Multimed Tools Appl (2017) 76:6111–6126
DOI 10.1007/s11042-016-3253-1

Recognizing key segments of videos for video annotation by learning from web image sets

Hao Song[1] · Xinxiao Wu[1] · Wei Liang[1] · Yunde Jia[1]

Received: 24 May 2015 / Revised: 5 November 2015 / Accepted: 6 January 2016 /
Published online: 1 February 2016

Abstract In this paper, we propose an approach of inferring the labels of unlabeled consumer videos and at the same time recognizing the key segments of the videos by learning from Web image sets for video annotation. The key segments of the videos are automatically recognized by transferring the knowledge learned from related Web image sets to the videos. We introduce an adaptive latent structural SVM method to adapt the pre-learned classifiers using Web image sets to an optimal target classifier, where the locations of the key segments are modeled as latent variables because the ground-truth of key segments are not available. We utilize a limited number of labeled videos and abundant labeled Web images for training annotation models, which significantly alleviates the time-consuming and labor-expensive collection of a large number of labeled training videos. Experiment on the two challenge datasets Columbia's Consumer Video (CCV) and TRECVID 2014 Multimedia Event Detection (MED2014) shows our method performs better than state-of-art methods.

Keywords Video annotation · Key segment · Image set · Transfer learning

✉ Xinxiao Wu
wuxinxiao@bit.edu.cn

✉ Wei Liang
liangwei@bit.edu.cn

Hao Song
songhao@bit.edu.cn

Yunde Jia
jiayunde@bit.edu.cn

[1] Beijing Laboratory of Intelligent Information Technology, School of Computer Science,
Beijing Institute of Technology, Beijing 100081, China

1 Introduction

Video annotation has become an important topic in computer vision. It has been widely used in video indexing and retrieval. Unlike the videos in some simple action datasets (e.g. KTH [26] and Weizman [4]), annotating consumer videos is increasingly challenging because of cluttered background, large intra-variations and complex camera motion in videos. Consumer videos are shot by consumers with mobile phone or video camera and uploaded to the Web.

Some video annotation methods [3, 8, 13, 21, 32] demonstrated promising result when there is a large amount of labeled training videos. Since annotating abundant consumer videos is time-consuming and the training examples is insufficient, domain adaptation [12, 20, 24] methods have been employed over a wide range of applications. Domain adaptation methods consist of two main portions, unsupervised and semi-supervised domain adaptation. We require several labeled image sets and a few labeled target videos, so our method belongs to semi-supervised domain adaptation.

Most researchers have focused on transfer learning in video annotation. The major difference is the resource type in the source domain. In contrast to videos, images can be easily queried from Web engines, and the images queried by the keywords contain more useful information than videos because the videos involves more variations on the Web. Therefore, several researchers tried to leverage loosely labeled Web Images for complex video annotation [10, 14]. In these methods, holistic video level features are aggregated over the entire videos to describe the complex events. A complex video usually contains a slice of certain key segments, this motivates us to build a joint framework to simultaneously infer the event label of the whole video and recognize the key segments of video. Each key segment is one shot clip in videos, and the clip contain its own semantic meaning corresponding to one concept.

In this paper, we propose an approach of simultaneously inferring the labels of unlabeled consumer videos and recognizing the key segments. We transfer the knowledge learning from the loosely labeled Web image sets (source domain) to the event videos (target domain). For example, a high-level complex event *Marriage Proposal* contains several key segments, *going down on one knee, people kissing, put ring on finger, people hugging, and so on*. We collect each image set by querying the concept-level keywords of the related key segments, as shown in Fig. 1.

Specifically, we train SVM classifiers using the images from each image set in the source domain. Then, we introduce an adaptive latent structure SVM that adapts the pre-learned SVM classifiers to the target domain. Due to the non-availability of the ground-truth of the key segments, the locations of key segments are treated as latent variables. In order to handle the noise knowledge of irrelevance with complex videos, we propose a regularization between the source domain and the target domain, assuming that the decision value of unlabeled videos in the target domain should be close to the predictions of the source classifiers. Additionally, we adopt weights of each image set to judge its contribution in transferring the knowledge to the key segment and develop a new iterative algorithm to jointly learn the weights and the target classifier. In our method, we can infer the labels of the unlabeled target videos and recognize the key segments of the videos at the same time. Figure 2 shows the framework of our method.

The main contributions of our method are three folds: 1) We propose a new approach of inferring the label of consumer videos and at the same time recognizing their key segments. 2) We transfer the knowledge learning from the image sets to divide complex videos into

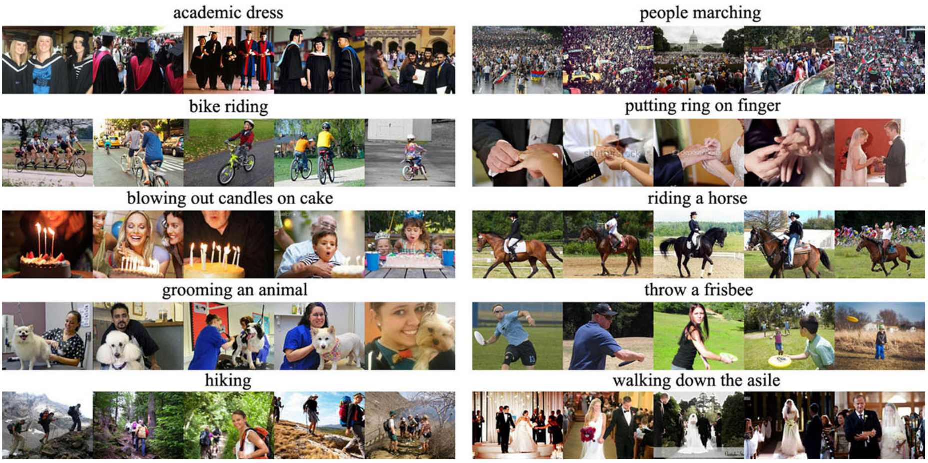


Fig. 1 Images queried by the concept-level keywords which are defined manually

several key segments. 3) We develop a new iterative algorithm to jointly learn the weights of each image set and the target classifier.

