A Multiple Image Group Adaptation Approach for Event Recognition in Consumer Videos

Dengfeng Zhang(✉), Wei Liang, Hao Song, Zhen Dong, and Xinxiao Wu

Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, People's Republic of China
{zhangdengfeng,liangwei,songhao,dongzhen,wuxinxiao}@bit.edu.cn

Abstract. Event recognition in the consumer videos is a challenging task since it is difficult to collect a large number of labeled training videos. In this paper, we propose a novel Multiple kernel Image Group Adaptation approach to divide the training labeled Web images into several semantic groups and optimize the combinations of each based kernel. Our method simultaneously learns a kernel function and a robust Support Vector Regression (SVR) classifier by minimizing both the structure risk of SVR with the smooth assumption and the distribution difference of weighted image groups and the consumer videos. Comprehensive experiments on the datasets CCV and TREATED 2014 demonstrate the effectiveness of our method for event recognition.

Keywords: Event recognition · Image group · Transfer learning · Kernel learning

1 Introduction

Event recognition in consumer videos is an increasingly important research in the computer vision due to its broad applications for automatic video retrieval and indexing. Unlike the simple action datasets (*e.g.* KTH [18]), consumer videos are usually captured by non-professionals using hand-held digital cameras. So it is a challenging task to annotate the events in consumer videos due to camera motion, cluttered background and large intra-class variations.

The most previous recognition methods [11] and [19–21] have demonstrated promising performances but need a large number of labeled training videos. However, annotating abundant consumer videos is expensive and time-consuming. The learned classifiers from a limited number of labeled training data are usually not robust and generalized. Since the image searching engines have become more mature and efficient and they can offer abundant web images with the loose

D. Zhang—The research was supported in part by the 973 Program of China under grant No. 2012CB720000the Natural Science Foundation of China (NSFC) under Grant 61203274, the Specialized Fund for Joint Building Program of Beijing Municipal Education Commission.

Y.-J. Zhang (Ed.): ICIG 2015, Part I, LNCS 9217, pp. 176–187, 2015.
DOI: 10.1007/978-3-319-21978-3_16

label, researchers are easily able to collect labeled Web images instead of manual annotation. Recently, several domain adaptation(also called cross-domain learning or transfer learning) algorithms [5,10,22,23] are proposed. These methods can learn a target classifier by transferring knowledge from the Web image sets (source domain) to the videos (target domain). Duan *et al.* [5] developed a multiple sources domain adaptation scheme for consumer videos by effectively leveraging web images from different sources. Ikizler-Cinbis *et al.* [10] employed incrementally collected web images to learn a target classifier for action recognition. Wang *et al.* [22] proposed to obtain knowledge for the consumer videos from the labeled Web images and a small amount of labeled videos.

In this paper, we divide the training labeled Web images into several semantic groups (image sets). Since the events in the videos are complex, an event is not able to be characterized by searching the single semantic images. An event corresponds to several event-related keywords. For instance, the related keywords of event "sports" are "football", "basketball", "baseball", *etc.* For each keyword, we collect a set of related images from the image searching engines regarded as an image group. It is inevitable that the feature distributions of samples from the image groups and the videos may change considerably in the terms of the statistical properties. We note that the kernel function plays a crucial role in the kernel methods (e.g., SVM). The single kernel can not well solve the problem of classification. We propose a new multiple kernel domain adaptation method, referred to as Multiple kernel Image Group Adaptation, in order to tackle the considerable variation in the feature distribution from the image groups and consumer videos.

Specifically, we first assign different weights to the different image groups, which is based on their relevances to the target video domain. Then, due to the different weights of each web image group, we employ a nonparametric criterion to minimize the mismatch of data distributions from weighted Web image groups and target domain. Finally, in order to utilize the unlabeled target domain videos, we introduce a regularizer into the objective of Support Vector Regression (SVR) using the ϵ-insensitive. The regularizer is based on the smooth assumption that the two nearby samples in a high-density region should share the similar decision values on the target domain. Our method can simultaneously learn an optimal kernel function and a robust SVR classifier by minimizing the structural risk functional with the smooth assumption and mismatch of data distribution between the two domains. To simplify kernel learning, we assume the form of kernel function is a linear combination of multiple base kernels. Moreover, our algorithm is under a unified convex optimization framework, and we can easily solve the linear combination coefficients of kernels and SVR classifier.

