IET Computer Vision

Research Article

IET Journals
The Institution of Engineering and Technology

Multi-group–multi-class domain adaptation for event recognition

ISSN 1751-9632
Received on 7th November 2014
Revised on 3rd April 2015
Accepted on 26th May 2015
doi: 10.1049/iet-cvi.2014.0405
www.ietdl.org

Yang Feng, Xinxiao Wu ✉, *Yunde Jia*

Beijing Laboratory of Intelligent Information Technology, Beijing Institute of Technology, Beijing 100081, People's Republic of China
✉ *E-mail: wuxinxiao@bit.edu.cn*

Abstract: In this study, the authors propose a multi-group–multi-class domain adaptation framework to recognise events in consumer videos by leveraging a large number of web videos. The authors' framework is extended from multi-class support vector machine by adding a novel data-dependent regulariser, which can force the event classifier to become consistent in consumer videos. To obtain web videos, they search them using several event-related keywords and refer the videos returned by one keyword search as a group. They also leverage a video representation which is the average of convolutional neural networks features of the video frames for better performance. Comprehensive experiments on the two real-world consumer video datasets demonstrate the effectiveness of their method for event recognition in consumer videos.

1 Introduction

In recent years, people have started to capture their own videos in daily life with the help of easy accessible digital cameras and mobile phone cameras. Automatically recognising events in these consumer videos have become an important research topic due to its usefulness in video management and retrieval. Consumer videos usually contain considerable camera motions and occlusions, which lead to large intra-class variations within the same type of event videos and make it very challenging to recognise event in consumer videos.

To tackle this challenge, many researchers collect a large number of training videos and label them, and then train a robust model from these data for recognition. Collecting a large number of labelled videos is labour expensive and time consuming. Noticing that domain adaptation [1] can make it possible to train classifiers using existing auxiliary data, researchers begin leveraging web videos for event recognition in consumer videos. As search engines have become increasingly mature and can offer abundant data with loose labels, many researchers collect training data by searching on the web instead of manual labelling. Duan *et al.* [2] proposed to transfer knowledge from YouTube videos to the consumer video domain, which minimises the structural risk function and the mismatch between the data distributions. Ikizler-Cinbis *et al.* [3] collected images from the web and used the knowledge from the images to automatically annotate actions in videos. Chen *et al.* [4] reported multi-domain adaptation with heterogeneous sources (MDA-HS) which integrates different source domains with different types of features to learn an optimal target classifier. Feng *et al.* [5] proposed a semi-supervised method named multi-group adaptation to transfer the knowledge in web videos to consumer videos, which can learn the optimal weights for the source groups by using both the labelled and unlabelled target domain data.

Most previous domain adaptation methods are based on binary classifiers and are extended to multi-class classification by using one versus the rest. Since the binary classifiers are usually trained independently, they cannot reflect the relationship among the classes. Additional problems with extending binary classifiers to multi-class problems include the difference of output scales of different binary classifiers and imbalanced training data.

To handle these problems, we propose a multi-group–multi-class domain adaptation framework which is extended from multi-class support vector machine (SVM) [6]. We introduce a novel data-dependent regulariser to adapt the event classifier learned on the web videos, which can enforce the event classifier to be smooth in the consumer video domain. We encourage the event classifiers to be smooth, mainly in the areas where their responses are relatively high because the high responses decide which class a sample belongs to. This framework can make full use of the label information in our web data and the distribution information in unlabelled consumer video data. By using the information in both domains, our method can transfer the label knowledge from web videos to consumer videos.

Since the events in consumer videos are complex and various, videos returned by searching with a single query word are not sufficient to describe the event. Following Wang *et al.* [7], we use several event-related keywords to query videos from the web. For example, we may associate the event 'parade' with the keywords 'march', 'demonstrate', 'procession' and so on. For each keyword, we collect a set of related web videos and regard them as a *group*. Several researchers also obtained training videos from YouTube, but they did not divide videos into groups. Luo *et al.* [8] also searched with several keywords for each concept, but the returned videos are mixed together to represent a concept. Jiang *et al.* [9] released a database called Columbia consumer video (CCV), containing 9317 web videos which were collected by searching a string 'MVI' with each of the category names. After downloading the web videos, we extract the features of video frames using the ImageNet 2012 winning model. Finally, a video is represented as the average of the features of its frames.

The rest of this paper is organised as follows. We briefly review the related work in Section 2. The convolutional neural network (CNN) feature for videos is introduced in Section 3. The domain adaptation framework is described in Section 4. The experimental results are reported in Section 5. Finally, we conclude this paper in Section 6.

2 Related work

Several examples have demonstrated the effectiveness of domain adaptation which can tackle the problem that there are no enough training data in the target domain, but there are sufficient data in related domains. In [10], Yang *et al.* proposed adaptive SVM (A-SVM) to learn a new classifier for video concept detection,

which is the sum of an existing classifier trained on the source domain and a perturbation for the labelled target data. Bruzzone and Marconcini [11] proposed domain adaptation support vector machine (DASVM) to iteratively label the unlabelled target domain data, removing some labelled source domain data at the same time. Hoffman *et al.* [12] proposed to learn a new representation to compensate the mismatch of source and target domains. A joint weight learning framework is introduced in [13] to fuse multiple source view action classifiers for recognition in the target view. These methods are based on binary classifiers and are extended to multi-class classification by approaches such as one versus the rest.