2 Related work

Domain adaptation method has been widely used in event video annotation. Chen et al. [6] proposed a MDA-HS method to learn an optimal target classifier by seeking the optimal weights for the different type features of data in source domains. Sun et al. [29] presented a two stage domain adaptation method which simultaneously takes account of the marginal probability differences and the conditional probability differences between source and target domain. Li et al. [18] transformed the two different domain into a common subspace and learned the feature mapping function to reduce the gap between both domains. Baktashmotlagh et al. [1] also minimized the empirical distributions of the source and target examples by utilizing the Domain Invariant Projection approach. Zhang et al. [34] leverage the abundant Web images to learn the noise-resistant classifiers to model the event-centric semantic concepts to promote event detection. The concepts are encoded in the knowledge base to narrow the semantic gap between complex events. Yang et al. [33] also apply the pre-cleansing Web images and a limited number of training videos to learn a robust transfer video indexing model which processes the domain gap between images and videos. For all the above methods, videos in source domain are required. To guarantee the recognition accuracy, the number of videos should be as large as possible. Despite the sources of videos from the Web have become rich, the images can be more easily and precisely collected.

Ikizler-Cinbis et al. [14] proposed to leverage increasingly loosely labeled web images to train action models to identify actions in videos. Duan et al. [10] proposed a multiple source domain adaption method by selecting the most relevant image source domains for annotating videos. Tang et al. [30] presented a novel self-paced domain adaptation framework to adapt the detector from images to videos. Wang et al. [31] introduced an Group-based Domain Adaptation learning framework to leverage different web images to adapt the group

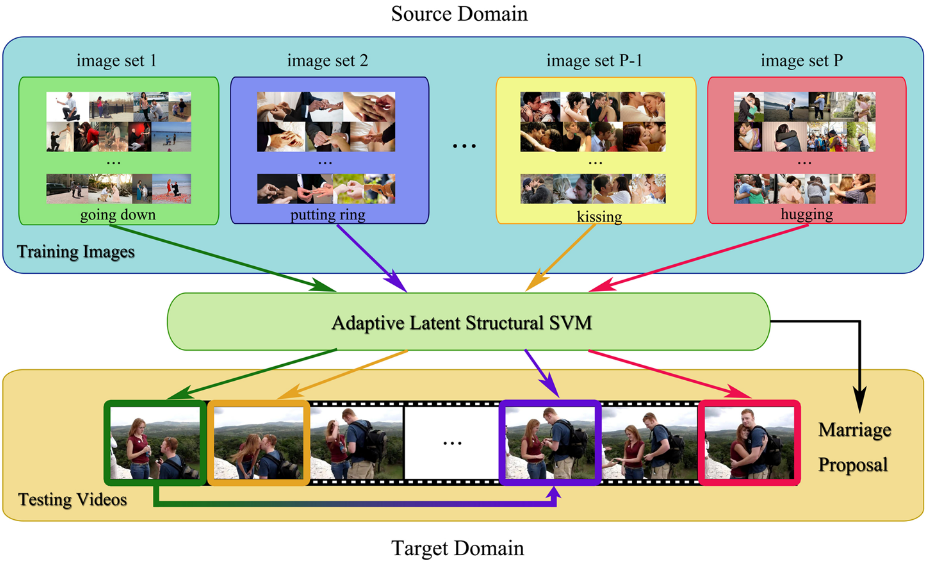


Fig. 2 A complex video contains several key segments. As an example, a high-level event *Marriage Proposal* contains *going down on one knee, people kissing, put ring on finger, people hugging, and so on*, the image sets are queried by the concept of the related key segments. The knowledge of each image set is transferred to the related key segment. P is the number of Web image sets, and the number of the key segments of videos

classifiers to the target domain with different weights. Fang et al. [12] proposed a multi-source transfer learning method. The method can extract the shared subspace between source and target domains. Schroff et al. [25] used textual features to rank the initial set of images and then utilized the first image set as the training set, expecting a support vector machine (SVM) to handle the rest of the noise. Ni et al. [23] leveraged data collected from the Web for age estimation in facial images. For age estimation, the visual details of the face (such as wrinkles, etc.) are important; whereas, for pose evaluation and action recognition, the variability in the articulation of the human body brings different challenges. In these methods, the event video annotation is achieved by estimating the label of the whole video. In this paper, complex videos consist of a slice of key segments and our method can simultaneously infer the labels of the unlabeled consumer videos and recognize their key segments.

Recently, a few researchers also focused on detecting the segments of videos for video annotation. Sun et al. [28] proposed an evidence location model to discover the video segments for events classification and recounting. And they also leveraged the detected oriented discriminative segments in videos and the descriptions of segments for event detection and recounting [27]. Li et al. [19] presented a dynamic pooling method to quantize the attribute segments for activity recognition. Cheng et al. [7] developed an automatic recognition tool to segment the videos into sequences for wedding ceremony event recognition. Bhattacharya et al. [2] divided videos into several segments, and discovered the minimally needed segments in each event to judge the presence of the events in complex videos. Different from these methods, our method recognizes a key segment by transferring the knowledge of the Web images sets queried by the concept-level keyword of the key segments without collecting a large number of labeled videos of time-consuming and labor-expensive.

3 The proposed framework

3.1 Formulation

We use image sets to recognize the key segments of consumer videos, and manually define a collection of semantic concepts $\mathbf{C} = \{C_1, C_2, \cdots, C_N\}$, where C_i is the i-th concept in the collection. For example, we define 55 concepts and 188 concepts for the CCV dataset and the MED2014 dataset respectively, including action-related concepts (e.g. "human waving", "person laughing"), object-related concepts(e.g. "baseball", "horse" and "dog"), scene-related concepts(e.g. "outside", "kitchen"). We assume that each key segment of videos only has a semantic concept. The images in each image set are queried by a concept-level keyword from the Web image search engine. The knowledge of each image set is transferred to the related key segment of videos.

