Video Annotation by Incremental Learning from Grouped Heterogeneous Sources

Han Wang, Hao Song, Xinxiao Wu, Yunde Jia

Beijing Lab of Intelligent Information Technology and the School of Computer Science, Beijing Institute of Technology, 100081, China.
wanghan@bjfu.edu.cn, {songhao, wuxinxiao, jiayunde}@bit.edu.cn

Abstract. Transfer learning has shown promising results in leveraging loosely labeled Web images (source domain) to learn a robust classifier for the unlabeled consumer videos (target domain). Existing transfer learning methods typically apply source domain data to learn a fixed model for predicting target domain data once and for all, ignoring rapidly updating Web data and continuously changes of users requirements. We propose an incremental transfer learning framework, in which heterogeneous knowledge are integrated and incrementally added to update the target classifier during learning process. Under the framework, images (image source domain) queried from Web image search engine and videos (video source domain) from existing action datasets are adopted to provide static information and motion information of the target video, respectively. For the image source domain, images are partitioned into several groups according to their semantic information. And for the video source domain, videos are divided in the same way. Unlike traditional methods which measure relevance between the source group and the whole target domain videos, the group weights in this paper are treated as latent variables for each target domain video and learned automatically according to the probability distribution difference between the individual source group and target domain videos. Experimental results on the two challenging video datasets (*i.e.*, CCV and Kodak) demonstrate the effectiveness of our proposed method.

1 Introduction

The rise of personal hand-held cameras and video sharing websites such as YouTube has resulted in massive amounts of consumer videos online. The ability to rapidly analyze and annotate the event from these unconstrained videos is a challenging computer vision task due to three main issues. First, these videos are generally captured by mobile devices at random and thus containing considerable camera motion, occlusion, cluttered background, and large intraclass variation, making the videos within the same type of event appear different and less discriminant. Second, the labels of these videos are usually meaningless due to users' random noting and subjective understanding, posing a great challenge to traditional learning methods which requires sufficient labeled videos to learn robust event classifiers. Third, the data on the internet updated every second, and fixed model trained on the pre-defined data may not work well for predicting new coming data. How to acquire sufficient knowledge while freeing the labor from burdensome annotation process is an important problem for event annotation in consumer videos.

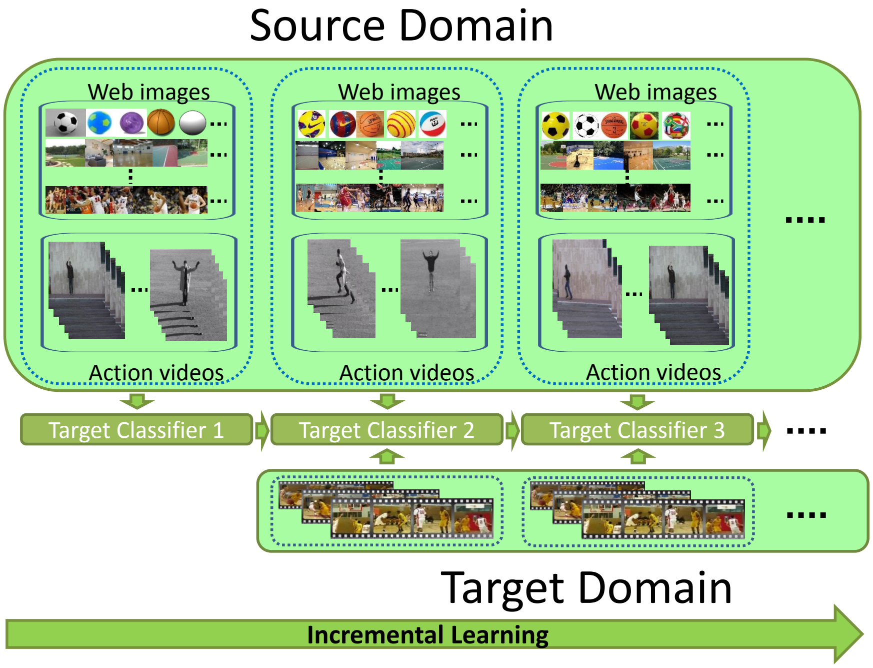

Fig. 1. Illustration of our framework.

Many researchers have tried to seek other sources of labeled data and transfer the related knowledge from these data to videos [1][2][3][4]. Most of previous work focuses on learning a fix model from a set of predefined images or videos (source domain) to predict the events in the complex videos (target domain). It is natural to ask if such fixed models would work well in the scenario that the Web data changes with each passing day. To deal with the fast updating of the source data, we propose a novel incremental learning framework for consumer video annotation. Under the framework, the classifiers learned on the source domains can be incrementally updated to capture the changes of both source and target domains and thereby facilitate the annotation task of real-world videos.

