Multimed Tools Appl (2015) 74:10439–10456
DOI 10.1007/s11042-014-2175-z

Cross-domain structural model for video event annotation via web images

Han Wang · Xiabi Liu · Xinxiao Wu · Yunde Jia

Received: 26 November 2013 / Revised: 3 June 2014 / Accepted: 30 June 2014 /
Published online: 30 July 2014

Abstract Annotating events in uncontrolled videos is a challenging task. Most of the previous work focuses on obtaining concepts from numerous labeled videos. But it is extremely time consuming and labor expensive to collect a large amount of required labeled videos for modeling events under various circumstances. In this paper, we try to learn models for video event annotation by leveraging abundant Web images which contains a rich source of information with many events taken under various conditions and roughly annotated as well. Our method is based on a new discriminative structural model called Cross-Domain Structural Model (CDSM) to transfer knowledge from Web images (source domain) to consumer videos (target domain), by jointly modeling the interaction between videos and images. Specifically, under this framework we build a common feature subspace to deal with the feature distribution mismatching between the video domain and the image domain. Further, we propose to use weak semantic attributes to describe events, which can be obtained with no or little labor. Experimental results on challenging video datasets demonstrate the effectiveness of our transfer learning method.

Keywords Video annotation · Knowledge transfer · Video analysis

1 Introduction

Following recent advances in data capturing, storage, and communication technologies, tasks in video understanding have shifted from classifying simple motions and actions [13, 24] to annotating complex events and activities in consumer videos [8]. Automatically annotating

H. Wang · X. Liu (✉) · X. Wu · Y. Jia
Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, China
e-mail: liuxiabi@bit.edu.cn

H. Wang
e-mail: wanghan@bit.edu.cn

X. Wu
e-mail: wuxinxiao@bit.edu.cn

Y. Jia
e-mail: jiayunde@bit.edu.cn

visual events in personal videos is a challenging problem due to various camera motion, occlusion, and cluttered background. To deal with automatic annotation, conventional methods [25] usually build a classifier to detect the presence of events in a video clip. In order to obtain satisfactory classification results, we usually require enough training examples. But it is difficult, if not impossible, to produce a large corpus of labeled example videos.

In this paper, we resort to Web image search engines for dealing with the problem of obtaining enough labeled video data and then transfer knowledge from images to videos. The main observation behind this idea is that increasingly mature image search engines (e.g., Google and Yahoo!) offer an abundance of images that can be harvested for training classifiers to annotate consumer videos [3, 19]. It is convenient and reasonable to import knowledge from Web images returned by image search engines, which has two advantages: 1) the knowledge can be easily obtained; 2) we can easily expand the vocabulary of events by simply adding query words. Since the event in most videos can be identified by a single keyframe, the knowledge discovered from Web images can be, to a large extent, used to infer the events of a video. As for those videos that could not be simply identified by a single frame, the motion features of the videos play an important role in annotating (such as "standing up" and "sitting down"). According to the discussion above, our goal in this paper is to annotate consumer videos using the knowledge from both labeled Web images and unlabeled consumer videos.

Web image data and consumer video data fall in two different feature spaces. Usually, we cannot expect a fine image classifier can work well on videos either. It is crucial to design a translator linking these two heterogeneous feature spaces and use it to transfer the knowledge from Web images (source domain) to consumer videos (target domain). In this paper, we link these two heterogeneous feature spaces by using Canonical Correlation Analysis (CCA) [11] to learn a common feature space. In CCA, two projection matrices are learned to translate features from two different domains into a common feature subspace. By this means, image features (e.g. SIFT) from the source domain and video features (e.g. STIP) from the target domain can be projected into a common subspace, in which the classifiers learned on the source domain can be adapted to the target domain.

Besides the difference between the domains, the semantic gap between the videos' semantic contents and their low-level descriptors also need to be considered. Although the low-level visual feature has been proven to perform well in many recognition tasks such as object recognition and scene classification, it is less effective to describe complex consumer videos due to the semantic gap between the low-level feature representation and the meaning of events. To tackle this problem, we apply a kind of mid-level feature lying between these two levels, called attributes, to describe events. Most of existing work uses a lot of pre-labeled attributes to learn the attribute classifiers, limiting their performance and scalability. In this work, we consider to obtain attributes containing weak semantics with the help of public available visual classifiers such as Classemes [21], instead of manually labeling the attributes of each image or video. We refer to such attributes as the "weak semantic", because they do not directly relate to the event class but to some extent have semantic meaning. Moreover, the attribute in our work is a continuous value which reflects both presents and strength of the corresponding semantic, instead of only indicating the absence of the semantic as traditional discrete attributes.

In this paper, we propose a new method for annotating consumer videos by leveraging information provided by a large number of Web images. A discriminative structural model, called Cross-Domain Structural Model (CDSM), is introduced to jointly capture the correlation between image features and video features, as well as the relationship

among different image attributes. Specifically, image features from Web images (keyframes) and video features from videos are used together to make them mutually beneficial. Moreover, these two different types of visual features are projected onto a common subspace by using CCA. Besides low-level features, we treat weak-semantic attributes of an event as continuous value in our model and capture the correlations among these attributes. Consequently, the source domain and the target domain are integrated and learned jointly in a unified framework. Fig. 1 shows the flowchart of the proposed video annotation method based on CDSM. We evaluate our method on three challenging video datasets: CCV [14], Kodak [8], and YouTube [8], and show that our CDSM can boost the event video annotation performance, without using any labeled consumer videos as training data.

The rest of the paper is organized as follows. Section 2 reviews the related work. Section 3 presents our method of transferring knowledge from Web images to consumer videos. In Section 4, the domain representation is described, together with the details about common subspace feature. In Section 5, we discuss the experimental results. We conclude the paper in Section 6.

