Transfer Learning for Visual Categorization: A Survey

Ling Shao, *Senior Member, IEEE*, Fan Zhu, *Student Member, IEEE*, and Xuelong Li, *Fellow, IEEE*

***Abstract*— Regular machine learning and data mining techniques study the training data for future inferences under a major assumption that the future data are within the same feature space or have the same distribution as the training data. However, due to the limited availability of human labeled training data, training data that stay in the same feature space or have the same distribution as the future data cannot be guaranteed to be sufficient enough to avoid the over-fitting problem. In real-world applications, apart from data in the target domain, related data in a different domain can also be included to expand the availability of our prior knowledge about the target future data. Transfer learning addresses such cross-domain learning problems by extracting useful information from data in a related domain and transferring them for being used in target tasks. In recent years, with transfer learning being applied to visual categorization, some typical problems, e.g., view divergence in action recognition tasks and concept drifting in image classification tasks, can be efficiently solved. In this paper, we survey state-of-the-art transfer learning algorithms in visual categorization applications, such as object recognition, image classification, and human action recognition.**

***Index Terms*— Action recognition, image classification, machine learning, object recognition, survey, transfer learning, visual categorization.**

I. Introduction

IN THE past few years, the computer vision community has witnessed a significant amount of applications in video search and retrieval, surveillance, robotics, and so on. Regular machine learning approaches [1]–[7] have achieved promising results under the major assumption that the training and testing data stay in the same feature space or share the same distribution. However, in real-world applications, due to the high price of human manual labeling and environmental restrictions, sufficient training data belonging to the same feature space or the same distribution as the testing data may not always be available. Typical examples are [8]–[11], where only one action template is provided for each action class for training, and [12], where training samples are captured from a different viewpoint. In such situations, regular machine learning techniques are very likely to fail. This reminds us of the capability of the human vision system. Given the gigantic geometric and intraclass variabilities of objects, 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 [13]. It is estimated that there are about 10–30 thousands object classes in the world [14] and children can learn 4–5 object classes per day [13]. Due to the limitation of objects that a child can see within a day, learning new object classes from large amounts of corresponding object data is not possible. Thus, it is believed that the existing knowledge gained from previous known objects assists the new learning process through their connections with the new object categories. For example, assuming we did not know what a watermelon is, we would only need one training sample of watermelons together with our previous knowledge on melons-circular shapes, the green color, and so on, to remember the new object category watermelon. Transfer learning mimics the human vision system by making use of sufficient amounts of prior knowledge in other related domains when executing new tasks in the given domain. In transfer learning, both the training data and the testing data can contribute to two types of domains: 1) the target domain and 2) the source domain. The target domain contains the testing instances, which are the task of the categorization system, and the source domain contains training instances, which are under a different distribution with the target domain data. In most cases, there is only one target domain for a transfer learning task, while either single or multiple source domains can exist. For example, in [15], action recognition is conducted across data sets from different domains, where the KTH data set [16], which has a clean background and limited viewpoint and scale changes, is set as the source data set, and the Microsoft research action data set[1] and the TRECVID surveillance data [17], which are captured from realistic scenarios, are used as the target data set. In [18], the source and target data sets are chosen from different TV program channels for the task of video concept detection.

Transfer learning can be considered as a special learning paradigm where partial/all training data used are

Manuscript received December 29, 2012; revised October 17, 2013, January 30, 2014, and May 26, 2014; accepted June 3, 2014. Date of publication July 1, 2014; date of current version April 15, 2015. This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2012CB316400, in part by the University of Sheffield, Sheffield, U.K., and in part by the National Natural Science Foundation of China under Grants 61125106 and 61072093.

L. Shao is with the College of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China, and also with the Department of Electronic and Electrical Engineering, University of Sheffield, Sheffield S1 3JD, U.K. (e-mail: ling.shao@sheffield.ac.uk).

F. Zhu is with the Department of Electronic and Electrical Engineering, University of Sheffield, Sheffield S1 3JD, U.K. (e-mail: fan.zhu@sheffield.ac.uk).