A few multi-class classifier-based domain adaptation methods [14–16] have been proposed. Lee and Jang [14] trained a model on the source domain as prior information and built the target model based on it. In [15], Xu *et al.* modified the delta loss of the source domain data in multi-class SVM. Our domain adaptation is based on smooth assumption and is different from [14–16]. There are some closely related semi-supervised multi-class learning methods [17–20] which are mainly based on smoothness, cluster or manifold assumptions. In [17], Valizadegan *et al.* designed an objective function that measures the consistency between the predicted class labels and the similarity matrix of unlabelled examples and enforced higher consistency in their framework. Tanha *et al.* [18] also proposed to maximise the consistency between similarity and predicted labels, but their formulation was different from [17], which minimises the difference of the sum of responses of two similar samples. Liu *et al.* [19] proposed positiveness exclusive regularisation to ensure that one example receives one positive response among all the classifiers.

Our method is much related to [17, 21], which are based on the smoothness assumption. Our method is to transfer the knowledge from one domain to another domain by leveraging data in both domains, while [17] is designed for semi-supervised learning to improve the classification performance with the help of unlabelled data. We have multiple groups in each event class. If we regard each group as an independent class, the regulariser in our method regresses to the same as [17]. The difference between our method and [21] is that their framework is based on the binary classifier. They summed the binary regularisers up for multi-class domain adaptation, while we design a special regulariser for multi-class domain adaptation.

The problem solved in [5] and the one in this paper are similar, but the methods used are completely different. Feng *et al.* [5] sought the optimal weight for each group by leveraging a few labelled and plenty of unlabelled target domain data. The main novelty of [5] is that the mean of the decision values of the negative and positive target samples is introduced to the objective for calculating the weights. In this paper, we do not require any labelled target domain data and we use multi-class SVM instead of binary SVM used in [5]. As mentioned above, the method in this paper can overcome the problems of imbalanced training data and different output scale. By adding a data-dependent regulariser, our method can make full use of the information in our data from both the source and the target domain.

3 Video representation

Karpathy *et al.* [22] collected a large sports video dataset and trained several CNN for video classification on the dataset. Their spatio-temporal networks obtained very good performance, but these networks required high computation cost. They also found that the best spatio-temporal networks performed only a little better than the single-frame models. So the appearance information in frames can represent videos well. Motivated by their work, we use a single-frame CNN for video feature extraction, which can greatly reduce the computation cost and keep good performance. The single-frame CNN is chosen as the ImageNet challenge winning model [23]. Instead of training the single-frame CNN on Sports-1M [22], we train it on ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC) because the image dataset is much easier to obtain. ILSVRC-2012 is a rich image dataset and the CNN model can learn good appearance representation on it. We can extract this image features easily with the help of Caffe [24], which implements the training and testing of CNN neatly. After extracting the features of frames, we average the features of selected frames in a video as the video feature.

4 Proposed framework

Following the terminology of domain adaptation, we refer to the web video domain as the source domain in which we have abundant searched videos, and the consumer video domain as the target domain in which we have plenty of unlabelled videos. To obtain the training videos for one event, we search with several keywords associated with the event. We refer to the videos returned by one keyword search as a group. Consequently, the videos retrieved from the search engine for one event are divided into different groups according to their corresponding search keywords. We define the gth group of the cth event as $D^{cg} = (\boldsymbol{x}_i^{cg}, c)|_{i=1}^{n_{cg}}$, $c \in \{1, \ldots, C\}$, $g \in \{1, \ldots, G\}$, where n_{cg} represents the number of videos in this group and C and G are the total number of events and groups per event, respectively. There are also plenty of unlabelled target data $D^{\mathrm{T}} = \boldsymbol{x}_i^{\mathrm{T}}|_{i=1}^{n_T}$. The goal of our work is to make full use of the label information in the source domain and the distribution information in the target domain to build a robust event classifier in consumer videos.

In the remainder of this paper, the symbol $'$ represents the transpose operation and the bold italic letters represent a vector or a matrix. The superscript T represents that a sample comes from the target domain.

4.1 Similarity-basedregulariser

Belkin *et al.* [25] proposed Laplacian regulariser which enforces that the decision function is smooth on the data manifold, namely, similar instances in a high-density region should have similar decision values. Although it was originally proposed for semi-supervised learning, several researchers applied the smooth assumption in domain adaptation and obtained promising results [26, 27]. If we have a binary target domain classifier $f(\boldsymbol{x}^{\mathrm{T}}) = \boldsymbol{w}'\Phi(\boldsymbol{x}^{\mathrm{T}})$, we can write the manifold-based regulariser defined in [25] as

$$\Omega_{\mathrm{b}}(f) = \boldsymbol{f}'\boldsymbol{L}\boldsymbol{f} = \frac{1}{2}\sum_{i}^{n_T}\sum_{j}^{n_T} W_{ij}(f(\boldsymbol{x}_i^{\mathrm{T}}) - f(\boldsymbol{x}_j^{\mathrm{T}}))^2, \quad (1)$$

where $\boldsymbol{f} = [f(\boldsymbol{x}_1^{\mathrm{T}}), \ldots, f(\boldsymbol{x}_{n_T}^{\mathrm{T}})]'$ and $\boldsymbol{L}$ is the graph Laplacian matrix constructed on D^{T}: $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$, where $\boldsymbol{W}$ is the edge weights between the samples in the target domain, and $\boldsymbol{D}$ is the diagonal matrix given by $D_{ii} = \sum_{j=1}^{n_T} W_{ij}$.