Formally, for each event, we collected P Web image sets $(x^s_{p,i}, y^s_{p,i})|^{N_p}_{i=1}$, $p \in \{1, ..., P\}$, where N_p is the total number of images in the p-th image set from the source domain $\mathcal{D}^S$. $x^s_{p,i}$ is the i-th training image with its label $y^s_{p,i} \in \{1, -1\}$ in the p-th image set. Each pre-learned source classifier $f^s(x^s_p) = w^{s\prime}_p \Phi(x^s_p)$ is learned using the images in an image set. where w^s_p is the template trained by the p-th image set. $\Phi(\cdot)$ is the feature mapping function. The labeled consumer videos from the target domain $\mathcal{D}^T_l$ are denoted by $(x^t_i, y^t_i)|^{N^l_t}_{i=1}$, where N^l_t is the number of labeled videos. We define $\mathcal{D}^T_u$ as the set of unlabeled videos in the target domain.

3.2 Model

Firstly, we train the source-domain classifiers w^s using the image sets collecting from the Web search engine. Then, we adapt a set of w^s to the target domain to generate an optimal target classifier $\boldsymbol{w}^t$ using a relatively small number of labeled target-domain samples. The target classifier is composed of several key segment classifiers. The label of a video is determined by its key segment classifier jointly.

We define the collection of the SVM classifiers in the source domain as $\boldsymbol{w}^S = [w^s_1, w^s_2, \cdots, w^s_P]$, where P is the number of image sets in the source domain. Note that w^s_p is an image set classifier trained by the images from p-th image set. To this end, we learn the following target classifier f^t for any consumer video sample x^t_i by fusing the decisions from different key segments:

$$\begin{aligned} f^t(x^t_i) &= \boldsymbol{w}^t \cdot \Phi(x^t_i) \\ &= \arg\max_{h_{i,p}} \sum_{p=1}^{P} w^{t\prime}_p \cdot \Phi(x^t_i, h_{i,p}), \end{aligned} \tag{1}$$

where w^t_p is a classifier for the p-th key segment of the videos. $h_{i,p}$ is the location of the p-th key segment in the video x^t_i. $\Phi(x_i, h_{i,p})$ is the feature mapping function for the test video x^t_i at the position $h_{i,p}$. Our model can simultaneously infer the label of the whole test video with $f^t(x^t_i)$ and detect the location $h_{i,p}$ of the key segments.

3.3 Learning

The SVM classifier in the source domain is defined as w^s, we leverage the classifier to predict the label of data through the decision function, $y = f^s(x) = w^s \cdot \Phi(x)$. Our goal is to learn an optimal target classifier $f^t(x)$ to accurately classify the data in target domain. In A-SVM, this is implemented by a "delta function" in the form of $\delta f(x) = \Delta w \cdot \Phi(x)$:

$$\begin{aligned} f^t(x) &= f^s(x) + \Delta f(x) \\ &= w^s \cdot \Phi(x) + \Delta w \cdot \Phi(x) \end{aligned} \tag{2}$$

We also introduce the weight $\boldsymbol{\beta}$ to describe the different contributions of different image sets in transferring the knowledge of images to the key segments, $\boldsymbol{\beta} = [\beta_1, \cdots, \beta_P]'$. In order to learn these adaptation weights, we add a regularization term $\|\boldsymbol{\beta}\|$, and use a scalar parameter λ_2 to control the relative penalty to the hyperplane parameter regularization term. We try to find a new target classifier $\boldsymbol{w}^t$ that can be close to the original hyperplane $\boldsymbol{w}^s$ using the term $\Delta \boldsymbol{w}$. $\Delta \boldsymbol{w}$ enforces the target model $\boldsymbol{w}^t$ to be relatively close to the original model $\boldsymbol{w}^s$ using the coefficient vector $\boldsymbol{\beta}$. $\Delta \boldsymbol{w} = [\Delta w_1, \cdots, \Delta w_P]'$, where $\Delta w_p = w_p^t - \beta_p w_p^s$, and $p \in 1, 2, \cdots, P$. In order to learn a newly target hyperplane for the target video x_i^t, we introduce the latent structural SVM to design an objective function as follows:

$$\begin{aligned} &\min_{\boldsymbol{w}^t, \boldsymbol{\beta}} \tfrac{1}{2}\left(\lambda_1 \|\Delta \boldsymbol{w}\|^2 + \lambda_2 \|\boldsymbol{\beta}\|^2\right) + C \sum_{i=1}^{N_l^t} \xi_i \\ s.t.\ & \\ &\max_{h_i} \boldsymbol{w}^t \Phi(x_i^t, h_i) - \max_{h^t} \boldsymbol{w}^t \Phi(x_i^t, h^t) \\ &\geq \mathbf{L}(y_i^t, y^t) - \xi_i, \\ &\sum_{p=1}^{P} \beta_p = 1, \\ &\xi_i \geq 0 \qquad \forall (x_i^t, y_i^t) \in D_l^T, \\ &\beta_p > 0, \forall p \in 1, 2, \cdots, P, \end{aligned} \tag{3}$$

where λ_1, λ_2, C are trade-off parameters, and $\mathbf{L}(y_i^t, y^t)$ is $0 - 1$ loss function defined by $\mathbf{L}(y_i^t, y^t) = 0$ if $y^t = y_i^t$ and 1 otherwise. y_i^t is the ground truth label, and y^t is the predicted label of the sample. $\boldsymbol{w}^t \Phi(x_i^t, h_i)$ is defined as

$$\begin{aligned} \boldsymbol{w}^t \Phi(x_i^t, h_i) &= [w_1^t, ..., w_p^t]' \Phi(x_i^t, h_{i,p}) \\ &= \sum_{p=1}^{P} w_p^{t\,\prime} \cdot \Phi(x_i^t, h_{i,p}), \end{aligned} \tag{4}$$

$h_{i,p}$ is the i-th sample ground truth location for the p-th key segment classifier. $\boldsymbol{w}^t \Phi(x_i^t, h^t)$ is given by

$$\begin{aligned} \boldsymbol{w}^t \Phi(x_i^t, h^t) &= [w_1^t, ..., w_p^t]' \Phi(x_i^t, h_p) \\ &= \sum_{p=1}^{P} w_p^{t\,\prime} \cdot \Phi(x_i^t, h_p), \end{aligned} \tag{5}$$

where h_p is the sample's prediction location for the p-th key segment classifier. $\Phi(x_i^t, h_p)$ is the feature mapping function of labeled video x_i^t at the location h_p for the p-th key segment classifier of the event.