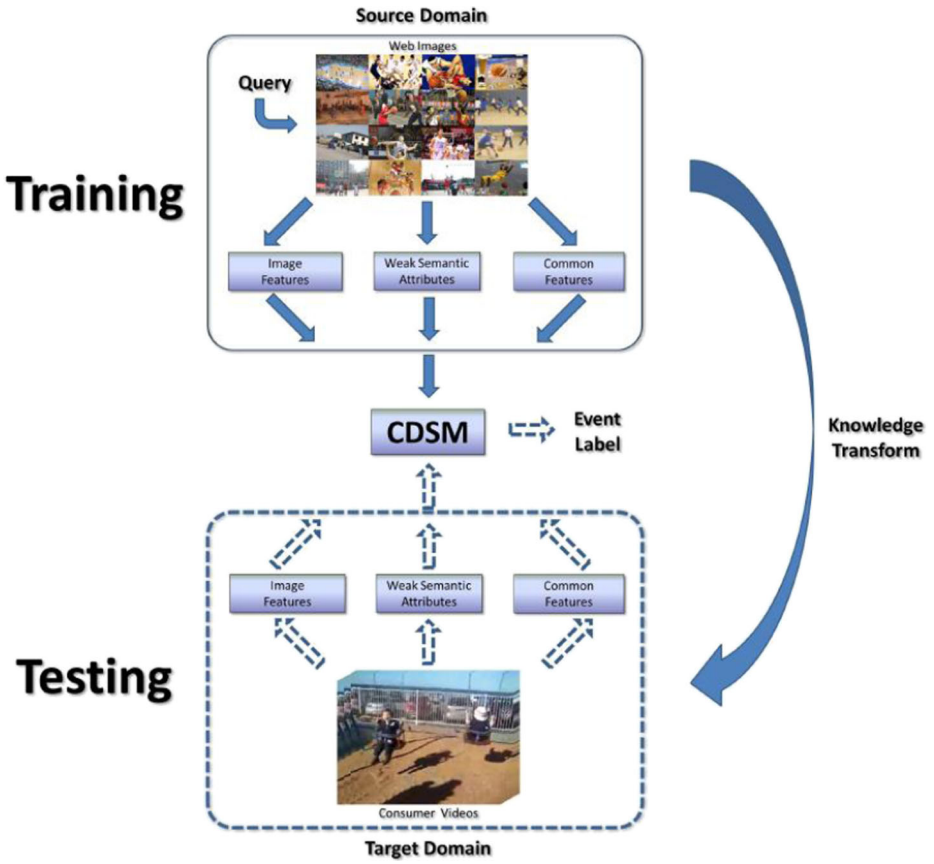

Fig. 1 Illustration of our framework on transferring knowledge for video event annotation

2 Related work

The core of our method is to transfer knowledge from Web images to consumer videos, which involves three key problems: (1) the issues of transfer learning; (2) heterogeneous feature space adaptation (3) weak semantic attributes. We review the related approaches in the following, respectively.

2.1 Knowledge transfer

Various methods [6, 16] of transferring knowledge across different domains have been developed. The settings of transfer learning are categorized to inductive transfer learning, transductive transfer learning and unsupervised transfer learning, based on different situations between the source and target domains and tasks [16]. Duan et al. [8] proposed to learn the target SVM classifier by minimizing a structural risk function. Bruzzone and Marconcini [4] used Domain Adaptation Support Vector Machine (DASVM) to learn the target classifier by iteratively removing some labeled source domain samples and adding some unlabeled target domain samples to the classifier simultaneously. Ikizler-Cinbis et al. [13] learned transfer models from loosely labeled Web images. These works only consider single feature type to learn classifiers, without utilizing the additional features which only exit in the target domain. Recently, Wang et al. [23] proposed to obtain knowledge from both labeled Web images and a small amount of labeled videos, and Duan et al. [9] developed a consumer video events recognition approach with auxiliary Web images. In their work, video features and image features were integrated into a target decision function to jointly determine events in videos. For most of previous work, however, features in different spaces are used separately, ignoring the potential connections among different feature spaces.

2.2 Heterogeneous feature space adaptation

For adapting classifiers in heterogeneous feature spaces, a reasonable approach is to translate all the training data into a target feature space, then the learning process can be done within a single feature space. This method has been proven successful in several applications in cross-lingual text classification [1]. But for more general translation learning problem, this method is not so suitable because machine translation between different feature spaces is very difficult to accomplish in many non-natural language cases, such as translating documents to images [19]. And hereby, the learning of the "feature space translator" cross different feature spaces becomes a challenge issue. Canonical Correlation Analysis (CCA) [11, 19] has been widely used to capture the relationship between texts and images. In [12], the author proposed to use Kernel CCA to learn a description that exploits the relations between the ordered tag words and the images visual descriptors, and compute similarities across the two views. In this paper, we apply CCA to translate image features and video features to a common feature space. Thus, classifiers learned on this common space can be adapted in both domains.

2.3 Weak semantic attributes

Attributes were introduced as describable properties of an object or event, and are well motivated in [10] as a kind of concept based image representation method, and exploited in recent literature. Attribute based approaches are also shown their strength for video and image search [5, 20, 22]. However, training attribute classifiers, on one hand, is a

burdensome process for labeling tremendous data. On the other hand, these kinds of attribute should be defined beforehand. This leads to the limitation of number of attributes, for only a small set of words can be chosen. Such a small amount of attributes are far from sufficient in forming an expressive describing, especially for searching a large consumer video corpus of diverse content. There is some work aiming to define attributes automatically: Berg et al. [2] try to find attributes through the Web, Parikh et al. [17] propose a semi-automatic attribute discovering method with the help of manual intervention, and they also try to explore new attribute types in [18]. Different from the previous work, the attribute used in our work is automatically obtained, which requires no or little human interruption.

3 Model formulation

We denote a domain as $D=(\chi,P(X))$, where χ is the feature space and $P(X)$ correspond to its distribution [16]. Following $P(X)$, we can sample instances $X\subset\chi$, each of which is assigned a label $y\in Y$ in a label space Y. Let $D^s=(\chi^s,P(X^s))$ and $D^t=(\chi^t,P(X^t))$ be a source domain and a target domain, respectively. In this paper, the source domain contains abundant labeled Web images, and the target domain consists of a large number of unlabeled consume videos.

Given a source domain D^s and its corresponding learning task T^s, and a target domain and its D^t corresponding leaning task T^t. Here the domain $D^s\neq D^t$ but the task $T^s=T^t$. There are unlabeled samples in the target domain but some of these samples are available at training time. Our aim is to learn a classifier $f_w(\cdot)$ that can effectively predict the target domain data using the knowledge in both the source and target domains. In our annotation scheme, to make the best use of the target domain data we assume some unlabeled target samples can be seen during the training stage. Under such setting, the unlabeled target data can be used to optimize the predictive function. Thus the function learned in the source domain can be more adapted in the target domain through these unlabeled target domain data. Notice that this cross-domain learning process is also called transductive learning [16].

