Event Recognition in Videos by Learning from Heterogeneous Web Sources

Lin Chen[1], Lixin Duan[2], Dong Xu[1]
[1]School of Computer Engineering, Nanyang Technological University
[2]SAP Research, Singapore
{chen0631, dongxu}@ntu.edu.sg, lxduan@gmail.com

Abstract

*In this work, we propose to leverage a large number of loosely labeled web videos (*e.g.*, from YouTube) and web images (*e.g.*, from Google/Bing image search) for visual event recognition in consumer videos without requiring any labeled consumer videos. We formulate this task as a new multi-domain adaptation problem with heterogeneous sources, in which the samples from different source domains can be represented by different types of features with different dimensions (*e.g.*, the SIFT features from web images and space-time (ST) features from web videos) while the target domain samples have all types of features. To effectively cope with the heterogeneous sources where some source domains are more relevant to the target domain, we propose a new method called Multi-domain Adaptation with Heterogeneous Sources (MDA-HS) to learn an optimal target classifier, in which we simultaneously seek the optimal weights for different source domains with different types of features as well as infer the labels of unlabeled target domain data based on multiple types of features. We solve our optimization problem by using the cutting-plane algorithm based on group-based multiple kernel learning. Comprehensive experiments on two datasets demonstrate the effectiveness of MDA-HS for event recognition in consumer videos.*

1. Introduction

There is an increasing interest in developing new event recognition techniques for searching and indexing the explosively growing consumer videos. However, visual event recognition in consumer videos is a challenging task because of cluttered backgrounds, complex camera motion and large intra-class variations.

In [4], Chang *et al.* developed a multi-modal system to fuse visual and audio features for consumer video classification. For action recognition in YouTube videos, different strategies were exploited in [24] to effectively integrate motion and static features, and a multi-instance learning approach was proposed in [16] to fuse different features. To improve the recognition accuracies for YouTube videos, Wang *et al.* [30] proposed a new descriptor by describing each video with dense trajectories. In these works, a sufficient labeled training videos are required to learn robust classifiers for event recognition.

However, collecting of labeled training videos based on human annotation is time-consuming and expensive. Observing that keyword (or tag) based search can be readily used to collect a large number of relevant web images or web videos without human annotation [9], researchers recently proposed new approaches that exploit the rich and massive web data for the event recognition task. With the aid of YouTube videos, Duan *et al.* [7] proposed a domain adaptation approach by reducing the data distribution mismatch between web and consumer videos. In [9], Duan *et al.* developed a multi-domain adaptation scheme by leveraging web images from different sources. In [17], classifiers are learnt for action recognition by using incrementally collected web images. However, temporal information was not used in [9, 17], so both works [9, 17] cannot distinguish events like “sitting down” and “standing up” [7].

In this work, we propose to leverage a large number of freely available web videos (*e.g.*, from YouTube) and web images (*e.g.*, from Google/Bing image search) for event recognition in consumer videos (see Fig. 1), where there are no labeled consumer videos. We propose to additionally use web images for event recognition, which is based on two motivations [7]: 1) there are more web images available with loose labels than web videos; 2) the learnt classifiers using relevant web images are also useful because web images are generally associated with more accurate tags than web videos. Motivated by [7, 9], we formulate this task as a new multi-domain adaptation problem with heterogeneous sources, in which samples from different source domains can be represented by different types of features with different dimensions (*e.g.*, SIFT features from web images and Space-time (ST) features from web videos[1]) while the

[1] For the ease of representation, we assume the samples from each source domain are only represented by one type of feature in this work. If the samples from one source domain have two types of features, it will

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.344

IEEE computer society

target domain samples have all types of features.

Observing that some source domains are more relevant to the target domain, in Section 3, we propose a new method called Multi-domain Adaptation with Heterogeneous Sources (MDA-HS) to effectively cope with heterogeneous sources. Specifically, we seek the optimal weights for different source domains with different types of features and also infer the labels of unlabeled target domain data based on all types of features. For each source domain, we propose to learn an adapted classifier based on the pre-learnt source classifier with data distribution mismatch, for which we minimize the distance between the two classifiers in terms of their weight vectors. We introduce a new regularizer by summing the weighted distances from all the source domains and combine all the weighted adapted classifiers as a new target classifier. We also propose a new ρ-SVM based objective function by using the new regularizer and target classifier for domain adaptation. We develop an iterative optimization method by using the cutting plane method and solving a group-based multiple kernel learning (MKL) problem. In Section 4, we conduct comprehensive experiments using two benchmark consumer video datasets as the target domain, and the results demonstrate that our method MDA-HS outperforms the existing multi-domain adaptation methods for event recognition.

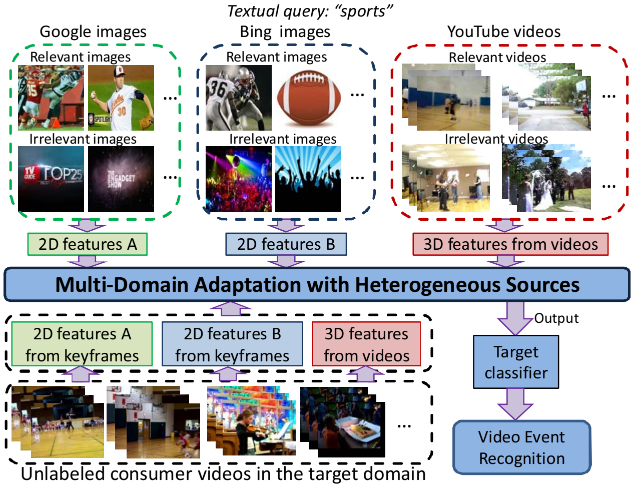


Figure 1. Overview of our proposed framework for visual event recognition in consumer videos. The source domain data contain both web images from Google/Bing and web videos from YouTube. The target domain contains unlabeled consumer videos.