Donahue *et al.* [21] applied manifold regularisation in semi-supervised domain adaptation. When there are multiple classes, they used a one-versus-all manner, in which every class has its own binary decision function. In their objective function, all the regularisers of binary classifiers are summed up to enforce smoothness in the whole feature space. We propose a different approach to regularise the classifiers which enforces a classifier to be smooth mainly in part of the feature space. In multi-class classification, a sample is assigned to the class of a classifier which gives the highest decision value. If a classifier gives a very low decision value to a sample compared with other classifiers, this classifier will not change which class the sample belongs to. Regularising a classifier in the areas where its response is relatively low may become a burden for domain adaptation. So we regularise the classifiers to be smooth, mainly in the areas where their responses are relatively high. Fig. 1 shows an example of three classes. According to the decision function, we can figure out that the rectangles belong to class c1, the triangles belong to class c2 and the circles belong to c3. The decision values from the grey classifier for the triangles are significantly lower than the

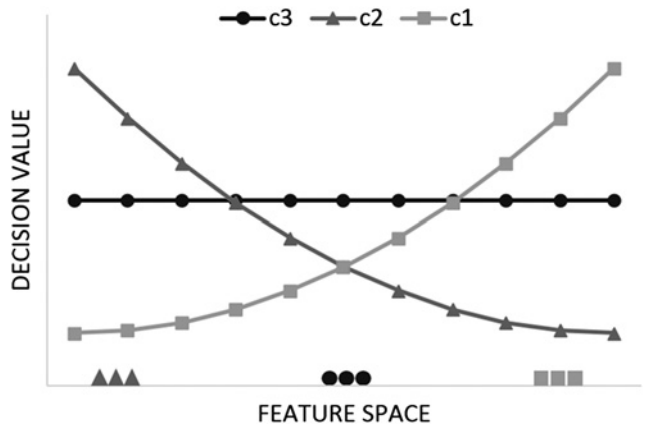

Fig. 1 *Illustration of multi-class smooth regularisation*

decision values from the other two classifiers. Thus, when the grey classifier changes a little in the triangle area will not affect the classification performance. We need not enforce the grey classifier to be smooth in the triangle area.

Our novel regulariser is written as

$$\Omega_m = \sum_{i=1}^{n_T}\sum_{j=1}^{n_T} W_{ij}\delta_{ij}, \tag{2}$$

where W_{ij} is the element in edge weights matrix $\boldsymbol{W}$ and δ_{ij} is the dissimilarity between the ith and the jth sample according to their decision values, which is similar to $(f(\boldsymbol{x}_i^{\mathrm{T}}) - f(\boldsymbol{x}_j^{\mathrm{T}}))^2$ in (1). δ_{ij} is given by

$$\delta_{ij} = \frac{1}{\sum_{c=1}^{C} p(y=c|\boldsymbol{x}_i^{\mathrm{T}})p(y=c|\boldsymbol{x}_j^{\mathrm{T}})}, \tag{3}$$

where $p(y=c|\boldsymbol{x}_i^{\mathrm{T}})$ is the probability that the ith sample belongs to the cth class. The dissimilarity is the reciprocal of the cosine similarity of $p(y=c|\boldsymbol{x}_i^{\mathrm{T}})$ and $p(y=c|\boldsymbol{x}_j^{\mathrm{T}})$. By minimising (3), we can encourage two samples to be classified to the same class. Let us denote the decision function of the gth group of the cth class as $f_{cg}(\boldsymbol{x}^{\mathrm{T}}) = \boldsymbol{w}'\Phi(\boldsymbol{x}^{\mathrm{T}}, c, g)$. Then $p(y=c|\boldsymbol{x}^{\mathrm{T}})$ is written as

$$p(y=c|\boldsymbol{x}^{\mathrm{T}}) = \frac{\exp\left(\sum_{g=1}^{G} f_{cg}(\boldsymbol{x}^{\mathrm{T}})/G\right)}{\sum_{k=1}^{C}\exp\left(\sum_{g=1}^{G} f_{kg}(\boldsymbol{x}^{\mathrm{T}})/G\right)}. \tag{4}$$