2 Related Work

Our work focuses on annotating consumer videos by levering a large amount of labeled Web images, in which training and testing data come from different domains. Several domain adaptation(also called cross-domain learning or transfer learning) algorithms have shown the effectiveness and been used for some

applications [16], such as text classification and WIFI location. Recently, several domain adaptation methods have been presented and achieved good results in the computer vision. Bruzzone and Marconcini [2] not only proposed Domain Adaptation Support Vector Machine (DASVM) but also exploited a circular strategy to validate the result of domain adaptation classifiers. Cross-Domain SVM (CD-SVM) proposed by Jiang *et al.* [13] uses the K neighbors from target domain to assign weights for each sample in each source domain, then utilizing the new re-weight samples in the source domains and the labeled samples of the target domain to obtain a SVM classifier. Yang *et al.* [24] proposed Adaptation SVM (A-SVM) to adapt the existing classifiers for video concept detection. The target decision function is defined as the sum of the existing classifiers from the source domain and the perturbation function from the source and target domain. The methods in [2,13,24] are used to solve the single source domain. However, when there exists several source domain, the researchers proposed multiple source domain adaptation methods such as [3,5,8,9,23]. Domain Selection Machine with a data-dependent regularizer proposed by Duan *et al.* [5] to determine the most relevant source domains for domain adaptation without any labeled data from the target domain. Chattopadhyay *et al.* [3] proposed a weighting scheme based on smoothness assumption on the and the target classifier was learned by using the weighted combination of source classifier decision values. Feng *et al.* [8] expanded the smoothness assumption in [3], which enforces that similar consumer videos have the similar decision values and positive labeled videos have higher decision values than the negative labeled videos. Hammoud *et al.* [9] developed a novel concept of graphical fusion from video and text data to enhance activity analysis from aerial imagery.

In this work, we develop a multiple domain adaptation method called Multiple Kernel Image Group Adaptation by leveraging the loosely labeled web image groups. Our work is mainly related to two multiple source domain adaptation methods including Conditional Probability Multi-Source Domain Adaptation (CPMDA) [3] and Domain Selection Machine (DSM) [5]. In CPMDA, the weights of source domains is optimized by the smoothness assumption. In DSM, Duan *et al.* introduce a new data-dependent regularizer to select relevant source domain, which enforces the target classifier to share decision values with the selected source classifiers. The two methods fail to consider the mismatch of the distribution between the multiple source domains and the target domain. We introduce a nonparametric criterion into the objection of the SVR using the ϵ-insensitive with smoothness regularization, which minimizes the mismatch of data distributions from the multiple source domains and the target domain. As [2,3,6,13,17,24] assumed the source and target domains have the same type of feature, we employ the CNNs feature [12] of web images in the source groups and videos in the target domain.

3 Proposed Framework

We regard the loosely web images from different groups as the multiple source domains and the consumer videos as the target domain. Our goal is learning

a robust classifier for the target domain where there is a few labeled patterns and lots of unlabeled patterns. To obtain the multiple source domains for one event, we search several keywords related the event and refer to the images from one keyword search as a *group*. $D^g = (x_i^g, y_i^g)|_{i=1}^{n_g}, g \in \{1, \ldots, G\}$ denoted the g-group of the event, and n_g represents the number of images in the $g-$group, G is the total number of groups. D^S represents the total samples of the all groups. $n_S = \sum g = 1^G n_g$ denotes the number of images in all groups. We define the labeled training videos and unlabeled videos in the target domain as $D_l^T = (x_i^T, y_i^T)|_{i=1}^{n_l}$ and $D_u^T = x_i^T|_{i=n_l+1}^{n_l+n_u}$, respectively, where y_i^T is the label of x_i^T, and $D^T = D_l^T \bigcup D_u^T$ is the data set from the target domain with the size $n_T = n_l + n_u$. The transpose of vector/maxtrix is denoted by the superscript $'$ and the trace of matrix A is represented as $tr(A)$. We denote the identity matrix, the zero vector and the vector of all ones as $\mathbf{I}$, $\mathbf{0}$ and $\mathbf{1}$. Moreover, the matrix $A \succeq 0$ means the matrix A is symmetric and positive semidefinite.

3.1 Multiple-group Weighting

In our problem, some groups related to the target domain, while some groups may not. To reduce the negative transfer, we assign each of group to a weight while evaluate the relevance or similarity between the *s-th* group and target domain. We represent β_g as the weight of *g-th* group. f_i^g and f_i^T denote the decision value of *g-th* group classifier and target classifier on the target domain data x_i^T, respectively. We estimate the decision value($\widetilde{y}_i$) of unlabeled target domain data x_i^T based on a weighted combination of the g group classifiers:

$$\widetilde{y}_i = \sum_{s=1}^{G} \beta_s f_i^g = F_i^g \beta, \tag{1}$$

where $\boldsymbol{\beta} = [\beta_1, \cdots, \beta_G]'$ and $F_i^S = [f_i^1, \cdots, f_i^G]$. We use Chattopadhyay *et al.* [3] to estimate the weight vector β based on the smoothness assumption that two nearby points in the target domain have the similar decision value. Specifically, the optimal vector β minimize the following problem:

$$\begin{aligned} &\arg\min_{\boldsymbol{\beta}} \sum_{i,j=n_l+1}^{n_T} (F_i^S \boldsymbol{\beta} - F_j^S \boldsymbol{\beta}) W_{ij} \\ &s.t. \quad \boldsymbol{\beta} \geq \mathbf{0}, \boldsymbol{\beta}' \mathbf{1} = 1, \end{aligned} \tag{2}$$

where $F_i^S \boldsymbol{\beta}$ and $F_j^S \boldsymbol{\beta}$ are the predicted labels for the target domain x_i^T and x_j^T , respectively, and W_{ij} is the edge weight between x_i^T and x_j^T patterns. We can rewrite Eq. 2 with normalized graph Laplacian:

$$\arg\min_{\boldsymbol{\beta}} \boldsymbol{\beta}' (F_u^S)' L_u F_u^S \boldsymbol{\beta}, \tag{3}$$

where $F_u^S = [(F_{n_l+1}^S)' \cdots (F_{n_l+n_u}^S)'] \in \Re^{n_u \times G}$ is a $n_u \times G$ matrix of predicted decision value of unlabeled data in the target domain D_u^T and L_u is a normalized graph Laplacian given by $L_u = I - D_u^{-0.5} W D_u^{-0.5}$,where W is the

adjacency graph defining edge weights and D_u is a diagonal matrix given by $D_{ii} = \sum_{j=n_l+1}^{n_l+n_u} W_{ij}$. Equation 3 can be solved by a existing standard quadratic programming solvers. After obtaining $\boldsymbol{\beta}$, the estimated decision value (pseudo labels) of D_u can be computed by Eq. 1.

3.2 Reducing Mismatch of Data Distribution

In the domain adaptation learning, it is vital to reduce mismatch of data distribution between source domain and target domain. Duan *et al.* [7] proposed Adaptive Multiple Kernel Learning (A-KML) to simultaneously reduce the difference of data distribution between the auxiliary and the target domain and learn a target decision function. The mismatch is measured by a nonparametric criterion called MMD [1], which is based on the distance between the means of samples from the source domain D^A and the target domain D^T in the Reproducing Kernel Hilbert Space (RKHS), namely,

$$DISK_k(D^A, D^T) = \left\| \frac{1}{n^A} \sum_{i=1}^{n_A} \varphi(x_i^A) - \frac{1}{n^T} \sum_{i=1}^{n_T} \varphi(x_i^T) \right\|_H^2, \tag{4}$$

where the kernel function k is induced from the nonlinear feature mapping $\varphi(\cdot)$, $k(x_i, x_j) = \varphi(x_i)'\varphi(x_j)$, and x_i^A and x_i^T are the data from the source and target domains, respectively. However, the Eq. 4 can only apply to the single source domain. We propose a re-weight MMD for evaluating the difference of distribution between the multiple source domains and the target domain, i.e., the mean of samples from multiple source domains is calculated by adding the weights $\boldsymbol{\beta}$ to the mean of each source domain. The weight $\boldsymbol{\beta}$ can be obtained by Multiple-Group Weighting (in Sect. 3.1). The Eq. 4 can be rewritten as following form:

$$DISK_k(D^S, D^T) = \left\| \sum_{g=1}^{G} \frac{\beta_g}{n_g} \sum_{i=1}^{n_g} \varphi(x_i^g) - \frac{1}{n^T} \sum_{i=1}^{n_T} \varphi(x_i^T) \right\|_H^2. \tag{5}$$

We define a column vector s_g and s_T with n_g and n_T entries, respectively. The all entries in the s_g and s_T are set as β/n_g and $-1/n_T$ respectively. Let $\mathbf{s} = [s_1', \cdots, s_G', s_T']'$ be the vector with $N = n_S + n_T$. Let $\Phi = [\varphi(x_1^1), \cdots, \varphi(x_{n_1}^1), \cdots, \varphi(x_1^G), \cdots, \varphi(x_{n_G}^G), \varphi(x_1^T), \cdots \varphi(x_{n_T}^T)]$ be the data matrix from the group and target domain after feature mapping. Thus, the re-weight MMD criterion in Eq. 5 can be simplified as

$$DISK_k(D^S, D^T) = \|\Phi\mathbf{s}\|^2 = tr(\mathbf{KS}), \tag{6}$$

where $\mathbf{S}{=}\mathbf{s}\mathbf{s}' \in \Re^{N\times N}$, and $\mathbf{K} = \begin{bmatrix} \mathbf{K}^{S,S} & \mathbf{K}^{S,T} \\ \mathbf{K}^{T,S} & \mathbf{K}^{T,T} \end{bmatrix} \in \Re^{N\times N}$, and $\mathbf{K}^{S,S} \in \Re^{n_S\times n_S}$, $\mathbf{K}^{S,T} \in \Re^{n_S\times n_T}$ and $\mathbf{K}^{T,T} \in \Re^{n_T\times n_T}$ are the kernel matrices defined for the multiple source domains, the cross-domain from the multiple domains to the target domain and the target domain, respectively.