2. Related Work

Recently, domain adaptation has attracted increasing attention in computer vision because of its successful applications in object recognition [12, 13, 14, 20, 26], event recognition [7, 9] and video concept detection [8]. Most existing approaches focus on the setting with a single source domain. For example, a few SVM based methods [2, 7, 32] were recently developed. Saenko *et al.* [26] and Kulis *et al.* [20] proposed to learn the feature transformations for domain adaptation. Gopalan *et al.* [14] and Gong *et al.* [13] proposed new domain adaptation methods by interpolating new subspaces to bridge the two domains.

Multi-domain adaptation methods [9, 15, 10, 5, 27, 28] were also proposed when training data come from multiple source domains. Duan *et al.* [9] proposed a domain selection method to select the most relevant source domains. Based on the discovered latent source domains, Hoffman *et al.* [15] extended [20] for multi-domain adaptation by learning multiple transformations. In [5], Chattopadhyay *et al.* proposed a two-step approach to first learn the weight for each source domain and then learn the target classifier by using the learnt source domain weights. However, most existing multi-domain adaptation methods assume all the training samples from the source and target domains are represented by the same type of feature. Therefore, they cannot effectively handle our setting with labeled single-view data from heterogeneous source domains and unlabeled multi-view data from the target domain. These existing methods can only fuse the decisions from multiple models (with each model learnt by using one type of feature) in a late-fusion fashion [27, 10, 5, 28] or concatenate multiple features into one lengthy vector as the feature representation for target domain data in an early fusion fashion [10, 5, 28, 9]. In contrast, our work, MDA-HS, can simultaneously learn the optimal target classifier and seek the optimal weights for different source domains with different types of features.

Our work is also different from heterogeneous domain adaption (HDA) methods [20, 11], in which the samples from the source and the target domains are represented by different types of features. In contrary, in our work the samples from each pair of source and target domains share the same type of feature because we assume the target domain samples are represented by all types of features. In the existing HDA methods [20, 11], labeled target domain samples must be provided. In contrast, we do not require any labeled target domain samples.

Our work is also different from existing multi-view learning approaches including multi-view based domain adaptation methods [33, 6]. These works [33, 6] generally assume all the data in the source and target domains have multiple types of features, which is different from our setting (see Fig. 2). Moreover, these works cannot be used to learn the weights for different source domains, which is the key challenging issue in multi-domain adaptation.

be treated as two source domains, in which different features are extracted from the same set of samples.

3. Consumer Video Recognition using Heterogeneous Data Sources

In this paper, our task is to recognize visual events in consumer videos by leveraging a large number of loosely labeled web images and web videos, where there are no labeled consumer videos available. Specifically, we represent each web image by using 2D visual features (*e.g.*, SIFT features [25]) and each web video by using 3D visual features (*e.g.*, space-time features [21]). And every consumer video is described based on both 2D and 3D features. As different domains (*i.e.*, the consumer video domain and the web domain) have different data distributions and the samples from different source domains (*i.e.*, web image domain and web video domain) are in different feature spaces, we aim to cope with the *unsupervised domain adaptation problem with heterogeneous sources* in this work.

Following the terminology of domain adaptation, we refer to the consumer video domain as the target domain, as well as consider the web image and video domains as the heterogeneous source domains. Note that the target data have multiple views of features, while the data from each source domain has only one view of feature. Our goal is to learn a robust target classifier by using the loosely labeled single-view data from the heterogeneous source domains and the unlabeled multi-view data from the target domain. In this work, we assume that we have S heterogeneous source domains and focus on the binary classification problem. For each class, we are given a set of labeled single-view data $\{(\mathbf{x}_i^s, y_i^s)|_{i=1}^{n_s}\}$ from the s-th source domain, where n_s is the total number of samples from the s-th source domain and each sample $\mathbf{x}^s$ is drawn from a fixed but unknown data distribution $\mathcal{P}_s$, $y^s \in \{-1, 1\}$ is the label of $\mathbf{x}^s$, and $s = 1, \ldots, S$. And we are also provided with a set of unlabeled multi-view data $\{\mathbf{z}_i|_{i=1}^{n_T}\}$ from the target domain, where n_T is the total number of target domain samples and each sample $\mathbf{z}$ has S views (*i.e.*, $\mathbf{z} = \left(\mathbf{z}^{[1]}, ..., \mathbf{z}^{[S]}\right)$) and the s-th view $\mathbf{z}^{[s]}$ (drawn from $\mathcal{P}_T^{[s]}$) is the same view as $\mathbf{x}^s$, namely, $\mathbf{z}^{[s]}$ and $\mathbf{x}^s$ share the same type of feature with the same dimension (see Fig. 2 for the feature correspondences). Note that for our domain adaptation setting with heterogeneous sources, we have $\mathcal{P}_i \neq \mathcal{P}_j, \mathcal{P}_T^{[i]} \neq \mathcal{P}_T^{[j]}$ and $\mathcal{P}_i \neq \mathcal{P}_T^{[i]}$ ($\forall i, j = 1, ..., S$ and $i \neq j$). The goal of our work is to simultaneously cope with multiple data distribution mismatches between each pair of web domain and consumer domain and assign higher weights to the most relevant source domains.

In the remainder of this paper, we denote the transpose of a vector or matrix by using the superscript $'$. Also, we define $\mathbf{0}_n$ and $\mathbf{1}_n$ as $n \times 1$ vectors of all zeros and all ones, respectively. Let us denote $\odot$ as the element-wise product between two vectors or two matrices. Moreover, $\mathbf{a} \leq \mathbf{b}$ represents $a_i \leq b_i, \forall i$. $\mathbf{I}_n$ is defined as a $n \times n$ identity matrix and $\mathbf{O}_{n\times m}$ is defined as a $n \times m$ matrix of all zeros.