We adopt a non-convex cutting plane method proposed in [9] to solve (3). We simplify (3) as follows:

$$\min_{\boldsymbol{w}^t, \beta} \mathbf{L}(\boldsymbol{w}^t, \boldsymbol{\beta}) = \frac{1}{2}(\lambda_1 \|\Delta \boldsymbol{w}\|^2 + \lambda_2 \|\boldsymbol{\beta}\|^2) + C \sum_{i=1}^{N_t^l} \mathbf{R}_i(\boldsymbol{w}^t), \tag{6}$$

the loss function $\mathbf{R}_i(\boldsymbol{w}^t)$ can be represented as

$$\begin{aligned}\mathbf{R}_i(\boldsymbol{w}^t) = &\big(\mathbf{L}(y_i^t, y^t) + \max_{h^t} \boldsymbol{w}^t \Phi(x_i^t, h^t)\big) \\ &- \max_{h_i} \boldsymbol{w}^t \Phi(x_i^t, h_i).\end{aligned} \tag{7}$$

where

$$\begin{aligned} h_i &= \arg\max_{h_i} \boldsymbol{w}^t \Phi(x_i^t, h_i), i \in 1, 2, \cdots, N_t^l \\ y^t &= \arg\max_{h^t} \boldsymbol{w}^t \Phi(x_i^t, h^t).\end{aligned} \tag{8}$$

The non-convex cutting plane method aims to iteratively build an increasingly accurate piecewise quadratic approximation of $\mathbf{R}_i(\boldsymbol{w}^t)$ based on its sub-gradient $\partial_{\boldsymbol{w}^t}\mathbf{R}_i(\boldsymbol{w}^t)$. Now the key issue is how to compute the sub-gradient $\partial_{\boldsymbol{w}^t}\mathbf{R}_i(\boldsymbol{w}^t)$.

Then sub-gradient $\partial_{\boldsymbol{w}^t}\mathbf{R}_i(\boldsymbol{w}^t)$ can be given by:

$$\partial_{\boldsymbol{w}^t}\mathbf{R}_i(\boldsymbol{w}^t) = \lambda_1|\Delta \boldsymbol{w}| + C(\Phi(x_i, h^t) - \Phi(x_i, h_i)). \tag{9}$$

Given the sub-gradient $\partial_{\boldsymbol{w}^t}\mathbf{R}_i(\boldsymbol{w}^t)$ computed according to (9), we can minimize $\mathbf{L}(\boldsymbol{w}^t, \boldsymbol{\beta})$ using the method with fixed $\boldsymbol{\beta}$. Then, with fixed $\boldsymbol{w}^s$ and $\boldsymbol{w}^t$, the objective function in (3) is easily reduced to a quadratic programming problem with some constraints of $\boldsymbol{\beta}$, and $\boldsymbol{\beta}$ can be solved by the traditional QP method. The detailed algorithm is summarized in Algorithm 1.

Algorithm 1 Algorithm for Event Video Annotation

```
Input: $(x_{p,i}^s, y_{p,i}^s)|_{i=1}^{N_p}$: image sets from source domain, $p \in 1, 2, \cdots, P$;
       $(x_i^t, y_i^t)|_{i=1}^{N_t^l}$: the labeled target domain training videos;
Output: $\boldsymbol{w}^t$: the updated target classifiers;
        $\beta$: the best weight for $\boldsymbol{w}^s$ in transfer learning process.
train image set classifier $w_p^s$ using Web images from image sets
repeat
    Step 1:compute $\mathbf{w}^t$ with fixed $\beta$
    for $i = 1$ to $N_t^l$ do
        $h_i \leftarrow \arg\max_{h_i} \boldsymbol{w}^t \Phi(x_i, y_i, h_i)$
        $y^t \leftarrow \arg\max_{y^t} \boldsymbol{w}^t \Phi(x_i, h^t)$
    end
    Compute $\partial_{\boldsymbol{w}^t}\mathbf{R}_i(\boldsymbol{w}^t)$ according to Eq.9
    Update $\mathbf{w}^t$ using the cutting plane method proposed in [9]
    Step 2:compute $\beta$ with fixed $\mathbf{w}^t, \mathbf{w}^s$
    Compute $\beta$ according to Quadratic Programming method.
until Convergence of objective function (3) cannot be decreased below tolerance $\delta$;
```

4 Experiment

4.1 Datasets

We evaluate our method on two datasets: the Columbia's consumer video CCV [16] and the TRECVID 2014 Multimedia Event Detection dataset [22]. We use the mean of Average Precision(mAP) [31] over all the event classes for performance evaluation.

CCV dataset 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 annotation, we do

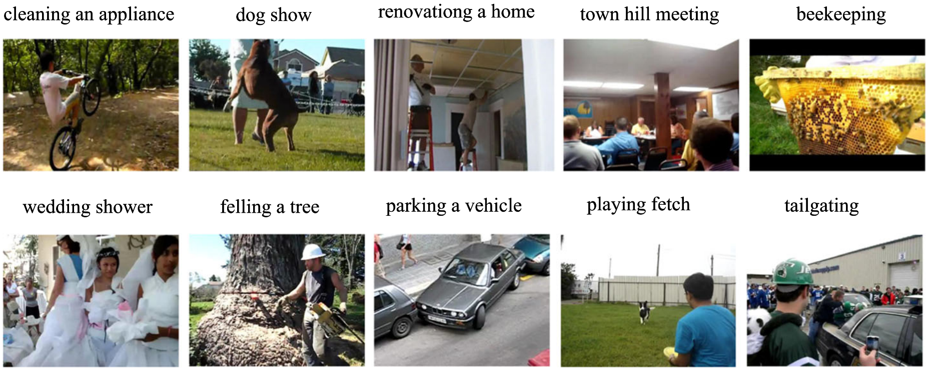


Fig. 3 The frames on the MED2014 dataset

not consider the non-event categories(e.g. “bird”, “beach”, “cat”, “dog” and “playground”). Also, we combine the events of “wedding ceremony”, “wedding reception” and “wedding dance” into the event of “wedding”. The events of “music performance” and “non-music performance” into the event of “show”. Finally, the dataset includes twelve event categories: “baseball”, “basketball”, “biking”, “birthday”, “graduation”, “ice-skating”, “show”, “parade”, “skiing”, “soccer”, “swimming” and “wedding”.