3.3 Multiple Kernel Image-Group Adaptation

Motivated by [3,5], we propose a new domain adaptation learning method, referred to as Multiple Kernel Image-Group Adaptation(MKIGA). Our method can learn a target classifier adapted from minimizing the mismatch of the distribution between two domains as well as a decision function which is based on multiple base kernels $k_m's$. The kernel function k is a linear combination of the based kernel $k_m's$, i.e., $k = \sum_{m=1}^{M} d_m k_m$, where kernel function k_m is induced from the nonlinear feature mapping function $\varphi_m(\cdot)$, *i.e.*, $k_m(x_i, x_j) = \varphi(x_i)'\varphi(x_j)$ and d_m is the linear linear combination coefficient of the kernel function k_m. In the MKIGA, the first objective is to reduce the mismatch in data distribution of the multiple groups and the target domain. Equation 6 can been rewritten as [4]

$$DISK_k(D^S, D^T) = \Omega(\boldsymbol{d}) = tr(\boldsymbol{KS}) = tr(\sum_{m=1}^{M} d_m \boldsymbol{K}_m \boldsymbol{S}) = \boldsymbol{h}'\boldsymbol{d}, \tag{7}$$

where $\mathbf{d} = [d_1, \ldots, d_M]'$ and the feasible set of $\mathbf{d}$ as $\boldsymbol{M} = \{\mathbf{d} \in \Re^M | \mathbf{1}'_M \mathbf{d} = 1, \mathbf{d} \geq 0_M\}$, $h = [tr(\mathbf{K}_1\mathbf{S}), \ldots, tr(\mathbf{K}_M\mathbf{S})]'$ and $\mathbf{K}_m = [\varphi(\mathbf{x})'\varphi(\mathbf{x})] \in \Re^{N \times N}$ is the m-th based kernel matrix defined on the samples from multiple groups and target domain. The second objective of the MKIGA is to minimize the risk functional [3]:

$$\min_{f^T \in H_K} \|f^T\|^2_{H_K} + C_l \Omega_l(f^T) + C_u \Omega_u(f^T) + r_m \Omega_m(f^T). \tag{8}$$

In Eq. 8, the first term is a regularizer to control the complexity of the classifier f^T in the Reproducing Kernel Hilbert Space(H_K), and the second and third terms are both the empirical error of the target classifier f^T on the few labeled target domain data D_l^T and the plenty of unlabeled target domain data D_u^T, respectively. In our method ,the empirical error is employed the ϵ-insensitive loss, *i.e.*, $\ell_\epsilon(t) = \begin{cases} |t| - \epsilon & \text{if } |t| > \epsilon \\ 0 & \text{otherwise} \end{cases}$. The fourth term is the manifold regularization which is enforced to be smooth on the data, namely, the two nearby samples in a high-density region should share the similar decision values. The manifold regularizer is defined as

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

where $\boldsymbol{f}^T = [f^T(\boldsymbol{x}_1^T), \ldots, f^T(\boldsymbol{x}_{n_T}^T)]'$ is the decision values of the target domain D^T, and L is the graph Laplacian matrix constructed on D^T. The r_A, C_u, C_l and r_m are penalty factors.

Recall that the use of Support Vector Regression (SVR) with the ϵ-insensitive loss function can usually lead to a sparse representation of target decision function. Therefore, to obtain the sparse solution, we introduce the SVR in Eq. 8, the target domain classifier $f(\boldsymbol{x}) = \sum_{m=1}^{M} d_m \boldsymbol{w}_m' \varphi_m(\boldsymbol{x}) + b$. By the representation theory, the Eq. 8 can be rewritten as