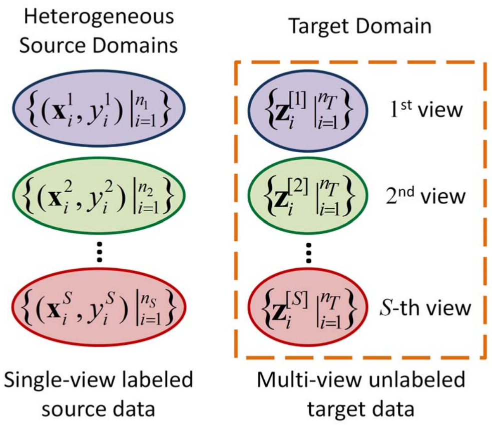

Heterogeneous Source Domains
Target Domain
$\{(\mathbf{x}_i^1, y_i^1)|_{i=1}^{n_1}\}$
$\{(\mathbf{x}_i^2, y_i^2)|_{i=1}^{n_2}\}$
$\{(\mathbf{x}_i^S, y_i^S)|_{i=1}^{n_S}\}$
$\{\mathbf{z}_i^{[1]}|_{i=1}^{n_T}\}$
1st view
$\{\mathbf{z}_i^{[2]}|_{i=1}^{n_T}\}$
2nd view
$\{\mathbf{z}_i^{[S]}|_{i=1}^{n_T}\}$
S-th view
Single-view labeled source data
Multi-view unlabeled target data

Figure 2. Illustration of our multi-domain adaptation setting with heterogeneous sources, which consists of the single view source data and multi-view target data.

3.1. Proposed Formulation

Motivated by multiple kernel learning (MKL) [31], we propose to learn the following target classifier f^T for the prediction of any test sample $\mathbf{z}$ from the target domain, which fuses the decisions from multiple views of data:

$$f^T(\mathbf{z}) = \sum_{s=1}^{S} d_s \mathbf{w}_s' \phi_s(\mathbf{z}^{[s]}), \tag{1}$$

where $\mathbf{w}_s$ is the weight vector for the s-th view of target data; ϕ_s is the feature mapping function for the target data $\mathbf{z}^{[s]}$ of the s-th view; and $d_s \geq 0$ is a combination coefficient.

Recall that there are no labeled data in the target domain. The existing domain adaptation methods [27, 13] are not specifically designed for our setting with multiple heterogeneous domains, so these methods generally cannot work well. Recently, Aytar and Zisserman [1] investigated the single source domain adaptation problem and tried to utilize the pre-learnt source classifier $\mathbf{u}'\phi(\mathbf{x})$ to learn the target classifier by regularizing $\mathbf{u}$ and the weight vector $\mathbf{w}_T$ of the target classifier, *i.e.*, $\|\mathbf{w}_T - \gamma\mathbf{u}\|_2^2$, where the parameter γ controls the amount of knowledge transferred from the source domain to the target domain. Inspired by their method [1], we make use of a set of pre-learnt source classifiers $f^s(\mathbf{x}^s) = \mathbf{u}_s'\phi_s(\mathbf{x}^s)$'s trained by using the training data from each individual source domain and propose a new regularizer as follows for multiple heterogeneous source domains by linearly combining the distances between the target classifier and pre-learnt source classifiers in terms of their weight vectors from all views:

$$\sum_{s=1}^{S} d_s \left\|\mathbf{w}_s - \gamma_s \mathbf{u}_s\right\|_2^2, \tag{2}$$

which is to be minimized in our proposed optimization problem. It is worth noting that we also use the same d_s in (1) as the weight in the above regularizer. The explanation is as follows. If the target model in the s-th view is closer to the source model, d_s will be larger. In this case, we also expect the classifier from the s-th view will have a bigger contribution on the final prediction in the target classifier (1).

Note that $\mathbf{d} = [d_1, \ldots, d_S]'$ is usually constrained based on either L1 or L2 norm [31]. In this work, we assume that $\|\mathbf{d}\|_2^2 = 1$. In order to learn the target classifier in (1) as well as simultaneously infer the labels y_i^T of unlabeled training data from the target domain, we introduce our new regularizer in (2) and our target classifier in (1) into a ρ-SVM based objective function as follows:

$$\min_{\mathbf{d}\in\mathcal{D},\mathbf{y}_T} \min_{\substack{\mathbf{w}_s,\gamma_s,\\ \rho,\xi_i^s,\xi_i^T}} \frac{1}{2}\left(\sum_{s=1}^{S} d_s \left\|\mathbf{w}_s - \gamma_s \mathbf{u}_s\right\|_2^2 + \theta\gamma_s^2\right) - \rho$$

$$+\frac{1}{2}\left(C_S \sum_{s=1}^{S}\sum_{i=1}^{n_s} \xi_i^{s2} + C_T \sum_{i=1}^{n_T} \xi_i^{T2}\right), \quad (3)$$

$$\text{s.t. } y_i^T \in \{\pm 1\},\ y_i^T \sum_{s=1}^{S} d_s \mathbf{w}'_s \phi_s(\mathbf{z}_i^{[s]}) \geq \rho - \xi_i^T, \quad (4)$$

$$y_i^s d_s \mathbf{w}'_s \phi_s(\mathbf{x}_i^s) \geq \rho - \xi_i^s,\ s = 1, ..., S, \quad (5)$$

where $\theta, C_S, C_T > 0$ are regularization parameters, $\mathcal{D} = \left\{\mathbf{d} \mid \|\mathbf{d}\|_2^2 = 1, \mathbf{d} \geq \mathbf{0}_S\right\}$ is the domain of $\mathbf{d}$, $\mathbf{y}_T = [y_1^T, ..., y_{n_T}^T]'$ is the label vector of the target training samples, and ξ_i^s, ξ_i^T are slack variables of the training samples in s-th source domain and the target domain, respectively. In the above formulation, we penalize the L2 norm of γ_s's in (3) to avoid overfitting. Note that we enforce the target model of the s-th view to have good classification performance on the corresponding labeled source data. We argue that such supervision is very important for our multi-domain adaptation problem. The reason is two-fold: 1) There is a certain amount of overlap between the s-th source domain and the target domain when using the s-th view of features, so it is very possible that a well trained model using the labeled source domain data would not perform poorly on the target domain; 2) we do not have any labeled data in the target domain, so the performance of our model will become much worse without having the constraints in (5) (see our experimental results in Section 4). It is also worth mentioning that this problem is a mixed integer programming (MIP) problem.

