represents the number of videos in the s-th group. There are a few labeled data $D_l^T = (\boldsymbol{x}_i^T, y_i^T)|_{i=1}^{n_l}$ in the target domain where n_l is the numbers of labeled target domain samples. These labeled data consist of positive samples $D_+^T = (\boldsymbol{x}_i^T, 1)|_{i=1}^{n_+}$ and negative samples $D_-^T = (\boldsymbol{x}_i^T, -1)|_{i=n_++1}^{n_++n_-}$ where n_+ and n_- are the numbers of positive samples and negative samples, respectively, $D_l^T = D_+^T \bigcup D_-^T$ and $n_l = n_+ + n_-$. There are also plenty of unlabeled target data $D_u^T = \boldsymbol{x}_i^T|_{i=n_l+1}^{n_l+n_u}$ available, where n_u is the number of unlabeled target samples, $D^T = D_l^T \bigcup D_u^T$ and $n_T = n_l + n_u$.

A. Motivation

In many domain adaptation problems, a few labeled target data are available. If we fully exploit the labeled target data, the performance may improve. In DAM [15] and CP-MDA [17], the labeled target data are only used when training the final target classifier. In the weighting stage, DAM and CP-MDA apply MMD [18] and smoothness assumption to calculate the weight of each source domain respectively. We propose to use both the label information and smoothness assumption to determine the weights for all the groups. In our problem, some groups may closely relate to the target domain, while others may not. We can evaluate whether a group is related to the target domain by the decision values of the labeled target data given by the classifier learned from the group. For example, if the classifier of one group outputs much higher decision values on the negative samples than the positive samples, we can believe this group is not a good one and we assign low weight to the group.

The target decision function in CP-MDA is non-sparse and involves the matrix inversion, making it very inefficient when the number of target data is large. Using the ϵ-insensitive loss function in Support Vector Regression (SVR) can usually lead to a sparse representation of the decision function. Therefore, we use ϵ-insensitive loss function in our framework and build a formulation which can be solved efficiently.

B. Smoothness Assumption for Multi-Group Weighting

In Manifold Regularization [19], Belkin et al. proposed Laplacian regularizer which enforces that the decision function is smooth on the data manifold, namely, similar instances in a high-density region should have similar decision values. Chattopadhyay et al. [17] applied this assumption to estimate the relevance between the source domain and the target domain. Let us define γ_s as the weight for measuring the relevance between the s-th group and the target domain. The weights are normalized and no negative value is allowed, that is, $\sum_s \gamma_s = 1, \gamma_s \geq 0$. Let $f_i^s = f^s(\boldsymbol{x}_i^T)$ be the decision values of the s-th group classifier on target domain data $\boldsymbol{x}_i^T$. We use a weighted combination of the source group classifiers to estimate the target classifier. Specifically, the estimated decision value ($\hat{y}_i$) of the unlabeled target data $\boldsymbol{x}_i^T$ based on the group classifiers $f^1 \cdots f^g$ is given by

$$\hat{y}_i = \sum_{s=1}^{g} \gamma_s f_i^s = \boldsymbol{F}_i^s \boldsymbol{\gamma}, \ i = n_l + 1 \cdots n_T, \tag{1}$$

where $\boldsymbol{F}_i^s = [f_i^1 \cdots f_i^g]$ and $\boldsymbol{\gamma} = [\gamma_1 \cdots \gamma_g]'$. Based on the smoothness assumption, Chattopadhyay et al. [17] proposed a weighting framework which solves the following problem:

$$\begin{aligned} \min_{\boldsymbol{\gamma}} & \sum_{i=n_l+1}^{n_T} \sum_{j=n_l+1}^{n_T} (\boldsymbol{F}_i^s \boldsymbol{\gamma} - \boldsymbol{F}_j^s \boldsymbol{\gamma})^2 W_{ij} \\ s.t. \ & \boldsymbol{\gamma} \geq 0, \boldsymbol{\gamma}' \mathbf{1} = 1, \end{aligned} \tag{2}$$

where W_{ij} is the edge weight between the i-th and the j-th samples. Eq. (2) can be rewritten as the graph Laplacian form:

$$\begin{aligned} \min_{\boldsymbol{\gamma}} & \boldsymbol{\gamma}' \left(\boldsymbol{F}_u^s\right)' \boldsymbol{L}_u \boldsymbol{F}_u^s \boldsymbol{\gamma} \\ s.t. \ & \boldsymbol{\gamma} \geq 0, \boldsymbol{\gamma}' \mathbf{1} = 1, \end{aligned}$$

