Int J Comput Vis (2014) 109:42–59
DOI 10.1007/s11263-014-0703-y

# Weakly-Supervised Cross-Domain Dictionary Learning for Visual Recognition

**Fan Zhu** · **Ling Shao**

Received: 15 March 2013 / Accepted: 14 February 2014 / Published online: 12 March 2014

**Abstract** We address the visual categorization problem and present a method that utilizes weakly labeled data from other visual domains as the auxiliary source data for enhancing the original learning system. The proposed method aims to expand the intra-class diversity of original training data through the collaboration with the source data. In order to bring the original target domain data and the auxiliary source domain data into the same feature space, we introduce a weakly-supervised cross-domain dictionary learning method, which learns a reconstructive, discriminative and domain-adaptive dictionary pair and the corresponding classifier parameters without using any prior information. Such a method operates at a high level, and it can be applied to different cross-domain applications. To build up the auxiliary domain data, we manually collect images from Web pages, and select human actions of specific categories from a different dataset. The proposed method is evaluated for human action recognition, image classification and event recognition tasks on the UCF YouTube dataset, the Caltech101/256 datasets and the Kodak dataset, respectively, achieving outstanding results.

Communicated by Dr. Trevor Darrell.

F. Zhu · L. Shao (✉)
College of Electronic and Information Engineering,
Nanjing University of Information Science and Technology,
Nanjing 210044, China
e-mail: ling.shao@ieee.org

F. Zhu · L. Shao
Department of Electronic and Electrical Engineering,
The University of Sheffield, Sheffield S1 3JD, UK
e-mail: ling.shao@sheffield.ac.uk

F. Zhu
e-mail: fan.zhu@sheffield.ac.uk

**Keywords** Visual categorization · Image classification · Human action recognition · Event recognition · Transfer learning · Weakly-supervised dictionary learning

## 1 Introduction

In the past few years, along with the explosion of online image and video data (Flickr[1], YouTube[2]), the computer vision community has witnessed a significant amount of applications in content-based image/video search and retrieval, human–computer interaction, sport events analysis, etc. These applications are built upon the development of several aspects of classical computer vision tasks, such as human action recognition, object localization and image classification, which, however, remain challenging in real-world scenarios due to cluttered background, view point changes, occlusion, and geometric and photometric variations of the target (Su and Jurie 2012; Yao et al. 2012; Wang and Mori 2011, 2009; Jégou et al. 2010; Junejo et al. 2011; Duchenne et al. 2009; Marszalek et al. 2009). These issues result in either imposing irrelevant information to the target introduced by, e.g., cluttered background, or producing very different representations for the same target caused by, e.g., geometric and photometric changes. Many previous methods that manage to deal with these issues are proposed and state-of-the-art approaches include semantic attributes (Su and Jurie 2012), estimated pose features (Yao et al. 2012), and mined hierarchical features (Gilbert et al. 2011). The conventional framework applies a robust classifier using human annotated training data, but it makes the assumption that the testing data stay in the same feature space or share the same distribution with

[1] http://www.flickr.com/

[2] http://www.youtube.com/

the training data. However, in real-world applications, due to the high price of human manual annotation and environmental restrictions, sufficient training data that stay in the same feature space or share the same distribution with the testing data cannot always be guaranteed, in which case insufficient training data can limit the potential discriminability of the trained model. Typical examples are Cao et al. (2013), Gao et al. (2011), and Orrite et al. (2011), where only one action template is provided for each action class for training, and Liu et al. (2011), where training samples are captured from a different viewpoint. In these situations, obtaining more labeled data is either impossible or expensive, while seeking for an alternative way of using data from other domains as compensation can be seen as a possible and economic solution.

Our work is inspired by two facts of the human vision system. The first fact is that humans are able to learn tens of thousands of visual categories in their life, which leads to the hypothesis that humans achieve such a capability by accumulated information and knowledge (Fei-Fei 2006), as shown in Fig. 1. Another fact is human's visual impressions towards the same action or the same object comes from a wide range, e.g., an action seen from 2D static images versus the same action seen from 3D dynamic movies or an object seen from real-world scenes versus the same object seen from low-resolution online images. However, the human vision system is still able to correctly distinguish such actions or objects regardless of their visual diversities, which, in other words, can be explained in the computer vision language that the human vision system possesses the ability to span the intra-class diversity of the original training data. In a similar way, we argue that the computer-based visual categorization system can also gain more discriminative power by spanning the coverage of training samples' intra-class variations, as shown in Fig. 2.

Motivated by the above two facts, we introduce a new visual categorization framework that utilizes weakly labeled data from other domains as the source data (motivated by the first fact) to span the intra-class diversity of the original learning system (motivated by the second fact). Following the classical single-task cross-domain learning setup (Pan and Yang 2010), our aim is to complete the visual categorization task in the target domain. In addition to the manually labeled training data in the target domain, the source domain data are utilized as extensions of category prototypes in the target domain. Based on the recent success of dictionary learning methods in solving computer vision problems, we present a weakly-supervised cross-domain dictionary learning method to learn a reconstructive, discriminative and domain-adaptive dictionary pair and an optimal linear classifier simultaneously. In order to demonstrate the effectiveness of our method, we gather supportive evidence by evaluating our method on action recognition, image classification and event recognition tasks. The UCF YouTube dataset (Liu et al. 2009), the Caltech101 dataset (Fei-Fei et al. 2007), the Caltech 256 dataset (Griffin et al. 2007) and the Kodak consumer video dataset (Loui et al. 2007) are used as the target domain data in our experiments, while selected actions in the HMDB51 dataset (Kuehne et al. 2011) and some indexed Web images or YouTube videos are used as the source domain data in our experiments. The preliminary results of our method have been presented in Zhu and Shao (2013).

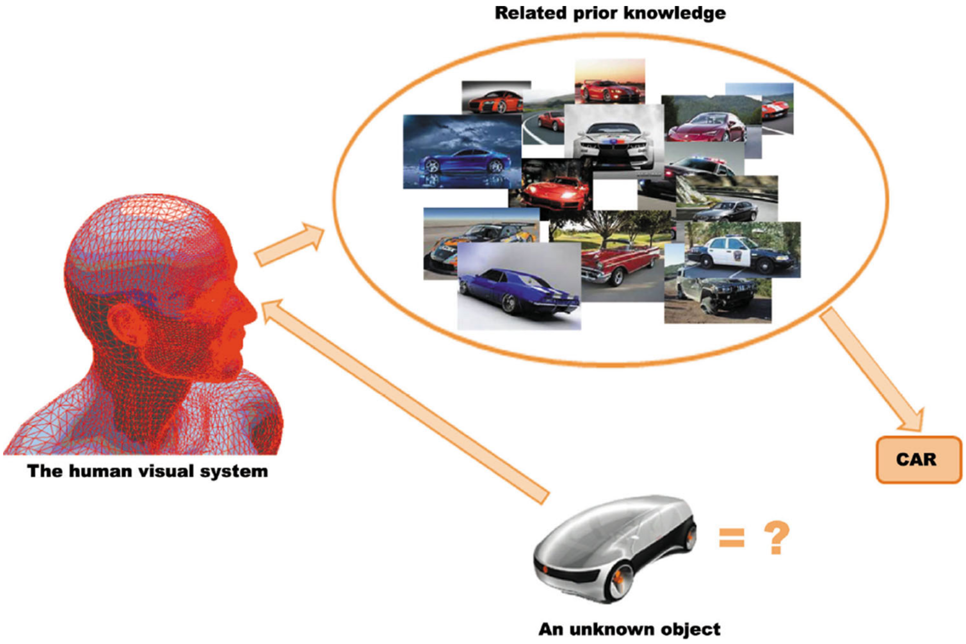

**Fig. 1** Illustration of how a new object is accumulated to the human visual system as prior knowledge for future usage. The given unknown object is a future car, which is unacquainted to the viewer. Since the viewer's prior knowledge towards cars spans a wild coverage of target samples, shared information (e.g., car shape and wheels) between the new object and prior knowledge is easily discovered.

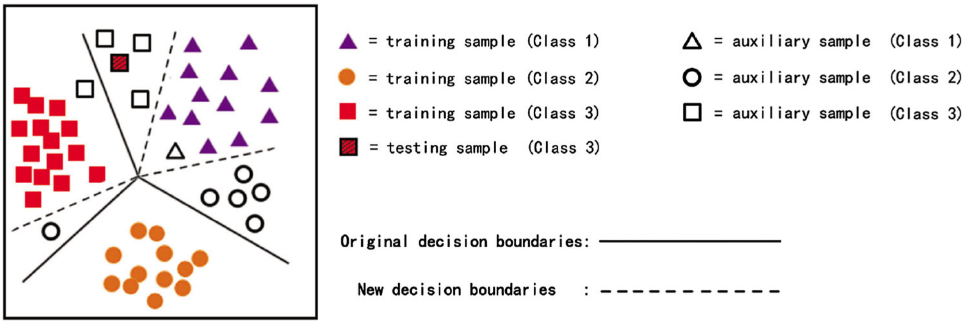
▲ = training sample (Class 1)
● = training sample (Class 2)
■ = training sample (Class 3)
▨ = testing sample (Class 3)
△ = auxiliary sample (Class 1)
○ = auxiliary sample (Class 2)
□ = auxiliary sample (Class 3)
Original decision boundaries:
New decision boundaries :

**Fig. 2** Illustration of how the categorization system can gain more discriminative power through the collaboration with the source domain data in the 2-dimensional feature space. The *purple triangles*, the *orange circles* and *the red squares* denote the training samples from Classes 1, 2 and 3 respectively, and the corresponding *hollow shapes* denote the auxiliary training samples from Classes 1, 2 and 3. Original decision boundaries are represented by the *solid lines* and the new decision boundaries are represented by the *dashed lines*. The testing sample, which is denoted as a *red square* with *black borders*, is misclassified as Class 1 according to the original decision boundaries. Proper auxiliary samples lead to more rational decision boundaries, so that the coverage of Class 1 spans against the centre of Class 2. Thus, the testing sample can be correctly labeled (Color figure online)

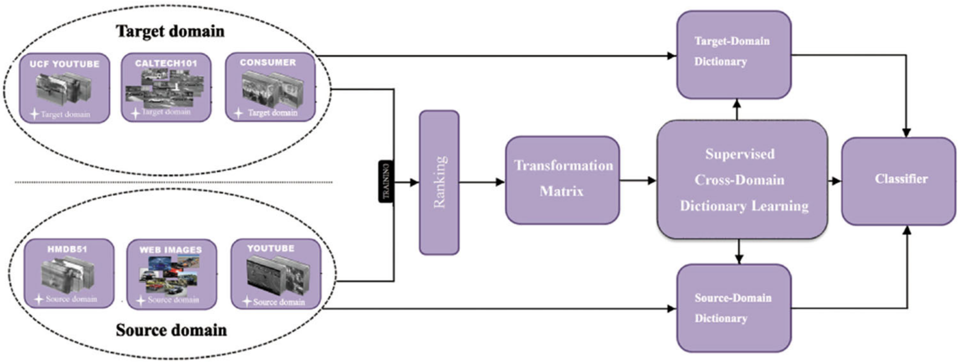

**Fig. 3** Flowchart of the proposed approach. The target domain data are split into the training part and the testing part, where the training data are used as a set of queries to rank the source domain data within selected categories. A pre-defined number of most relevant source samples are chosen to construct the transformation matrix, which describes the connections between the source domain data and the target domain data. With the label information of the target domain training samples, weakly-supervised cross-domain dictionary learning is performed. A reconstructive, discriminative and domain-adaptive dictionary pair is learned together with the target classifier. The target domain testing data can be encoded with the learned target domain dictionary, following which the labels can be predicted by feeding the new representations into the learned classifier

Our proposed method is illustrated in Fig. 3 and it offers the following two main contributions. Firstly, it attempts to make use of as much as possible existing knowledge by a novel weakly-supervised visual categorization framework. An efficient manifold ranking method is applied to the source domain for the selection of a pre-defined number of most relevant instances per category according to the target domain training data, following which correspondences connecting the source domain and the target domain are established based on the selected source domain data and the target domain training data. Secondly, we propose a new cross-domain dictionary learning method to cope with the feature distribution mismatch problem across the source domain and the target domain. Specifically, we perform dictionary learning upon the correspondences built from both domains so that the projections of data from different domains can obey the same distribution when limited by the learning function. In addition to the dictionary, classifier parameters are learned jointly during the discriminative dictionary function learning process. Thus, knowledge transfer of the proposed framework is accomplished through both the feature level and the classifier level. As the samples from the source domains are weakly labeled rather than being manually (correctly) labeled, we call our algorithm “Weakly-Supervised Cross-Domain Dictionary Learning”(WSCDDL).