Using (2) as a regulariser has two advantages. First, it can alleviate the aforementioned smooth area problem because the probabilities $p(y=c|\boldsymbol{x}^{\mathrm{T}})$ are mainly affected by the decision functions with relatively high decision values. The decision functions are adjusted and encouraged to be smooth, mainly in the areas where their responses are relatively high. Second, it will penalise a lot if two near samples have very different decision values because δ_{ij} will become very large when $\sum_{c=1}^{C} p(y=c|\boldsymbol{x}_i^{\mathrm{T}})p(y=c|\boldsymbol{x}_j^{\mathrm{T}})$ is small. We notice that our regulariser is similar to [17] but different from it because there are different groups in our regulariser. The groups consist of the videos returned by automatic queries in YouTube, so they are noisy. We average the decision values of the group classifiers belonging to one event to alleviate the effect of the noise. If we regard each group as an independent class and do $C*G$ classes classification, we can directly use the regulariser in [17]. However, this simple employment of [17] ignores the relation between the groups belonging to one event. It has the problem that the difference of two samples may become large when they are classified into different groups within the same event. For example, if two similar videos are classified into two groups 'march' and 'demonstrate' and their ground truth event label is 'parade', then both of them are correctly classified but the difference between them will be large using the similarity measure in [17]. We call this problem *group inconsistency*. By averaging the decision values of the group classifiers, our regulariser can overcome this problem.

4.2 Multi-group–multi-class domain adaptation

We extend the multi-class SVM [6] to a domain adaptation formulation by adding a consistency regulariser

$$\begin{aligned} &\min_{\boldsymbol{w},\xi_i^{cg}} \frac{\lambda}{2}\boldsymbol{w}'\boldsymbol{w} + \sum_{c=1}^{C}\sum_{g=1}^{G}\sum_{i=1}^{n_{cg}} \xi_i^{cg} + \mu\Omega_m \\ &\text{s.t.} \quad \boldsymbol{w}'\Phi(\boldsymbol{x}_i^{cg}, c, g) - \boldsymbol{w}'\Phi(\boldsymbol{x}_i^{cg}, y, h) \geq \Delta(c,y) - \xi_i^{cg},\ \xi_i^{cg} \geq 0 \\ &\forall c, \forall g, \forall i, \forall y, \forall h, \end{aligned} \tag{5}$$

where λ and μ are balancing parameters, and $\Delta(c, y) = 1$ if $c \neq y$, and 0 otherwise. This constrained optimisation problem can be rewritten as an unconstrained problem

$$\begin{aligned} &\min_{\boldsymbol{w}} \frac{\lambda}{2}\boldsymbol{w}'\boldsymbol{w} + \sum_{c=1}^{C}\sum_{g=1}^{G}\sum_{i=1}^{n_{cg}} \left(\max_{y,h}\left(\Delta(c,y) + \boldsymbol{w}'\Phi(\boldsymbol{x}_i^{cg}, y, h)\right)\right. \\ &\left. - \boldsymbol{w}'\Phi(\boldsymbol{x}_i^{cg}, c, g)\right) + \mu\sum_{i=1}^{n_T}\sum_{j=1}^{n_T} W_{ij}\delta_{ij}. \end{aligned} \tag{6}$$

By optimising the three terms in (6) together, we can use both the label information in the source domain and the structural information in the target domain, and thus solve the domain adaptation problem. We employ the non-convex bundle optimisation technique in [28] to solve (6) which is the same fashion as solving SVM in primal. $\boldsymbol{w}$ is initialised by the solution of multi-class SVM.

To leverage the kernel trick, we refer to the idea in [29]. In [29], the data are mapped to a n-dimensional representation by using kernel PCA, where n is the number of examples in the training set. The new representation makes sense because the dot product of two samples in the new representation equals to the value of the kernel mapping of them [i.e. $\Phi(\boldsymbol{x}_i)\cdot\Phi(\boldsymbol{x}_j) = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$]. In our paper, the solution to (6) can be written as $\boldsymbol{w} = \sum_{c=1}^{C}\sum_{g=1}^{G}\sum_{i=1}^{n_{cg}} \alpha_{cgi}\Phi(\boldsymbol{x}_i^{cg}, c, g)$. So only the dot product of two samples is used in (6) and we can use the method in [29] to incorporate kernels. Instead of using eigendecomposition in [29], we use singular value decomposition because it is faster and more accurate. Specifically, we first put all the samples in both the source and target domain together. Let us denote $\boldsymbol{X}^{cg} = [\boldsymbol{x}_1^{cg}, \ldots, \boldsymbol{x}_{n_{cg}}^{cg}]$, $\boldsymbol{X}^{\mathrm{T}} = [\boldsymbol{x}_1^{\mathrm{T}}, \ldots, \boldsymbol{x}_{n_T}^{\mathrm{T}}]$ and all the samples can be written as $\boldsymbol{X} = [\boldsymbol{X}^{11}, \ldots, \boldsymbol{X}^{CG}, \boldsymbol{X}^{T}]$. Then we calculate the kernel matrix $\boldsymbol{K}$. Any kind of kernel can be used here. We calculate the singular value decomposition of $\boldsymbol{K}$ as $[\boldsymbol{U}, \boldsymbol{D}, \boldsymbol{V}] = \mathrm{SVD}(\boldsymbol{K})$. At last, we represent the data as $\boldsymbol{X}_{\mathrm{new}} = \boldsymbol{U}\mathrm{sqrt}(\boldsymbol{D})\boldsymbol{U}'$, where sqrt is the element-wise square root operation. Every column is a new representation of a sample and we can verify that $\boldsymbol{X}'_{\mathrm{new}}\boldsymbol{X}_{\mathrm{new}} = \boldsymbol{K}$. The new representation is used to solve (6).