where $\boldsymbol{F}_u^s = [\boldsymbol{F}_{n_l+1}^{s\prime} \cdots \boldsymbol{F}_{n_T}^{s\prime}]'$ and $\boldsymbol{L}_u$ is graph Laplacian of the unlabeled target domain data, which can be defined as $\boldsymbol{L}_u = \boldsymbol{D}_u - \boldsymbol{W}_u$, where $\boldsymbol{W}_u = (W_{ij})$ and $\boldsymbol{D}_u$ is a diagonal matrix given by $D_{ii} = \sum_{j=1}^{n} W_{ij}$.

Although the unlabeled target data are used in the above weighting framework, the precious labeled target data are not used. The few label information may improve the performance if properly used. So we propose a new weighting scheme with the objective function:

$$\begin{aligned} \min_{\boldsymbol{\gamma}} & \boldsymbol{\gamma}' \left(\boldsymbol{F}^s\right)' \boldsymbol{L} \boldsymbol{F}^s \boldsymbol{\gamma} + \lambda n_l (\boldsymbol{M}^- - \boldsymbol{M}^+) \boldsymbol{\gamma} \\ s.t. \ & \boldsymbol{\gamma} \geq 0, \boldsymbol{\gamma}' \mathbf{1} = 1, \end{aligned} \tag{3}$$

where $\boldsymbol{L}$ is graph Laplacian of all the target domain data, $\boldsymbol{F}^s = [\boldsymbol{F}_1^{s\prime} \cdots \boldsymbol{F}_{n_T}^{s\prime}]'$, $\boldsymbol{M}^- = [M_1^- \cdots M_g^-]$ and $\boldsymbol{M}^+ = [M_1^+ \cdots M_g^+]$ are 1×g vectors and λ is a trade off parameter. $M_s^- = \frac{1}{n_-} \sum_{i=n_++1}^{n_l} f_i^s$ and $M_s^+ = \frac{1}{n_+} \sum_{i=1}^{n_+} f_i^s$ are the mean of the decision values of negative and positive samples by the s-th group classifier, respectively. The second term in Eq. (3) encourages that the negative samples have lower decision values relative to positive samples. We notice that Li et al. [20] proposed meanS3VM, which also uses the label means. In meanS3VM, the label means are calculated on the unlabeled data and the objective is to maximize the margin between the label means. While in our method, label means are calculated on the labeled target data and our objective is to determine the weights of different groups. Eq. (3) can be solved using a standard quadratic programming solver. After $\boldsymbol{\gamma}$ is obtained, the estimated decision value of the unlabeled target data can be computed by Eq. (1).

C. Multi-Group Adaptation

Motivated by [15][17], we formally present the objective of our method as follows:

$$\min_{f^T \in \mathcal{H}_k} \frac{1}{2} \|f^T\|_{\mathcal{H}_k}^2 + C_l \Omega_l(f^T) + C_u \Omega_u(f^T) + \frac{\mu}{2} \Omega_r(f^T), \tag{4}$$

where $C_l, C_u, \mu > 0$ are parameters balancing different terms. The details of each term in Eq. (4) are described in the following. The first term is a regularizer to control the complexity of the target classifier f^T in the Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_k$. $\Omega_l(f^T)$ is the empirical error of the target classifier f^T on the labeled target data, defined by

$$\Omega_l(f^T) = \sum_{i=1}^{n_l} \ell_\epsilon(f^T(\boldsymbol{x}_i^T) - y_i^T),$$

where ℓ_ϵ is ϵ-insensitive loss: $\ell_\epsilon(t) = \begin{cases} |t| - \epsilon, & \text{if} |t| > \epsilon; \\ 0, & \text{otherwise.} \end{cases}$

$\Omega_u(f^T)$ is the empirical error on the unlabeled target data, defined by

$$\Omega_u(f^T) = \sum_{i=n_l+1}^{n_T} \ell_\epsilon(f^T(\boldsymbol{x}_i^T) - \hat{y}_i),$$

where $\hat{y}_i$ is the estimated decision value of the i-th target data. The forth term is the manifold regularizer:

$$\Omega_r(f^T) = \boldsymbol{f}^{T\prime} \boldsymbol{L} \boldsymbol{f}^T,$$

where $\boldsymbol{f}^T = [f^T(\boldsymbol{x}_1^T) \cdots f^T(\boldsymbol{x}_{n_T}^T)]'$ is target decision values of all the target data.