The remainder of this paper is organized in the following way. Related works are reviewed in Sect. 2. In Sects. 3 and 4, we first extensively discuss related dictionary learning techniques and then introduce the proposed cross-domain dictionary learning method. Experimental results on human action recognition, image classification and event recognition are comprehensively presented in Sect. 5. Finally, the conclusion of this work is given in Sect. 6.

## 2 Background Work

A considerable number of methods have been proposed to address visual categorization problems (Maji et al. 2013; Ji et al. 2013; Zafeiriou et al. 2012; Liwicki et al. 2012; Xiang et al. 2012; Liu et al. 2012). Reasonable results are achieved using traditional machine learning approaches without considering the data distribution mismatch among the training data and the testing data when training data are abundant. Transfer learning (a.k.a., cross-domain learning, domain transfer, domain adaptation) approaches begin to attract increasing interests in the computer vision community in recent years due to the data explosion on the Internet and the growing demands for visual computational tasks. In Cao et al. (2010), action detection is conducted across datasets from different visual domains, where the KTH dataset (Schuldt 2004), which has a clean background and limited viewpoint and scale changes, is set as the source domain, and the Microsoft Research Action Dataset[3] and the TRECVID surveillance data (Dikmen et al. 2008), which are captured from realistic scenarios, are used as the target domain. Yang et al. (2007) and Duan et al. (2012a) addressed the problem of video concept detection using domain transfer approaches. The former one utilized the Adaptive Support Vector Machine (A-SVM) to adapt one or more existing classifiers of any type to a new dataset, and the latter proposed a Domain Transfer Multiple Kernel Learning (DTMKL) method to simultaneously learn a kernel function and a robust SVM classifier by minimizing both the structural risk function of SVM and the distribution mismatch of labeled and unlabeled data in different domains. Liu et al. (2011) and Li and Zickler (2012) constructed cross-domain representations to cope with the cross-view action recognition problem, where the divergences across domains are caused by view-point changes. Liu et al. (2011) built a bipartite graph via unsupervised co-clustering to measure the visual-word to visual-word relationship across the target view and the source view so that a high-level semantic feature that bridges the semantic gap between the two vocabularies can be filled. Similarly, Li and Zickler (2012) captured the conceptual idea of “virtual views”to represent an action descriptor continuously from an observer’s viewpoint to another. Duan et al. (2012b) considered to leverage large amounts of loosely labeled web videos for visual event recognition using the Adaptive Multiple Kernel Learning (A-MKL) to fuse the information from multiple pyramid levels and features and cope with the considerable variation in feature distributions between videos across two domains.

Recently, dictionary learning for sparse representation has attracted much attention. It has been successfully applied to a variety of computer vision tasks, e.g., face recognition (Wright et al. 2009) and image denoising (Zhou et al. 2009). Using an over-complete dictionary, sparse modeling of signals can approximate the input signal by a sparse linear combination of items from the dictionary. Many algorithms (Lee et al. 2007; Wang et al. 2010; Wright et al. 2009) have been proposed to learn such a dictionary according to different criteria. The K-Singular Value Decomposition (K-SVD) algorithm (Aharon et al. 2006) is a classical dictionary learning algorithm that generalizes the K-means clustering process for adapting dictionaries to efficiently learn an over-complete dictionary from a set of training signals. The K-SVD method focuses on the reconstructive ability, however, since the learning process is unsupervised, the discriminative capability is not taken into consideration. Consequently, methods that incorporate the discriminative criteria into dictionary learning were proposed in Zhang and Li (2010), Yang et al. (2010), Mairal et al. (2008a, 2008b, 2009), Boureau et al. (2010). In addition to the discriminative capability of the learned dictionary, other criteria designed on top of the prototype dictionary learning objective function include multiple dictionary learning (Zhang et al. 2009), category-specific dictionary learning (Yang et al. 2008), etc. Different from most dictionary learning methods, which learned the dictionary and the classifier separately, Zhang and Li (2010) and Jiang et al. (2011) unified these two learning procedures into a single supervised optimization problem and learned a discriminative dictionary and the corresponding classifier simultaneously. Taking a step further, Qiu et al. (2012) and Zheng et al. (2012) designed dictionaries for the situations that the present training instances are different from the testing instances. The former presented a general joint optimization function that transforms a dictionary learned from one domain to the other, and applied such a framework to applications such as pose alignment, pose and illumination estimation and face recognition. The latter achieved promising results on the cross-view action recognition problem with pairwise dictionaries constructed using correspondences between the target view and the source view. To make use of some data that may not be relevant to the target domain data, Raina et al. (2007) proposed a method that applies sparse coding to unlabeled data to break the tremendous amount of data in the source domain into basic patterns (e.g., edges in the task of image classification) so that knowledge can be transferred through the bottom level to a high level representation.

Our approach differs from the above approaches in such aspects that it more comprehensively learns pairwise dictionaries and a classifier while considering the capacity of the dictionaries in terms of reconstructability, discriminability and domain adaptability. Additionally, corresponding observations across the domains are not required in our framework. While most previous knowledge transfer algorithm focus on the situations where the target domain is incomplete, but have not attempted to utilize other domain data as an aide for enhancing present categorization systems, in our approach,

[3] http://research.microsoft.com/~zliu/ActionRecoRsrc

the learned classifier in the target domain becomes more discriminative against intra-class variations as a result of the learning process that integrates with source domain data.

## 3 Dictionary Learning

### 3.1 Reconstruction

Let $y \in \Re^n$ denote an $n$-dimensional input signal, and suppose it can be reconstructed by the linear transformation of an $N$-dimensional projection coefficient $x \in \Re^N$ via a projection dictionary $D \in \Re^{n \times N}$. Considering the reconstruction error, the transformation can be formulated as:

$$y = Dx + E(x). \tag{1}$$

where we use $E(x)$ to represent the reconstruction error, then the optimal dictionary and coefficient can be obtained by minimizing $E(x)$. We quantitatively measure $E(x)$ using:

$$E(x) = \|y - Dx\|_2^2. \tag{2}$$

It is worth to point out that if the dimension of the projection coefficient $x$ is larger than the dimension of input signal $y$, i.e., $N > n$, the solution to the unconstrained optimization problem in Eq. (2) is not unique, thus it leads to the overfitting problem.

### 3.2 Sparsity Constraints

The sparsity constraints for dictionary learning attract more attention recently, and applications that can benefit from sparsity include compression, regularization in inverse problems, etc. The commonly used sparsity constraints are $l_0$-norm and $l_1$-norm.

#### *3.2.1 Dictionary Learning with $l_0$-Norm*

$l_0$-norm is the lowest normalization form, and it indicates the solution with fewest non-zero entries. When learning a dictionary with the $l_0$-norm sparse constraint, Eq. (2) can be formulated as:

$$E(x) = \|y - Dx\|_2^2, \quad s.t. \|x\|_0 \leq T, \tag{3}$$

where $T$ is the sparsity constraint factor that limits the number of non-zero elements in the sparse codes, so that the number of items in the decomposition of each $x$ is less than $T$. Updating both $x$ and $D$ simultaneously is generally NP-hard; however, we can manage to seek an improved $D$ when fixing $x$, or seek an optimal $x$ when fixing $D$. Thus, the construction of dictionary $D$ is achieved through iteratively minimizing the reconstruction error and learning a reconstructive dictionary for sparse representations (Aharon et al. 2006). Given $D$, the computation of the sparse code $x$ is generally NP-hard under the sparsity constraint, thus one has to seek alternative methods to approximate the solution, e.g., the greedy algorithms Matching Pursuit (MP) (Mallat and Zhang 1993) and Orthogonal Matching Pursuit (OMP) (Pati et al. 1993), which sequentially select the dictionary atoms. More details on optimizing the objective function under the $l_0$-norm constraint are given in Sect. 3.3.

#### *3.2.2 Dictionary Learning with $l_1$-Norm*

The Basis Pursuit (BP) (Chen et al. 1993) suggests an alternative sparse solution by relaxing the $l_0$-norm with the higher order $l_1$-norm. The dictionary learning problem in Eq. (3) can be reformulated as follows with the $l_1$-norm constraint:

$$E(x) = \|y - Dx\|_2^2, \quad s.t. \|x\|_1 \leq T. \tag{4}$$

Again, such a problem can be solved iteratively by alternatingly optimizing $D$ or the sparse code $x$ while fixing the other. When the dictionary $D$ is fixed, the optimization problem is equivalent to a linear regression problem with $l_1$-norm regularization on the coefficients, which can be solved by the feature-sign search algorithm (Lee et al. 2006). When the sparse code $x$ is fixed, the problem is reduced to a Least square problem with quadratic constraints, so that it can be solved by the Lagrange dual as in Lee et al. (2006).

### 3.3 Classification via Dictionary Learning

A classifier $f(x)$ can be directly employed to the sparse representation $x$ for classification, and the classifier can be obtained by satisfying:

$$W = arg \min_W \mathcal{L}\{h, f(x, W)\} + \lambda \|W\|_F^2, \tag{5}$$

where $\mathcal{L}$ is the classification loss function, e.g., quadratic loss function and hinge loss function, $h$ indicates the label of $x$, $W$ denotes the classifier parameters and $\lambda$ is a regularization parameter for preventing overfitting. However, separating the dictionary learning stage from the classification procedure might lead to a suboptimal $D$. Previous approaches (Zhang and Li 2010; Yang et al. 2010; Mairal et al. 2008a,b, 2009; Jiang et al. 2011) attempt to jointly learn a dictionary and a classifier for classification tasks. In this case, the dictionary learning problem can be formulated as:

$$\begin{aligned} < D, W, x > = arg \min_{D,W,x} \|y - Dx\|_2^2 + \mathcal{L}\{h, f(x, W)\} \\ + \lambda \|W\|_2^2, \quad s.t. \|x\|_0 \leq T. \end{aligned} \tag{6}$$

An extra classification term can encourage the data to be smooth. However, if we deal with data from two domains, the classification term can only guarantee the local smoothness in each respective domain. Thus, we introduce a new term to seek the global smoothness across both domains.