3.1 Cross-domain structural model

Following the above setting, we introduce our Cross-Domain Structural Model (CDSM) to transfer knowledge from Web images (source domain) to consumer videos (target domain) by jointly modeling the interaction between videos and images.

In this paper, we let each sample X consists of two components: X_{img} and X_{com}, where X_{img} represents low-level features extracted from images in the source domain or those extracted from keyframes of videos, and X_{com} denotes the common feature subspace that links the two heterogeneous feature spaces. The attribute feature of X is denoted by a K-dimensional vector $\mathrm{h}=[h_1,h_2,\ldots,h_K]$, where $h_k\in H$ indicates the weight of the k-th attribute and K is the total number of attributes. We consider a discriminative structural target function

$$f_{\mathrm{w}}: X\times Y\rightarrow\mathbb{R} \tag{1}$$

to annotate consumer videos, where $\mathrm{w}=[\alpha_y;\beta_j;\gamma_y;\mu_{j,k};\theta_{j,y}]$ is the parameter vector providing a weight for each feature component of a sample X.

During testing, f_w is used to predict the event label y^* of a consumer video x, namely

$$y^* = argmax_{y*\in Y} f_w(x, h, y) \tag{2}$$

We assume that $f_w(x,h,y)$ takes the following form:

$$f_w(x, h, y) = w^T \Phi(x, h, y), \tag{3}$$

where $\Phi(x,h,y)$ is a feature vector related to the sample x, its attribute h, and its class label y.

Given the labeled Web images (source domain) and unlabeled videos (target domain), the target classifier $w^T\Phi(x,h,y)$ is defined as

$$w^T \Phi(x, h, y) = \alpha_y^T \phi(x_{img}) + \sum_{j\in V} \beta_j^T \varphi_j(x_{img}) + \gamma_y^T \upsilon(x_{com}) + \sum_{j,k\in\varepsilon} \mu_{j,k}^T \psi(h_j, h_k) + \sum_{j\in V} \theta_{j,y} h_j, \tag{4}$$

The details of each term in Eq. (4) are described below.

Event model on the image feature: $\alpha^T_y\phi(x_{img})$:

This term provides the score measuring how well the image feature x_{img} matches the event class without considering attributes. $\phi(x_{img})$ represents the image/keyframe low-level feature vector of the sample. If we ignore other potential functions in Eq. (4) and only consider the object class model, the parameters $\alpha_{y\in Y}$ can be obtained by training a standard multi-class linear SVM.

Global attribute model: $\beta^T_j\varphi_j(x_{img})$

This term reflects the influence of a single attribute, which is used to indicate the relative presence of an attribute in the image/keyframe without considering its event class or other attributes. β_j is a weight for predicting the j-th attribute and $\varphi_j(x)=\phi(x)\cdot h_j$, where $\phi(x)$ is the image/keyframe low-level feature vector and h_j represents the weight of the j-th attribute. The attribute we used here is a continuous value. So if we only consider this model as our predictive function, the parameter $\{\beta_j\}_{j\in K}$ can be learned from a standard regression problem.

Event class model on the common feature $\gamma_y^T\upsilon(x_{com})$:

This term provides the score measuring how well the common subspace feature matches the event class. γ^T_y represents the weight of the common feature vector of the sample. x_{com} denotes the projection of x onto the common space. At training stage, x_{com} is represented by $x_{com}=x_{img}w_I$, where x_{img} indicates image features in the source domain. At testing stage, x_{com} is represented by $x_{com}=x_{vid}w_V$, where x_{vid} indicates the video features in the target domain. The matrices w_I and w_V are projection matrices that project different features onto a common feature subspace. The learning process of W_I and w_V will be discussed in Sec. 4.2.

The motivations of this function are two-fold. First, video features are integrated in the decision function at training stage with no supervised information in the target domain. This connects the image feature and the video feature in an indirect way. Second, the appearance of a same object might appear differently across different events.

Attribute-attribute interaction model $\mu_{j,k}\psi(h_j,h_k)$:

We believe that there are certain dependencies between different attributes. This

dependency potential

$$\psi(h_j, h_k) = h_j \cdot h_k \tag{5}$$

captures the relationship of the j-th attribute and the k-th attribute. Here μ is a symmetric matrix with the size of $|K|\times|K|$. Each element $\mu_{j,k}$ indicates the dependencies between attributes h_j and h_k. For example, since the attributes "cake" and "candle" tend to significantly often co-occur, the entries of $\mu_{j,k}$ will probably have large values.

Event-attribute interaction model $\theta_{j,y} \cdot h_j$:

This is scalar evaluates the relationship between the event class label y and the j-th attribute. For instance, let y correspond to the event "birthday" and the j-th attribute be "cake", then $\theta_{j,y}$ will probably have a large value since "birthday" co-occurs frequently with "cake".

3.2 Learning objective

Even though we can obtain the attributes automatically, we do not want to put too much trust on these weak semantic attributes. Since we do not have ground-truth attribute labels during training and testing, the problem cannot be solved through a standard classification problem. Therefore, we assume the attributes of keyframes are unobservable during testing stage and are treated as latent variables, a natural way to learn the parameter vector w is to use the latent SVM by the following objective function:

$$\min \lambda \|\mathrm{w}\|^2 + \sum_{n=1}^{N} \xi^{(n)} \text{ s.t.} \mathrm{w}^T \Phi\left(\mathrm{x}^{(n)}, \mathrm{h}^{(n)}, y^{(n)}\right) - \max_{\mathrm{h}} \mathrm{w}^T \Phi\left(\mathrm{x}^{(n)}, \mathrm{h}, y\right) \geq \Delta\left(y, y^{(n)}\right), \forall n, \forall y, \tag{6}$$

where λ controls the penalty factor, and $\Delta(y,y^{(n)})$ is a loss function indicating the cost of empirical error $y^{(n)}$,. Here we typically use the 0–1 loss $\Delta_{0/1}(\mathrm{y},\mathrm{y}^{(n)})$:

$$\Delta_{0/1}\left(y, y^{(n)}\right) = \begin{cases} 1 & y \neq y^{(n)} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