3.2. A Dual Perspective

Before solving the optimization problem in (3), let us define $\Phi_s = [\phi_s(\mathbf{x}_1^s), \ldots, \phi_s(\mathbf{x}_{n_s}^s)]$, $\Phi_T^{[s]} = [\phi_s(\mathbf{z}_1^{[s]}), \ldots, \phi_s(\mathbf{z}_{n_T}^{[s]})]$ as the nonlinear mapping function for the data in the s-th source domain and the target domain in the s-th view respectively, and denote $\mathbf{f}_s = [f^s(\mathbf{x}_1^s), \ldots, f^s(\mathbf{x}_{n_s}^s)]'$ and $\mathbf{f}_T^{[s]} = [f^s(\mathbf{z}_1^{[s]}), \ldots, f^s(\mathbf{z}_{n_T}^{[s]})]'$ as the decision values after using the pre-learnt source classifiers $f^s(\mathbf{x})$, $s = 1, \ldots, S$. Moreover, let us denote h_s as the dimension of $\phi_s(\mathbf{x}^s)$ and $N(p, q) = \sum_{s=p}^{q} n_s$ as the total number of samples from the p-th source domain to the q-th source domain ($q \geq p$). Base on Φ_s and $\Phi_T^{[s]}$, we then define $\Phi^{[s]}$ over all the samples by setting the columns not related to the samples from the s-th source domain and the target domain as zeros, namely:

$$\Phi^{[s]} = \left[\mathbf{O}_{h_s\times N(1,s-1)}, \Phi_s, \mathbf{O}_{h_s\times N(s+1,S)}, \Phi_T^{[s]}\right]. \quad (6)$$

Based on $\mathbf{f}_s$ and $\mathbf{f}_T^{[s]}$, we can similarly define $\mathbf{f}^{[s]}$ as

$$\mathbf{f}^{[s]} = \left[\mathbf{0}'_{N(1,s-1)}, \mathbf{f}'_s, \mathbf{0}'_{N(s+1,S)}, \mathbf{f}_T^{[s]\prime}\right]'. \quad (7)$$

Note when $s = 1$ (*resp.* $s = S$), $\mathbf{O}_{h_s\times N(1,s-1)}$ and $\mathbf{0}_{N(1,s-1)}$ (*resp.* $\mathbf{O}_{h_s\times N(s+1,S)}$ and $\mathbf{0}_{N(s+1,S)}$) in (6) and (7) become an empty matrix or an empty vector.

We solve (3) by first taking the dual form of the inner optimization problem with respect to the primal variables $\rho, \mathbf{w}_s, \gamma_s, \xi_i^s$ and ξ_i^T, where $s = 1, \ldots, S$. Specifically, by introducing the Lagrange multipliers α_i's for the inequality constraints in (4) and (5), we then have the Lagrangian $\mathcal{L}$ of inner optimization problem in (3). By setting the derivatives of the Lagrangian to zeros with respect to the primal variables ρ, $\mathbf{w}_s$'s, γ_s's, ξ_i^s's and ξ_i^T's, and substituting the resultant equalities back into $\mathcal{L}$, we can rewrite the optimization problem in (3) as follows by replacing the inner problem with its dual form:

$$\min_{\mathbf{d}\in\mathcal{D},\, \mathbf{y}\in\mathcal{Y}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} -\frac{1}{2}\boldsymbol{\alpha}'\left(\sum_{s=1}^{S} d_s \tilde{\mathbf{K}}^{[s]} \odot \mathbf{y}\mathbf{y}' + \tilde{\mathbf{I}}\right)\boldsymbol{\alpha}, \quad (8)$$

where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]'$ is a vector of dual variables with $n = \sum_{s=1}^{S} n_s + n_T$ (note its first $N(1, S)$ elements are for the constraints in (5) of all source domain samples and its last n_T elements are for the constraints in (4) of the target domain samples); $\mathcal{A} = \{\boldsymbol{\alpha}|\boldsymbol{\alpha}'\mathbf{1}_n = 1, \boldsymbol{\alpha} \geq \mathbf{0}_n\}$ is the domain of $\boldsymbol{\alpha}$; $\mathbf{y}$ is the label vector of all training samples, which takes the values from the feasible set of $\mathcal{Y} = \{\mathbf{y}|\mathbf{y} = [\mathbf{y}'_1, \ldots, \mathbf{y}'_S, \mathbf{y}'_T]', \mathbf{y}_T \in \{-1, 1\}^{n_T}\}$ (note $\mathbf{y}_s = [y_1^s, \ldots, y_{n_s}^s]'$ is the given label vector for the s-th source domain data), and $\mathbf{y}$ is also represented as $\mathbf{y} = [y_1, \ldots, y_n]'$ with its elements in the same order; $\tilde{\mathbf{I}} = \text{diag}\left\{\left[\mathbf{1}'_{n_1}/C_S, \ldots, \mathbf{1}'_{n_S}/C_S, \mathbf{1}'_{n_T}/C_T\right]'\right\}$ is a diagonal matrix; $\tilde{\mathbf{K}}^{[s]}$ is the transformed kernel matrix defined over all the samples but only based on the s-th source domain data and the s-th view of target data, namely:

$$\tilde{\mathbf{K}}^{[s]} = \mathbf{K}^{[s]} + \frac{1}{\theta}\mathbf{f}^{[s]}\mathbf{f}^{[s]\prime}, \quad (9)$$

where $\mathbf{K}^{[s]} = \Phi^{[s]\prime}\Phi^{[s]}$ with $\Phi^{[s]}$ and $\mathbf{f}^{[s]}$ as defined in (6) and (7), respectively.