D. Detailed solution

Putting everything together, we arrive at the following optimization problem:

$$\begin{aligned} \min_{f^T \in \mathcal{H}_k} & \frac{1}{2} \|f^T\|_{\mathcal{H}_k}^2 + C_l \sum_{i=1}^{n_l} \ell_\epsilon(f^T(\boldsymbol{x}_i^T) - y_i) \\ & + C_u \sum_{i=n_l+1}^{n_T} \ell_\epsilon(f^T(\boldsymbol{x}_i^T) - \hat{y}_i) + \frac{\mu}{2} \boldsymbol{f}^{T\prime} \boldsymbol{L} \boldsymbol{f}^T. \end{aligned} \tag{5}$$

By the representer theorem, the optimal solution to the problem above is given by a liner expansion of the kernel function K over the target domain data:

$$f^T(\boldsymbol{x}) = \sum_{i=1}^{n_T} \alpha_i K(\boldsymbol{x}_i^T, \boldsymbol{x}). \tag{6}$$

After substituting Eq. (6) back to Eq. (5), relaxing the ϵ-insensitive loss and introducing Lagrange multipliers, Eq. (5) can be transformed to the following dual:

$$\begin{aligned} \min_{\boldsymbol{\beta},\boldsymbol{\beta}^*} & \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)'\boldsymbol{Q}(\boldsymbol{\beta}-\boldsymbol{\beta}^*) + \epsilon \sum_{i=1}^{n_T}(\beta_i+\beta_i^*) \\ & + \sum_{i=1}^{n_l} y_i(\beta_i-\beta_i^*) + \sum_{i=n_l+1}^{n_T} \hat{y}_i(\beta_i-\beta_i^*) \\ s.t. & \sum_{i=1}^{n_T}(\beta_i-\beta_i^*) = 0, \\ & 0 \le \beta_i, \beta_i^* \le C_l, i = 1 \cdots n_l, \\ & 0 \le \beta_i, \beta_i^* \le C_u, i = n_l+1 \cdots n_T, \end{aligned} \tag{7}$$

where $\boldsymbol{Q}$ is the transformation of kernel matrix $\boldsymbol{K}$:

$$\boldsymbol{Q} = \boldsymbol{K}(\boldsymbol{I} + \mu \boldsymbol{L}\boldsymbol{K})^{-1}.$$

Eq. (7) has the same form with standard ϵ-SVR except that the kernel matrix is replaced by $\boldsymbol{Q}$, and labeled target data and unlabeled target data are given different penalize parameters. After $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$ are solved, the final solution is obtained by solving the linear system $\boldsymbol{\alpha} = (\boldsymbol{I} + \mu \boldsymbol{L}\boldsymbol{K})^{-1}(\boldsymbol{\beta}^* - \boldsymbol{\beta})$. The final decision function is written as

$$f^T(\boldsymbol{x}) = \sum_{i=1}^{n_T} \alpha_i K(\boldsymbol{x}_i^T, \boldsymbol{x}) + b,$$

where b is the unregularized bias term. When the dataset is quite large, we can employ LapESVR [21] for efficiency which leads to a sparse solution of $\boldsymbol{\alpha}$ and avoids the matrix inversion.

IV. Experiments

We compare our work with the baseline method SVM, the existing single source domain adaptation algorithm (A-SVM [6] and DASVM [14]) , as well as the existing multi-domain adaptation methods (CP-MDA [17], DAM [15], DSM [22] and MDA-HS [8]). For all the methods, we use the Average Precision (AP) for performance evaluation and report the mean AP (mAP) over all events.

A. Datasets and Features

1) Kodak Dataset: The Kodak dataset was collected by Kodak [10] from about 100 real users over one year. We use the features provided by [6], which contains 195 consumer videos from six event classes (i.e., "birthday", "parade", "picnic", "show", "sports" and "wedding").

2) YouTube Dataset: The YouTube dataset we use is also provided by [6], which was collected using keywords based search from YouTube. In this dataset, there are 906 videos from six event classes (i.e., "birthday", "parade", "picnic", "show", "sports" and "wedding").

Figure 2. Examples in multi-group video dataset.