5 Experiments

We compare the proposed domain adaptation framework with the baseline method multi-class SVM, multi-class SVM plus each of the smooth regularisers proposed in [21, 17, 18] and geodesic flow kernel (GFK) [30] with multi-class SVM. We also compare it with the binary SVM and binary classifier-based domain adaptation methods MDA-HS [4] and DASVM [11]. While mean average precision is widely used in measuring the performance of event recognition methods, we cannot use it because our method is based on the multi-class SVM, in which the classification is made according to the decision values from all the event classifiers. The

IET Comput. Vis., 2016, Vol. 10, Iss. 1, pp. 60–66

Table 1 Query keywords for each event

Event	Keywords				
birthday	birthday	anniversary	celebration	birthday party	birthday cake
parade	parade	march	demonstrate	procession	Walk
picnic	picnic	eat	breakfast	dinner	Lunch
show	show	dance	display	nightlife	exhibit
sports	sports	athletic	football	baseball	basketball
wedding	wedding	marriage	nuptial	wedding ceremony	marriage ceremony

decision value of a sample from one event classifier cannot determine how likely this sample belongs to the event alone. So we use the recognition accuracy for performance evaluation.

5.1 Datasets

5.1.1 Kodak dataset: The Kodak dataset was collected by Kodak [8] from about 100 real users over 1year. We use the features provided by Duan *et al.*[2], which contain 195 consumer videos from six event classes (i.e. 'birthday', 'parade', 'picnic', 'show', 'sports' and 'wedding').

5.1.2 Columbia consumer video dataset: This is a consumer video dataset collected by Columbia University [9]. It contains a training set of 4659 videos and a test set of 4658 videos which are annotated to 20 semantic categories. Since our work focuses on event analysis, we only use the videos from the event-related categories (i.e. 'basketball', 'baseball', 'soccer', 'iceskating', 'skiing', 'swimming', 'biking', 'birthday', 'wedding reception', 'wedding ceremony', 'wedding dance', 'music performance', 'non-music performance' and 'parade').

5.1.3 Multi-group video dataset: We collect a large number of videos by searching on YouTube as the source domain data. We first need to define the associational keywords for each event. It is not a trivial task to choose the search keywords. We refer to WordNet [31] to reduce the effect of personal experience. WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms, each expressing a distinct concept. We view sister term, hyponym and paraphrase of an event word and choose five closely related words as keywords. Table 1 lists the associational keywords for each event in our experiment. Each row shows the five groups of an event. If we directly search with the keywords defined above, the returned results will contain many TV commercial videos. To avoid these videos, queries are formed by combining a string 'MVI' with each of the keyword as suggested in [9]. We download the top ranked 100 videos from YouTube searches for each keyword. Some examples of our multi-group video dataset are shown in Fig. 2.

5.2 Features

To extract our CNN feature, we select N_f frames (i.e. $N_f = 100$) on the whole video with a fixed step, which is set to the total number of frames/N_f. If a video does not have N_f frames, we use all the frames in the video. Then we feed these frames into the ImageNet model in Caffe and use the values in the 'fc7' layer as the feature. The details of how to extract image features using Caffe is given in the home page of Caffe. After obtaining the features of the frames in one video, we average them to represent the video.

The scale invariant feature transform (SIFT) features [32] of the CCV dataset and web videos are extracted in the following way. For each video clip, we sample at a rate of 2 frames per second to extract keyframes. For each keyframe, we extract 128-dimensional SIFT features from salient regions. Then, we cluster the SIFT features extracted from all the keyframes of the training videos into 2000 words by using k-means clustering. Each keyframe is represented by a 2000-dimensional term frequency (TF) feature based on the bag-of-words representation. Each video is represented by the average of the TF features over all the keyframes within it.

5.3 Experimental setups

The multi-group video dataset is used as the source domain, and the Kodak or the CCV dataset is used as the target domain. Since the videos in the Kodak dataset are not available to the public, we only use the SIFT feature for Kodak. Following [33], we merge 'wedding ceremony', 'wedding reception' and 'wedding dance' as 'wedding'; 'non-music performance' and 'music performance' as 'show'; and 'baseball', 'basketball', 'biking', 'ice skating', 'skiing', 'soccer' and 'swimming' as 'sports'. Finally, there are 2502 videos from the training part of the CCV dataset. They belong to five event classes (i.e. 'birthday', 'parade', 'show', 'sports' and 'wedding'). In our settings there is no labelled data in the target domain. So we train a multi-class SVM in the source domain and use it to classify target domain data directly in the multi-class SVM baseline. Each group has its own classifier. If the event of the classified group of a test sample is the same as the ground truth event, it is regarded as a correct classification. For example, if two videos are classified into two groups 'march' and 'demonstrate' and their ground truth event label is 'parade', then both of them are correctly classified. In the experiments for regulariser comparing, our regulariser is replaced by one of the smooth regularisers in [17, 18, 21]. Each group is regarded as a class in these three regularisers because they are originally designed for multi-class learning without groups. So there are $\tilde{C} = C*G$ classes in these three experiments. The three regularisers and our regulariser are listed in Table 2.