We convert the constrained optimization problem Eq. (6) into an equivalent unconstrained problem:

$$\min_{w} L(\mathrm{w}) = \lambda \|\mathrm{w}\|^2 + \sum_{n=1}^{N} R^n(\mathrm{w}), \tag{8}$$

where

$$R^n(\mathrm{w}) = \max_{y} \left(\max_{h} \mathrm{w}^T \Phi\left(\mathrm{x}^{(n)}, \mathrm{h}, y\right) + \Delta\left(y, y^{(n)}\right) \right) - \mathrm{w}^T \Phi\left(\mathrm{x}^{(n)}, \mathrm{h}^{(n)}, y^{(n)}\right). \tag{9}$$

We adopt a convex cutting plane method [7] to solve the optimization problem in Eq. (8).

This method aims to iteratively build an increasingly accurate piecewise quadratic approximation of $L(w)$ based on its sub-gradient:

$$\partial_w L(w) = 2\lambda \cdot w + \sum_{n=1}^{N} \Phi\left(\mathrm{x}^{(n)}, \mathrm{h}^{*}, y\right) - \Phi\left(\mathrm{x}^{(n)}, \mathrm{h}^{(n)}, y^{(n)}\right), \tag{10}$$

where $y^{*}=\max_y(\max_{\mathrm{h}} \mathrm{w}^T \Phi(\mathrm{x}^{(n)},\mathrm{h},y)+\Delta(y,y^{(n)}))$.

In order to solve the inference problem $\max(\max \mathrm{w}^T \Phi(\mathrm{x}^{(n)},\mathrm{h},y)+\Delta(y,y^{(n)}))$ we enumerate all the possible video class labels to find the optimal y and infer h by quadratic optimization under the fixed video class y.

4 Feature representation

Since the domains of image feature and video feature are different, we cannot use a single feature to fit the problem of both domains. Moreover, the target domain (consumer videos) has additional features (video features) of which we can take advantage. In this section, we discuss how to make the best use of the information from both domains.

4.1 Single domain feature representation

Low-level features: We extract SIFT feature [15] as image feature for both source domain images and target domain keyframes. As for videos in target domain, we extract space-time (ST) features, including Histograms of Oriented Gradient (HOG) and Histograms of Optical Flow (HOF) [8].

Mid-level features: We propose to use weak-semantic attributes as mid-level feature, which are comprised of classifier scores extracted automatically from Web images (i.e. scores from Classemes semantic classifiers [21]). We call these scores as weak semantic attributes for the reason that they are automatically obtained and can represent certain mid-level semantic meanings of an event. Each dimension of attribute corresponds to the score of a classifier trained on images returned by image search engines using the corresponding attribute as the query keyword. Different from traditional human labeling attributes, all of our weak semantic attributes are generated automatically. Moreover, the attribute in our work is a continuous value rather than binary pattern, which not only represents the absence of a semantic but also reflects the strength of the semantic. 2,659 classifiers trained on images returned by search engines of corresponding query words/phrases. There are 2,659 weak-semantic attributes in total. Since the attribute number is relatively large, the size of the attribute interaction parameter can be several million which makes the learning very difficult. We adopt Principal Component Analysis (PCA) to reduce the dimensionality of the attributes.

4.2 Cross domain feature representation

To effectively utilize the heterogeneous features from both domains, we introduce a common subspace of the source and target data for our task, in which the heterogeneous features from two domains can be compared. So that classifiers learned on the source domain can be used in the target domain by translating the features to the common subspace. CCA is a technique of joint dimensionality reduction across two (or more) feature spaces that provide heterogeneous

representation of the same instance. Formally, in our work, we need N samples of paired data $\{(I_1, V_1),\ldots,(I_N, V_N\}$, where each $I_i \in \mathbb{R}^m$ *and* $V_i \in \mathbb{R}^n$ denote the sample of image data and video data, respectively.

The goal is to learn two projection matrices $w_I \in \mathbb{R}^{d\times m}$ and $w_V \in \mathbb{R}^{d\times n}$ to maximize the canonical correlation:

$$\max_{w_I, w_v} \frac{\widehat{E}[\langle I, w_I\rangle\langle V, w_V\rangle]}{\sqrt{\widehat{E}\left[\langle I, w_I\rangle^2\right]\widehat{E}\left[\langle V, w_V\rangle^2\right]}} = \max_{w_I, w_V} \frac{{w_I}^T C_{IV} w_V}{\sqrt{{w_I}^T C_{II}} {w_V}^T C_{VV^w V}} \tag{12}$$

where $\widehat{E}$ denotes the empirical expectation of image data I and video data V, C_{IV} denotes the between-sets covariance matrix, and C_{II} and C_{VV} denote the within-sets covariance matrices for image and video data, respectively.

Since CCA is a supervised method, while the videos in the target domain are unlabeled, we cannot directly use CCA to connect features from these two domains. Fortunately, we notice that there is a correspondence between keyframes and video itself. Thus, we can use CCA to learn two projection matrixes w_I and w_V for image feature and video feature, respectively. And a common subspace can be built by projecting both feature spaces according to w_I and w_V. With this common feature space, knowledge can be transferred to across different feature spaces.

The solution of CCA can be found via a generalized eigenvalue problem [11], which is introduced briefly in the following.

The first d canonical components $\{w_I\}^d$ and $\{w_V\}^d$, where $d=\min(rank(I), rank(V))$, define a basis for projecting image feature space (e.g. SIFT) as well as video feature space (e.g. ST features) to a common feature space.

In training stage, each training image sample $\mathbf{x}$ is projected onto the common feature subspace according to $\{w_I\}^d$. Then we have a transform from image feature space to the common subspace: $\mathbf{X}_{img} \rightarrow \langle W_I, \mathbf{X}_{img}\rangle$ Similarly, we have a transform from video feature space to the common subspace: $\mathbf{X_{vid}} \rightarrow \langle W_V, \mathbf{X_{vid}}\rangle$ We denote this common subspace as $\mathbf{X_{com}}$. For a training sample, $\mathbf{X_{com}}$ is represented by $\mathbf{X_{com}} = W_I \mathbf{X_{img}}$. While $\mathbf{X_{com}}$ is represented by $\mathbf{X_{com}} = W_I \mathbf{X_{vid}}$ at testing period. Therefore, we can obtain a common subspace of two heterogeneous spaces through CCA embedding.