3) Columbia Consumer Video Dataset: This is a consumer video dataset collected by Columbia University [9]. 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 focuses on event analysis, we only use the videos from the event related categories. Following [22], we merge "wedding ceremony", "wedding reception" and "wedding dance" as "wedding", "non-music performance" and "music performance" as "show", and "baseball", "basketball", "biking", "ice skating", "skiing", "soccer", "swimming" as "sports". We use the training set defined by [9] in our experiments. Finally, there are 2,502 videos from five event classes (i.e., "birthday", "parade", "show", "sports" and "wedding").

4) Multi-group video dataset: We collect a large number of videos by keywords search from YouTube as our source domain data. We first need to define the associational keywords for each event. It is not a trivial task to choose the search keywords. To reduce the effect of personal experience, we refer to WordNet [23]. WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms, each expressing a distinct concept. We view sister term, hyponym and paraphrase of a event word and choose five close related words as keywords. Table I lists the associational keywords for each event in our experiment. Each row shows the five groups of an event. If we directly search with the keywords defined above, the results we get will contain many TV commercial videos. To ensure that only consumer videos are retrieved, queries are formed by combining a string "MVI" with each of the keyword as

Table I
QUERY KEYWORDS FOR EACH EVENT.

Event	Keywords				
Birthday	birthday	anniversary	celebration	birthday party	birthday cake
Parade	parade	march	demonstrate	procession	walk
Picnic	picnic	eat	breakfast	dinner	lunch
Sports	sports	athletic	football	baseball	basketball
Show	show	dance	display	nightlife	exhibit
Wedding	wedding	marriage	nuptial	wedding ceremony	marriage ceremony

suggested in [9]. We download the top ranked 100 videos from YouTube searches for each keyword. Some examples of our multi-group video dataset are shown in Fig. 2.

5) Features: For each video clip, we sample at a rate of 2 frames per second to extract keyframes. For each keyframe, we extract 128-dimensional SIFT features from salient regions, which are detected by Difference-of-Gaussian (DoG) interest point detector [24]. Then, we cluster the SIFT features extracted from all the keyframes of the training videos into 2,000 words by using k-means clustering. Each keyframe is represented by a 2,000-dim TF feature based on the bag-of-words representation. Each video is represented by the average of the TF features over all the keyframes within it.

B. Experimental Setups

In our experiments, the source training data consist of six events and each event has five groups. So we have 30 groups in total. We first train a pre-learned classifier for each group using the videos in the group as positive samples and randomly select samples from groups of other events as negative samples. The negative samples are selected from all the samples in 25 groups with equal probability $P_n = 0.05$. For Kodak and YouTube, we randomly sample one video from each event (6 videos in total) as the labeled training videos in the target domain, the rest videos are used as unlabeled data. For CCV, we randomly sample 15 videos per event (75 videos in total) as the labeled target videos. We sample the training videos for ten times and report the means and standard deviations of mAPs for each method. For the baseline SVM algorithm, we report the results for three cases: 1) in SVM_S, the samples in 5 groups are put together for SVM learning; 2) in SVM_A, we equally fuse the decision values of 5 pre-learned classifier; 3) in SVM_T, only the labeled target domain samples are used for SVM learning. A-MKL and DASVM are single domain adaptation frameworks, so we also put the samples in 5 groups together as the source training data. CP-MDA, DAM, DSM and MDA-HS can handle multiple source domains, so each group is regarded as a source domain. DASVM and MDA-HS are originally designed to infer the labels of target domain samples when there is no labeled data in target domain. In our settings, there are a few labeled target domain samples, so we fix these samples with their ground truth labels in DASVM and MDA-HS.

As suggested in [25], we use the non-linear χ^2 distance to measure the distance between two videos. We evaluate all the methods by training one-vs-rest SVMs with the Laplacian kernel (i.e., $K(i,j) = \exp(-\frac{1}{\sqrt{A}} D(\boldsymbol{x}_i, \boldsymbol{x}_j))$) for good results reported in [6], where A is the mean value of square distances between all training samples. The adjacency matrix $\boldsymbol{W}$ is set as "binary" type based on the N nearest neighbors with $N = 20$. We set the trade off parameters $\lambda = 100, C_l = 10, C_u = 1, \mu = 0.002$ in our experiments. To investigate the effects of each term in Eq. (3), we do two additional experiments, setting $\lambda = 0$ and $\lambda = \infty$. The results when setting $C_l = 1$ are also reported.