To demonstrate that it is helpful to divide the web videos into groups, we merge the samples in different groups of one event together and refer to this method as ours-ng. The regulariser of ours-ng is also defined by R2 by replacing the $\tilde{C}$ with C. In the binary SVM-based methods, we use the one-versus-the-rest strategy to extend them for multi-class classification and manually balance the training data by selecting part of negative samples randomly. MDA-HS can handle multiple source domains while DASVM and GFK are single source domain adaptation methods. So we treat each group as a source domain in MDA-HS and merge the samples in the groups of an event together in DASVM and GFK.

As suggested in [34], we use the non-linear χ^2 distance to measure the distance between two videos when using the bag-of-words of SIFT features. It is defined by

$$D(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\sum_{i=1}^{d}\frac{(x_i - y_i)^2}{x_i + y_i},$$

where d is the dimensionality of $\boldsymbol{x}$ or $\boldsymbol{y}$. Then we choose the Laplacian kernel (i.e. $K(i, j) = \exp(-1/\sqrt{A}\, D(\boldsymbol{x}_i, \boldsymbol{x}_j))$) for the good results reported in [2], where A is the mean value of square distances between all training samples. Since the CNN feature is not a bag-of-words representation, we use the general Gaussian kernel [i.e. $K(i, j) = \exp(-1/A\, D(\boldsymbol{x}_i, \boldsymbol{x}_j))$] and Euclidean distance for it. The adjacency matrix $\boldsymbol{W}$ is set as 'binary' type based on the N nearest neighbours with $N = 6$. The dimension of the subspace in GFK is chosen automatically using the method provided in [30]. The other parameters of different methods are chosen in the following way: we first define a set of parameter values and report the best results of different methods.

Fig. 2 *Examples in multi-group video dataset*

IET Comput. Vis., 2016, Vol. 10, Iss. 1, pp. 60–66

Table 2 Different regularisers

R1	$\sum_{c=1}^{\tilde{C}} \sum_{i=1}^{n_T} \sum_{j=1}^{n_T} W_{ij} (f_c(\boldsymbol{x}_i^T) - f_c(\boldsymbol{x}_j^T))^2$
R2	$\sum_{i=1}^{n_T} \sum_{j=1}^{n_T} W_{ij} \frac{\left(\sum_{c=1}^{\tilde{C}} \exp f_c)\boldsymbol{x}_i^T)\right)\left(\sum_{c=1}^{\tilde{C}} \exp f_c(\boldsymbol{x}_j^T)\right)}{\sum_{c=1}^{\tilde{C}} \exp(f_c(\boldsymbol{x}_i^T)) \exp f_c(\boldsymbol{x}_i^T))}$
R3	$\sum_{i=1}^{n_T} \sum_{j=1}^{n_T} W_{ij} \exp\left(\left(\sum_{c=1}^{\tilde{C}} f_c(\boldsymbol{x}_i^T) - \sum_{c=1}^{\tilde{C}} f_c(\boldsymbol{x}_j^T)\right)/(\tilde{C}-1)\right)$
ours	$\sum_{i=1}^{n_T} \sum_{j=1}^{n_T} W_{ij} / \left(\sum_{c=1}^{C} p(y=c\|\boldsymbol{x}_i^T) p(y=c\|\boldsymbol{x}_j^T)\right)$

Table 3 Accuracies of all methods on the Kodak dataset

Method	Kodak + SIFT, %
binary SVM	47.2
DASVM [11]	55.4
MDA-HS [4]	47.7
GFK [30] + multi-class SVM	47.7
multi-class SVM	53.8
multi-class SVM + R1	54.9
multi-class SVM + R2	59.5
multi-class SVM + R3	54.4
ours-ng	62.6
ours	63.6

5.4 Results

Tables 3 and 4 show the recognition accuracies of all the methods on the Kodak dataset and the CCV dataset, respectively. From the results, we can observe that:

i. The result of R1 is the same as the baseline multi-class SVM on the CCV dataset, because the parameter μ is given a very small value. If we give a larger value to μ, the performance of R1 will degrade.
ii. The regulariser R2 can improve the results on both datasets. However, because the web data are noisy and group inconsistency, the improvement is less than our method. The improvement of R3 is less than our method probably because R3 regulates all the classifiers in an ensemble means. When two similar samples are classified differently, all the classifiers are adjusted in the same way.
iii. Our methods are better than the binary SVM, DASVM and MDA-HS, which demonstrates that using multi-class classification directly is more suitable in our problem.
iv. Compared with multi-class SVM, the improvements of our method on the Kodak dataset, the CCV dataset with SIFT feature and the CCV dataset with CNN feature are 9.8, 13.1 and 13.9%, respectively. The improvement on the CCV dataset is higher. One explanation is that there are more videos in the CCV dataset and these data provide more distribution information of the target domain.
v. The performance of the CNN feature is constantly better than the SIFT feature, which demonstrates that the CNN feature can capture the appearances in videos better than the SIFT feature.
vi. Our method achieves better results than ours-ng because dividing the web video into groups can better represent the information in them. The improvement on Kodak dataset is quite modest because the Kodak dataset is smaller and less diverse. The results of our method are the best on two datasets, which demonstrates the effectiveness of our method for event recognition in consumer videos by making full use of the information in available data.