In this paper, we propose to acquire the source knowledge from the increasingly mature Web image search engines as well as existing labeled action datasets (e.g. KTH [5] and Weizmann [6]), which is based on the following observations: (1) the duration of videos on the Web are relatively long, so it will take more time and labor to analyze a video than an image. Obviously, it is more efficient to query images from the Web than directly search videos from the Web; (2) Besides the static information provided by Web images, temporal information provided by action videos is also beneficial for recognizing some key actions in consumer videos; (3) The action videos in the two datasets are relatively simple (the time spend on these videos is relatively less) but can provide basic human action information (e.g., running, waving) for complex social event; (4) All the action video datasets are well labeled by researchers and do not need additional labeling efforts.

Though it is beneficial to learn from Web images and action datasets, noise knowledge of little relevance with consumer videos still exists due to random noting and subjective understanding. To handle this negative transfer, we propose to organize the source samples in groups, and each group stands for one event-related semantic concept. Given the groups for each event class, we can leverage these groups by assigning different weights to different groups according to their relevance to target domain data. Besides irrelevant source domain data, the intra-class variation of the target domain videos also must not be overlooked. In other words, one piece of knowledge which is irrelevant to one video may be useful to identify another video, even through both videos belonging to the same event class. To deal with above problem, the relevance between the source and the target should be measured not only according to the event class variation but also to the video itself. In this paper, we measure the relevance between a set of the source groups and an individual target domain sample, the relevance are described by using group weights. Instead of fix the group weight for every target domain sample, we treat the weights as latent variables and try to optimize the group weights and the group templates simultaneously in a latent structural learning framework.

As mentioned before, the knowledge on the Web is updating rapidly, one cannot learn a fixed model for consumer video annotation once and for all. The emerging images and videos make the fixed model hard to be well generalized. Besides changes of the Web data, continuously changes of users' needs also require updating the classifiers for annotating videos. An incremental transfer learning work is introduced by acquiring new knowledge of new added data from both the source and target domain while retaining the knowledge learned before. Our incremental transfer learning work is based on a latent structural model which minimizes the difference in a marginal probability measure between the new added source and target domain data. To make the learned model more stable on target domain data, we biasing the new target classifier close to the hyperplanes of old ones during incremental process. At the same time, smooth assumptions on two regularizers and different groups are imposed to enhance the target classifier more adaptable to the target domain data.

Fig. 1 illustrates the framework of our method. The contributions of this paper are three folds. (1) We develop a principle framework for annotating consumer videos by incrementally updating the model using heterogenous sources. (2) We propose a latent structural model by treating the groups weights as latent variables to capture the relevance between the source domain groups and the target domain samples. (3) We introduce two constraints in our incremental learning process by biasing the hyperplane on the target domain close to those learn earlier.

1.1 Related Work

Recently, applying domain adaptation to multimedia content analysis has attracted more attentions [1–4]. Yang *et al.* [7] proposed an Adaptive 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.* [8] 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 [9] to add the pre-learned classifiers as the prior. Their methods mainly focus on the single source domain setting.

To utilize numerous labeled source domain data, multiple source domain adaptation methods [4, 10–12] 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 propose leveraging different groups of source domain training data according to their semantic meanings. We insure that the data in each group are of the same concept, and different groups within the same event are correlated to each other.

Several recent methods have been proposed to investigate the knowledge transform from Heterogeneous domain adaptation methods. In [13], Web images are incrementally collected to learn classifiers for action video recognition. Tang *et al.* [14] introduced a novel self-paced domain adaptation algorithm to iteratively adapt the detector from source images to target videos. Duan *et al.*[4] developed a domain selection method to select the most relevant source domains. In these existing works, the pre-learned classifiers are primarily using training data from different source domains and then the target classifiers are learned from pre-learned classifiers in a late-fuse fashion. In contrast, our work can simultaneously learn the optimal classifiers and weights of different source-domain groups to construct the target classifier in an incremental way. [15] is closely related to our work, which deals with heterogeneous feature spaces and aims at transferring knowledge from labeled source domain images and videos to unlabeled target domain videos. However, in our method group weights are treated as latent variables which can explicitly describe the contribution of different groups for different target domain samples.