TRECVID 2014 Multimedia Event Detection dataset with 40 categories of events. It is well known that TRECVID is the largest public available video recognition corpus. The event categories are quite diverse, including life, instructional and sport events, *etc.*. We use a part of the whole video dataset. The partition of “*Background*” contains 4,983 background videos which do not belong to any target event. This set of videos can be used as negative samples in the training procedure. The partition of “*10EX*” contains totally 10 positive videos for each of the pre-defined 20 event classes. The partition of “*MEDTest*” contains 29,200 videos, in which there are about 25 positive samples for each event class and 26717 negative videos, they are not belonging to any event. Figure 3 shows several event class samples.

The Image Dataset is constructed by the images of each image set. The images are queried from the Web using the Google search engine. For each image set, we select the top ranked 200 images and the corrupted images with invalid URLs are discarded. We ensure that the photos we used are with full color. About 35,828 images are collected in our image dataset. In the experiment, almost 150 positive and 400 negative samples are selected to train the basic image set source classifiers. Table 1 shows the keyword of the concepts we set up in several event classes in detail. In our method, we utilize a limited number of labeled videos and abundant labeled Web images simultaneously.

4.2 Experiment Setup

We extract 4096-dimensional feature vectors of images and frames from videos by using the Caffe[15] implementation of the CNNs described by Krizhevsky [17]. Each target video is divided into a sequence of short clips and each clip is represented by several frames which are randomly selected from the corresponding clip using a sliding window. Features are

Table 1 The keyword of the concepts in some event classes

event exist concept	*Basketball*	*beekeeping*	*Soccer*	*playing fetch*	*Skiing*	*dog show*	*Picnic*	*Birthday*	*Graduation*	*Wedding*	*Parade*	*bike trick*	*Swimming*	*rock climbing*	*marriage proposal*	*winning a race*	*Baseball*
academic address									×								
arm pull													×				
baseball field																	×
baseball pitcher																	×
basketball court	×																
basketball kids	×																
basketball shots	×																
calling an animal				×													
bike riding												×					
blowing candles								×									
cheering											×						
honey combs		×															
chorus												×					
clapping									×		×						
eating							×										
dancing										×							
calling an animal				×													
ball shot	×																
hugging									×						×		
waving	×		×													×	
jumping	×				×							×		×			
kissing										×					×		
football field			×														
dog				×		×											
indoor soccer			×														
laughing							×	×									
singing							×	×									
running	×				×		×									×	×
swimming pool													×				
throw				×													×
dog run						×											
bee keeper		×															
slalom					×												
walking down the asile										×							
burning candles								×									
graduation ceremony									×								
basketball gym	×																
kicking			×														
picnic food							×										
marching											×						
bee		×															
outside		×										×				×	
swimwear													×				
going down on knee															×		
putting ring on finger															×		
walking											×			×			
bride & groom										×							
pulling on leash						×											
person petting						×											
throw a frisbee				×													
snow field					×												
rock mountain														×			
backpacker														×			
swimming ring													×				
passing water																×	

The top row shows some event categories.

The left-most row shows the detailed concepts we set up in the experiment.

computed by forward propagating a mean-subtracted 227 × 227 RGB image through five convolutional layers and two fully connected layers. We refer readers to [15, 17] for more network architecture details.

We compare our annotation approach with state-of-art methods, including the standard basic SVM (*B_SVM*), the target domain SVM (*SVM_T*), the single domain adaptation method of Domain Adaptive SVM (DASVM) [5], the multi-domain adaptation methods of Domain Adaptation Machine (DAM) [11], and the Domain Selection Machine (DSM) [10]. For DASVM, it is a semi-supervised leaning method, all the concepts belonging to the same event category are combined as the single source domain, a few of labeled target videos are

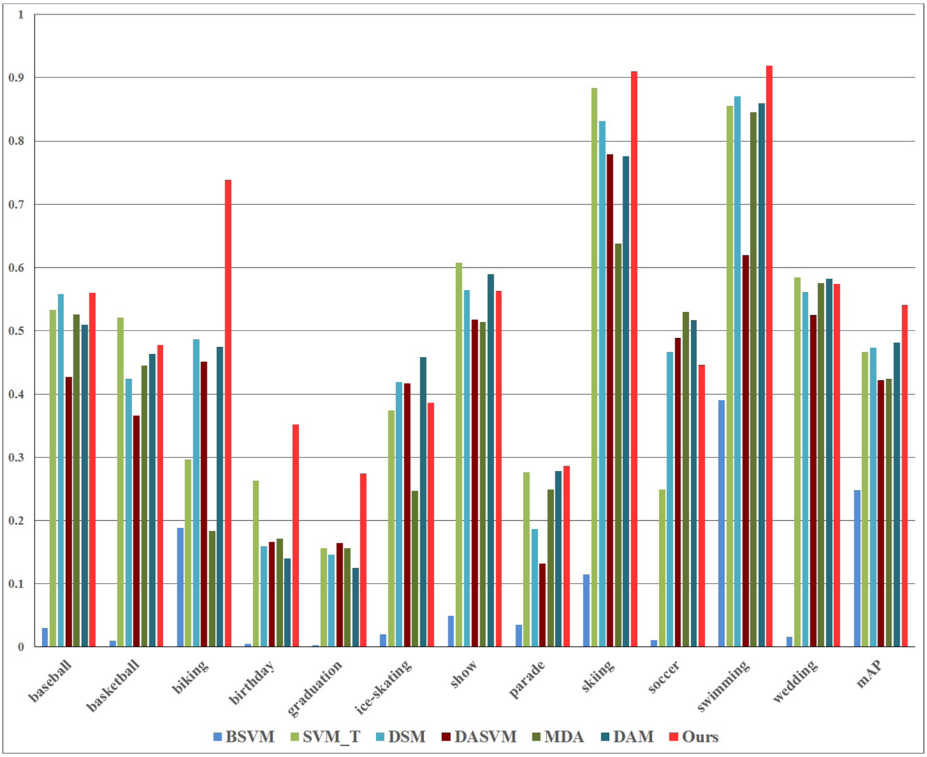


Fig. 4 The results of all the methods on the CCV dataset

also used in this method. For DSM and DAM, these methods are multi-source domain adaptation framework, in comparison with our method, we treat every concept belonging to the same event as a source domain. The SVM parameter C is set the same in all the methods. Because we have different settings from these methods, we re-implemented all the above state-of-the-art methods used for comparison on our image and video dataset.