## 4 Domain Adaptation via Dictionary Learning

We denote $Y_t$ as $L$ $n$-dimensional target domain instances, and $Y_s$ as $M$ source domain $n$-dimensional instances, i.e., $Y_t = [y_t^1, \ldots, y_t^L] \in \Re^{n \times L}$ and $Y_s = [y_s^1, \ldots, y_s^M] \in \Re^{n \times M}$. Learning a reconstructive dictionary pair while pursuing the global smoothness can be accomplished by solving the following optimization problems:

$$\begin{aligned} < D_t, D_s, X_t, X_s > = arg \min_{D_t, D_s, X_t, X_s} & \|Y_t - D_t X_t\|_2^2 \\ & + \|Y_s - D_s X_s\|_2^2 + \Phi([X_t\ X_s]) \\ & s.t. \forall i, [\ \|x_t^i\|_0, \|x_s^i\|_0\ ] \leq T, \end{aligned} \tag{7}$$

where $\Phi(\cdot)$ is designed to measure the distances of similar cross-domain instances of the same category, $D_t = [d_t^1, \ldots, d_t^N] \in \Re^{n \times N}$ is the learned target domain dictionary, $X_t = [x_t^1, \ldots, x_t^L] \in \Re^{N \times L}$ is the set of target domain sparse codes, $D_s = [d_s^1, \ldots, d_s^N] \in \Re^{n \times N}$ is the learned source domain dictionary and $X_s = [x_s^1, \ldots, x_s^M] \in \Re^{N \times M}$ is the set of source domain sparse codes, respectively. The number of dictionary items $N$ is set to be larger than either $L$ or $M$ to ensure that the dictionaries are over-complete. To define $\Phi(\cdot)$, we aim to force the sparse codes that possess the same class label to be close to each other, and thus geometrically simple decision boundaries are preferred. To this end, Zheng et al. (2012) presented a strategy that manually sets up a set of correspondence training instances for cross-view action recognition, where the same action pair performed in different views are encouraged to share the same representation when being projected onto the cross-view dictionary pair. Inspired by such a strategy, we measure the cross-domain divergence by constructing virtual correspondences across both domains through a transformation matrix $\mathbb{A}$. Given $\Phi([X_t\ X_s]) = \|X_t^T - \mathbb{A} X_s^T\|_2^2$, Eq. (7) can be written as:

$$\begin{aligned} < D_t, D_s, X_t, X_s > = arg \min_{D_t, D_s, X_t, X_s} & \|Y_t - D_t X_t\|_2^2 \\ & + \|Y_s - D_s X_s\|_2^2 + \|X_t^T - \mathbb{A} X_s^T\|_2^2 \\ & s.t. \forall i, [\ \|x_t^i\|_0, \|x_s^i\|_0\ ] \leq T. \end{aligned} \tag{8}$$

However, in our case, rather than cross-view action pairs, the data we are dealing with come from different datasets, so that setting up correspondence instances is not possible. We turn to seek an alternative solution to building up such correspondences. For each category, we introduce a transformation matrix $\mathbb{A}_c$. The general sense of $\mathbb{A}_c$ is that it maps the most similar source domain instance to a target domain instance of the same category. We adopt a fuzzy category-specific searching method to compute each $\mathbb{A}_c$. Considering that $Y_t^c$ and $Y_s^c$ are the $c$-th category data from both domains, we first compute the Gaussian distances between each pair of data between $Y_t^c$ and $Y_s^c$, and store the result in a matrix $\mathbb{G}_c$. Then $\mathbb{A}_c$ can be computed by preserving the maximum element in each column of $\mathbb{G}_c$ while discarding the remain elements, i.e., we only ensure a one-to-one correspondence for each source domain instance:

$$\mathbb{A}_c(i, j) = \begin{cases} 1, & if \quad \mathbb{G}_c(i, j) = max(\mathbb{G}_c(:, j)) \\ 0, & otherwise. \end{cases} \tag{9}$$

Once the set of transformation matrices for all the $C$ categories are computed, the global transformation matrix $\mathbb{A} \in \Re^{L \times M}$ can be obtained by filling all the category-specific sub-matrices into $\mathbb{A}$:

$$\mathbb{A} = \begin{pmatrix} \mathbb{A}_1 & & & \\ & \mathbb{A}_2 & & \\ & & \ddots & \\ & & & \mathbb{A}_C \end{pmatrix}, \tag{10}$$

where all the blank elements are set to 0, so that $\mathbb{A}$ is a binary matrix. Since $\mathbb{A}$ is computed in a category-specific manner, target domain training samples can only be connected to those source domain samples of the same category. Thus, overall smoothness across both domains can be guaranteed after such a transformation. Assuming $\mathbb{A}$ leads to a perfect mapping across the sparse codes $X_t$ and $X_s$ and each matched pair of samples in different domains possesses an identical representation after encoding, then $\|X_t^T - \mathbb{A} X_s^T\|_2^2 = 0$. Since these two terms are computed with $l_2$ normalization, if they equal to zero, we can obtain $X_t^T = \mathbb{A} X_s^T$, i.e., $X_t = X_s \mathbb{A}^T$. By transforming the source domain data to match the target domain data, we formulate the new objective function as:

$$\begin{aligned} & < D_t, D_s, X_t, X_s > \\ & = arg \min_{D_t, D_s, X_t, X_s} \|Y_t - D_t X_t\|_2^2 + \|(Y_s - D_s X_s)\mathbb{A}^T\|_2^2 \\ & = arg \min_{D_t, D_s, X_t} \|Y_t - D_t X_t\|_2^2 + \|Y_s \mathbb{A}^T - D_s X_s \mathbb{A}^T\|_2^2 \\ & = arg \min_{D_t, D_s, X_t} \|Y_t - D_t X_t\|_2^2 + \|Y_s \mathbb{A}^T - D_s X_t\|_2^2 \\ & s.t. \forall i, \ \|x_t^i\|_0 \leq T. \end{aligned} \tag{11}$$

Following Zhang and Li (2010), Mairal et al. (2008a; 2008b, 2009), and Jiang et al. (2011), we include a label consistency regularization term and the classification error of a linear predictive classifier $f(x)$ into the objective function to further enhance the global smoothness. Thus, the new objective function for cross-domain dictionary learning is updated as:

$$\begin{aligned} & < D_t, D_s, X_t, A, W >= arg \min_{D_t, D_s, X_t, A, W} \|Y_t - D_t X_t\|_2^2 \\ & + \|Y_s \mathbb{A}^T - D_s X_t\|_2^2 + \alpha \|Q - \vartheta X_t\|^2 + \beta \|H - W X_t\|_2^2 \\ & s.t. \forall i, \ \|x_t^i\|_0 \leq T, \end{aligned} \tag{12}$$

where $W$ are the coefficients of the linear classifier $f(x)$, $H$ are the class labels of target domain data, $\vartheta$ is a linear transformation matrix that maps the the original sparse codes to be in correspondence with the target discriminative sparse codes

$Q = [q_1, q_2, \ldots, q_L] \in \Re^{L\times L}$ of the input signal $Y_t$. Specifically, $q_i = [q_i^1, q_i^2, \ldots, q_i^K]^T = [0, \ldots, 1, 1, \ldots, 0]^T \in \Re^{L\times 1}$, and the non-zeros occur at those indices where $y_t^i \in Y_t$ and $X_t^k \in X_t$ share the same class label. Given $X_t = [x_1, x_2, \ldots, x_6]$ and $Y_t = [y_1, y_2, \ldots, y_6]$, and assuming $x_1$, $x_2$, $y_1$ and $y_2$ are from class 1, $x_3$, $x_4$, $y_3$ and $y_4$ are from class 2, $x_5$, $x_6$, $y_5$ and $y_6$ are from class 3, $Q$ is then defined with the following form:

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}, \tag{13}$$

and $H = [h_1, h_2, \ldots, h_L] \in \Re^{C\times L}$ are the class labels of $Y_t$, where the non-zero element indicates the class of an input signal within each column $h_i = [0, \ldots, 1, \ldots, 0]^T \in \Re^{C\times 1}$. Following the same example in (13), $H$ can be defined as:

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}. \tag{14}$$

Scalers $\alpha$ and $\beta$ are set to control the relative contribution of the terms $\|Q - \vartheta X_t\|^2$ and $\|H - WX_t\|_2^2$. By solving the optimization problem in Eq. (12), the reconstructive, discriminative and domain-adaptive dictionary pair $D_t$ and $D_s$ as well as the optimal classifier parameter $W$ can be obtained.

## 4.1 Optimization

### *4.1.1 Solving WSCDD with the K-SVD algorithm*

We rewrite Eq. (12) as:

$$< D_t, D_s, X_t, W >= arg \min_{D_t, D_s, X_t, W} \left\| \begin{pmatrix} Y_t \\ Y_s \mathbb{A}^T \\ \sqrt{\alpha} Q \\ \sqrt{\beta} H \end{pmatrix} - \begin{pmatrix} D_t \\ D_s \\ \sqrt{\alpha}\vartheta \\ \sqrt{\beta} W \end{pmatrix} X_t \right\|_2^2, \; s.t. \forall i, \; \|x_t^i\|_0 \le T, \tag{15}$$

To make it clear, we write the left side of Eq. (15) as $Y = (Y_t^T, (Y_s\mathbb{A}^T)^T, \sqrt{\alpha}Q^T, \sqrt{\beta}H^T)^T$ and the right side of Eq. (15) as $D = (D_t^T, D_s^T, \sqrt(\alpha)\vartheta^T, \sqrt(\beta)W^T)^T$, where column-wise $l_2$ normalization is applied to $D$, so that optimizing Eq. (15) is cast as optimizing Eq. (16):

$$< D, X_t >= arg \min_{D, X_t} \|Y - DX_t\|_2^2, \quad s.t. \forall i, \; \|x_t^i\|_0 \le T. \tag{16}$$

Such an optimization problem can be solved using the K-SVD (Aharon et al. 2006) algorithm. Specifically, Eq. (16) can be solved in an iterative manner through both dictionary updating stage and sparse coding stage. In the dictionary updating stage, each dictionary element is updated sequentially to better represent the original data in both the source domain and the target domain as well as the discriminative property along with the training data. When pursuing a better dictionary $D$, the sparse codes $X_t$ are frozen, and each dictionary element is updated through a straightforward solution which tracks down a rank-one approximation to the matrix of residuals. Following K-SVD, the $k$th element of the dictionary $D$ and its corresponding coefficients, i.e. the $k$th row in the coefficient matrix $X_t$, are denoted as $d_k$ and $x_k$ respectively. Let $S_k = Y - \sum_{j\neq k} d_j x_t^j$ and we further denote $\widetilde{x}_k$ and $\widetilde{S}_k$ as the results we obtain when all zero entries in $x_k$ and $S_k$ are discarded, respectively. Thus, each dictionary element $d_k$ and its correspondingly non-zero coefficients $\widetilde{x}_k$ can be computed by

$$< d_k, \widetilde{x}_k >= arg \min_{d_k, \widetilde{x}_k} \|\widetilde{S}_k - d_k\widetilde{x}_k\|_F^2. \tag{17}$$

The approximation in Eq. (17) is achieved through performing Singular Value Decomposition (SVD) on $\widetilde{S}_k$:

$$\begin{aligned} SVD(\widetilde{S}_k) &= U\Sigma V^T \\ d_k &= U(:, 1) \\ \widetilde{x}_k &= \Sigma(1, 1)V(1, :), \end{aligned} \tag{18}$$

where $U(:, 1)$ indicates the first column of $U$ and $V(1, :)$ indicates the first row of $V$.