C. Results

Table II shows the mAPs of all the methods. In Table II, "NG", "NT" and "G" are short for "No Grouping", "No Transfer" and "Grouping", respectively. In Table III, the mAPs of our method with different parameters are shown. From the results, we can observe that:

1. DASVM outperforms the other two no grouping methods SVM_S and A-MKL. Moreover, there is no consistent winner among SVM_A, CP-MDA, DAM, DSM and MDA-HS in terms of mAPs. These results show that knowledge transfer from Web videos to consumer videos is a quite challenging task.
2. SVM_A achieves relative high performance on YouTube, but the performance of SVM_T is low. An explanation is that our Web videos and YouTube videos are collected in a similar manner and at least 20% of the videos in YouTube are incorrect labeled according to a study in [6]. On YouTube, we have only one labeled video from each event and the label of this video may be incorrect.
3. Generally speaking, several grouping methods can achieve better results than no grouping methods, which shows that dividing the Web videos into groups is beneficial to domain adaptation.
4. When $\lambda = 0$, the weights are calculated by the Laplacian and $\lambda = \infty$ means that we only choose the classifier with the minimal $M_s^- - M_s^+$. Both results of these two settings are inferior to the results of $\lambda = 100$, which shows that putting the two terms in Eq. (3) together can obtain better performance.
5. When $C_l = C_u = 1$, the empirical errors on labeled target data and unlabeled target are equally penalized. The performance is also inferior to the performance of assigning a larger value to C_l. This shows that the ground truth labels

Table II
MEANS AND STANDARD DEVIATIONS (%) OF MAPS OF ALL METHODS. "NG", "NT" AND "G" ARE SHORT FOR "NO GROUPING", "NO TRANSFER" AND "GROUPING", RESPECTIVELY.

	Method	Kodak	YouTube	CCV
NG	SVM_S	34.9 ± 0.5	32.8 ± 0.1	38.1 ± 0.3
	A-MKL [6]	35.5 ± 0.9	33.1 ± 0.3	39.3 ± 0.4
	DASVM [14]	39.3 ± 1.3	33.8 ± 1.7	40.8 ± 2.1
NT	SVM_T	41.2 ± 6.7	24.9 ± 4.4	42.6 ± 2.3
G	SVM_A	33.0 ± 3.1	34.6 ± 2.0	36.5 ± 2.0
	CP-MDA [17]	41.0 ± 5.1	35.0 ± 1.7	37.5 ± 1.4
	DAM [15]	42.0 ± 4.3	34.8 ± 2.0	38.1 ± 2.8
	DSM [22]	42.8 ± 4.3	33.9 ± 2.4	43.8 ± 2.8
	MDA-HS [8]	36.8 ± 3.3	35.1 ± 2.1	39.2 ± 2.2
	Ours	44.7 ± 2.3	36.3 ± 2.2	46.5 ± 1.8

Table III
MEANS AND STANDARD DEVIATIONS (%) OF MAPS WITH DIFFERENT PARAMETERS.

Parameter	Kodak	YouTube	CCV
$\lambda = 0$	41.5 ± 2.4	35.6 ± 1.7	37.1 ± 1.3
$\lambda = \infty$	42.9 ± 5.7	34.7 ± 4.0	45.6 ± 2.3
$C_l = 1$	43.7 ± 2.5	35.4 ± 2.3	45.3 ± 1.7
Ours	44.7 ± 2.3	36.3 ± 2.2	46.5 ± 1.8

of labeled target data are more reliable than the estimated labels of unlabeled target data.

6. On the Kodak dataset, our method can achieve the relative improvements of 28.1%, 25.6%, 13.7%, 8.5%, 35.4%, 9.0%, 6.3%, 4.5% and 21.3% over SVM_S, A-MKL, DASVM, SVM_T, SVM_A, CP-MDA, DAM, DSM and MDA-HS, respectively. On the YouTube dataset (resp., the CCV dataset), the relative improvement of our method over the best existing method is 3.4% (resp., 6.2%). Our method achieves the best results on three datasets, which demonstrates the effectiveness of our method for event recognition in consumer videos by making full use of the information in available data.

V. CONCLUSION

In this paper, we have proposed a novel Multi-Group Adaptation (MGA) method to utilize a large number of Web videos to annotate consumer videos. In our framework, we divide the Web videos into different semantic groups by querying different keywords in YouTube and referring the videos returned by one keyword search as a group. We further assign different weights to the groups in a joint manner which takes correlations among the source groups, as well as the correlations between the source and the target into account. At last, we encourage the final classifier to be smooth on the target domain. Comprehensive experiments on Kodak, YouTube and CCV datasets demonstrate the effectiveness of our method for event recognition in consumer videos.