2 Problem Setting and Definitions

To obtain the Web images of the image source domain, we first manually define a semantic concept collection as $\mathcal{C} = \{C_1, C_2, ..., C_G\}$, where C_i represents one event-related concept. In this paper, we use 73 semantic concept keywords, including event names (e.g. "play basketball"), action related concepts (e.g. "waving"), object related concepts (e.g. "ball"), and scene related concepts (e.g. "basketball court"). For each concept, a group of images are collected by querying a keyword to the Web image search engine. And for action videos in the video source domain, videos are clustered into groups according to their action labels in the corresponding datasets (e.g. "waving", "running", etc.). The image source domain and video source domain form the source domain. Following this grouping strategy, and G groups of heterogeneous data including web images and action videos consist the source domain, and each group is represented by one type of features (i.e. image feature for the image source domain or motion feature for the video source domain). As for the target domain, each consumer video is represented by two types of feature: motion features (i.e. STIP features) extracted from the whole video, as well as image features (i.e. SIFT features) extracted from the keyframes.

Formally, for each event class, we are given a set of groups $\{(x_{g_i}^s, y_{g_i}^s)|_{i=1}^{N_g}\}, g \in \{1, ..., G\}$ including both images and videos from the source domain $\mathcal{D}^s$, where N_g is the total number of samples in the g-th group and $x_{g_i}^s$ is the i-th sample in the group with its label $y^s_{g_i} \in \{-1, 1\}$. A set of pre-learned source classifiers $f^s_g(x^s_g) = \widetilde{w'_g}\varphi_g(x^s_g)$ are learned by using the training data from each individual group. φ_g is the feature mapping function for the g-th group. Also, we are provided with a set of unlabeled consumer videos $\{x^t_i|^{N_t}_{i=1}\}$ from the target domain $\mathcal{D}^t$.

In our setting of incremental learning, there are two types of information in the source domain. First, we have a set of group classifiers f^s_g that are obtained from initial G groups of images in the source domain. Since in incremental learning there is no access to the samples used to train the initial source classifiers, we encoded these source models as a set of G hyperplanes represented in a matrix form as $\widetilde{W} = [\widetilde{w_1}, \widetilde{w_2}, ..., \widetilde{w_G}]$. Second, for every incremental stage, we are given a set of new Web images or action videos from the source domain and a small set of consumer videos in the target domain. Our goal is to use the newly given source domain data to boost the annotation performance on the newly given target domain videos. It should be noted that the newly given source domain data may belong original groups or new groups, which means new concept can be learned during the incremental transfer learning process.

3 Incremental Learning from Heterogeneous Sources

Unlike traditional multi-source adaptation methods, our method treats the combination coefficients (group weights) of different groups as latent variable rather than fixes them for each incoming target video. To this end, we learn the following target classifier f^t for any consumer video sample x^t_i, which fuses the decisions from multiple sources according to the latent weights:

$$f^t(x^t_i) = \sum_{g=1}^{G} w'_g \phi(\theta_{i_g}, \varphi_g(x^t_i)), \tag{1}$$

where w_g is the template for the g-th group data; θ_{i_g},$i \in \{1, ..., N_t\}$ and $g \in \{1, ..., G\}$ is the g-th latent group weight for consumer video x^t_i, and φ_g is the feature mapping function for the target video x^t_i on the g-th group.

Once there are new target domain videos available, we update the target classifier to make it more adaptable for these newly coming videos. In other words, the aim of our incremental approach is to find a new combination (i.e. $\Theta_i = [\theta_{i_1}, \theta_{i_2}, ..., \theta_{i_G}]$) of a new set of hyperplanes (i.e. $W = [w_1, w_2, ..., w_G]$), such that (1) performance on newly coming target data improves by transferring knowledge from both the original source models and new source domain data, (2) efficiency of learning additional information improves without any access to the original data used to train the existing classifiers, and (3) the model is able to accommodate new event class introduced with new data. Thanks to the development of the Internet, we can easily obtain new labeled source domain data to incrementally update knowledge.

3.1 Learning

Although the explosion of Web data can bring new knowledge, the random noting and subjective understanding of images make the noise images unavoidable. To prevent negative transfer brought by the newly coming data, we enforce the new learned hyperplanes W to remain close to the original hyperplanes $\widetilde{W}$ using the term $\| W - \beta\widetilde{W} \|^2$.

This term enforces the target model W to be relatively close to the original model $\widetilde{W}$, using coefficient vector $\beta = [\beta_1, ..., \beta_G]^T$. Besides the constraints on the hyperplane, we also enforce a smooth assumption on the single group decision value, i.e., different group classifiers belonging to the same event should have similar decision values on the target domain data. In our work, this constraint is implemented using the regularizer $\sum_{g=1}^{G} \theta_{i_g} \sum_{k \neq g}^{G} \|w_g f_g^s(x_i^t) - w_k f_k^s(x_i^t)\|^2$. For example, if the g-th group and the k-th group represent different concepts of the same event, we ensure that $f_s^k(x)$ should be close to $f_s^g(x)$. Actually, we introduce this term to penalize those groups far from major event-related groups. For domain adaptation, we similarly assume that the pre-learned classifiers in the source domain should have similar decision values on the unlabeled samples in the target domain.