At the sparse coding stage, we compute the "best matching" projections $X_t$ of the multidimensional training data onto the updated dictionary $D$ using an appropriate pursuit algorithm. As introduced above, given the fixed $D$, the optimization of Eq. (16) remains NP-hard under the $l_0$-norm constraint. Therefore the OMP algorithm is adopted to approximate the solution in a computationally efficient way. The proposed cross-domain dictionary learning method is summarized in Algorithm 1.

### *4.1.2 Initialization*

To initialize $D_t$ and $D_s$, we run the K-SVD algorithm several times on both of them within each category, and then combine all K-SVD outputs in each respective domain. To initialize $\vartheta$ and $W$, we employ the multivariate ridge regression model (Golub et al. 1999) with $l_2$-norm regularization as follows:

$$\begin{aligned} \vartheta &= arg \min_{\vartheta} \|Q - \vartheta X_t\|^2 + \varphi_1\|\vartheta\|_2^2, \\ W &= arg \min_{W} \|H - WX_t\|^2 + \varphi_2\|W\|_2^2, \end{aligned} \tag{19}$$

which yields the following solutions:

$$\begin{aligned} \vartheta &= QX_t^T(X_tX_t^T + \varphi_1 I)^{-1}, \\ W &= HX_t^T(X_tX_t^T + \varphi_2 I)^{-1}, \end{aligned} \tag{20}$$

where $X_t$ can be computed given the initialized $D_t$.

```
Input : Input signals $Y_t$ and $Y_s$, discriminative sparse code $Q$,
        target domain class label $H$, sparsity constraint
        parameter $T$, balancing parameters $\alpha$ and $\beta$, dictionary
        size $N$ and maximum iteration $Max.iter$.
Output: Cross-domain dictionary pair $\tilde{D}_t$ and $\tilde{D}_s$, transformation
        matrixes $\mathbb{A}$ and $\tilde{\vartheta}$, and linear classifier parameter $\tilde{W}$.
Compute $\mathbb{A}$ by combining each transformation matrix $\mathbb{A}_c$ for all
$C$ classes;
Initialize $D_t$, $D_s$, $\vartheta$ and $W$;
Reformulate $Y = \begin{pmatrix} Y_t \\ Y_s\mathbb{A}^T \\ \sqrt{\alpha}Q \\ \sqrt{\beta}H \end{pmatrix}$ and $D = \begin{pmatrix} D_t \\ D_s \\ \sqrt{\alpha}\vartheta \\ \sqrt{\beta}W \end{pmatrix}$;
$D^0 \leftarrow D$;
for $i \leftarrow 1$ to $Max.iter$ do
    Sparse coding stage:
    Compute $X_t$ using OMP according to:
        $E(x) = \|Y - D^{(i-1)}X_t\|_2^2, s.t.\forall i, \|x_t^i\|_0 \le T,$
    Dictionary updating stage:
    for $k \leftarrow 1$ to $N$ do
        Compute $S_k = Y - \sum_{j \neq k} d_j x_t^j$;
        Discard all zero entries in $x_k$ and $S_k$, and obtain $\widetilde{x}_k$ and $\widetilde{S}_k$;
        Apply a Singular Value Decomposition (SVD) operation
        on $\widetilde{S}_k$:
        $SVD(\widetilde{S}_k) = U\Sigma V^T$,
        $d_k \leftarrow U(:,1), \widetilde{x}_k \leftarrow \Sigma(1,1)V(1,:)$
    end
    $D^i \leftarrow D^{updated}$
end
Decompose $D$ to obtain $D_s$, $D_t$, $\vartheta$ and $W$;
Compute $\tilde{D}_t$, $\tilde{D}_s$, $\tilde{\vartheta}$ and $\tilde{W}$ according to Equation (21).
```

**Algorithm 1**: Weakly-Supervised Cross-Domain Dictionary Learning.

#### *4.1.3 Convergence Analysis*

The convergence proof of the proposed WSCDD method can be given similarly as the K-SVD algorithm (Aharon et al. 2006). At the dictionary updating stage, each dictionary element and its corresponding coefficients are updated by minimizing quadratic functions, and the remaining dictionary elements are updated upon the previous updates. Consequently, the MSE of the overall reconstruction error is monotonically decreasing with respect to the dictionary updating iterations. At the sparse coding stage, computation of the "best matched" coefficients under the $l_0$-norm constraint also leads to a reduction in MSE conditioned on the success of the OMP algorithm. Finally, since MSE is non-negative, the optimization procedure is monotonically decreasing and bounded by zero from below, thus the convergence of the proposed dictionary learning method is guaranteed. The typical strategy to avoid the optimization procedure getting stuck in a local minimum is to initialize the dictionary with a few different random matrices in several runs. Such a strategy is applied in our approach.

### 4.2 Classification

Since $D_t$, $D_s$, $\vartheta$ and $W$ are jointly normalized in the optimization procedure, they cannot be directly applied to construct the classification framework. Also, since $W$ is obtained with the un-normalized $D$, simply re-normalizing $D$ is not applicable. According to the lemma in Zhang and Li (2010), $\tilde{D}_t$, $\tilde{D}_s$, $\tilde{\vartheta}$ and $\tilde{W}$ can be computed as:

$$\begin{aligned} \tilde{D}_t &= \left\{ \frac{d_t^1}{\|d_t^1\|_2}, \frac{d_t^2}{\|d_t^2\|_2}, \ldots, \frac{d_t^K}{\|d_t^K\|_2} \right\} \\ \tilde{D}_s &= \left\{ \frac{d_s^1}{\|d_s^1\|_2}, \frac{d_s^2}{\|d_s^2\|_2}, \ldots, \frac{d_s^K}{\|d_s^K\|_2} \right\} \\ \tilde{\vartheta} &= \left\{ \frac{\vartheta^1}{\|\vartheta^1\|_2}, \frac{\vartheta^2}{\|\vartheta^2\|_2}, \ldots, \frac{\vartheta^K}{\|\vartheta^K\|_2} \right\} \\ \tilde{W} &= \left\{ \frac{w^1}{\|w^1\|_2}, \frac{w^2}{\|w^2\|_2}, \ldots, \frac{w^K}{\|w^K\|_2} \right\} \end{aligned} \tag{21}$$

Given a target domain query sample $y_t^i$, its sparse representation $x_t^i$ can be computed through $(\tilde{D})_t$. With the linear classifier $f(x)$, the label $l$ of $y_t^i$ can be predicted as:

$$l = arg \min_j (l_j = \tilde{W}x_t^i). \tag{22}$$

## 5 Experiments

### 5.1 Experimental Data Preparation

To demonstrate the effectiveness of the proposed method, we evaluate it on action recognition, image classification and event recognition tasks. For event recognition and action recognition, the source domain data are obtained from an existing dataset or selected categories of an existing dataset. For image classification, the source domains are constructed by choosing 20 image categories (chosen according to the ascending alphabetic order) and use the first 100 results returned from Google Image Search for each chosen category as the source domain data, where the indexing procedure is performed by simply searching the category names. Since the retrieved images are very noisy, we apply the manifold ranking (Zhou et al. 2004a,b) algorithm as a pre-processing stage for the source domain data. As the source domain data are weakly labeled, we allow 5 samples per category as labeled in the source domain. The average ranking scores of the unlabeled source domain data are obtained by treating both the target domain data and the labeled source domain data as queries, and rank the unlabeled source domain data. We keep

**Fig. 4** Example images from video sequences in the UCF YouTube dataset

the first 20−30 % instances from the ranked source domain data for each image category, and filter out the remaining retrieved data. The same ranking procedure is applied to the action recognition and the event recognition task, where we keep the most highly ranked 30 instances from the source domain dataset for the former, and the most highly ranked 80 % instances from the source domain dataset for the latter. We denote both scenarios of the proposed WSCDDL method when manifold ranking is utilized or not as WSCDDL-MR and WSCDDL-EU respectively in image classification, action recognition and event recognition experiments.

### 5.2 Action Recognition

The UCF YouTube dataset and the HMDB51 dataset are used for the action recognition task, where the UCF YouTube dataset is used as the target domain and the HMDB51 dataset is used as the source domain. The UCF YouTube dataset (shown in Fig. 4) is a realistic dataset that contains camera shaking, cluttered background, variations in actors' scale, variations in illumination and view point changes. There are 11 actions including cycling, diving, golf swinging, soccer juggling, jumping, horse-back riding, basketball shooting, volleyball spiking, swinging, tennis swinging and walking with a dog, and these actions are performed by 25 actors. The HMDB51 dataset (shown in Fig. 5) contains video sequences which are extracted from commercial movies as well as YouTube, and it represents a fine multifariousness of light conditions, situations and surroundings in which actions can appear, different recording camera types and viewpoint changes. Since the HMDB51 dataset is a more challenging dataset, our case closely resembles real-world scenarios, where the source domain data can contain a wide range of noise levels. In correspondence with the target domain action categories, we choose 7 body movements from the HMDB51 dataset, including ride bike, dive, golf, jump, kick ball, ride horse and shoot ball.

We adopt the dense trajectories (Wang et al. 2011) as the low-level action video representation to distinguish the motion of interest. To leverage the motion information in the dense trajectories, a set of local descriptors are computed within space-time volumes around the trajectories at multiple spatial and temporal scales, and these features include the HOGHOF (Laptev et al. 2008), the optical flow (Ikizler-Cinbis and Sclaroff 2010) and the Motion Boundary Histogram (MBH) (Dalal et al. 2006). Specifically, the HOGHOF feature is a combination of appearance information (captured by HOG Dalal and Triggs 2005) and local motion probabilities (captured by Histogram of Optical Flow (HOF)). Since motion is the most important cue for analyzing actions, the optical flow works effectively by computing the relative motion between the observer and the scene. MBH represents the gradient of the optical flow by separately computing the derivatives for the horizontal and vertical components of the optical flow, so that relative motion between pixels is encoded. Changes in the optical flow field being preserved and constant motion information being suppressed, the MBH descriptor can effectively eliminate noise caused by background motion compared with video stabilization (Ikizler-Cinbis and Sclaroff 2010) and motion compensation (Uemura et al. 2008) approaches (Wang et al. 2011). Despite its powerful capability of describing action motions,