Convex relaxation. Note that we need to determine the label vector $\mathbf{y}_T$ for the unlabeled target domain data by solving the MIP problem in (8), which is NP hard. We thus relax (8) to be a convex optimization problem which is the lower bound of (8) as shown in the following Proposition 1:

Proposition 1 *The objective value of the mixed integer programming (MIP) problem in (8) is lower bounded by the optimal value of the following group-based multiple kernel learning (MKL) problem:*

$$\begin{aligned}\min_{\mathbf{D}}\max_{\boldsymbol{\alpha}\in\mathcal{A}} \quad & -\frac{1}{2}\boldsymbol{\alpha}'\left(\sum_{s=1}^{S}\sum_{o:\,\mathbf{y}^o\in\mathcal{Y}} D_{so}\mathbf{Q}^{so} + \tilde{\mathbf{I}}\right)\boldsymbol{\alpha}, \\ s.t. \quad & \|\mathbf{D}\|_{2,1} = 1, D_{so} \geq 0\ \forall s\,, \forall o,\end{aligned} \tag{10}$$

where $\mathbf{Q}^{so}$ is a base label-kernel defined as $\mathbf{Q}^{so} = \tilde{\mathbf{K}}^{[s]} \odot (\mathbf{y}^o\mathbf{y}^{o\prime})$ with $\mathbf{y}^o$ being the o-th feasible labeling candidate for $\mathbf{y}$, $\mathbf{D} = [D_{so}] \in \mathbb{R}^{S\times|\mathcal{Y}|}$ is the kernel coefficient matrix (note $|\mathcal{Y}|$ is the size of $\mathcal{Y}$), and $\|\mathbf{D}\|_{2,1} = \sum_{o=1}^{|\mathcal{Y}|}\sqrt{\sum_{s=1}^{S} D_{so}^2}$ is the mixed $L_{2,1}$ norm.

Proof According to the theoretical results in [23], the objective value of (8) is lower bounded by the optimal value of the following optimization problem:

$$\min_{\mathbf{d},\boldsymbol{\mu}}\max_{\boldsymbol{\alpha}\in\mathcal{A}} \; -\frac{1}{2}\boldsymbol{\alpha}'\left(\sum_{s=1}^{S}\sum_{o:\mathbf{y}^o\in\mathcal{Y}} d_s\mu_o\tilde{\mathbf{K}}^{[s]}\odot\mathbf{y}^o\mathbf{y}^{o\prime} + \tilde{\mathbf{I}}\right)\boldsymbol{\alpha}, \tag{11}$$

$$s.t. \quad \|\mathbf{d}\|_2^2 = 1, \mathbf{d} \geq \mathbf{0},\ \|\boldsymbol{\mu}\|_1 = 1, \boldsymbol{\mu} \geq \mathbf{0}, \tag{12}$$

where $\boldsymbol{\mu} = [\mu_1, ..., \mu_{|\mathcal{Y}|}]'$. By setting $D_{so} = d_s\mu_o$, we have $\|\mathbf{D}\|_{2,1} = 1$. And (11) is converted into the convex optimization problem in (10), and its optimal objective value is the lower bound of the objective value of (11).

3.3. Detailed Algorithm

As the size of $\mathcal{Y}$ increases exponentially with the number of unlabeled target data, making the optimization of the objective function in (10) still computationally expensive when there exist a large number of unlabeled target data. Fortunately, we can employ the cutting-plane method to iteratively select a small number of most violated labeling candidates (*i.e.*, $\mathbf{y}^o$'s) which are good enough to approximate the optimal solution [19]. The details are listed in Algorithm 1.

Algorithm 1 Cutting-plane algorithm for MDA-HS

```
1: Initialize $\mathbf{y}^1$ based on the outputs from the source classifiers and set $o = 1$, $\mathcal{Y}^o = \{\mathbf{y}^1\}$
2: repeat
3:     Update $\boldsymbol{\alpha}$ and $\mathbf{D}$ in the group-based MKL problem (10) with $\mathcal{Y} = \mathcal{Y}^o$ by using Algorithm 2
4:     Find the most violated labeling candidate $y^{o+1}$ by solving (13)
5:     Set $\mathcal{Y}^{o+1} = \mathcal{Y}^o \cup \{\mathbf{y}^{o+1}\}$
6:     $o \leftarrow o + 1$
7: until The objective of (10) converges
```

Finding the most violated $\mathbf{y}^o$. At each iteration in Algorithm 1, when $\mathbf{D}$ and $\boldsymbol{\alpha}$ are fixed after Step 3, we find the most violated $\mathbf{y}^o$ by solving the following optimization problem for each s:

$$\max_{\mathbf{y}^o\in\mathcal{Y}} \boldsymbol{\alpha}'\left(\tilde{\mathbf{K}}^{[s]} \odot \mathbf{y}^o{\mathbf{y}^o}'\right)\boldsymbol{\alpha} = \max_{\mathbf{y}^o\in\mathcal{Y}} \|\mathbf{U}^{[s]\prime}\boldsymbol{\alpha} \odot \mathbf{y}^o\|_2^2, \tag{13}$$

where $\tilde{\mathbf{K}}^{[s]} = \mathbf{U}^{[s]}\mathbf{U}^{[s]\prime}$ is decomposed by using eigenvalue decomposition. Motivated by [23, 22], we develop an efficient algorithm to solve the integer programming problem in (13) by relaxing the L_2 norm into L_∞ norm:

$$\max_{\mathbf{y}^o\in\mathcal{Y}}\left\|\mathbf{U}^{[s]\prime}\boldsymbol{\alpha}\odot\mathbf{y}^o\right\|_\infty = \max_{j=1,\ldots,n}\left(\max_{\mathbf{y}^o\in\mathcal{Y}}\left|\sum_{i=1}^{n}\alpha_i y_i^o U_{ij}^{[s]}\right|\right), \tag{14}$$

where $U_{ij}^{[s]}$ is the element in the i-th row and j-th column of $\mathbf{U}^{[s]}$. The integer programming problem in (14) can be efficiently solved by simply sorting the coefficients $\alpha_i y_i^o$'s. Note we only need to infer the labels of unlabeled target domain data (*i.e.*, $\mathbf{y}_T \in \{-1, 1\}^{n_T}$), because the source label vectors $\mathbf{y}_s$'s are already given.