Table 4 Accuracies of all methods on the CCV datasets

Method	CCV + SIFT, %	CCV + CNN, %
binary SVM	44.5	49.0
DASVM [11]	41.9	48.1
MDA-HS [4]	52.2	57.1
GFK [30] + multi-class SVM	40.8	45.0
multi-class SVM	43.7	50.2
multi-class SVM + R1	43.7	50.2
multi-class SVM + R2	47.8	54.8
multi-class SVM + R3	44.5	50.3
ours-ng	54.3	61.8
ours	56.8	64.1

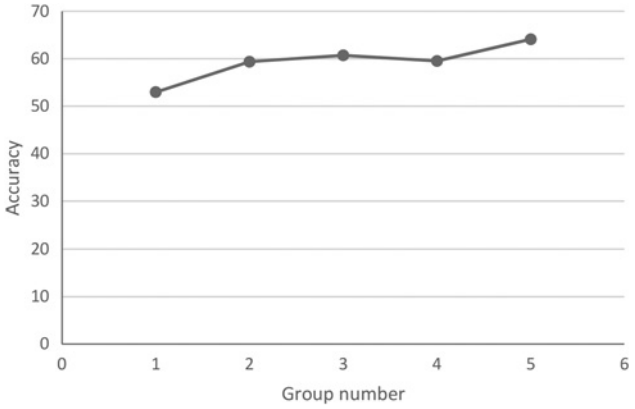

Fig. 3 *Accuracies using different number of groups*

We also evaluated the performance of our domain adaptation method when using different number of groups on the CCV dataset with the CNN feature. We select the groups from the left columns to the right. For example, when there are three groups, we use the left three columns as the keywords. The result is shown in Fig. 3. We can find that the results of multiple groups are consistently better than single group. This indicates that using more keywords to search web videos is helpful to event recognition in consumer videos.

6 Conclusion

In this paper, we have proposed a novel multi-group–multi-class domain adaptation method to leverage a large number of web videos to recognise events in consumer videos. To obtain good video representation, we have used the ImageNet 2012 winning model to extract video features. In our framework, we divide the web videos into different semantic groups by querying different keywords in YouTube and referring the videos returned by one keyword search as a group. To make full use of the information in both the source domain and the target domain, we add a novel data-dependent regulariser to multi-class SVM, which can encourage the target classifier to be smooth and consistent in the target domain. Comprehensive experiments on the Kodak and CCV datasets demonstrate the effectiveness of our method for event recognition in consumer videos.

In the future, we intend to use more kinds of source knowledge from the web. Some top ranking videos from YouTube search are very relevant to the search keyword, but the latter ones may not be relevant to the search keyword. To further improve the results, we can exploit the knowledge in other source domain, for example we can use the images returned by Google or Flicker.

7 Acknowledgments

The research was supported in part by the Natural Science Foundation of China (NSFC) under grant 61203274, the Specialised Fund for Joint Building Program of Beijing Municipal Education Commission, the Excellent Young Scholars Research Fund of Beijing Institute of Technology and the Beijing Key Laboratory of Advanced Information Science and Network Technology.