In our method, we divided video into several short clips, we can recognize the semantic short clips which have high probability in the corresponding concept classifier. With the concept classifiers, each shot clip which has been detected should be treated as a key segment. The most time consuming part of our algorithm is the computation of $\boldsymbol{w}^t$. Our optimization speed relies on the initial solution. We set the initial solution as all 1 in our experiment. The average time of the procedure of our optimization method in one event class consumes about 753.2 (seconds). The proposed approach is implemented in MATLAB on a Intel Core 3.4GHz processor with 16GB RAM.

Table 2 The results on comparison of mAP (%) between our proposed method and several other methods on the CCV and MED2014 datasets

Method	B_SVM	SVM_T	DASVM	DAM	DSM	Ours
CCV	47.07	48.57	42.1	48.08	47.26	53.27
MED2014	3.46	7.31	7.58	7.66	8.22	11.19

Table 3 The results of mAP(%) on MED2014 dataset

Events	B_SVM	SVM_T	DSM	DASVM	DAM	Ours
attempting a bike trick	5.50	3.00	4.07	2.80	5.04	10.83
cleaning an appliance	1.82	0.95	1.06	1.60	2.78	2.73
dog show	11.82	18.85	11.96	13.66	10.9	20.42
giving directions to a location	0.56	0.43	0.61	0.89	0.81	0.32
marriage proposal	2.04	0.32	0.51	1.21	8.21	13.16
renovating a home	0.87	1.95	1.82	1.03	1.81	0.67
rock climbing	9.54	4.90	4.36	5.92	4.75	13.82
town hill meeting	2.08	3.5	2.95	4.23	0.91	8.83
winning a race without a vehicle	1.24	11.5	10.27	15.23	18.02	13.72
working on a metal crafts project	0.41	1.06	2.42	1.33	3.43	0.84
bee keeping	5.40	39.04	43.24	30.80	37.13	50.47
wedding shower	0.47	1.57	10.48	10.36	8.49	28.22
non-motorized vehicle repair	3.46	24.78	3.71	14.89	13.64	5.93
fixing musical instrument	1.41	3.31	1.50	2.45	1.54	2.08
horse riding competition	5.75	11.94	14.66	17.27	8.53	24.75
felling a tree	2.83	1.50	4.98	3.62	6.01	5.97
parking a vehicle	10.67	8.21	11.47	10.74	16.31	17.54
playing fetch	0.55	1.82	6.27	5.17	4.27	4.93
tailgating	0.54	5.60	9.97	9.39	10.58	3.4
tuning a musical instrument	2.20	2.04	5.22	5.63	1.25	4.95
mAP	3.46	7.31	7.58	7.66	8.22	11.19

4.3 Results

Figure 4 illustrates the per-event AP of all the methods on the CCV dataset. Table 2 and Table 3 show the mAP results of all the transfer learning methods on the dataset CCV and MED2014. From the results, it is obvious that our method performs better than other transfer learning methods. We get a better result in event “attempting a bike trick”, “dog show”, “marriage proposal” and so on. These events are complex, and we select more proper concepts for them. For the result displayed in other events, our method may not work well, the events such as “giving directions to a location”, “renovating a home”, “working on a metal crafts project” may be abstract and the queried image sets in the source domain also may influence the final result. Moreover, as shown in Fig. 4, in the per-event AP, there is no consistent winner among all the methods on per-event AP. This explains that the irrelevance of the data between source and target domains may influence the performances of transfer learning methods. Our method performs much better in more complex events, such as “biking”, “birthday”, “marriage proposal”, “bee keeping” and “graduation”, which clearly

Table 4 Result of mAP(%) on CCV and MED2014 dataset with equal $\boldsymbol{\beta}$ and different $\boldsymbol{\beta}$ in our method.

Types of $\boldsymbol{\beta}$	equals	different(ours)
CCV	48.67	53.27
MED2014	9.82	11.19