Since the group weights are treated as latent variables, our goal is to learn a prediction rule of the following form:

$$\begin{aligned} f^t(x) &= \arg\max_{\Theta, y} F(x, y, \Theta) \\ &= \arg\max_{\Theta, y} W \cdot \Phi(x, y, \Theta), \end{aligned} \tag{2}$$

where $\Phi(x, y, \Theta)$ is a joint feature vector that describes the relationship among the input consumer video x, output event class label y, and latent group weights Θ.

In order to learn the group weights Θ_i for each newly coming target video x_i^t and simultaneously update the group templates W, we introduce above constraints into a latent structural objective function as follows:

$$\min_{W} \frac{1}{2} \parallel W - \beta\widetilde{W} \parallel^2 + \lambda_1 \sum_{i=1}^{N_t} \xi_i + \lambda_2 \sum_{j=1}^{N_t} \zeta_j, \tag{3}$$

$$\text{s.t. } \xi_i = l(\max_{\Theta_i} F(x_i^t, y_i^t, \Theta_i) - \max_{\widetilde{\Theta}_i, \widetilde{y}_i} \widetilde{F}(x_i^t, y_i^t, \Theta_i)), \tag{4}$$

$$\zeta_j = \sum_{g=1}^{G} \theta_{j_g} \sum_{k \neq g}^{G} \|f_g^s(x_j^t) - f_k^s(x_j^t)\|^2, \tag{5}$$

$$\sum_{g=1}^{G} \theta_{i_g} = 1, \tag{6}$$

where λ_1, λ_2 are tradeoff parameters. Here $l(t)$ is the hinge loss function defined by $l(t) = max(0, 1 - t)$. We use this loss function to enforce the decision value of the newly learned target classifier not far away from that of the original classifier. We argue that such supervision is very important for our incremental adaptation problem. The reason is two-fold: (a) There is a certain amount of overlap between the updated target classifier and the original target classifier, so it is very possible that decisions on these two types of classifier would not be too far from each other; (b) We do not have any labeled data in the target domain, so the performance of updated classifier will become much worse without having the constraints in Eq. 4. Our experiments demonstrate the strength of this constraint.

The optimization problem in Eq. 3 can be solved in many different ways. In our implementation, we adopt a non- convex cutting plane method proposed in [16]. First, it is easy to show that Eq. 3 is equivalent to $\min_W L(w) = \frac{1}{2} \parallel W - \beta\widetilde{W} \parallel^2 + \sum_{i=1}^{N_t} R(W)$ where $R(W)$ is a loss function defined as

$$\begin{aligned} R(W) =& \lambda_1 l(\max_{\Theta_i} F(x_i^t, y_i^t, \Theta_i) - \max_{\widetilde{\Theta}_i, \widetilde{y}_i} \widetilde{F}(x_i^t, y_i^t, \Theta_i)) \\ & + \lambda_2 \sum_{g=1}^{G} \theta_{g_i} \sum_{k \neq g}^{G} \|f_g^s(x_i^t) - f_k^s(x_i^t)\|^2. \end{aligned} \tag{7}$$

The non-convex cutting plane method [16] aims to iteratively build an increasingly accurate piecewise quadratic approximation of $L(W)$ based on its sub-gradient $\partial_W L(W)$. The key issue here is how to compute the sub-gradient $\partial_W L(W)$. We define

$$\begin{aligned} \Theta_i^* &= \arg\max_{\Theta} \widetilde{F}(x_i^t, y_i^t, \Theta_i), \forall y \in \mathcal{Y}, \\ y^{t*} &= \arg\max_{y} \widetilde{F}(x_i^t, y_i^t, \Theta_i^*). \end{aligned} \tag{8}$$

The inference problem in Eq. 8 will be described in Sec. 4.2. It is easy to show that $\partial L(W)$ can be calculated as follows:

$$\begin{aligned} \partial_W L(W) = W - \beta\widetilde{W} + \sum_{i=1}^{N_t} \Theta_i \Phi(x_i^t, y_i^t, \Theta_i) \\ - \sum_{i=1}^{N_t} \Theta_i^* \Phi_(x_i^t, y_i^{t*}, \Theta_i^*) + \sum_{i=1}^{N_t} \Theta_i \Omega_i. \end{aligned} \tag{9}$$

Here

$$\Omega_i = \{\Omega_i^1, ..., \Omega_i^g, ..., \Omega_i^G\} \tag{10}$$

and

$$\Omega_i^g = \sum_{k=1, k \neq g}^{G} \phi(x_i^t)(f_g^s(x_i^t) - f_k^s(x_i^t)) \tag{11}$$

Given the sub-gradient $\partial_W L(W)$ according to Eq. 9, we can minimize $L(W)$ using the method in [16].