Solving $\boldsymbol{\alpha}$ and $\mathbf{D}$. After finding $\mathbf{y}^o$, we fix $\mathcal{Y} = \mathcal{Y}^o$ and solve the remaining group-based MKL problem in (10) by alternatively updating $\boldsymbol{\alpha}$ and $\mathbf{D}$. Specifically, when we fix $\mathbf{D}$, (10) reduces to a standard SVM and $\boldsymbol{\alpha}$ is updated by efficiently solving the standard SVM with existing softwares such as LIBSVM [3]. When $\boldsymbol{\alpha}$ is fixed, after re-formulating (10) in its primal form as well as dropping the irrelevant terms, $\mathbf{D}$ can be updated by solving the following optimization problem[2]:

$$\min_{\mathbf{D}\in\mathcal{M}} \quad \frac{1}{2}\sum_{s=1}^{S}\sum_{o=1}^{|\mathcal{Y}|}\frac{\|\mathbf{v}_{so}\|_2^2}{D_{so}}, \tag{15}$$

where $\mathcal{M} = \{\mathbf{D}\big|\|\mathbf{D}\|_{2,1} = 1, D_{so} \geq 0\ \forall s, o\}$ is the domain of $\mathbf{D}$, $\|\mathbf{v}_{so}\|_2 = D_{so}\sqrt{\boldsymbol{\alpha}'\mathbf{Q}^{so}\boldsymbol{\alpha}}, \forall s, o.$ With simple derivation, (15) has the following close-form solution:

$$D_{so} \quad = \quad \frac{\|\mathbf{v}_{so}\|_2^{2/3}\left(\sum_{l=1}^{S}\|\mathbf{v}_{lo}\|_2^{4/3}\right)^{1/4}}{\sum_{o=1}^{|\mathcal{Y}|}\left(\sum_{l=1}^{S}\|\mathbf{v}_{lo}\|_2^{4/3}\right)^{3/4}}. \tag{16}$$

We list the algorithm to solve the group-based MKL problem in (10) in Algorithm 2.

[2] The group-based MKL problem in (10) is a special case of the composite kernel learning problem in [29], when setting $p = 0, q = 1, d_\ell = 1$ in (7a)–(7c) in [29].

Algorithm 2 The algorithm of group-based MKL

```
Initialize $\mathbf{D}^1$ by using uniform values such that $\|\mathbf{D}^1\|_{2,1} = 1$ and set $\tau = 1$
repeat
    With fixed $\mathcal{Y}$, update $\boldsymbol{\alpha}$ by solving the standard SVM problem with $\mathbf{D}^\tau$ in (10)
    Update $\mathbf{D}^{\tau+1}$ by using (16)
    $\tau \leftarrow \tau + 1$
until The objective of (10) converges with fixed $\mathcal{Y}$
```

Target classifier. After finding the optimal $\boldsymbol{\alpha}, \mathbf{D}$ and $\mathbf{y}^o$'s, the target classifier in (1) can be rewritten as

$$f^T(\mathbf{z}) = \sum_{s=1}^{S} \boldsymbol{\beta}_s' \left(\Phi^{[s]\prime} \phi_s(\mathbf{z}^{[s]}) + \frac{1}{\theta} \mathbf{f}^{[s]} f^s(\mathbf{z}^{[s]}) \right),$$

where $\boldsymbol{\beta}_s = \boldsymbol{\alpha} \odot \left(\sum_{o=1}^{|\mathcal{Y}|} D_{so} \mathbf{y}^o \right)$.

4. Experiments

We compare our work with the baseline method SVM, the existing single source domain adaptation algorithms Geodesic Flow Kernel (GFK)[3] [13] and Domain Adaptive SVM (DASVM) [2], as well as the existing multi-domain adaptation methods Domain Adaptation Machine (DAM)[10], Conditional Probability based Multi-source Domain Adaptation (CPMDA) [5], Maximal Margin Target Label Learning (MMTLL) [28] and Domain Selection Machine (DSM) [9].

We also report the results of two simplified versions (referred to as MDA-HS_sim1 and MDA-HS_sim2) of our method MDA-HS in order to validate our new regularizer in (2) and the constraint in (5). In MDA-HS_sim1, we set the parameter $\theta = \infty$. In this case, we have $\gamma_s = 0$ in (3) and our regularizer in (2) becomes $\sum_{s=1}^{S} d_s \|\mathbf{w}_s\|_2^2$, so the pre-learnt source classifiers will not be employed when calculating the kernel (See Eq. (9)). In order to demonstrate it is beneficial to employ the source domain data in MDA-HS, we compare our work with MDA-HS_sim2 which excludes the constraints in (5).

4.1. Datasets and Features

We evaluate all the methods on two benchmark consumer video datasets (*i.e.*, Kodak [7] and CCV [18]), which are also used in [9]. To construct the heterogenous sources, we use the YouTube dataset in [7] and additionally collect two datasets by using Google/Bing image search. Note the labels of source domain data are noisy because we do not spend additional efforts to annotate the YouTube dataset and the two web image datasets. The detailed information of the five datasets is introduced below.

Google/Bing image dataset: We download the top ranked 200 images for each event class by using related keywords as queries (*e.g.*, we use “wedding ceremony”, “wedding reception” and “wedding dance” for the event class “wedding”) and we enforce the returned images to be *photo with full color* by using the advanced options provided by Google and Bing image search. We do not download the corrupted images or the images with invalid URLs. Finally, we have collected 1049 images and 1134 images from Google and Bing respectively.

YouTube dataset: The YouTube dataset was used as the source domain data in [7], which consists of 906 videos from six event classes (*i.e.*, “birthday”, “picnic”, “parade”, “show”, “sports” and “wedding”). The videos were collected by using keyword based search from YouTube. According to the study in [7], at least 20% of the videos in this dataset are with incorrect labels.

Kodak dataset: The Kodak dataset was used in [7, 9], which contains 195 consumer videos from six event classes (*i.e.*, “birthday”, “picnic”, “parade”, “show”, “sports” and “wedding”).

CCV dataset: The CCV dataset [18] collected by Columbia University was also used in [9]. It consists of a training set of 4,659 videos and a test set of 4,658 videos from 20 semantic categories. Following [9], we only use the videos from the event related categories and we also merge “wedding ceremony”, “wedding reception” and “wedding dance” as “wedding”, “non-music performance” and “music performance” as “show”[4], and “baseball”, “basketball”, “biking”, “ice skating”, “skiing”, “soccer”, “swimming” as “sports”. Finally, there are 2440 videos from five event classes[5] (*i.e.*, “birthday”, “parade”, “show”, “sports” and “wedding”).