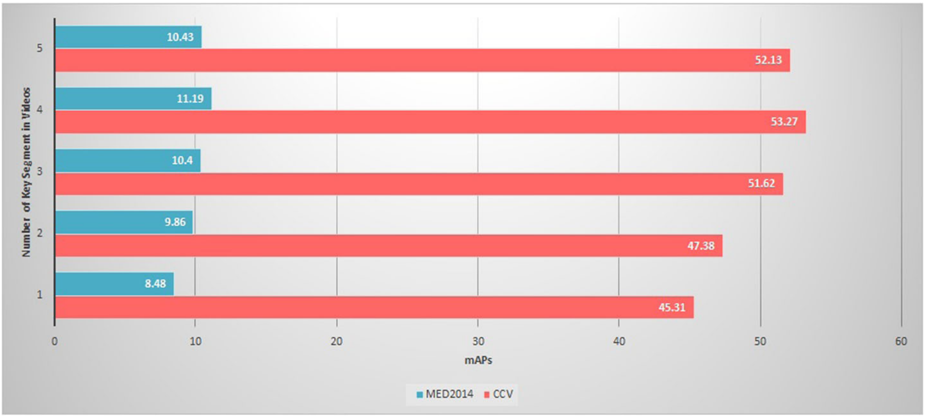


Fig. 5 The results of different numbers of key segments in videos

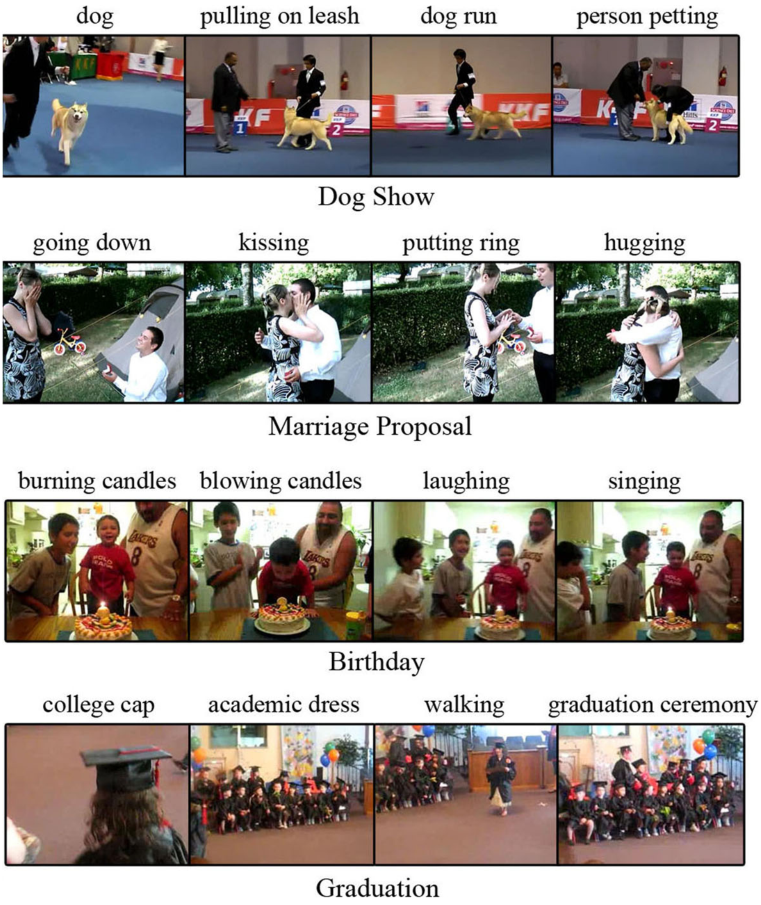


Fig. 6 Some examples of key segments recognized from videos by our method

validates that our method is able to annotate more challenging and difficult event videos when exploring the key segments in the videos. In the simple condition, videos can be represented as a singular entirety. Therefore, our method may perform a little worse than several transfer learning methods, such as the event “show”, “soccer”, “ice-skating”.

Table 4 shows that the mAP degrades when the weights of different concepts belonging to the same event are treated equally. This demonstrates that it is beneficial to assign different weights to different concepts. In our method, each weight can describe the contribution of transferring the knowledge of an image set to the related key segment.

Motivated by the results of Fig. 4, for the complex videos, our method performs better than several transfer learning ones. However, in the simple cases, our performance is about flat. Thus, we test our method on the different number of key segments in videos. We display the results in Fig. 5. This explains that the importance of the image sets. It is obvious that the key segments number increases, the model may not get better. In case, a less relevant image set append to an event model, the performance of the model may turn down. In Fig. 5, we increase an image set to make the number of key segments change from 4 to 5, nevertheless, the accuracy does not increase.

In our method, we can not only effectively infer the label of the unlabeled test videos, but can also recognize the key segments of videos. Figure 6 shows the main frames of the recognized key segments in several events, where each frame represents the corresponding key segment with its semantic concept. Take the “Dog Show” as an example, four key segments annotated with the concepts of “dog”, “pulling on leash”, “dog run” and “person petting” are recognized from the video and each key segment is represented by a frame.

5 Conclusion

In this paper, we proposed an approach of recognizing the key segments of videos for consumer video annotation by leveraging loosely labeled Web image sets. We introduced an adaptive latent structural SVM model to adapt the pre-learned classifiers to an optimal target classifier, the locations of key segments are modeled as latent variables since the groundtruth of the key segments are not available. Also, we developed an iterative algorithm to simultaneously learn the weights of image sets and the target classifier. Our method can recognize the key segments of the test videos and at the same time infer their labels. Experimental results showed that our method performs better than state-of-art methods on the challenging dataset Columbia Consumer Video (CCV) and TRECVID2014 Multimedia Event Detection (MED2014).

Acknowledgements This work was supported in part by the 973 Program of China under grant No. 2012CB720000, the Natural Science Foundation of China(NSFC) under Grant No. 61375044 and 61472038, the Specialized Research Fund for the Doctoral Program of Higher Education of China (20121101120029), the Specialized Fund for Joint Building Program of Beijing Municipal Education Commission, and the Excellent young scholars Research Fund of BIT (2013).

References

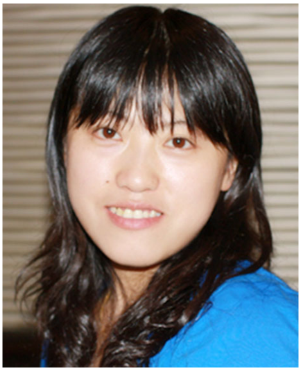
1. Baktashmotlagh M, Harandi MT, Lovell BC, Salzmann M (2013) Unsupervised domain adaptation by domain invariant projection. In: Computer Vision (ICCV), 2013 IEEE International Conference on, pp. 769–776. IEEE