ACKNOWLEDGMENT

This work was supported in part by the Natural Science Foundations of China (NSFC) under Grant No.61203274, the Specialized Fund for Joint Building Program of Beijing Municipal Education Commission, and the Key Laboratory of Advanced Information science and Network Technology.

REFERENCES

[1] H. Izadinia and M. Shah, "Recognizing complex events using large margin joint low-level event model," in *ECCV*. Springer, 2012, pp. 430–444. 1

[2] K. Tang, L. Fei-Fei, and D. Koller, "Learning latent temporal structure for complex event detection," in *CVPR*. IEEE, 2012, pp. 1250–1257. 1

[3] Z. Ma, Y. Yang, Z. Xu, S. Yan, N. Sebe, and A. Hauptmann, "Complex event detection via multi-source video attributes," in *CVPR*. IEEE, 2013, pp. 2627–2633. 1

[4] A. Vahdat, K. Cannons, G. Mori, S. Oh, and I. Kim, "Compositional models for video event detection: A multiple kernel learning latent variable approach," in *ICCV*. IEEE, 2013. 1

[5] S. J. Pan and Q. Yang, "A survey on transfer learning," *KDE*, vol. 22, no. 10, pp. 1345–1359, 2010. 1

[6] L. Duan, D. Xu, I. W.-H. Tsang, and J. Luo, "Visual event recognition in videos by learning from web data," *T-PAMI*, vol. 34, no. 9, pp. 1667–1680, 2012. 1, 4, 5, 6

[7] N. Ikizler-Cinbis, R. G. Cinbis, and S. Sclaroff, "Learning actions from the web," in *ICCV*. IEEE, 2009, pp. 995–1002. 1

[8] L. Chen, L. Duan, and D. Xu, "Event recognition in videos by learning from heterogeneous web sources," in *CVPR*. IEEE, 2013, pp. 2666–2673. 1, 4, 6

[9] 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 *ICMR*. ACM, 2011, p. 29. 1, 4, 5

[10] A. Loui, J. Luo, S.-F. Chang, D. Ellis, W. Jiang, L. Kennedy, K. Lee, and A. Yanagawa, "Kodak's consumer video benchmark data set: concept definition and annotation," in *MIR*. ACM, 2007, pp. 245–254. 1, 4

[11] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *CVPR*. IEEE, 2012, pp. 2066–2073. 2

[12] J. Yang, R. Yan, and A. G. Hauptmann, "Cross-domain video concept detection using adaptive svms," in *Proceedings of the 15th international conference on Multimedia*. ACM, 2007, pp. 188–197. 2

[13] X. Wu, H. Wang, C. Liu, and Y. Jia, "Cross-view action recognition over heterogeneous feature spaces," in *ICCV*. IEEE, 2013, pp. 609–616. 2

[14] L. Bruzzone and M. Marconcini, "Domain adaptation problems: A dasvm classification technique and a circular validation strategy," *T-PAMI*, vol. 32, no. 5, pp. 770–787, 2010. 2, 4, 6

[15] L. Duan, D. Xu, and I. W. Tsang, "Domain adaptation from multiple sources: A domain-dependent regularization approach," *T-NNLS*, vol. 23, no. 3, pp. 504–518, 2012. 2, 3, 4, 6

[16] H. Wang, X. Wu, and Y. Jia, "Video annotation via image groups from the web," *TMM*, 2014. 2

[17] R. Chattopadhyay, Q. Sun, W. Fan, I. Davidson, S. Panchanathan, and J. Ye, "Multisource domain adaptation and its application to early detection of fatigue," *KDD*, vol. 6, no. 4, p. 18, 2012. 2, 3, 4, 6

[18] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006. 2

[19] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *JMLR*, vol. 7, pp. 2399–2434, 2006. 3

[20] Y.-F. Li, J. T. Kwok, and Z.-H. Zhou, "Semi-supervised learning using label mean," in *ICML*. ACM, 2009, pp. 633–640. 3

[21] L. Chen, I. W.-H. Tsang, and D. Xu, "Laplacian embedded regression for scalable manifold regularization," *T-NNLS*, vol. 23, no. 6, pp. 902–915, 2012. 4

[22] L. Duan, D. Xu, and S.-F. Chang, "Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach," in *CVPR*. IEEE, 2012, pp. 1338–1345. 4, 6

[23] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995. 4

[24] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004. 5

[25] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *CVPR*. IEEE, 2008, pp. 1–8. 5