3.2 Inference

Given the group templates W, we need to solve the following inference problem for each target domain sample x_i^t:

$$\Theta_i = \arg\max_{\theta_{i_g}} \sum_{g=1}^{G} w_g' \phi(\theta_{i_g}, \varphi_g(x_i^t)) \qquad \forall y \in \mathcal{Y}. \tag{12}$$

As we know, the key issue of transfer learning approach is to measure the relevance between the source domain data and the target domain data. Motivated by MMD [17][18][19], we infer the group weights $\Theta_g = [\Theta_i, ...\Theta_{N_t}]$ for the target domain samples by measuring the marginal probability distribution difference between two sets of samples:

$$\Theta_g = \|\frac{1}{N_g} \sum_{j=1}^{N_g} \phi(x_{g_j}^s) - \frac{1}{N_t} \sum_{i=1}^{N_t} \theta_{g_i} \phi(x_i^t)\|_{\mathcal{H}}^2. \tag{13}$$

Algorithm 1 Incremental Heterogeneous Domain Adaptation.

```
Require:
    $\{X^g\}_{g=1}^G$: the set of source domain groups;
    $X^t$ : unlabeled target videos;
    $\widetilde{W}$: original source domain hyperplane set;
Ensure:
    $W$: Updated target classifiers;
repeat
    Calculate $F(\cdot)$ and $\widetilde{F(\cdot)}$ for $x_i^t$
    Compute group smooth constraint $\Omega_i$
    Infer the latent group weight $\Theta_i$ for $x_i$ according to Eq. (13)
until All target domain samples are involved
Use cutting plane method to minimize Eq. (3) to update $W$
return $W$
```

We stress that the criterion above is defined according to source domain groups which are a subset of the source domain, as the sample mean is computed only on the semantic related instances. This is much different from the other MMD approaches that have used similar nonparametric techniques for comparing distributions. There they make stronger assumptions that all data points in the source domain need to be collectively distributed similarly to the target domain. Furthermore, in our inference problem, different weights are assigned to different target domain samples. Our results below will show that these differences are crucial to the success of our approach. Our incremental learning method is summarized in Algorithm 1.

3.3 Datasets

We evaluate our method on two consumer video datasets: CCV [20] and Kodak [21].
CCV dataset 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, twelve event categories: "basketball", "baseball", "soccer", "iceskating", "biking", "swimming", "skinning", "graduation", "birthday", "wedding", "show", and "parade" are conducted in our experiment.
Kodak dataset is collected by Kodak 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").

To construct clearly labeled source domain videos, we apply two widely used action video datasets (i.e. KTH [5] and Weizman [6]).
KTH action video dataset contains six types of human actions: walking, jogging, running, boxing, hand waving and hand clapping. These actions are performed several

times by twenty-five subjects in four different scenarios.
Weizmann action video dataset consists of about 90 low-resolution video sequences showing nine different subjects, each performing 10 actions including bending, jumping, running, skipping, galloping, walking and waving.

Web image dataset covers thirteen events: “basketball”, “baseball”, “soccer”, “iceskating”, “biking”, “swimming”, “graduation”, “birthday”, “wedding”, “skinning”, “show”, “parade” and “picnic”. In our experiment, we use the Google image search engine to collect images, and for each input keyword, the top ranked 300 images are downloaded and the corrupted images with invalid URLs are discarded. Finally, total 76 event-related semantic keywords are used to query images from Web image search engine and $16,708$ images are collected.

3.4 Experimental Setup

For videos in both domains, we extract 162-dimensional 3D Space-Time Interest Point (STIP) in which 72-dimensional Histograms of Oriented Gradient (HOG) and 90-dimensional Histograms of Optical Flow (HOF) are extracted by using the online tool from [22]. For consumer videos in the target domain, we additionally extract image features by randomly sampling five frames from each video as its keyframe and extracting 128-dimensional SIFT features from salient regions on each frame detected by the Difference of Gaussians (DoG) detectors [23]. The bag-of-words representation is used for both image and video features. Specifically, we cluster the SIFT descriptors extracted from all the training Web images and keyframes, 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. Similarly, we cluster the STIP features extracted from consumer videos and action videos into 2000 words using k-means, and the motion feature of each video is then represented by a 2000-dimensional token frequency feature. Finally, two types of feature is used for videos in the target domain, and one type of feature is sued for images and videos in the source domain, respectively.