Features: We extract the 128-dim SIFT features using Difference of Gaussians (DoG) interest point detector [25] for each image in the Google and Bing datasets. Then, each image is represented by a 4000-dim token frequency (TF) feature using the bag-of-word (BoW) representation, in which the codebook is constructed by using k-means to cluster all the SIFT features from the images. For each video in the Kodak and CCV datasets, we sample one keyframe per two seconds and then extract the SIFT features from each keyframe. Then, each keyframe is represented by a 4000-dim TF feature based on the BoW representation by using the same codebook. Finally, we average the TF features over all the keyframes within each video as the final feature representation for the videos in the Kodak and CCV datasets

[3] http://www-scf.usc.edu/~boqinggo/domain_adaptation/GFK_v1.zip

[4] We observe that the videos from “non-music performance” and “music performance” in the CCV dataset and those from “show” in the YouTube/Kodak dataset describe similar semantic concepts, so we merge them as “show” in this work.

[5] Note that there are only five common event classes (*i.e.*, “birthday”, “parade”, “show”, “sports” and “wedding”) between the YouTube and CCV datasets, so we only report the results from these five event classes when using CCV as the target domain.

when using the SIFT features.

For each video in the Kodak, YouTube and CCV datasets, we extract three types of space-time (ST) features (*i.e.*, 96-dim Histogram of Oriented Gradient, 108-dim Histogram of Optical Flow and 192-dim Motion Boundary Histogram) by using the source codes provided in [30], in which we set the trajectory length as 50, the sampling stride as 16, and all the other parameters as their default values. We also use the BoW representation for each type of ST features, in which the codebook is constructed by using k-means to cluster the ST features from all videos in the YouTube dataset into 2000 clusters. Finally, each video is represented as a 6000-dim feature by concatenating the 2000-dim TF feature from each type of ST feature.

4.2. Experimental Setups

In our experiments, the Google and Bing image datasets and the YouTube dataset are used as $S = 3$ heterogeneous source domains, and Kodak/CCV dataset is used the target domain.

Note we do not have any labeled consumer videos in the target domain. We refer to the baseline SVM as SVM_A in which S independent SVM classifiers (*i.e.*, f^s's) are trained based on the training data from each individual source domain and further used to predict the test data from the target domain using the same feature. And the final prediction of each test sample is obtained by averaging the predictions from all the classifiers. We also employ the same late fusion strategy for the single source domain adaptation methods GFK and DASVM. The traditional multiple source domain adaptation methods CPMDA, DAM, DSM and MMTLL can not directly deal with our setting with single-view source data and multi-view target data, so we use S pre-learnt source classifiers and also averagely fuse the kernels from all views as the kernel for the target domain data.

We evaluate all the methods by training one-vs-rest SVMs with the Gaussian kernel $k_s(\mathbf{x}_i, \mathbf{x}_j) = \phi_s(\mathbf{x}_i)'\phi_s(\mathbf{x}_j) = \exp\left(-\frac{1}{\nu}\|\mathbf{x}_i - \mathbf{x}_j\|_2^2\right)$, in which we set the bandwidth parameter ν as the mean of the square distances between all pairs of training samples when using s-th view of features. We also set the regularization parameters C_S and C_T as 1 and 10 respectively, since the target domain data is more important than the source domain data. We fix $\theta = 0.1$ for our method MDA-HS and its simplified cases MDA-HS_sim1 and MDA-HS_sim2. As in [7, 9], we use the non-interpolated average precision (AP) for performance evaluation and report the mean AP (MAP) over all the event classes.

4.3. Results

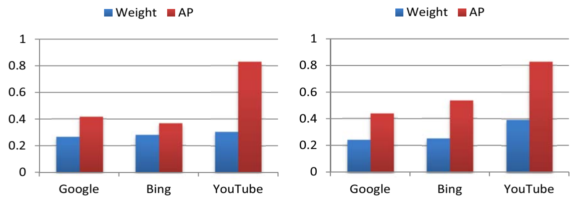
Table 1 shows the MAPs of all methods. We have the following observations from these results:

Table 1. MAP (%) of all methods on the Kodak and CCV datasets.

Method	Kodak	CCV
SVM_A	44.80	40.84
CPMDA [5]	25.72	30.89
DASVM [2]	43.49	41.55
DAM [10]	44.21	38.56
DSM [9]	46.21	43.44
GFK [13]	26.76	34.38
MMTLL [28]	44.09	35.98
MDA-HS_sim2	31.22	30.74
MDA-HS_sim1	44.68	42.38
MDA-HS	**49.61**	**44.52**

1) SVM_A outperforms the existing domain adaptation methods GFK, CPMDA, DAM and MMTLL. Moreover, there is no consistent winner between SVM_A and DASVM on both datasets. These results show that it is a quite challenging task to conduct domain adaptation from heterogeneous source domains. The existing domain adaptation methods cannot always achieve good performances in this application because these methods cannot effectively cope with the heterogeneous sources. Moreover, the promising performance from SVM_A indicates that the data distributions of the s-th source domain and the target domain when using the s-the view of features overlap to some extent.
2) DSM is beter than SVM_A. An explanation is that DSM can select the more relevant source domains.
3) MDA-HS and MDA-HS_sim1 are both much better than MDA-HS_sim2, which demonstrates it is beneficial to train the target weight vector $\mathbf{w}_s$'s from multiple views of features by using the labeled source domain samples. Again, the explanation is there is a certain amount of overlap between each source domain and the target domain when using the same view of features. Moreover, MDA-HS is also better than MDA-HS_sim1, which demonstrates the effectiveness of our new regularizer by leveraging the pre-learnt source classifiers.
4) Our method MDA-HS achieves the best performance among all methods on both datasets, which clearly demonstrates the effectiveness of our MDA-HS for event recognition in consumer videos by utilizing our new regularizer and the new target classifier.