5 Experiments

5.1 Datasets and features

We collected the initial candidate image set by submitting a keyword per event to an image search engine (Google in our experiments). For each event, we keep the top 120 images returned by image search engine. We manually re-filter the candidate image set by removing irrelevant images to construct a less noisy source domain in our experiments. Fig. 2 shows some examples of the candidate Web image set. The left-most column represents the query keyword for each event. We use the following three video datasets as target data for performance evaluation.

Kodak dataset [8]. This dataset contains 195 consumer videos. Each video belongs to one of the six event classes: “birthday”, “picnic”, “parade”, “show”, “sports” and “wedding”).

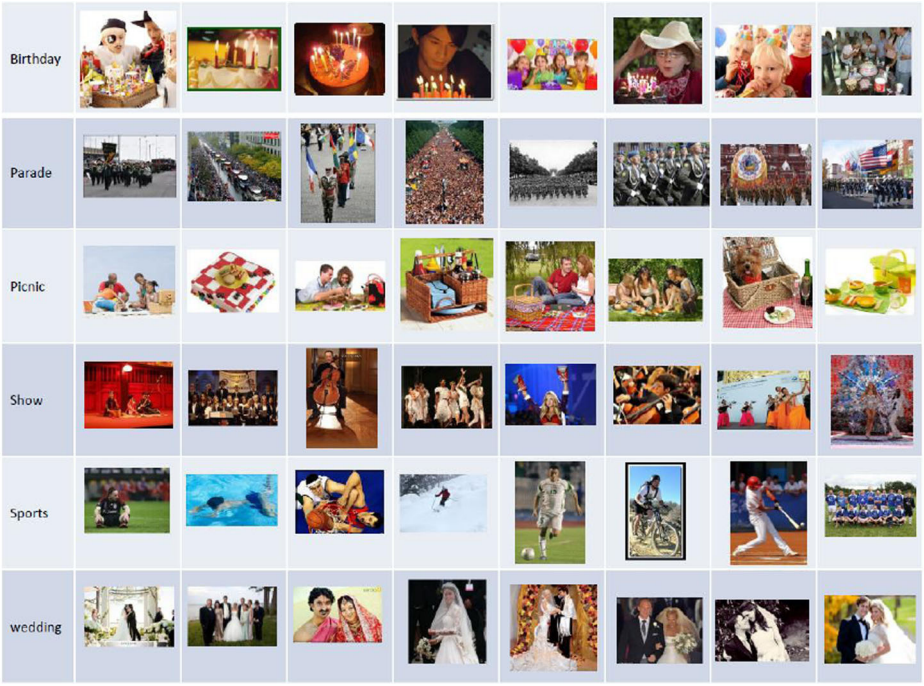

Fig. 2 Example images collected from web

YouTube dataset [8]. This dataset contains 561 consumer videos from YouTube with labels of the same six event classes as in the Kodak dataset.

CCV dataset [14]. This dataset is a consumer video dataset released by Columbia University, and contains 9,317 consumer videos from YouTube. All the 9,317 videos are divided to 4,659 training videos and 4,658 testing videos with 20 semantic categories. We exclude five non-event categories (i.e.,"playground", "bird", "beach", "cat" and "dog"). We merge "wedding ceremony", "wedding reception", and "wedding dance" into one event as "wedding" for the convenience of keyword based image searching. Finally, we evaluate different algorithms using 2,700 videos from the 13 event classes (i.e., "baseball", "basketball", "biking", "birthday", "graduation", "iceskating", "performance", "parade", "show", "skiing", "soccer", "swimming", and "wedding").

For each Web image, we extract SIFT features as well as weak-semantic attributes. Each attribute is the score of a Classemes semantic classifiers [21], and there are 2,659 weak-semantic attributes in total. Since the attribute number is relatively large, the size of the attribute interaction parameter can be several million which makes the learning very difficult. We adopt PCA to reduce the dimensionality of the attributes, and use the following three video datasets as target data for performance evaluation.

For each video, we randomly select one frame as its keyframe. For all sampled keyframes, we extract the static SIFT feature as their low-level feature. Moreover, ST features are also extracted from each video. We directly use the local ST features (i.e., HOG & HoF) provided by [8] for Kodak and YouTube datasets and by [14] for CCV dataset.

We build a codebook of 2,000 visual words by applying K-means on the SIFT feature extracted from the source domain. Then, each image of the source domain and

each keyframe of the target domain is represented as a 2,000- dimensional token frequency (TF) feature, by quantizing its SIFT features with respect to the visual codebook. For common feature subspace, we apply CCA on 4,659 videos from CCV database. Consequently, we finally get 2,000-dimensional common features for the keyframes and videos.

5.2 Results

5.2.1 Comparison of different methods

We compare the proposed CDSM method with the standard SVM (SVM_A), Domain Adaptation SVM (DASVM) [4], and Domain Selection Machine (DSM) [9], as these methods can work when all the target domain data are unlabeled. Since the standard SVM and DASVM can only handle the data that represent in the same type for both the source domain and the target domain, we only use the SIFT feature in these two methods to learn classifiers for the target domain. The standard SVM is learned by using the labeled training data only from the source domain. We learn the DASVM classifier with both the labeled images from the source domain and unlabeled keyframes from the target domain. For DSM, we only use its parametric target decision function $f(x)=\beta f^s(x)+w'\varphi(x)+b$ to train classifier, since our work focuses on how to transfer knowledge instead of what to transfer. For performance evaluation, we use the mean Average Precision (mAP) as the mean of APs over all events.