To pre-learn a classifier for each group of each event, the positive samples are constructed by the samples belonging to the corresponding group in the corresponding event class and the negative samples consist of randomly selected 300 samples in the same type of any other groups. At the training stage, for the CCV dataset the training set defined by [20] is used as the unlabeled target domain. For the Kodak dataset, all the 195 target domain videos are used as unlabeled training data. Consequently, the training data includes the heterogeneous groups from the source domain and unlabeled videos from the target domain.

We compare our transfer learning method with several state-of-the-art methods, including the standard SVM (S_SVM), the single domain adaptation methods of Domain Adaptive SVM (DASVM) [24], the multi-domain adaptation methods of Domain Adaptation Machine (DAM) [25], Conditional Probability based Multi-source Domain Adaptation (CPMDA) [26], Domain Selection Machine (DSM) [4] and Multi-domain Adaptation with Heterogeneous Sources (MDA-HS) [15]. Since the S_SVM can only handle data from a single group, we merge the training samples represented by same type of features into one source to train SVM classifier. Consequently, two types of SVM classifiers based on SIFT and STIP features are obtained for Web images and action videos, respectively. The final S_SVM classifier is obtained by equally fusing these two types of source classifiers. For DASVM, which is semi-supervised learning method and also cannot handle the multi-group setting, the target classifiers are trained using the labeled source domain samples and the keyframes of unlabeled videos from the target domain. Similar to S_SVM, we employ the same fusing strategy to obtain the target classifier for DASVM. The traditional multi-source adaptation methods CPMDA, DAM and DSM can not directly handle the heterogeneous sources problem, so we perform these multi-source leveraging strategies on single feature type and then average the decision values to obtain the final decision.

In our incremental learning setting, the source domain data is partitioned into two parts: an initial set used for learning the initial source group classifiers and the remaining sets added successively for updating. More specifically, about 3000 source domain samples are used for updating the target model at each incremental stage. And for the testing data, we evaluate the annotation performance on all the input target domain videos (including the new videos in the current step and the videos used before). 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.

3.5 Results

We first compare our method with existing approaches and report the per-event APs of all the methods on the CCV and Kodak datasets in Fig.2 and Fig.3, respectively. We also show the mAPs of all methods on these datasets in Table 1.

Table 1. Comparison of mAPs (%) between our method and other methods on the CCV and Kodak datasets.

Method	S_SVM	DASVM [24]	CPMDA [26]	DAM [25]	DSM [4]	MDA [15]	**Ours**
CCV	7.43	8.08	8.46	11.59	11.10	10.47	**17.05**
Kodak	12.95	17.78	21.58	30.25	23.38	25.18	**35.63**

From the results, we notice that:

- Our method achieves the best results on both datasets, which shows that our incremental weighting strategy is beneficial to positive transform. Multi-source adaptation methods (i.e. CPMDA, DAM, DSM, MDA and our method) generally outperform the single source methods (i.e. S_SVM and DASVM), which clearly reveals that it is helpful to weight different sources for knowledge transfer. The contribution of different sources may be different, by this means, the weighting strategy becomes particularly important.
- Our method is better than MDA, which illustrates the benefit of using latent video-specific weights for domain adaptation. A possible explanation is that the events in real-world vary dramatically, so fixed group weights can not capture the relevance information between different groups in various situations in every situation.

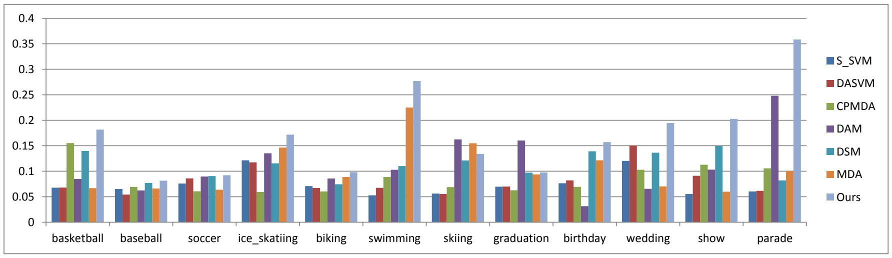

Fig. 2. Per-event Average Precisions (APs) (%) of all methods on the CCV dataset.