$$
\begin{aligned}
J(\boldsymbol{d}) = \min_{\boldsymbol{w}_m, b, \xi, \xi^*, \boldsymbol{f}^T} & \frac{1}{2} \sum_{m=1}^{M} d_m \|\boldsymbol{w}_m\|^2 + C \sum_{i=1}^{n_T} (\xi_i + \xi_i^*) \\
& + \frac{1}{2} \left(C_u \|\boldsymbol{f}_u^T - \widetilde{\boldsymbol{y}_u}\|^2 + C_l \|\boldsymbol{f}_l^T - \boldsymbol{y}_l\|^2 \right) + r_m \boldsymbol{f}^{T\prime} \boldsymbol{L} \boldsymbol{f}^T \\
s.t. & \sum_{m=1}^{M} d_m \boldsymbol{w}_m' \varphi_m(\boldsymbol{x}_i) + b - f_i^T \leq \epsilon + \xi_i, \xi_i \geq 0, i = 1, \ldots, n_T \\
& f_i^T - \sum_{m=1}^{M} d_m \boldsymbol{w}_m' \varphi_m(\boldsymbol{x}_i) - b \leq \epsilon + \xi_i^*, \xi_i^* \geq 0, i = 1, \ldots, n_T,
\end{aligned} \tag{9}
$$

where $\boldsymbol{f}_u^T = [f_1^T, \ldots, f_{n_u}^T]'$ and $\boldsymbol{f}_l^T = [f_{n_u+1}^T, \ldots, f_{n_T}^T]'$ are the vectors of the target decision function on the unlabeled samples D_u^T and labeled samples D_l^T from the target domain,$\widetilde{\boldsymbol{y}_u} = [\widetilde{y_1}, \ldots, \widetilde{y_{n_u}}]'$ and $\boldsymbol{y}_l = [y_{n_u+1}, \ldots, y_{n_T}]'$ are the vectors of pseudo labels and true labels in the target domain D_u^T and D_l^T, respectively. The optimization problem in MKIGA is minimizing the combination of the distance between the data distributions of multiple groups and target domain, as well as the risk loss function of kernel. Putting the Eqs. 5 and 9 together, we are arriving at the formulation as follows:

$$
\min_{\boldsymbol{d} \in \boldsymbol{M}} G(\boldsymbol{d}) = \frac{1}{2} \Omega(\boldsymbol{d})^2 + \theta J(\boldsymbol{d}). \tag{10}
$$

Let us define $\boldsymbol{v}_m = d_m \boldsymbol{w}_m$. The optimization in Eq. 9 can be rewritten as

$$
\begin{aligned}
J(\boldsymbol{d}) = \min_{\boldsymbol{v}_m, b, \boldsymbol{\xi}, \boldsymbol{\xi}^*, \boldsymbol{f}^T} & \frac{1}{2} \sum_{m=1}^{M} \frac{\|\boldsymbol{v}_m\|^2}{d_m} + C \sum_{i=1}^{n_T} (\xi_i + \xi_i^*) \\
& + \frac{1}{2} \left(C_u \|\boldsymbol{f}_u^T - \widetilde{\boldsymbol{y}_u}\|^2 + C_l \|\mathbf{f}_l^T - \mathbf{y}_l\|^2 \right) + r_m \mathbf{f}^{T\prime} \mathbf{L} \mathbf{f}^T \\
s.t. & \sum_{m=1}^{M} \boldsymbol{v}_m' \varphi_m(\boldsymbol{x}_i) + b - f_i^T \leq \epsilon + \xi_i, \xi_i \geq 0, i = 1, \ldots, n_T \\
& f_i^T - \sum_{m=1}^{M} \boldsymbol{v}_m' \varphi_m(\boldsymbol{x}_i) - b \leq \epsilon + \xi_i^*, \xi_i^* \geq 0, i = 1, \ldots, n_T.
\end{aligned} \tag{11}
$$

Note that the first term $\frac{1}{2}\Omega(\boldsymbol{d})$ in Eq. 10 is a quadratic term with respect to $\boldsymbol{d}$. The third and fourth terms is convex with respect to $\boldsymbol{f}^T$, since the the graph Laplacian matrix L is the positive semidefinite and the third term in Eq. 11 is a quadratic term with respect to $\boldsymbol{f}^T$. The other terms in Eq. 11 are the linear or convex except the term $\frac{1}{2}\sum_{m=1}^{M} \frac{\|\boldsymbol{v}_m\|^2}{d_m}$. As discussed in [4], the term $\frac{1}{2}\sum_{m=1}^{M} \frac{\|\boldsymbol{v}_m\|^2}{d_m}$ is also jointly convex with respect to $\boldsymbol{d}$ and $\boldsymbol{v}_m$. Therefore, the optimization problem in Eq. 10 is jointly convex with respect to $\boldsymbol{d}, \boldsymbol{v}_m, b, \boldsymbol{f}^T, \xi_i$ and ξ_i^*. The objective in Eq. 10 can reach its global minimum. By introducing the Lagrangian multipliers α_i and η_i (resp. α_i^* and η_i^*) for the constraints of Eq. 11, we obtain the dual form of the optimization problem in Eq. 11 as follows:

$$
\begin{aligned}
J(\boldsymbol{d}) = \min_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*} \frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)'\boldsymbol{Q}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) + (\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)'\boldsymbol{y}^* \\
+ \epsilon(\boldsymbol{\alpha}+\boldsymbol{\alpha}^*)'\mathbf{1} + const \\
\boldsymbol{s.t.} \qquad \boldsymbol{\alpha}'\mathbf{1} = \boldsymbol{\alpha}^{*\prime}\mathbf{1}, \quad 0 \le \boldsymbol{\alpha}, \boldsymbol{\alpha}^* \le C\mathbf{1},
\end{aligned}
\tag{12}
$$

where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_{n_T}]'$, $\boldsymbol{\alpha^*} = [\alpha_1^*, \ldots, \alpha_{n_T}^*]'$, $\boldsymbol{y} = [\widetilde{\boldsymbol{y}_u}', \boldsymbol{y}_l']'$, $\boldsymbol{y^*} = (\Lambda\boldsymbol{I} + \gamma_m L)^{-1}\Lambda\boldsymbol{y}$, $\boldsymbol{Q} = \sum_{m=1}^{M} d_m\boldsymbol{K}_m + (\Lambda\boldsymbol{I}+\gamma_m L)^{-1}$, $\Lambda = diag(C_u, \ldots, C_u, C_l, \ldots, C_l)$ includes the n_u entries of C_u and n_l entries of C_l, and the last term is a constant term that is irrelevant to he Lagrangian multipliers. Surprisingly, the optimization problem in Eq. 12 has the same form as the dual of the SVR except the kernel matrix and the labels. Thus we can exploit the existing SVR solvers such as LIBSVM to solve the optimization. Substituting Eqs. 7 and 12 back into Eq. 10, the final optimization formulation is given by

$$
\begin{aligned}
\min_{\boldsymbol{d},\boldsymbol{\alpha},\boldsymbol{\alpha}'} G = & \frac{1}{2}\boldsymbol{d}'\boldsymbol{h}\boldsymbol{h}'\boldsymbol{d} + \theta\left(\frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)'\boldsymbol{Q}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)\right) \\
& + \theta\left((\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)'\boldsymbol{y}^* + \epsilon(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)'\mathbf{1}\right) + const.
\end{aligned}
\tag{13}
$$

We employ the reduced gradient descent procedure to iteratively update the linear combination coefficient $\boldsymbol{d}$ and the dual variables $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}*$ in Eq. 13. Given the linear combination coefficient $\boldsymbol{d}$, we obtain the dual variable $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}*$ by utilizing the LIBSVM to solve the optimization. Suppose the dual variables $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}*$ is fixed, the second-order gradient descent method [7] is introduced to update the linear combination $\boldsymbol{d}$. After obtaining the optimal $\boldsymbol{d}$ and $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}*$, we rewrite the decision function as follows:

$$
f(\boldsymbol{x}) = \sum_{m=1}^{M} d_m \boldsymbol{w}_m^{'}\varphi_m(\boldsymbol{x}) + b = \sum_{i:\alpha_i-\alpha_i^*\neq 0} (\alpha_i - \alpha_i^*)\sum_{m}^{M} d_m k_m(x_i, x) + b.
$$

4 Experiments

In this section, we compare our method with the baseline method SVM, the existing single source domain adaptation methods of Domain Adaptive SVM (DASVM) [2] and Adaptive SVM (A-SVM) [24], as well as the multi-domain adaptive methods including Domain Adaptive Machine (DAM) [6], Conditional Probability Multi-Source Domain Adaptation (CP-MDA) [3] and Domain Selection Machine (DSM) [5]. We evaluate our method on two datasets: the Columbias Consumer video CCV [14] and the TRECVID 2014 Multimedia Event Detection dataset [15]. We use Average Precision (AP) to evaluate performance, and report the mean AP over all events.

4.1 Datasets and Features

(1) **CCV Dataset:** It contains a training set of 4,649 videos and a test set of 4,658 videos which are annotated to 20 semantic categories. Since our work

focuses on event analysis, we do not consider the non-event categories (*e.g.* "bird", "cat" and "dog"). We only use the videos from the event related categories. We also merge "wedding ceremony", "wedding reception" and "wedding dance" into the event of "wedding", "music performance" and "non-music performance" into "show". Finally, there are twelve event categories: "basketball", "baseball", "biking", "graduation", "ice-skating", "show", "parade", "skiing", "soccer", "swimming", "birthday" and "wedding".