Table 1 shows the comparison of mAP between our method and the methods above. From the results, we have the following observations:

- All methods show promising results on transferring knowledge from Web images to consumer videos. This leads us to believe that knowledge can be transformed from image to video by means of static image features.
- DSM and our method outperform the other two methods, which clearly demonstrates that it is helpful to employ video features for transferring knowledge from image to video.
- DASVM is better than SVM_A. A possible explanation behind it is that the low-level feature distribution of Web images sometimes cannot correctly model the true underlying distribution of the keyframes of videos.
- Our CDSM method outperforms DSM, showing that translating image feature and video feature into a common feature subspace is more effective than handling them separately.
- The values of the AP obtained in this paper are relatively low. This is caused by: 1) no labeled data exists in the target domain, and the learning process is unsupervised; 2) all the keyframes are randomly selected from the target videos without any human intervention; 3) all the source domain images are directly collected from the Web with little human labor.

We also plot the per-event APs of all methods on the CCV, Kodak, and YouTube datasets in Figs. 3, 4, and 5, respectively. From the results, we can observe that the method DSM performs

Table 1 Comparison of mAP between our method and other methods

Method	SVM_A	DASVM [4]	DSM [9]	CDSM
CCV	8.52	10.90	11.71	14.57
Kodak	23.29	26.62	29.51	31.54
YouTube	23.29	28.63	30.54	32.70

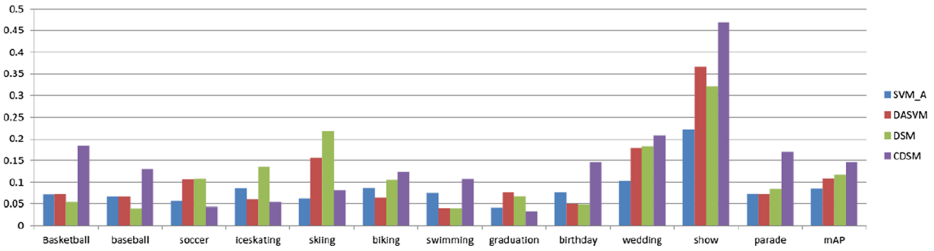

Fig. 3 Per-event Average Precission (APs) on CCV dataset

better on "skiing" and "iceskating" than other methods. This may because the status of these events is relatively simple and the DSM method can choose the directly related domain to transfer knowledge. However, when the status of the events becomes divers (e.g. parade and show), the performance of the DSM degrades. On the CCV dataset, all methods achieve the best performance on event "show". The reason for this phenomenon is that the videos of "show" are much more that those of "soccer" in the dataset, which also demonstrates the effectiveness of integrating the target videos in the learning process. This reasoning can also be proofed by Figs. 4 and 5 in which the videos of "sports" are much more than those of other events. In terms of mAPs, the performance on CCV datasets is worse than that on the Kodak and YouTube datasets. A possible explanation is that the categories we have in CCV are more complex than those in Kodak and YouTube, which makes the unsatisfactory performance.

5.2.2 Evaluation on different components

We evaluate the effect of the three feature components in our CDSM model on the CCV datasets which is more challenging than the other two datasets. We first exclude weak-semantic attributes and only use image feature (SIFT) and common subspace feature to learn the model (Eq. (13)). In this function, the two interaction models are also excluded since there is no attribute data in this model. The target function is given by

$$\mathbf{w}^T \Phi(x, h, y) = \alpha_y^T \phi(\mathbf{x}) + \gamma_y^T u(w(\mathbf{x})) \tag{13}$$

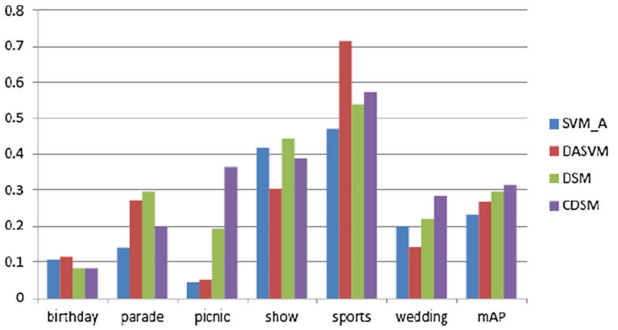

Fig. 4 Per-event Average Precission (APs) on Kodak Dataset

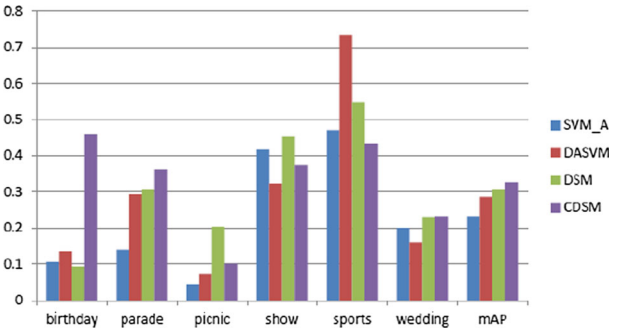

Fig. 5 Per-event Precission (APs) on youtube dataset

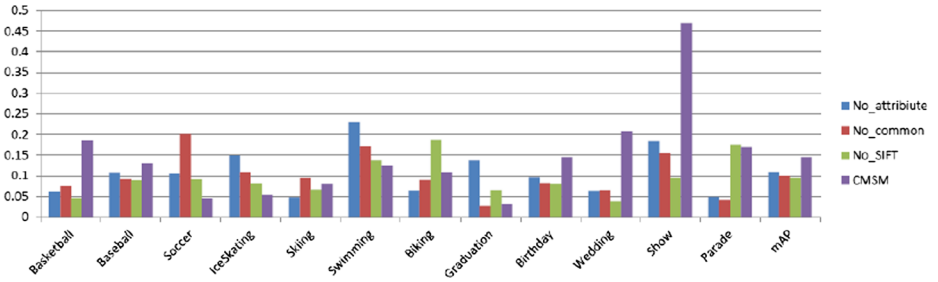

Fig. 6 Evaluation of different components in CDSM on the CCV dataset

Then weak-semantic attributes and image features are used to learn the following function:

$$\mathbf{w}^T\Phi(x,h,y) = \alpha_y^T\phi(\mathbf{x}) + \sum_{j\in V}\beta_j^T\varphi i(\mathbf{x}) + \sum_{j,k\in\varepsilon}\mu_{j,k}^T\psi(h_j,h_k) + \sum_{j\in V}\theta_{j,y}h_j \tag{14}$$