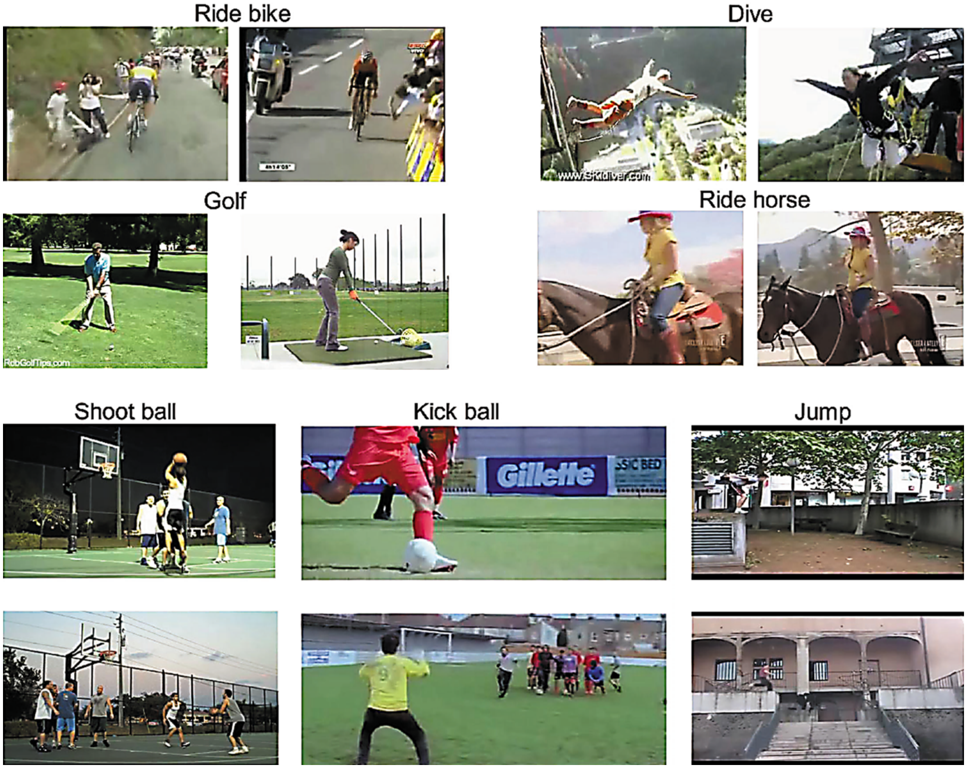

**Fig. 5** Example images from video sequences in the selected body movements of the HMDB51 dataset

the dense trajectories come with two weaknesses: (1) trajectories tend to drift from their initial locations during motion tracking, which is a common problem in tracking; (2) the large quantity of local trajectory descriptors leads to high computational complexity and memory consumption for the coding methods, such as VQ and SC. To cope with the first issue, the length of a trajectory is limited to a pre-defined number of frames. Taking the second issue into account, a Locality-constrained Linear Coding (LLC) (Wang et al. 2010) scheme is adopted instead of VQ and SC. LLC represents the low-level dense trajectories by multiple bases. In addition to achieving less quantization error, the explicit locality adaptor in LLC guarantees the local smooth sparsity.

Dense trajectories are extracted from raw action video sequences with 8 spatial scales spaced by a factor of $1/\sqrt{2}$, and feature points are sampled on a grid spaced by 5 pixels and tracked in each scale, separately. Each point at frame $t$ is tracked to the next frame $t+1$ by median filtering in a dense optical flow field. To avoid the drifting problem, the length of a trajectory is limited to 15 frames. HOGHOF and MBH are computed within a $32 \times 32 \times 15$ volume along the dense trajectories, where each volume is sub-divided into a spatio-temporal grid of size $2 \times 2 \times 3$ to impose more structural information in the representation. Considering both efficiency and the construction error, LLC coding scheme is applied to the low-level local dense trajectories features with 30 local bases, and the codebook size is set to be 4,000 for all training-testing partitions. To reduce the complexity, only 200 local dense trajectories features are randomly selected from each video sequence when constructing the codebook. We run our method on five different partitions of the UCF YouTube dataset, where we randomly choose all action categories performed by the number of 5/9/16/20/24 actors as the training actions while using the remaining actions as the testing actions for each partition. 30 most relevant actions are chosen from each of the 7 source domain categories using manifold ranking, and they are represented in the same manner as the target domain actions and coded with the same codebook. The weight $\alpha$ on the label constraint term and the weight $\beta$ on the classification error term are set as 4 and 2 respectively, and 50 iterations of SVD decomposition are executed during optimization (We use the same values of $\alpha$, $\beta$ and K-SVD maximum iteration for the image classification and event recognition tasks). To avoid over-fitting, the dictionary size is set to be larger when more training data are available at the training stage. The results are demonstrated in Table 1 for all five partitions, where we use the size of 200, 300, 500, 700 and 900 for each partition. We compare the performance of the baseline LLC, sparse coding methods K-SVD (Aharon et al. 2006) and LC-SVD (Jiang et al. 2011), and transfer learning methods FR (Daumé 2007) and A-SVM (Yang et al. 2007) with the proposed WSCDDL method. Results are reported on both scenarios where the source domain data are included or excluded in Tables 1 and 2 respectively. Comparing Tables 1 and 2, we can discover that for many cases, brute-forcing the knowledge from the

**Table 1** Performance comparison between the WSCDDL and other methods on the UCF YouTube dataset when the source domain data are only used by the WSCDDL

| Algorithm | LLC (Wang et al. 2010) | K-SVD (Aharon et al. 2006) | LC-KSVD (Jiang et al. 2011) | WSCDDL-EU | WSCDDL-MR |
|---|---|---|---|---|---|
| Dictionary learning | N/A | Unsupervised | Supervised | Supervised | Supervised |
| Source data | No | No | No | Yes | Yes |
| 24 actors (%) | 86.67 | 82.22 | 86.67 | 88.89 | **91.11** |
| 20 actors (%) | 75.42 | 68.75 | 75.42 | 77.50 | **78.30** |
| 16 actors (%) | 70.88 | 63.96 | 72.08 | 73.03 | **73.03** |
| 09 actors (%) | 61.41 | 55.70 | 65.25 | 66.31 | **66.05** |
| 05 actors (%) | 54.10 | 50.05 | 56.55 | 56.66 | **57.19** |

Best results are in bold

**Table 2** Recognition results on the UCF YouTube dataset when using the HMDB dataset as the source domain

| Algorithm | LLC (Wang et al. 2010) | K-SVD (Aharon et al. 2006) | LC-KSVD (Jiang et al. 2011) | FR (Daumé 2007) | A-SVM (Yang et al. 2007) | WSCDDL-EU | WSCDDL-MR |
|---|---|---|---|---|---|---|---|
| Dictionary learning | N/A | Unsupervised | Supervised | Supervised | Supervised | Supervised | Supervised |
| Source data | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| 24 actors (%) | 86.67 | 77.78 | 82.22 | 83.74 | 82.51 | 88.89 | **91.11** |
| 20 actors (%) | 70.21 | 72.08 | 75.42 | 74.88 | 79.05 | 77.50 | **78.30** |
| 16 actors (%) | 70.17 | 67.54 | 72.08 | 71.56 | 72.46 | 73.03 | **73.03** |
| 09 actors (%) | 61.80 | 59.15 | 64.72 | 62.77 | 61.65 | 66.31 | **66.05** |
| 05 actors (%) | 53.35 | 48.88 | 54.10 | 54.09 | 51.54 | 56.66 | **57.19** |

Best results are in bold

**Table 3** Performance comparison of the WSCDDL with state-of-the-art methods under the leave-one-actor-out setting on the UCF YouTube dataset

| Methods | Liu et al. (2009) | Ikizler-Cinbis and Sclaroff (2010) | BoF | WSCDDL-EU | WSCDDL-MR |
|---|---|---|---|---|---|
| Results (%) | 71.2 | 75.21 | 80.02 | 81.13 | **82.32** |

Best result is in bold

source domain into the target domain irrespective of their divergence can cause certain performance degeneration. On the other hand, the proposed WSCDDL method consistently leads to the best performance over all the partitions. Figure 7 shows the convergence analysis and performance of varying dictionary size of the WSCDDL-MR method. Figure 10 shows the confusion matrix comparisons between the LLC method and the WSCDDL-MR method for all five partitions. In order to compare the WSCDDL method with state-of-the-art methods, we further demonstrate its performance under the leave-one-actor-out setting in Table 3.

### 5.3 Image Classification

We utilize the Caltech101 dataset as the target domain and some collected Web images as the source domain for the image classification task. The Caltech101 image dataset (shown in Fig. 6) consists of 101 categories (e.g., accordion, cannon, and chair), and each category contains 30–800 images. The source domain data of the Caltech101 dataset are constructed by a set of images returned by Google Image Search (shown in Fig. 6) (Fig. 7).

For image representations, we choose the dense SIFT (Lowe et al. 2004) plus LLC (Wang et al. 2010) model. The SIFT descriptors are extracted from 16 × 16 pixel patches and densely sampled from each image on a grid with the step size of 8 pixels. We evaluate our method with both dictionary sizes 1024 and 4096. The same values of the weights $\alpha$, $\beta$ and K-SVD iterations are adopted as in the action recognition task. We compare the performance of the proposed WSCDDL approach and state-of-the-art approaches in Table 4. Results on six different numbers of training data are reported, and all the results are averaged over 5 times of different randomly selected training and testing images to guarantee the reliability. For the LLC (Wang et al. 2010), K-SVD (Aharon et al. 2006) and LC-SVD (Jiang et al. 2011) methods, we consider both scenarios of whether the source domain data are included. For fair comparisons, we choose both dictionary size 1,024 and 4,096 to test the proposed method. Figure 6 shows samples of 6 categories with high classification accuracies when using 30 training images per category. As shown in Fig. 8, the proposed WSCDDL method results in larger improvements over others when fewer samples are used for training, which demonstrates its effective-