(2) **TRECVID 2014 Multimedia Event Detection dataset:** It contains 40 categories of events: we use the *10EX*, *Background*, *MEDTest*. It contains 10 positive videos for each event in the *10EX*, 4,983 background videos which do not belong to any event category in the *Background* and 29200 videos in the *MEDTest*. Especially, the videos in the *MEDTest* contain about 25 positive samples for each event and 26717 negative videos which do not belong to any event event category. We only use the labeled training videos from the 21-th category to 40-th category. In these training videos, we randomly select a small number of videos as the labeled videos, the rest of videos as the unlabeled videos. Finally, there are 3483 videos in our experiment.

(3) **Web Image Dataset:** We collect a large number of images by keyword search from Google image search engine as our source domain. For each event category, we define five keywords related event. The top ranked 200 images are downloaded and we enforce the returned images to be photo with full color by using the advanced options provided by Google image search engine. We do not download the corrupted images or the images with invalid URLs. Finally, 26,708 images are collected. Some examples of multi-group image dataset are show in Fig. 1.

Fig. 1. Exemplar images from the Web image groups related the event "basketball", each row shows a image group.

(4) **Feature:** For each target domain video, we sample one keyframe per 2 seconds. For the each sampled keyframe ,we extract the 4096-dimensional feature vector

CNNs feature by using Caffe [12]. The fc7 layer of CNNs is used as features, and we use the method of max-pooling to obtain a video feature. We pool the extracted CNNs features from the video into a 4096-dimensional feature vector. Finally, we represent a video/image as a 4096-dimensional feature vector.

4.2 Experiment Setup

In the experiment, we construct five image groups for each event. We first train a pre-learned classifier for each image group using the images in the group as positive samples and randomly select equivalent number of samples from groups of other events as negative sample. Base kernels are predetermined for all methods. Specifically, we make use of four kernel types: Guassian kernel (*i.e.*, $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = exp(-A\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2)$), Laplacian kernel (*i.e.*, $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = exp(-\gamma\|\boldsymbol{x}_i - \boldsymbol{x}_j\|)$), inverse square distance kernel (*i.e.*, $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = -\frac{1}{\gamma\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2+1}$) and inverse distance kernel $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = -\frac{1}{\sqrt{\gamma}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|+1}$), We set $\gamma = 4^{n-1}\gamma_0$ where $n \in -2, -1, \ldots, 2, \gamma_0 = \frac{1}{A}$ and A is the mean value of square distances between all training samples. In total, we have 20 based kernels from kernel types and five kernel parameters. For our method, we empirically set $C_l = 1$, $C_u = 0.1$, $\theta = 10^{-5}$, $\epsilon = 10^{-5}$ and $\gamma_m = 0.002$ in our experiment. The SVM parameter C is set to 1 in all methods.

For CCV and MED2014, we randomly samples with 20 % per event as the labeled target videos. We sample target domain training videos ten times and report the means and standard deviations of mAPs for each methods. For the baseline SVM algorithm, we report the results for two cases: (1) in SVM_S, the samples in 5 groups are put together for SVM learning; (2) in SVM_A, we equally fuse the decision values of 5 pre-learned classifier. For the single domain method, we put the samples in five groups together as the source domain. For the multi-source domain, each group is regarded as a source domain.

4.3 Results

We show the MAPs of all methods on the two datasets in Table 1. From the results, we can have the following observations:

(1) SVM_A is better than SVM_S and DASVM, which indicates irrelevant images may be harmful for the classification performances in the target videos. However, the domain adaptation methods CPMDA, A_MKL, and DAM is better than SVM_S and SVM_A in the terms of MAPs, which demonstrates that the domain adaptation can successfully make use of the source domain to learn a better classifier for target domain.
(2) In the terms of MAPs, the performances of the multiple source domain adaptation methods CPMDA, DAM and DSM are better than the single source method DASVM, which demonstrates that it is effective to divide images into the multiple image groups. Moreover, multiple kernel learning method A_MKL shows a better performance.

(3) It is obvious that our method achieves the best results on the both datasets. We believe that the multiple image group adaptation can cope with noisy web images. On the CCV dataset (*resp.*, the MED14 dataset), the relative improvement of our method over the best existing method are 4.36 % (*resp.*, 4.29 %). It demonstrate that the distribution of the data between image groups and target domain videos influence the performance of knowledge transfer from the Web images to consumer videos.

Table 1. Mean Average Precisions MAPs (%) of all methods on CCV and MED14 datasets

Method	SVM_S	SVM_A	DASVM	DSM	DAM	CPMDA	A_MKL	Ours
CCV	44.49	48.75	47.25	47.91	50.99	53.24	52.22	57.60
MED14	30.01	33.56	31.30	33.06	32.79	36.31	36.02	40.60

5 Conclusion

In this paper, we have proposed a novel Multiple Kernel Image-Group Adaptation to explore to a large number of labeled Web images to recognize the events in consumer videos. We divide the images into several semantic groups and assign different weights to these groups. To reduce the mismatch of distribution between multiple image groups and target domain videos, as well as the a large number of unlabeled target domain videos, MKIGA minimizes the SVR structural function and the distribution mismatch between two domains. MKIGA simultaneously learns a kernel function and a target classifier which is smooth on the target domain.