Finally, the function

$$\mathbf{w}^T\Phi(x,h,y) = \gamma_y^T v(w(\mathbf{x})) + \sum_{j,k\in\varepsilon}\mu_{j,k}^T\psi(h_j,h_k) + \sum_{j\in V}\theta_{j,y}h_j \tag{15}$$

is learned when image features are excluded. As shown in Figs. 6, 7, and 8, there is no consistent winner among these three feature types. Specifically, the performance of "birthday" degrades when

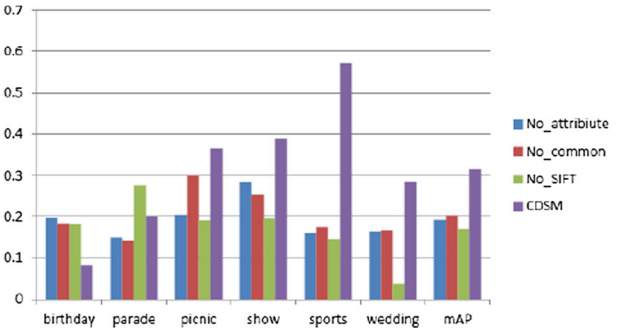

Fig. 7 Evaluation of different components in CDSM on the Kodak Dataset

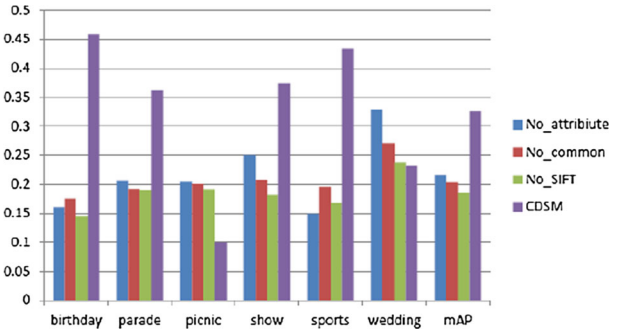

Fig. 8 Evaluation of different components in CDSM on the YouTube dataset

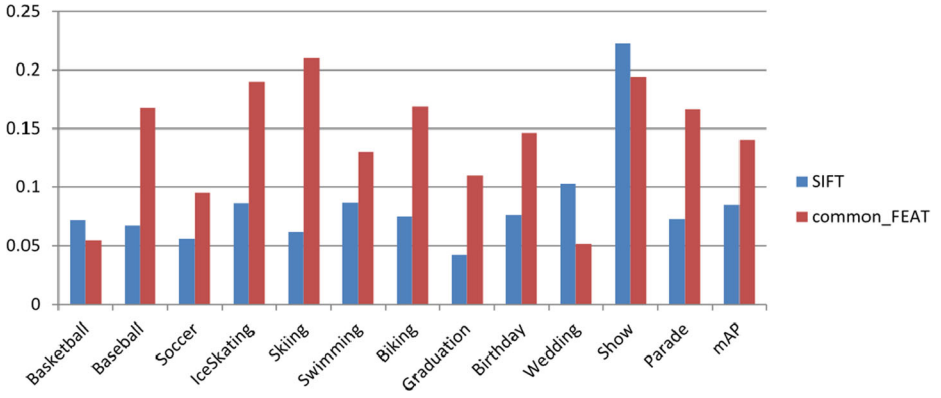

Fig. 9 Evaluation of common features on the CCV dataset

three types of information merges in Fig. 7, and the best performance is obtained when attributes are exclude from the model. This may because 1) the three types of information (e.g. SIFT, attributes and common feature) in the event "birthday" of the Kodak dataset is not complemented; 2) different attributes may be independent with each other in the event "birthday". For example, we can directly identify a birthday event simply by appearance of a birthday cake. From the mAPs over the whole dataset we can see that SIFT feature of images/keyframes affect the system more than the other two

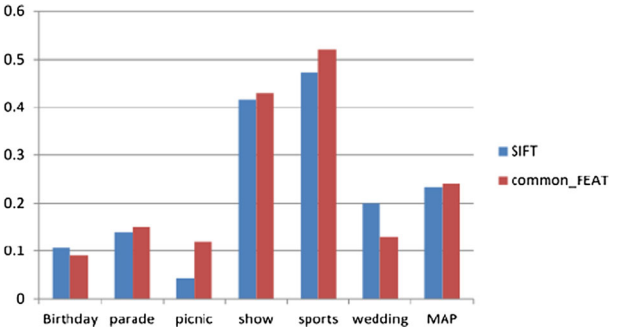

Fig. 10 Evaluation of common features on the Kodak dataset

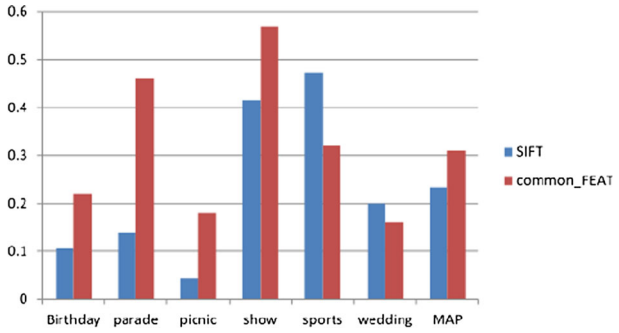

Fig. 11 Evaluation of Common features on the YouTube dataset

features (attributes and common feature). This validates the assumption we made earlier that most videos can be inferred from a single keyframe. When we excluded the two interaction models in our proposed method the problem is degraded into a standard SVM problem, so we didn't evaluate the performance of excluding the two interaction model alone. The best results appear when we integrate all the components in a unified learning process.

5.2.3 Evaluation on common feature

We further evaluate the proposed common feature subspace on the three datasets (Figs. 9, 10 and 11). The last bin of the histogram is the mAP on each dataset. In this experiment, only standard SVM learning method is applied to train classifiers on the common feature space, and the last bin of the histogram is the mAP on each dataset. We also plot SVM classification results with SIFT feature on these three datasets for comparison. In terms of mAPs, we can see that our common feature can produce positive influence on the knowledge transform. In most events, This also confirms that there are some connections between image feature and video feature in same video. For average performance on most events, we observe that the proposed common feature achieves comparable and stable performance.