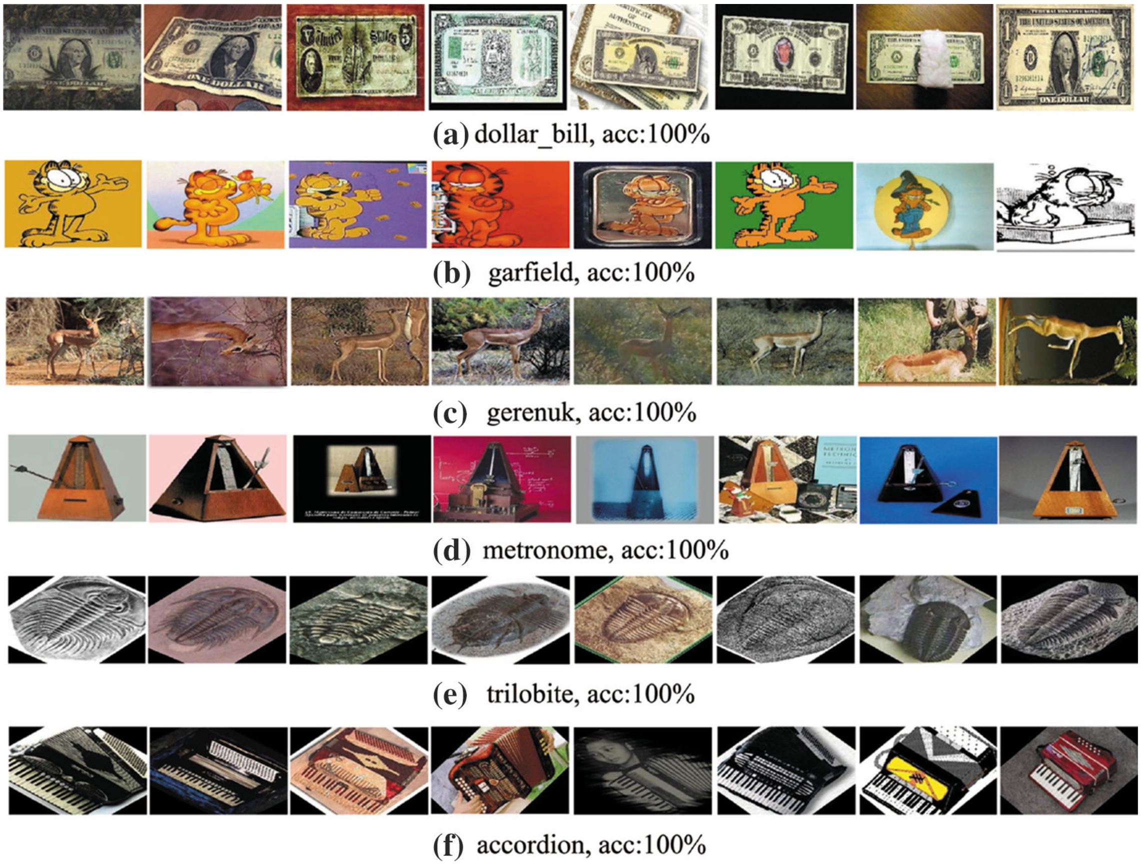

**Fig. 6** Example images from classes with high classification accuracy from the Caltech101 dataset

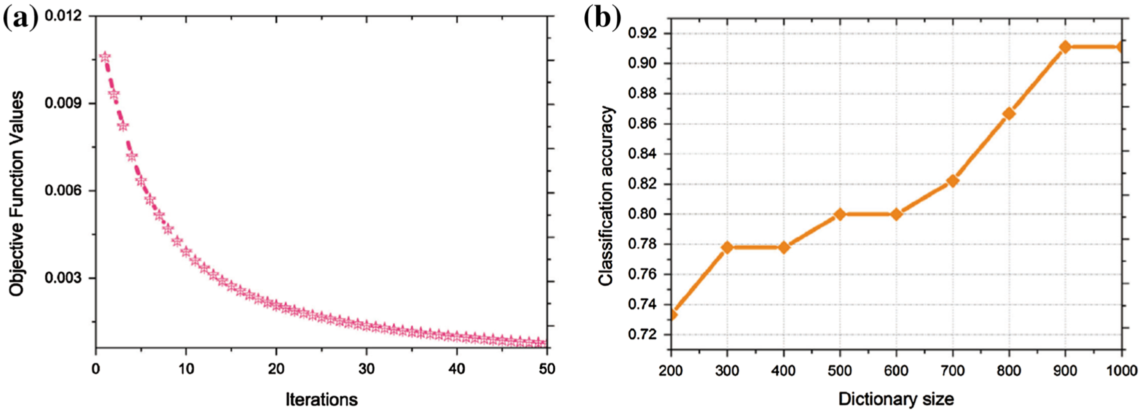

**Fig. 7** Performance analysis on the UCF YouTube dataset when actions performed by 24 actors are used in the training data. **a** The optimization process of the objective function for WSCDDL-MR with 50 iterations. **b** Performance when varying the dictionary size

ness in terms of utilizing the source domain data. Figure 9 demonstrates the performance of all the 101 image categories (Fig. 10).

We further evaluate our method on the more challenging Caltech 256 dataset (Griffin et al. 2007), which contains 30,607 images of 256 categories. Compared to the Caltech101 dataset, it is much more difficult due to the large variations on object location, pose, and size. Similar as the strategy adopted in constructing the source domain for the Caltech101 dataset, 400 images from 20 categories indexed by Google Images are used as the source domain. We evaluate our approach on both 15 and 30 training images per class, and set the dictionary size to 1,024 or 4,096 respectively. We compare our method with state-of-the-art approaches as

**Table 4** Comparison with the state-of-the-art methods on the Caltech101 dataset

| Number of training samples | 5 (%) | 10 (%) | 15 (%) | 20 (%) | 25 (%) | 30 (%) |
|---|---|---|---|---|---|---|
| Malik (Zhang et al. 2006) | 46.6 | 55.8 | 59.1 | 62.0 | – | 66.2 |
| Griffin (Griffin et al. 2007) | 44.2 | 54.5 | 59.0 | 63.3 | 65.8 | 67.6 |
| SRC (Wright et al. 2009) | 48.8 | 60.1 | 64.9 | 67.7 | 69.2 | 70.7 |
| LLC (Wang et al. 2010) | 51.15 | 59.77 | 65.43 | 67.74 | 70.16 | 73.44 |
| LLC (Wang et al. 2010) (source) | 21.43 | 36.37 | 51.11 | 60.02 | 67.55 | 72.17 |
| K-SVD (Aharon et al. 2006) | 39.63 | 50.30 | 58.82 | 64.73 | 67.92 | 71.04 |
| K-SVD (Aharon et al. 2006) (source) | 20.42 | 35.64 | 44.93 | 53.69 | 60.07 | 66.07 |
| LC-KSVD (Jiang et al. 2011) | 46.25 | 57.73 | 68.45 | 70.79 | 72.83 | 73.75 |
| LC-KSVD (Jiang et al. 2011) (source) | 48.95 | 62.71 | 67.14 | 70.17 | 73.39 | 75.05 |
| WSCDDL-EU ($N = 1{,}024$) | **60.62** | **67.81** | **70.09** | **72.98** | **76.17** | **77.30** |
| WSCDDL-MR ($N = 1{,}024$) | **61.31** | **68.69** | **71.59** | **74.73** | **76.82** | **78.44** |
| CRBM (Sohn et al 2011) ($N = 4{,}096$) | 56.7 | 66.7 | 71.3 | 74.2 | 76.2 | 77.8 |
| WSCDDL-EU ($N = 4{,}096$) | **63.47** | **68.90** | **70.88** | **74.01** | **77.54** | **78.68** |
| WSCDDL-MR ($N = 4{,}096$) | **64.05** | **69.31** | **72.39** | **75.22** | **78.40** | **79.02** |

Best results are in bold

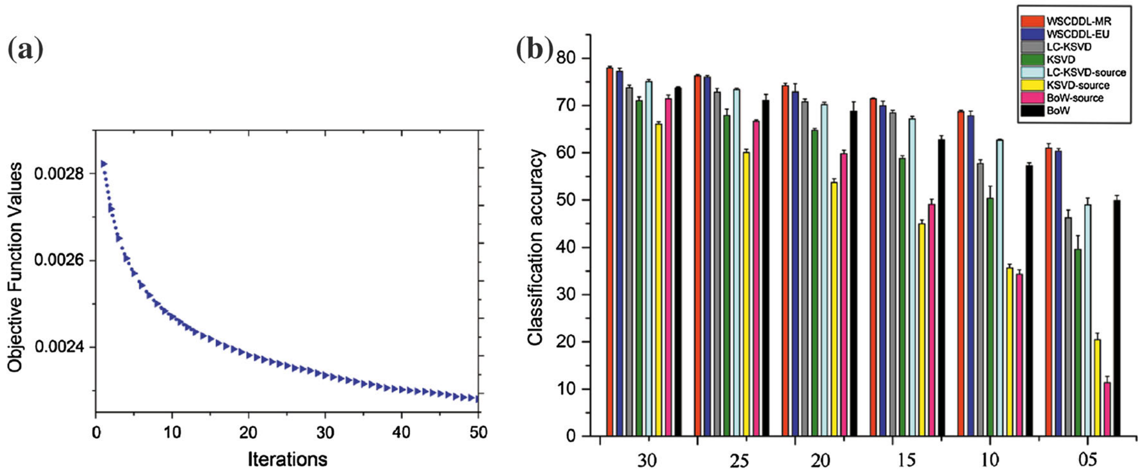

**Fig. 8** Performance analysis on the Caltech101 dataset. **a** The optimization process of the objective function for WSCDDL-MR with 50 iterations. **b** Means and standard deviations of different methods when the number of training samples per class varies from 5 to 30 (The dictionary size of WSCDDL-MR is set to 1,024)

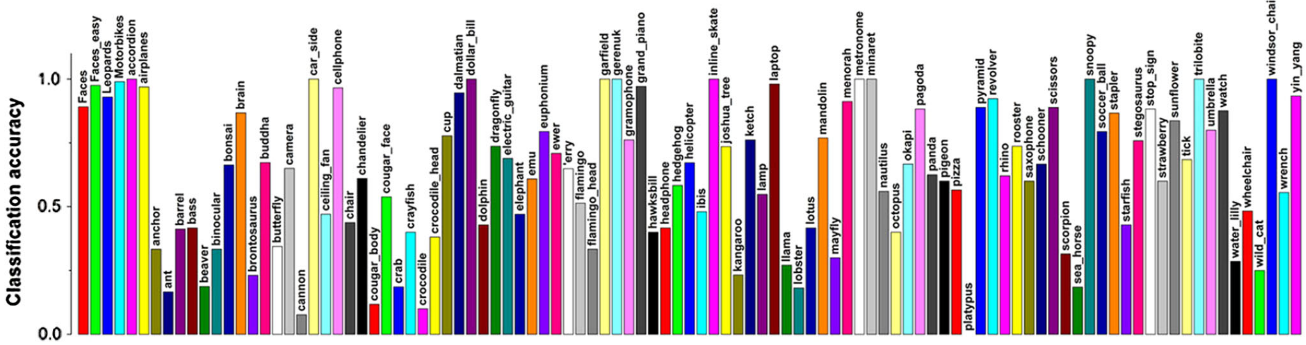

**Fig. 9** Performance on all the categories of the Caltech101 dataset achieved by the WSCDDL-MR (The dictionary size of WSCDDL-MR is set to 1,024) method when using 30 training images per category

**Fig. 10** Comparison of the confusion matrixes between the baseline ScSPM and the WSCDDL on five different data partitions of the UCF YouTube dataset