References

1. Borgwardt, K.M., Gretton, A., Rasch, M.J., Kriegel, H.P., Schölkopf, B., Smola, A.J.: Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics **22**(14), e49–e57 (2006)
2. Bruzzone, L., Marconcini, M.: Domain adaptation problems: a DASVM classification technique and a circular validation strategy. IEEE Trans. Pattern Anal. Mach. Intell. **32**(5), 770–787 (2010)
3. Chattopadhyay, R., Sun, Q., Fan, W., Davidson, I., Panchanathan, S., Ye, J.: Multisource domain adaptation and its application to early detection of fatigue. ACM Trans. Knowl. Discov. Data (TKDD) **6**(4), 18 (2012)
4. Duan, L., Tsang, I.W., Xu, D.: Domain transfer multiple kernel learning. IEEE Trans. Pattern Anal. Mach. Intell. **34**(3), 465–479 (2012)
5. Duan, L., Xu, D., Chang, S.F.: Exploiting web images for event recognition in consumer videos: a multiple source domain adaptation approach. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1338–1345. IEEE (2012)
6. Duan, L., Xu, D., Tsang, I.W.: Domain adaptation from multiple sources: a domain-dependent regularization approach. IEEE Trans. Neural Netw. Learn. Syst. **23**(3), 504–518 (2012)

7. Duan, L., Xu, D., Tsang, I.H., Luo, J.: Visual event recognition in videos by learning from web data. IEEE Trans. Pattern Anal. Mach. Intell. **34**(9), 1667–1680 (2012)
8. Feng, Y., Wu, X., Wang, H., Liu, J.: Multi-group adaptation for event recognition from videos. In: 2014 22nd International Conference on Pattern Recognition (ICPR), pp. 3915–3920. IEEE (2014)
9. Hammoud, R.I., Sahin, C.S., Blasch, E.P., Rhodes, B.J.: Multi-source multi-modal activity recognition in aerial video surveillance. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 237–244. IEEE (2014)
10. Ikizler-Cinbis, N., Cinbis, R.G., Sclaroff, S.: Learning actions from the web. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 995–1002. IEEE (2009)
11. Izadinia, H., Shah, M.: Recognizing complex events using large margin joint low-level event model. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 430–444. Springer, Heidelberg (2012)
12. Jia, Y.: Caffe: an open source convolutional architecture for fast feature embedding (2013). http://caffe.berkeleyvision.org
13. Jiang, W., Zavesky, E., Chang, S.F., Loui, A.: Cross-domain learning methods for high-level visual concept classification. In: 15th IEEE International Conference on Image Processing, 2008, ICIP 2008, pp. 161–164. IEEE (2008)
14. Jiang, Y.G., Ye, G., Chang, S.F., Ellis, D., Loui, A.C.: 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, p. 29. ACM (2011)
15. MED2014. http://www.nist.gov/itl/iad/mig/med14.cfm
16. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2010)
17. Schölkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge (2002)
18. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach.In: Proceedings of the 17th International Conference on Pattern Recognition, 2004, ICPR 2004, vol. 3, pp. 32–36. IEEE (2004)
19. Sefidgar, Y.S., Vahdat, A., Se, S., Mori, G.: Discriminative key-component modelsfor interaction detection and recognition. In: Computer Vision and Image Understanding (2015)
20. Trichet, R., Nevatia, R.: Video segmentation descriptors for event recognition. In:2014 22nd International Conference on Pattern Recognition (ICPR), pp. 1940–1945. IEEE (2014)
21. Vahdat, A., Cannons, K., Mori, G., Oh, S., Kim, I.: Compositional models for video-event detection: a multiple kernel learning latent variable approach. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 185–1192. IEEE (2013)
22. Wang, H., Wu, X., Jia, Y.: Annotating videos from the web images. In: 2012 21st International Conference on Pattern Recognition (ICPR), pp. 2801–2804. IEEE(2012)
23. Wang, H., Wu, X., Jia, Y.: Video annotation via image groups from the web. IEEE Trans. Multimed. **16**, 1282–1291 (2014)
24. Yang, J., Yan, R., Hauptmann, A.G.: Cross-domain video concept detection using adaptive SVMS. In: Proceedings of the 15th International Conference onMultimedia,pp. 188–197. ACM (2007)