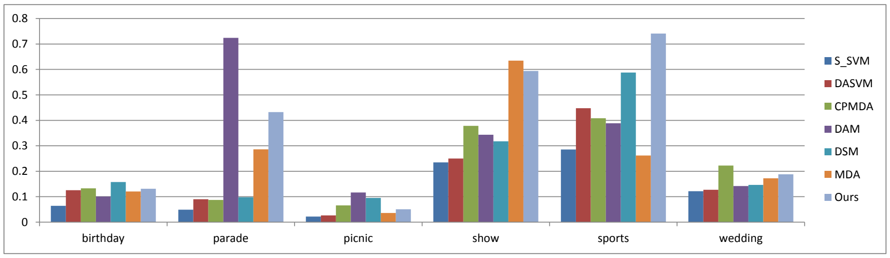

Fig. 3. Per-event Average Precisions (APs) (%) of all methods on the Kodak dataset.

- It is also interesting to notice that our method performs better than DSM, which indicates that the data from all groups querying by associational keywords can benefit understanding video events to some extend.
- In terms of per-event average precisions, there is no consistent winner among these methods. This indicates the existence of the irrelevant data which hinds these transfer learning methods to acquire good target classifiers. Our method achieves more stable performance, which demonstrates that latent weighting strategy can effectively cope with noisy data in the source domain.

We also investigate the effects of each constraint in our optimization function in Eq. (3) for learning knowledge from the source. The column of $\theta_{i_g} = \theta_{j_g}$ reports the performance of annotating when group weights for all the target video are equal. The objective function is given by

$$\begin{aligned} &\min_{W} \frac{1}{2} \parallel W - \beta\widetilde{W} \parallel^2 + \frac{1}{2} \parallel \Theta \parallel^2 + \sum_{l=1}^{N_s} \parallel \sum_{g=1}^{G} f_g^s(x_l^s)) - y_l^s \parallel^2 \\ &\quad + \sum_{j=1}^{N_t} \sum_{g=1}^{G} \theta_g \sum_{k \neq g}^{G} \|f_g^s(x_j^t) - f_k^s(x_j^t)\|^2, \\ &s.t. \sum_{g=1}^{G} \theta_g = 1, \end{aligned} \tag{14}$$

where θ_g stands for the group weights of the g-th group. In the objective function, group weights are treated as explicit variable and simultaneously optimized with the group templates W. As shown is the results, the annotation performance degrades dramatically when group weights are not treated as latent variables for each target domain video. A possible explanation is that large intra-class variations within the same type of events exist in the target domain videos, making their visual cues highly variable. The relevance between the different groups and the individual target video cannot be accurately represented by a unified group weight. The performance is degraded when all groups are treated equally($\theta_g = 1/G$), which demonstrates that the contributions of different groups are different to the target classifier. This further indicate the relevance between the source and target is crucial for processing positive knowledge transfer.

Method	$\theta_{i_g} = \theta_{j_g}$	$\theta_g = 1/G$	$\lambda_1 = 0$	$\lambda_2 = 0$	$Ours$
CCV	9.94	15.21	15.20	7.87	17.05
Kodak	32.81	15.09	29.3	28.72	35.63

Table 2. Evaluation on different components of the optimal function using mAPs (%).

Finally, we evaluate the efficiency of our incremental domain adaptation method. Fig. 4 and Fig. 5 give the per-event comparison of non-incremental results and incremental results on both datasets. Table 3 shows mAP and computational time in minutes of our incremental method. As shown in the table, the non-incremental method degrades a lot, especially on the CCV dataset. A possible explanation is that the videos in the CCV dataset are much more than those in the Kodak, which is more close to the real-world situation. This also confirms our claim that the incremental method is more suitable for modeling consumer videos.

	Kodak		CCV	
	mAP (%)	Time (min)	mAP (%)	Time (min)
Non-incremental	31.90	14.19	7.25	54.32
Incremental	35.63	11.34	17.05	38.21

Table 3. The efficiency of our incremental method on the Kodak and CCV dataset.

4 Conclusion

In this paper, we have presented a new framework for consumer video event annotation by leveraging a large number of freely available labeled sources(i.e. images from Google and action videos from lab). By introducing a new incremental learning method and a new latent weighting scheme, our method, called Incremental Learning with Latent Groups Weights, can simultaneously seek and update the optimal group weights and group templates by using data from both domains. Comprehensive experiments on two benchmark datasets demonstrate the effectiveness of our method for video event annotation without requiring any labeled consumer videos.

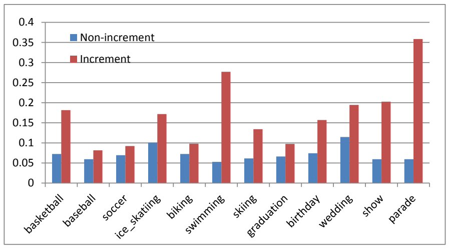

Fig. 4. Evaluation on the incremental efficiency on CCV dataset.