Analysis on the learnt source domain weights. We report the learnt weights of source domains in our MDA-HS and also report the per-event AP results of three SVMs with each trained by using the training data from one single source domain (*i.e.*, YouTube, Bing or Google). If the AP of one SVM is higher, the corresponding source domain is expected to be more relevant to the target domain when using the same view of features, namely, we have larger source domain weight. As in our method MDA-HS, we relax our original optimization problem in (8) to be a group-based MKL problem in (10), in this experiment we therefore an-



(a) Kodak dataset (b) CCV dataset

Figure 3. Illustration of three learnt weights of source domains for the event "sports". We show the learnt weights by using MDA-HS and the per-event APs of three SVMs with each trained using the training data from one source domain.

alyze $\mathbf{D}$ in (10) instead of $\mathbf{d}$ in (8). Specifically, we report the three coefficients of the column in $\mathbf{D}$ with the largest L_2 norm, and similar results can be observed for other columns. In Fig. 3, we take the event "sports" as an example to show the per-event AP results of three SVMs as well as the three source domain weights (*i.e*., the three learnt coefficients) on both datasets. From these results, we observe that MDA-HS can assign the highest weight to the most relevant source domain for which the per-event AP is also the highest. The results clearly show the effectiveness of the group-based MKL technique in MDA-HS for combining multiple heterogeneous source domains.

5. Conclusion

We have proposed a new framework for visual event recognition in consumer videos by leveraging a large number of freely available web videos (*e.g*., from YouTube) and web images (*e.g*., from Google/Bing image search). This task is formulated as a new multi-domain adaptation problem with heterogeneous sources. By introducing a new target classifier and a new regularizer based on the weights of heterogeneous source domains, our method called Multi-domain Adaptation with Heterogeneous Sources (MDA-HS) can simultaneously seek the optimal weights for different source domains with different types of features, infer the labels of unlabeled target domain data based on multiple types of features, and learn the optimal target classifier. Comprehensive experiments by using two benchmark consumer video datasets, Kodak and CCV, demonstrate the effectiveness of our method MDA-HS for event recognition without requiring any labeled consumer videos.

Acknowledgement. This research is partially supported by Multi-plAtform Game Innovation Centre (MAGIC) in Nanyang Technological University. MAGIC is funded by the Interactive Digital Media Programme Office (IDMPO) hosted by the Media Development Authority of Singapore.

References

[1] Y. Aytar and A. Zisserman. Tabula rasa: Model transfer for object category detection. In *ICCV*, 2011.

[2] L. Bruzzone and M. Marconcini. Domain adaptation problems: A dasvm classification technique and a circular validation strategy. *T-PAMI*, 32(5):770–787, 2010.

[3] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011.

[4] S.-F. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, E. C. Loui, and J. Luo. Large-scale multimodal semantic concept detection for consumer video. In *MIR workshop, ACM Multimedia*, 2007.

[5] R. Chattopadhyay, J. Ye, S. Panchanathan, W. Fan, and I. Davidson. Multi-source domain adaptation and its application to early detection of fatigue. In *KDD*, 2007.

[6] M. Chen, K. Q. Weinberger, and J. C. Blitzer. Co-training for domain adaptation. In *NIPS*, 2011.

[7] L. Duan, I. W. Tsang, D. Xu, and J. Luo. Visual event recognition in videos by learning from web data. *T-PAMI*, 34(9):1667–1680, 2012.

[8] L. Duan, I. W. Tsang, D. Xu, and S. J. Maybank. Domain transfer multiple kernel learning. *T-PAMI*, 34(3):465–479, 2012.

[9] L. Duan, D. Xu, and S.-F. Chang. Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach. In *CVPR*, pages 1338–1345, 2012.

[10] L. Duan, D. Xu, and I. W. Tsang. Domain adaptation from multiple sources: A domain-dependent regularization approach. *T-NNLS*, 23(3):504–518, 2012.

[11] L. Duan, D. Xu, and I. W. Tsang. Learning with augmented features for heterogeneous domain adaptation. In *ICML*, 2012.

[12] B. Gong, K. Grauman, and F. Sha. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *ICML*, 2013.

[13] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012.

[14] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, 2011.

[15] J. Hoffman, K. Saeko, B. Kulis, and T. Darrell. Discovering latent domains for multisource domain adaptation. In *ECCV*, 2012.

[16] N. Ikizler-Cinbis and S. Sclaroff. Object, scene and actions: Combining multiple features for human action recognition. In *ECCV*, 2010.

[17] N. Ikizler-Cinbis and S. Sclaroff. Web-based classifiers for human action recognition. *T-MM*, 14(4):1031–1045, 2012.

[18] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. C. Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *ICMR*, 2011.

[19] J. E. Kelley. The cutting plane method for solving convex programs. *SIAM Journal on Applied Mathematics*, 8(4):703–712, 1960.

[20] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR*, 2011.

[21] I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV*, 2003.

[22] W. Li, L. Duan, D. Xu, and I. W. Tsang. Text-based image retrieval using progressive multi-instance learning. In *ICCV*, pages 2049–2055, 2011.

[23] Y.-F. Li, I. W. Tsang, J. T.-Y. Kwok, and Z.-H. Zhou. Tighter and convex maximum margin clustering. In *AISTATS*, 2009.

[24] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". In *CVPR*, 2009.

[25] D. G. Lowe. Disctinctive image features from scale-invariance keypoint. *IJCV*, 60(2):91–110, 2004.

[26] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, 2010.

[27] G. Schweikert, C. Widmer, B. Schölkopf, and G. Rätsch. An empirical analysis of domain adaptation algorithms for genomic sequence analysis. In *NIPS*, 2009.

[28] C.-W. Seah, I. W. Tsang, and Y.-S. Ong. Healing sample selection bias by source classifier selection. In *ICDM*, 2011.

[29] M. Szafranski, Y. Grandvalet, and A. Rakotomamonjy. composite kernel learning. *Machine Learning*, 79(1-2):73–103, 2010.

[30] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, 2011.

[31] X. Xu, I. W. Tsang, and D. Xu. Soft margin multiple kernel learning. *T-NNLS*, 24(5):749–761, 2013.

[32] J. Yang, R. Yan, and A. Hauptmann. Cross-domain video concept detection using adaptive svms. In *ACM MM*, 2007.

[33] D. Zhang, J. He, Y. Liu, L. Si, and R. D. Lawrence. Multi-view transfer learning with a large margin approach. In *KDD*, 2011.