2. Bhattacharya S, Yu FX, Chang SF (2014) Minimally needed evidence for complex event recognition in unconstrained videos. In: Proceedings of International Conference on Multimedia Retrieval, International Conference on Multimedia Retrieval, pp. 105:105–105:112, numpages = 8
3. Bianco S, Ciocca G, Napoletano P, Schettini R (2015) An interactive tool for manual, semi-automatic and automatic video annotation. Comput Vis Image Underst 131:88–99
4. Blank M, Gorelick L, Shechtman E, Irani M, Basri R (2005) Actions as space-time shapes. In: International Conference on Computer Vision, vol. 2, pp. 1395–1402. IEEE
5. Bruzzone L, Marconcini M (2010) Domain adaptation problems: A dasvm classification technique and a circular validation strategy. Pattern Recogn Mach Intell 32(5):770–787
6. Chen L, Duan L, Xu D, Xu D (2013) Event recognition in videos by learning from heterogeneous web sources. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pp. 2666–2673. IEEE
7. Cheng WH, Chuang YY, Lin YT, Hsieh CC, Fang SY, Chen BY, Wu JL (2008) Semantic analysis for automatic event recognition and segmentation of wedding ceremony videos. IEEE Transactions on Circuits and Systems for Video Technology 1639–1650
8. Divakaran A, Javed O, Ali S, Sawhney H, Yu Q, Liu J, Cheng H, Tamrakar A (2013) Video event recognition using concept attributes. In: IEEE Winter Conference on Applications of Computer Vision, pp 339–346
9. Do TMT, Artières T (2009) Large margin training for hidden markov models with partially observed states. In: International Conference on Machine Learning, pp. 265–272. ACM
10. Duan L, Xu D, fu Chang S (2012) Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach. In: Computer Vision and Pattern Recognition
11. Duan L, Xu D, Tsang IWH (2012) Domain adaptation from multiple sources: A domain-dependent regularization approach. IEEE Trans Neural Netw Learn Syst 23(3):504–518
12. Fang M, Guo Y, Zhang X, Li X (2015) Multi-source transfer learning based on label shared subspace. Pattern Recogn Lett 51:101–106
13. Habibian A, Snoek CG (2014) Recommendations for recognizing video events by concept vocabularies. Comput Vis Image Underst 124:110–122
14. Ikizler-Cinbis N, Cinbis R, Sclaroff S (2009) Learning actions from the web. In: International Conference on Computer Vision, pp 995–1002
15. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: Convolutional architecture for fast feature embedding. arXiv preprint. arXiv:1408:5093
16. gang Jiang Y, Ye G, fu Chang S, Ellis D, Loui EC (2011) Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In: International Conference on Multimedia Retrieval, p. 29
17. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: NIPS, pp. 1097–1105
18. Li W, Duan L, Xu D, Tsang IW (2014) Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. IEEE Trans Pattern Anal Mach Intell 36(6):1134–1148
19. Li W, Yu Q, Sawhney H, Vasconcelos N (2013) Recognizing activities via bag of words for attribute dynamics. In: Computer Vision and Pattern Recognition, pp 2587–2594
20. Long M, Wang J, Ding G, Pan SJ, et al. (2014) Adaptation regularization: A general framework for transfer learning. IEEE Trans Knowl Data Eng 26(5):1076–1089
21. Mazloom M, Gavves E, van de Sande K, Snoek C (2013) Searching informative concept banks for video event detection. In: Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval, International Conference on Multimedia Retrieval, pp 255–262
22. MED2014: http://www.nist.gov/itl/iad/mig/med14.cfm
23. Ni B, Song Z, Yan S (2011) Web image and video mining towards universal and robust age estimator. IEEE Trans Multimedia 13(6):1217–1229
24. Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359. doi:10.1109/TKDE.2009.191
25. Schroff F, Criminisi A, Zisserman A (2011) Harvesting image databases from the web. IEEE Trans Pattern Anal Mach Intell 33(4):754–766
26. Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local svm approach. In: International Conference on Pattern Recognition, vol. 3, pp. 32–36. IEEE
27. Sun C, Burns B, Nevatia R, Snoek C, Bolles B, Myers G, Wang W, Yeh E (2014) Isomer: Informative segment observations for multimedia event recounting. In: Proceedings of International Conference on Multimedia Retrieval, International Conference on Multimedia Retrieval, pp. 241:241–241:248. ACM

28. Sun C, Nevatia R (2014) Discover: Discovering important segments for classification of video events and recounting
29. Sun Q, Chattopadhyay R, Panchanathan S, Ye J (2011) A two-stage weighting framework for multi-source domain adaptation. In: Advances in neural information processing systems, pp 505–513
30. Tang K, Ramanathan V, Fei-fei L, Koller D (2012) Shifting weights: Adapting object detectors from image to video. In: NIPS, pp. 647–655
31. Wang H, Wu X, Jia Y (2014) Video annotation via image groups from the web. IEEE Trans Multimedia 16(5):1282–1291
32. Yan Y, Yang Y, Meng D, Liu G, Tong W, Hauptmann AG, Sebe N (2015) Event oriented dictionary learning for complex event detection. IEEE Trans Image Process 24(6):1867–1878
33. Yang Y, Zha ZJ, Gao Y, Zhu X, Chua TS (2014) Exploiting web images for semantic video indexing via robust sample-specific loss. IEEE Trans Multimedia 16(6):1677–1689
34. Zhang X, Yang Y, Zhang Y, Luan H, Li J, Zhang H, Chua TS (2015) Enhancing video event recognition using automatically constructed semantic-visual knowledge base

Hao Song received the B.S. degree from North China Electric Power University (NCEPU), Baoding, China, in 2012. He is currently pursuing the Ph.D. degree at the Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, under the supervision of Prof. Y. Jia. His research interests include computer vision, machine learning and video retrieval.

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 an associate professor in the School of Computer Science at the Beijing Institute of Technology. Her research interests include machine learning, computer vision, and human action perception.



Wei Liang received a Ph.D. degree in computer science from Beijing Institute of Technology in 2005. She is currently an associate professor of the School of Computer Science at Beijing Institute of Technology. Her research interests include computer vision, events understanding and HCI.

Yunde Jia is Professor of Computer Science at BIT, and serves as the Director of the Beijing Laboratory of Intelligent Information Technology. He received the B.S., M.S., and Ph.D. degrees in Mechatronics from the Beijing Institute of Technology (BIT) in 1983, 1986, and 2000, respectively. He has previously served as the Executive Dean of the School of Computer Science at BIT from 2005 to 2008. He was a Visiting Scientist at Carnegie Mellon University from 1995 to 1997, and a Visiting Fellow at the Australian National University in 2011. His current research interests include computer vision, media computing, and intelligent systems.