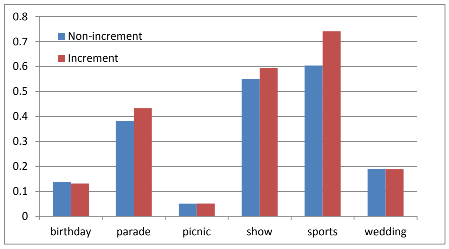

Fig. 5. Evaluation on the incremental efficiency on Kodak dataset.

5 Acknowledgements

The research 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.

References

1. Bergamo, A., Torresani, L.: Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. Advances in Neural Information Processing Systems (NIPS) (2010)
2. Gopalan, R., Li, R., Chellappa, R.: Domain adaptation for object recognition: An unsupervised approach. In: ICCV. (2011) 999–1006
3. Kulis, B., Saenko, K., Darrell, T.: What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In: CVPR. (2011) 1785–1792
4. Duan, L., Xu, D., Tsang, Chang, S.F.: Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach. In: CVPR. (2012) 1959–1966
5. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. Volume 3., IEEE (2004) 32–36
6. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on. Volume 2., IEEE (2005) 1395–1402

7. Yang, J., Yan, R., Hauptmann, A.: Cross-domain video concept detection using adaptive svms. In: International Conference on Multimedia. (2007) 188–197
8. Duan, L., Tsang, I., Xu, D., Maybank, S.: Domain transfer svm for video concept detection. In: CVPR. (2009) 1375–1381
9. Duan, L., Xu, D., Tsang, I., Luo, J.: Visual event recognition in videos by learning from web data. In: CVPR. (2010) 1959–1966
10. Wang, H., Wu, X., Jia, Y.: Video annotation via image groups from the web. In: IEEE Transactions on Multimedia. Volume 16. (2014) 1282–1291
11. Doretto, G., Yao, Y.: Boosting for transfer learning with multipple auxiliary domains. In: CVPR. (2010)
12. Schweikert, G., Widmer, C., Schölkopf, B., Rätsch, G.: An empirical analysis of domain adaptation algorithms for genomic sequence analysis. In: NIPS. (2009)
13. Ikizler-Cinbis, N., Cinbis, R., Sclaroff, S.: Learning actions from the web. In: CVPR. (2009) 995–1002
14. Tang, K., Ramanathan, V., Fei-Fei, L., Koller, D.: Shifting weights: Adapting object detectors from image to video. In: Advances in Neural Information Processing Systems. (2012) 647–655
15. Chen, L., Duan, L., Xu, D.: Event recognition in videos by learning from heterogeneous web sources. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE (2013) 2666–2673
16. Do, T.M.T., Artières, T.: Large margin training for hidden markov models with partially observed states. In: Proceedings of the 26th Annual International Conference on Machine Learning, ACM (2009) 265–272
17. Borgwardt, K.M., Gretton, A., Rasch, M.J., Kriegel, H.P., Schölkopf, B., Smola, A.J.: Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics **22** (2006) e49–e57
18. Gretton, A., Borgwardt, K., Rasch, M.J., Scholkopf, B., Smola, A.J.: A kernel method for the two-sample problem. NIPS (2008)
19. Sriperumbudur, B.K., Gretton, A., Fukumizu, K., Schölkopf, B., Lanckriet, G.R.: Hilbert space embeddings and metrics on probability measures. The Journal of Machine Learning Research **99** (2010) 1517–1561
20. Jiang, Y., Ye, G., Chang, S., Ellis, D., Loui, A.: Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In: ICMR. (2011) 29
21. Loui, A., Luo, J., Chang, S., Ellis, D., Jiang, W., Kennedy, L., Lee, K., Yanagawa, A.: Kodak's consumer video benchmark data set: concept definition and annotation. In: Workshop on Multimedia Information Retrieval. (2007) 245–254
22. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR. (2008) 1–8
23. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV **60** (2004) 91–110
24. Bruzzone, L., Marconcini, M.: Domain adaptation problems: A dasvm classification technique and a circular validation strategy. PAMI **32** (2010) 770–787
25. Duan, L., Xu, D., Tsang, W.H.: Domain adaptation from multiple sources: A domain-dependent regularization approach. IEEE Transactions on Neural Networks and Learning Systems **23** (2012) 504–518
26. Chattopadhyay, R., Sun, Q., Fan, W., Davidson, I., Panchanathan, S., Ye, J.: Multisource domain adaptation and its application to early detection of fatigue. ACM Transactions on Knowledge Discovery from Data (TKDD) **6** (2012) 18