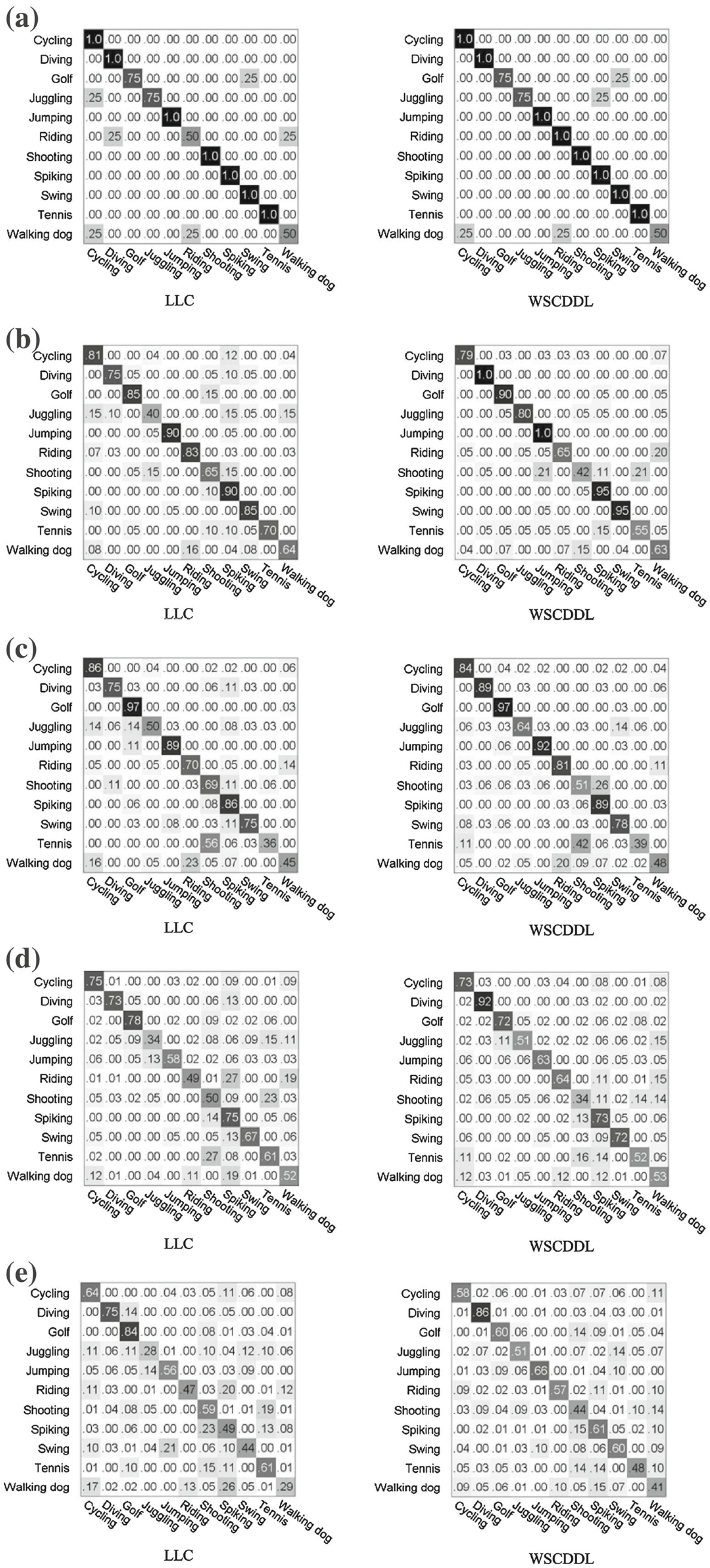

**Table 5** Recognition results on the Caltech256 dataset

| Number of training samples | 15 (%) | 30 (%) |
|---|---|---|
| Griffin et al. (2007) | 28.3 | 34.10 |
| Yang et al. (2009) | 27.73 | 34.02 |
| K-SVD (Aharon et al. 2006) | 25.33 | 30.62 |
| SRC (Wright et al. 2009) | 27.86 | 33.33 |
| WSCDDL-EU ($N = 1{,}024$) | 29.68 | 35.78 |
| WSCDDL-MR ($N = 1{,}024$) | **31.89** | **36.86** |
| LLC (Wang et al. 2010) ($N = 4{,}096$) | 34.36 | 41.19 |
| CRBM (Sohn et al 2011) ($N = 4{,}096$) | 35.09 | 42.05 |
| WSCDDL-EU ($N = 4{,}096$) | 36.21 | 42.33 |
| WSCDDL-MR ($N = 4{,}096$) | **37.42** | **42.80** |

Best results are in bold

shown in Table 5, where our approach consistently leads to the best performance. Figure 11 shows samples from 5 categories with high classification accuracies when using 30 images per category.

### 5.4 Event Recognition

We compare our proposed method WSCDDL with state-of-the-art transfer learning methods on the event recognition task using the Kodak Consumer Videos and a set of additional videos. The Kodak consumer video benchmark dataset was collected by Kodak from about 100 real users over the period of one year, and it includes two video subsets from two different sources, where the first part contains Kodak's video data which includes 1,358 video clips contributed by involved users and the second part contains 1,873 clips downloaded from the YouTube website after removing TV commercial videos and low-quality videos. Similarly, the additional videos collected by Duan et al. (2012b) contain two parts, which are the self-collected consumer videos and downloaded YouTube videos. To resemble the real-world scenario, the downloaded YouTube videos are not additionally annotated so that they can remain in a loosely labeled setting. Thus, only the self-collected consumer videos from the dataset used in Duan et al. (2012b) possess precise labels. The total numbers of consumer videos and YouTube videos are 195 and 906, respectively, and each video belongs to only one event category. Following the settings in Duan et al. (2012b), six events, namely "birthday", "picnic", "parade", "show", "sports" and "wedding" are chosen for experiments. The target domain is constructed using both the consumer videos from the Kodak dataset and additional self-collected consumer videos in Duan et al. (2012b). On the other hand, the second part of the Kodak dataset and the loosely labeled YouTube videos used in Duan et al. (2012b) constitute the source domain. In the target domain, three consumer videos from each event (18 videos in total) are randomly chosen as the labeled training videos and the remaining videos are

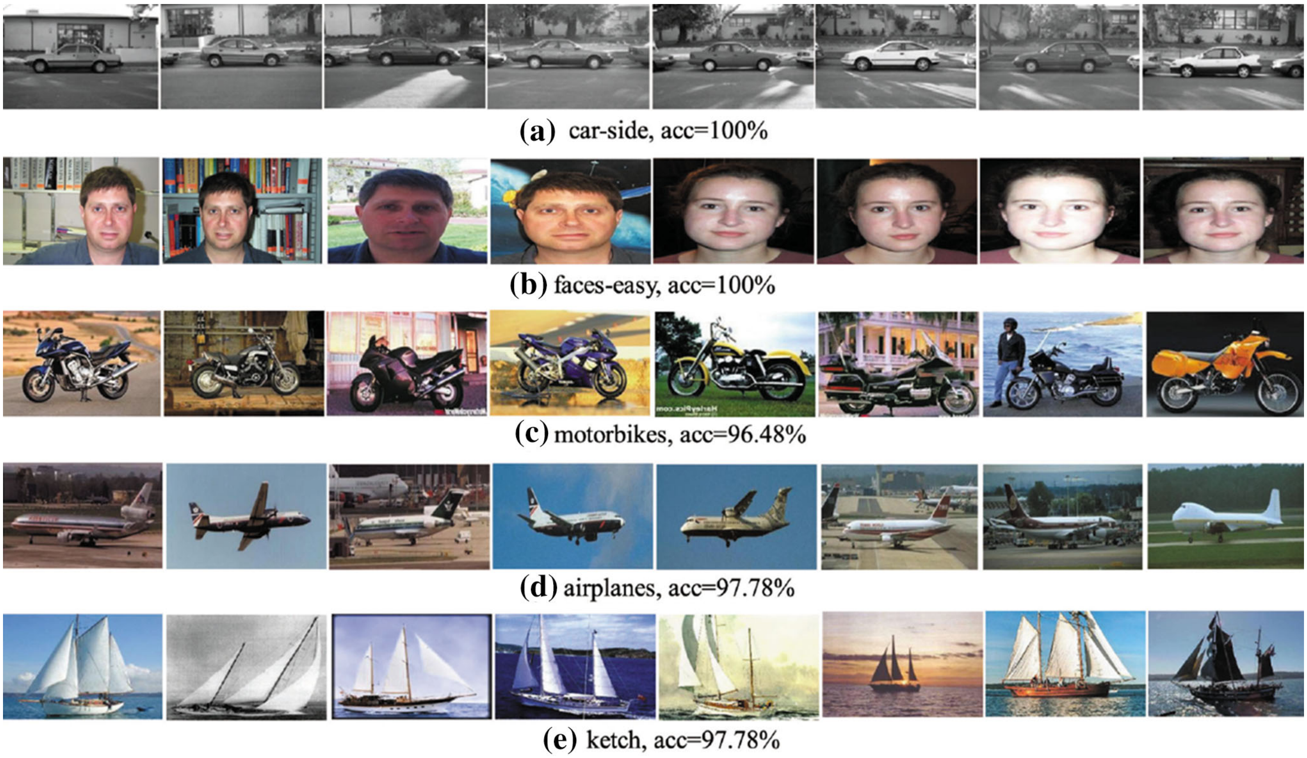

**Fig. 11** Example images of the categories with high classification accuracy from the Caltech256 dataset

**Table 6** Comparison with the state-of-the-art methods on the Kodak and YouTube dataset

| | SVM-T | SVM-AT | FR (Daumé 2007) | A-SVM (Yang et al. 2007) | MKL (Duan et al. 2009) | DTSVM (Duan et al. 2009) | A-MKL (Duan et al. 2012b) | WSCDDL-Eu | WSCDDL-MR |
|---|---|---|---|---|---|---|---|---|---|
| (a) | 42.32 ± 5.50 | 53.93 ± 5.58 | 49.98 ± 5.63 | 38.42 ± 7.93 | 47.19 ± 2.59 | 52.36 ± 1.88 | 47.14 ± 2.34 | 57.18 ± 0.84 | 58.42 ± 2.25 |
| (b) | 32.56 ± 2.08 | 24.73 ± 2.22 | 28.44 ± 2.61 | 24.95 ± 1.25 | 35.34 ± 1.55 | 31.07 ± 2.60 | 37.24 ± 1.58 | 37.80 ± 1.77 | 39.11 ± 2.76 |
| (c) | 42.00 ± 4.94 | 36.23 ± 3.37 | 44.11 ± 3.57 | 32.40 ± 4.99 | 46.92 ± 2.53 | 53.78 ± 2.99 | 58.20 ± 1.87 | 61.92 ± 2.89 | 62.60 ± 1.76 |

used as the test data. In order to set up a fair comparison in correspondence with the experimental results in Duan et al. (2012b), we use the same low-level features, which are SIFT features and ST features. For each sampled frame, which is sampled at the sampling rate of 2 frames per second, the 128-dimensional SIFT features are extracted from the salient regions, which are detected by the Difference-of-Gaussians (DoG) interest point detector (Lowe et al. 2004). The 162-dimensional local ST feature is the concatenation of the 72-dimensional HOG feature and the 90-dimensional HOF feature. We also conduct experiments in the same three cases as in Duan et al. (2012b): (a) dictionaries and classifiers are learned based on SIFT features, (b) dictionaries and classifiers are learned based on ST features and (c) dictionaries and classifiers are learned on both SIFT and ST features. Based on the same experimental settings as in Duan et al. (2012b), we compare our method WSCDDL with SVM-AT, SVM-T, FR (Daumé 2007), A-SVM (Yang et al. 2007), MKL (Duan et al. 2009), DTSVM (Duan et al. 2009) and A-MKL (Duan et al. 2012b), where SVM-AT denotes the case that labeled training samples are obtained from both the target domain and the source domain, and correspondingly SVM-T denotes the case that labeled training samples are only obtained from the target domain. Table 6 demonstrates the recognition results of the proposed WSCDDL method and other cross-domain methods. We can observe that SVM-T consistently outperforms SVM-AT in both scenarios of (b) and (c), which indicates that brutally including the ST features of source domain videos may degrade the recognition performance. The proposed WSCDDL method consistently outperforms other cross-domain methods in all three cases.