8 References

1 Pan, S.J., Yang, Q.: 'A survey on transfer learning', *IEEE Trans. Knowl. Data Eng.*, 2010, **22**, (10), pp. 1345–1359
2 Duan, L., Xu, D., Tsang, I.W., *et al*.: 'Visual event recognition in videos by learning from web data', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012, **34**, (9), pp. 1667–1680
3 Ikizler-Cinbis, N., Cinbis, R.G., Sclaroff, S.: 'Learning actions from the web'. IEEE Int. Conf. on Computer Vision, Kyoto, Japan, September 2009, pp. 995–1002
4 Chen, L., Duan, L., Xu, D.: 'Event recognition in videos by learning from heterogeneous web sources'. IEEE Conf. on Computer Vision and Pattern Recognition, Portland, USA, June 2013, pp. 2666–2673
5 Feng, Y., Wu, X., Wang, H., *et al*.: 'Multi-group adaptation for event recognition from videos'. Int. Conf. on Pattern Recognition, Stockholm, Sweden, August 2014, pp. 3915–3920
6 Crammer, K., Singer, Y.: 'On the algorithmic implementation of multiclass kernel-based vector machines', *J. Mach. Learn. Res.*, 2002, **2**, pp. 265–292
7 Wang, H., Wu, X., Jia, Y.: 'Video annotation via image groups from the web', *IEEE Trans. Multimed.*, 2014, **16**, (5), pp. 1282–1291
8 Loui, A., Luo, J., Chang, S., *et al*.: 'Kodaks consumer video benchmark data set: concept definition and annotation'. Proc. of the Int. Workshop on Multimedia Information Retrieval, Augsburg, Germany, 2007, pp. 245–254
9 Jiang, Y., He, G., Chang, S., *et al*.: 'Consumer video understanding: a benchmark database and an evaluation of human and machine performance'. Proc. of the First ACM Int. Conf. on Multimedia Retrieval, Trento, Italy, 2011, p. 29
10 Yang, J., Yan, R., Hauptmann, A.G.: 'Cross-domain video concept detection using adaptive SVMs'. Proc. of the 15th Int. Conf. on Multimedia, Augsburg, Germany, 2007, pp. 188–197
11 Bruzzone, L., Marconcini, M.: 'Domain adaptation problems: a DASVM classification technique and a circular validation strategy', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010, **32**, (5), pp. 770–787
12 Hoffman, J, Rodner, E., Donahue, J., *et al*.: 'Efficient learning of domain-invariant image representations', arXiv preprint, arXiv:1301.3224, 2013
13 Wu, X., Wang, H., Liu, C., *et al*.: 'Cross-view action recognition over heterogeneous feature spaces'. IEEE Int. Conf. on Computer Vision, Sydney, Australia, December 2013, pp. 609–616
14 Lee, C., Jang, M.G.: 'A prior model of structural SVMs for domain adaptation', *ETRI J.*, 2011, **33**, (5), pp. 712–719
15 Xu, J., Ramos, S., Vázquez, D., *et al*.: 'Cost-sensitive structured SVM for multi-category domain adaptation'. Int. Conf. on Pattern Recognition, Stockholm, Sweden, August 2014, pp. 3886–3891
16 Xu, J., Ramos, S., Vázquez, D., *et al*.: 'Domain adaptation of deformable part-based models', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2014, **36**, (12), pp. 2367–2380
17 Valizadegan, H., Jin, R., Jain, A.K.: 'Semi-supervised boosting for multi-class classification'. Proc. of the European Conf. on Machine Learning and Knowledge Discovery in Databases, Antwerp, Belgium, 2008, pp. 522–537
18 Tanha, J., Van Someren, M., Afsarmanesh, H.: 'Boosting for multiclass semi-supervised learning', *Pattern Recognit. Lett.*, 2014, **37**, pp. 63–77
19 Liu, X., Yuan, X., Yan, S., *et al*.: 'Multi-class semi-supervised SVMs with positiveness exclusive regularization'. IEEE Int. Conf. on Computer Vision, Barcelona, Spain, 2011, pp. 1435–1442
20 Saffari, A., Leistner, C., Bischof, H.: 'Regularized multi-class semisupervised boosting'. IEEE Conf. on Computer Vision and Pattern Recognition, Miami, USA, June 2009, pp. 967–974
21 Donahue, J., Hoffman, J., Rodner, E., *et al*.: 'Semi-supervised domain adaptation with instance constraints'. IEEE Conf. on Computer Vision and Pattern Recognition, Portland, USA, June 2013, pp. 668–675
22 Karpathy, A., Toderici, G., Shetty, S., *et al*.: 'Large-scale video classification with convolutional neural networks'. IEEE Conf. on Computer Vision and Pattern Recognition, Columbus, USA, 2014, pp. 1725–1732
23 Krizhevsky, A., Sutskever, I., Hinton, G.E.: 'Imagenet classification with deep convolutional neural networks'. Advances in Neural Information Processing Systems, Harrahs Lake Tahoe, USA, 2012, pp. 1097–1105
24 Jia, Y., Shelhamer, E., Donahue, J.: 'Caffe: convolutional architecture for fast feature embedding', arXiv preprint, arXiv:1408.5093, 2014
25 Belkin, M., Niyogi, P., Sindhwani, V.: 'Manifold regularization: a geometric framework for learning from labeled and unlabeled examples', *J. Mach. Learn. Res.*, 2006, **7**, pp. 2399–2434
26 Duan, L., Xu, D., Tsang, I.W.: 'Domain adaptation from multiple sources: a domain-dependent regularization approach', *IEEE Trans. Neural Netw. Learn. Syst.*, 2012, **23**, (3), pp. 504–518
27 Chattopadhyay, R., Sun, Q., Fan, W., *et al*.: 'Multisource domain adaptation and its application to early detection of fatigue', *ACM Trans. Knowl. Discov. Data*, 2012, **6**, (4), p. 18
28 Do, T.M.T., Arti'eres, T.: 'Large margin training for hidden Markov models with partially observed states'. Proc. of Int. Conf. on Machine Learning, Montreal, Canada, 2009, pp. 265–272
29 Zien, A., De Bona, F., Ong, C.S.: 'Training and approximation of a primal multiclass support vector machine', ASMDA, 2007
30 Gong, B., Shi, Y., Sha, F., *et al*.: 'Geodesic flow kernel for unsupervised domain adaptation'. IEEE Conf. on Computer Vision and Pattern Recognition, Providence, USA, 2012, pp. 2066–2073
31 Miller, G.A.: 'WordNet: a lexical database for English', *Commun. ACM*, 1995, **38**, (11), pp. 39–41
32 Lowe, D.G.: 'Distinctive image features from scale-invariant keypoints', *Int. J. Comput. Vis.*, 2004, **60**, (2), pp. 91–110
33 Duan, L., Xu, D., Chang, S.F.: 'Exploiting web images for event recognition in consumer videos: a multiple source domain adaptation approach'. IEEE Conf. on Computer Vision and Pattern Recognition, Providence, USA, 2012, pp. 1338–1345
34 Laptev, I., Marszalek, M., Schmid, C., *et al*.: 'Learning realistic human actions from movies'. IEEE Conf. on Computer Vision and Pattern Recognition, Anchorage, USA, 2008, pp. 1–8