6 Conclusion

In this paper, we have proposed a new knowledge transform method, called Cross-Domain Structural Model (CDSM) for annotating consumer videos by leveraging information provided by Web images. Our CDSM jointly learns low-level static features, weak-semantics, and common features in a unified framework. By introducing CCA to embed two different types of heterogeneous features into a common feature subspace, our work can adapt knowledge learned from the source domain to the target domain. Experiments on three real-world video datasets clearly demonstrate the effectiveness of our CDSM in transferring the most relevant knowledge. The evaluation on different components of the model indicates the strength of our weak semantics. Furthermore, the performance of the common feature validates the connection between images and videos.

In this paper, we focused on how to transform knowledge between different domains instead of what to transform. However, the inappropriate source domain images could deteriorate the system. In the future, we will study how to select less noisy source domain data to achieve more efficient transform. We will further consider to use non-linear common feature subspace to boost our annotation framework.

Acknowledgments This work was partially supported by National Natural Science Foundation of China (Grant no. 60973059, 81171407) and Program for New Century Excellent Talents in University of China (Grant no. NCET-10-0044).

References

1. Bel N, Koster C, Villegas M (2003) Cross-lingual text categorization. 7th European Conference on Research and Advanced Technology for Digital Libraries, Springer LNCS 2769:126–139
2. Berg T, Berg A, Shih J (2010) Automatic attribute discovery and characterization from noisy web data. ECCV 2010 1:663–676
3. Borth D, Ulges A, Breuel TM (2012) Dynamic vocabularies for web-based concept detection by trend discovery. In: Proceedings of the 20th ACM international conference on Multimedia, ACM, pp 977–980
4. Bruzzone L, Marconcini M (2010) Domain adaptation problems: a dasvm classification technique and a circular validation strategy. Pattern Anal Mach Intell, IEEE Trans on 32(5):770–787
5. Cai J, Zha Z-J, Zhou W, Tian Q (2012) Attribute-assisted re-ranking for web image retrieval. In: Proceedings of the 20th ACM international conference on Multimedia, ACM, pp 873–876
6. Cao L, Liu Z, Huang T (2010) Cross-dataset action detection. In: CVPR, IEEE, pp 1998–2005
7. Do T-M-T and Arti`eres T (2009) Large margin training for hidden markov models with partially observed states. In: Proceedings of the 26th Annual International Conference on Machine Learning. ACM, pp 265–272
8. Duan L, Xu D, Tsang I, Luo J (2010) Visual event recognition in videos by learning from web data. In: CVPR, IEEE, pp 1959–1966
9. Duan L, Xu D, Chang S-F (2012) Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach. In: Computer Vision and Pattern Recognition (CVPR), 2012 I.E. Conference on, IEEE, pp 1959–1966
10. Ferrari V, Zisserman A (2007) Learning visual attributes. Advances in Neural Information Processing Systems pp 433–440
11. Hardoon D, Szedmak S, Shawe-Taylor J (2004) Canonical correlation analysis: an overview with application to learning methods. Neural Comput 16(12):2639–2664
12. Hwang SJ and Grauman K (2010) Accounting for the Relative Importance of Objects in Image Retrieval. In: Proceedings of the British Machine Vision Conference (BMVC), Aberystwyth, UK
13. Ikizler-Cinbis N, Cinbis R, Sclaroff S (2009) Learning actions from the web. In: CVPR, IEEE, pp 995–1002
14. Jiang Y, Ye G, Chang S, Ellis D, Loui A (2011) Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In: Proceedings of the 1st ACM International Conference on Multimedia Retrieval, ACM, p 29
15. Lowe D (2004) Distinctive image features from scale-invariant keypoints. IJCV 60(2):91–110
16. Pan S, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359
17. Parikh D, Grauman K (2011) Interactively building a discriminative vocabulary of nameable attributes. In: Computer Vision and Pattern Recognition (CVPR), pp 1681–1688
18. Parikh D, Grauman K (2011) Relative attributes. In: ICCV, pp 1681–1688
19. Rasiwasia N, Costa Pereira J, Coviello E, Doyle G, Lanckriet G, Levy R, Vasconcelos N (2010) A new approach to cross-modal multimedia retrieval. In: Proceedings of the international conference on Multimedia, ACM, pp 251–260
20. Siddiquie B, Feris R, Davis L (2011) Image ranking and retrieval based on multiattribute queries, in: Computer Vision and Pattern Recognition (CVPR), pp 801–808
21. Torresani L, Szummer M, Fitzgibbon A (2010) Efficient object category recognition using classemes, Computer Vision–ECCV 2010 776–789
22. Vaquero D, Feris R, Tran D, Brown L, Hampapur A, Turk M (2009) Attribute based people search in surveillance environments. In: Applications of Computer Vision (WACV), 2009 Workshop on, IEEE, pp 1–8
23. Wang H, Wu X, Jia Y (2012) Annotating videos from the web images, in: International Conference on Pattern Recognition, IEEE, pp 2801–2804

24. Wu X, Jia Y (2012) View-invariant action recognition using latent kernelized structural svm. In: ECCV, pp 995–1002
25. Xu X-S, Jiang Y, Xue X, Zhou Z-H (2012) Semi-supervised multi-instance multi-label learning for video annotation task. In: Proceedings of the 20th ACM international conference on Multimedia, ACM, pp 737–740

Han Wang received the B.A. degree in computer science from National University of Defence Technology in 2008 and the Ph.D. degree in computer science from the Beijing Insititue of Technology in 2014. Her thesis is focus on the problem of video event analysis and retrieval, machine learning, computer vision, and image retrieval.

Xiabi Liu received the B.A. degree in computer science from the Huazhong University of Science and Technology in 1998 and the Ph.D. degree in computer science from the Beijing Institute of Technology in 2005. He is currently an associate professor at Beijing Institute of Technology. His research interests include machine learning, multimedia retrieval, pattern recognition, and computer vision.

Xinxiao Wu received the B.A. degree in computer science from the Nanjing University of Information Science and Technology in 2005 and the Ph.D. degree in computer science from the Beijing Institute of Technology in 2010. She is currently a lecturer in the School of Computer Science at the Beijing Institute of Technology. Her research interests include machine learning, computer vision, and human action perception.

Yunde Jia He is currently a professor in the School of Computer Science at the Beijing Institute of Technology. His research interests include multimedia analysis, machine learning and computer vision.