## 6 Conclusion

In this paper, we have presented a novel visual categorization framework using the weakly-supervised cross-domain dictionary learning algorithm. Auxiliary domain knowledge is utilized to span the intra-class diversities, so that the overall performance of the original system can be improved. The proposed framework only requires a small set of labeled samples in the source domain. By means of a transformation matrix, dictionary learning is performed on both the source domain data and the target domain data while no correspondence annotations between the two domains are required. Promising results are achieved on action recognition, image classification and event recognition tasks, where knowledge from either the Web or a related dataset is transferred to standard benchmark datasets. The proposed framework leads to an interesting topic for future investigation when large scale source and target domain data are available.

## References

Aharon, M., Elad, M., & Bruckstein, A. (2006). K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transaction on Signal Processing*, *54*(11), 4311–4322.

Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H. P., Schölkopf, B., & Smola, A. J. (2006). Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatices*, *22*, e49–e57.

Boureau, Y., Bach, F., LeCun, Y., & Ponce, J. (2010). Learning mid-level features for recognition. *CVPR*.

Cao, L., Liu, Z., & Huang, T. S. (2010). Cross-dataset action detection. *CVPR*.

Cao, X., Wang, Z., Yan, P., & Li, X. (2013). Transfer learning for pedestrian detection. *Neurocomputing*, *100*, 51–57.

Chen, S. S., Donoho, L. D., & Saunders, A. M. (1993). Atomic decomposition by basis pursuit. *IEEE Transaction on Signal Processing*, *41*(12), 3397–3415.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *CVPR*.

Dalal, N., Triggs, B., & Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. *ECCV*.

Daumé III, Hal, *Frustratingly easy domain adaptation, Proceedings of the Annual Meeting Association for Computational Linguistics*, pp. 256–263 (2007).

Dikmen, M., Ning, H., Lin, D. J., Cao, L., Le, V., Tsai, S. F., et al. (2008). Surveillance event detection. *TRECVID Video Evaluation Workshop*.

Dollár, P., Rabaud, V., Cottrell, G., & Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features, IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72 .

Duan, L., Tsang, I. W., & Xu, D. (2012). Domain transfer multiple kernel learning. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, *34*, 465–479.

Duan, L., Tsang, I. W., Xu, D., & Maybank, J. S. (2009). Domain transfer svm for video concept detection. *CVPR*.

Duan, L., Xu, D., Tsang, I. W., & Luo, J. (2012). Visual event recognition in videos by learning from web data. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, *34*, 1667–1680.

Duchenne, O., Laptev, I., Sivic, J., Bach, F., & Ponce, J. (2009). Automatic annotation of human actions in video. *ICCV*.

Fei-Fei, L. (2006). Knowledge transfer in learning to recognize visual objects classes. *ICDL*.

Fei-Fei, L., Fergus, R., & Perona, P. (2007). Learning generative visual models from few training examples. An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, *106*, 59–70.

Gao, X., Wang, X., Li, X., & Tao, D. (2011). Transfer latent variable model based on divergence analysis. *Pattern Recognition*, *44*, 2358–2366.

Gilbert, A., Illingworth, J., & Bowden, R. (2011). Action recognition using mined hierarchical compound features. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, *33*, 883–897.

Golub, G., Hansen, P., & O'Leary, D. (1999). Tikhonov regularization and total least squares. *Journal on Matrix Analysis and Applications*, *21*(1), 185–194.

Gregor, K., & LeCun, Y. (2010). *ICML: Learning fast approximations of sparse coding*. New York: Saunders.

Griffin, G., Holub, A., & Perona, P. (2007). Caltech-256 object category dataset, CIT Technical Report 1694.

Ikizler-Cinbis, N., Sclaroff, S. (2010). Object, scene and actions: Combining multiple features for human action recognition. *ECCV*.

Jégou, H., Douze, M., & Schmid, C. (2010). Improving bag-of-features for large scale image search. *International Journal of Computer Vision*, *87*, 316–336.

Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3D convolutional neural networks for human action recognition. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, *35*, 221–231.

Jiang, Z., Lin, Z., & Davis, L. S. (2011) Learning a discriminative dictionary for sparse coding via label consistent K-SVD. *CVPR*.

Junejo, I. N., Dexter, E., Laptev, I., & Pérez, P. (2011). View-independent action recognition from temporal self-similarities. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, *33*, 172–185.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, & T., Serre, T. (2011). HMDB: A large video database for human motion recognition. *ICCV*.

Kullback, S. (1987). The kullback-leibler distance. *The American Statistician*, *41*, 340–341.

Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld, B. (2008). Learning realistic human actions from movies. *CVPR*.

Laptev, I. (2005). On space-time interest points. *Internation Journal of Computer Vision*, *64*, 107–123.

Lazebnik, S., Schmid, C., & Ponce, J. (2006) Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *CVPR*.

Lee, H., Battle, A., Raina, R., & Andrew, Ng. (2007). Efficient sparse coding algorithms. *NIPS*.

Lee, H., Battle, A., Raina, R., & Ng, A. (2006). Efficient sparse coding algorithms. *NIPS*.

Li, R., & Zickler, T. (2012). Discriminative virtual views for cross-view action recognition. *CVPR*.

Liu, J., Luo, J., & Shah, M. (2009). Recognizing realistic actions from videos "in the wild". *CVPR*.

Liu, J., Shah, M., Kuipers, B., & Savarese, S. (2011). Cross-view action recognition via view knowledge transfer. *CVPR*.

Liu, L., Shao, L., & Rockett, P. (2012). Boosted key-frame selection and correlated pyramidal motion-feature representation for human action recognition. *Pattern Recognition*. doi:10.1016/j.patcog.2012.10.004.

Liwicki, S., Zafeiriou, S., Tzimiropoulos, G., & Pantic, M. (2012). Efficient online subspace learning with an indefinite kernel for visual tracking and recognition. *IEEE Transaction on Neural Networks and Learning Systems*, *23*, 1624–1636.

Loui, A., Luo, J., Chang, S., Ellis, D., Jiang, W., Kennedy, l., Lee, K., & Yanagawa, K. (2007). Kodak's consumer video benchmark data set: concept definition and annotation. *IWMIR*.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*, 91–110.

Lowe, D. G., Luo, J., Chang, S. F., Ellis, D., Jiang, W., Kennedy, L., et al. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Mairal, J., Bach, F., Ponce, J., Sapiro, G,. & Zisserman, A. (2008). Discriminative learned dictionaries for local image analysis. *CVPR*.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., & Zisserman, A. (2009). Supervised dictionary learning. *NIPS*.

Mairal, J., Leordeanu, M., Bach, F., Hebert, M., & Ponce, J. (2008) Discriminative sparse image models for class-specific edge detection and image interpretation. *ECCV*.

Maji, S., Berg, A., & Malik, J. (2013). Efficient classification for additive Kernel SVMs. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, *35*, 66–77.

Mallat, S. G., & Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Transaction on Signal Processing*, *41*(12), 3397–3415.

Marszalek, M., Laptev, I., & Schmid, C. (2009). Actions in context. *CVPR*.

Orrite, C., Rodríguez, M., & Montañés, M. (2011). One-sequence learning of human actions. *Human Behavior Unterstanding*, *7065*, 40–51.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transaction on Knowledge and Data Engineering*, *22*, 1345–1359.

Pati, Y., & Ramin, R. (1993). Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. *Asilomar Conference on Signals, Systems and Computers*, *4*, 40–44.

Qiu, Q., Patel, V. M., Turaga, P., & Chellappa, R. (2012). Domain adaptive dictionary learning. *ECCV*.

Raina, R., Battle, A., Lee, H., Packer, B., & Ng, A. Y. (2007). Self-taught learning: Transfer learning from unlabeled data. *ICML*.

Schuldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: A local svm approach. *ICPR*.

Sidenblada, H., & Black, M. J. (2003). Learning the statistics of people in images and video. *International Journal of Computer Vision*, *54*, 183–209.

Sohn, K., Jung, D., Lee, H., & Hero, A. (2011) Efficient learning of sparse, distributed, convolutional feature representations for object recognition. *ICCV*.

Su, Y., & Jurie, F. (2012). Improving image classification using semantic attributes. *International Journal of Computer Vision*, *100*, 1–19.

Uemura, H., Ishikawa, S., Mikolajczyk, K. (2008). Feature tracking and motion compensation for action recognition. *BMVC*.

Wang, H., Klaser, A., Schmid, C., Liu, C. (2011). Action recognition by dense trajectories. *CVPR*.

Wang, H., Ullah, M., Klaser, A., Laptev, I., Schmid, C. (2009). Evaluation of local spatio-temporal features for action recognition. *BMVC*.

Wang, J., Yang, J., Yu, K., Lv, F., huang, T., Gong, Y. (2010). Locality-constrained linear coding for image classification. CVPR.

Wang, Y., & Mori, G. (2009). Max-margin hidden conditional random fields for human action recognition. *CVPR*.

Wang, Y., & Mori, G. (2011). Hidden part models for human action recognition: Probabilistic versus max margin. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, *33*, 1310–1323.

Wright, J., Yang, Y. A., Ganesh, A., Sastry, S. S., & Ma, Y. (2009). *IEEE Transaction on Pattern Analysis and Machine Intelligence*, *31*, 210–227.

Xiang, S., Nie, F., Meng, G., Pan, C., & Zhang, C. (2012). Discriminative least squares regression for multiclass classification and feature selection. *IEEE Transaction on Neural Networks and Learning Systems*, *23*, 1738–1754.

Yang, L., Jin, R., Sukthankar, R., & Jurie, F. (2008). Unifying discriminative visual codebook generation with classifier training for object category recognition. *CVPR*.

Yang, J., Yan, R., & Hauptmann, A. G. (2007). Cross-domain video concept detection using adaptive SVMs. *ACM MM*.

Yang, J., Yu, K., Gong, Y., Huang, T. (2009). Linear spatial pyramid matching using sparse coding for image classification. *CVPR*.

Yang, J., Yu, K., & Huang, T. (2010). Supervised translation-invariant sparse coding. CVPR.

Yao, A., Gall, J., & Van, L. G. (2012). Coupled action recognition and pose estimation from multiple views. *International Journal of Computer Vision*, *100*, 16–37.

Zafeiriou, S., Tzimiropoulos, G., Petrou, M., & Stathaki, T. (2012) Regularized kernel discriminant analysis with a robust kernel for face recognition and verification. *NIPS*.

Zhang, H., Berg, C. A., Maire, M., & Malik, J. (2006) SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. CVPR.

Zhang, Q., & Li, B. (2010). Discriminative K-SVD for dictionary learning in face recognition. *CVPR*.

Zhang, W., Surve, A., Fern, X., & Dietterich, T. (2009). Learning non-redundant codebooks for classifying complex objects. *ICML*.

Zheng, J., Jinag, Z., Phillips,P. J., & Chellappa, R. (2012) Cross-view action recognition via a transferable dictionary pair. *BMVC*.

Zhou, D., Bousquet, O., Lal, T., Weston, J., Gretton, A., & Schölkopf, B. (2004). Learning with local and global consistency. *NIPS*.

Zhou, M., Chen, H., Paisley, J., Ren, L., Sapiro, G., & Carin, L. (2009). Non-parametric bayesian dictionary learning for sparse image representations. *NIPS*.

Zhou, D., Weston, J., Gretton, A., Bousquet, O., & Schölkopf, B. (2004). Ranking on data manifolds. *NIPS*.

Zhu, F., & Shao, L. (2013). Enhancing action recognition by cross-domain dictionary learning. *BMVC*.