X. Li is with the Center for OPTical IMagery Analysis and Learning, State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China (e-mail: xuelong_li@opt.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2014.2330900

[1] http://research.microsoft.com/∼zliu/ActionRecoRsrc

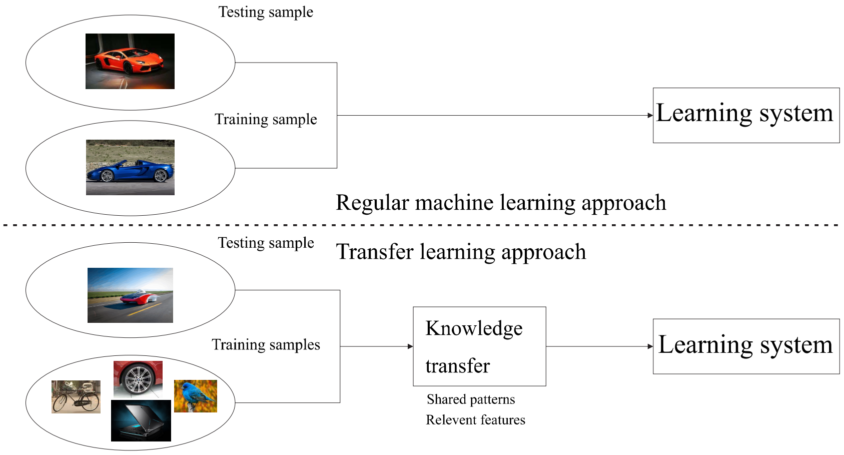


Fig. 1. Basic frameworks of traditional machine learning approaches and knowledge transfer approaches. For regular machine learning approaches, the learning system can only handle the situation that testing samples and training samples are under the same distribution. On the other hand, transfer learning approaches have to deal with the data distribution mismatch problem through specific knowledge transfer methods, e.g., mining the shared patterns from data across different domains.

under a different distribution with the testing data. To understand the significance of knowledge transfer in terms of visual learning problems, the literature, (see [19]–[21]) has concluded three general issues regarding the transfer process: 1) when to transfer; 2) what to transfer; and 3) how to transfer. First, when to transfer includes the issues whether transfer learning is necessary for specific learning tasks and whether the source domain data are related to the target domain data. In the scenarios of [22]–[24], where training samples are sufficient and impressive performance can be achieved, while being constrained in the target domains, including another domain as the source domain becomes superfluous. A variety of divergence levels exist across different pairs of source domain and target domain data, brute-forcing the knowledge from the source domain into the target domain irrespective of their divergence would cause certain performance degeneration, or, in even worse cases, it would break the original data consistency in the target domain. Second, the answer to what to transfer can be concluded in three aspects: 1) inductive transfer learning, where all the source domain instances and their corresponding labels are used for knowledge transfer; 2) instance transfer learning, where only the source domain instances are used; and 3) parameter transfer learning, in addition to the source domain instances and labels, some parameters of prelearned models from the source domain are utilized to help improve the performance in the target domain. Finally, how to transfer includes all the specific transfer learning techniques, and it is also the most important part that has been studied in the transfer learning literature. Many transfer learning techniques have been proposed, e.g., in [25]–[27], where knowledge transfer is based on the nonnegative matrix trifactorization framework, and in [28], where the transfer learning phase is via dimensionality reduction. We illustrate the basic frameworks of traditional machine learning approaches and knowledge transfer approaches in Fig. 1. For traditional machine learning approaches, the ideal choice of the training set to predict a testing instance car should contain cars. However, in the case of knowledge transfer, the training set can just contain some relevant categories rather than cars, e.g., wheels, which are similar to the wheels of cars; bicycles, which share the knowledge of wheels with the car wheels, or even some irrelevant objects, e.g., laptops and birds, which seem to have no connections with cars, but actually share certain edges or geometrical layouts with local parts of a car image.

As the age of big data has come, transfer learning can provide more benefits to solve the target problem with more relevant data. Thus, it is believed that more applications on transfer learning will emerge in future research. This survey aims to give a comprehensive overview of transfer learning techniques on visual categorization tasks, so that readers could potentially use the analysis and discussions in this survey to understand how transfer learning can be applied to visual categorization tasks or to solve their problem with a suitable transfer learning method. The visual categorization tasks possess some unique characteristics due to certain visual properties that can be potentially used in the training process, e.g., the appearance or shape of an object part, the local symmetries of an object, and the structural. All these unique properties can be employed when designing transfer learning algorithms, which makes our work different from that of [19] and [29], where the former focuses on classification, regression and clustering problems related to data mining tasks and the latter focuses on reinforcement learning, which addresses problems with only limited environmental feedback rather than correctly labeled examples.

The remaining part of this survey is structured as follows. An overview is given in Section II. In Sections III and IV, two transfer learning categories, which execute knowledge transfer through feature representations and classifiers, are discussed in detail, respectively, answering the problems of what to transfer and how to transfer. In Section V, the model selection methods from multiple source domains, i.e., when to transfer, are discussed. Evaluation, analysis, and discussions on the stated transfer learning methods are given in Section VI. Finally, the conclusions are drawn in Section VII.

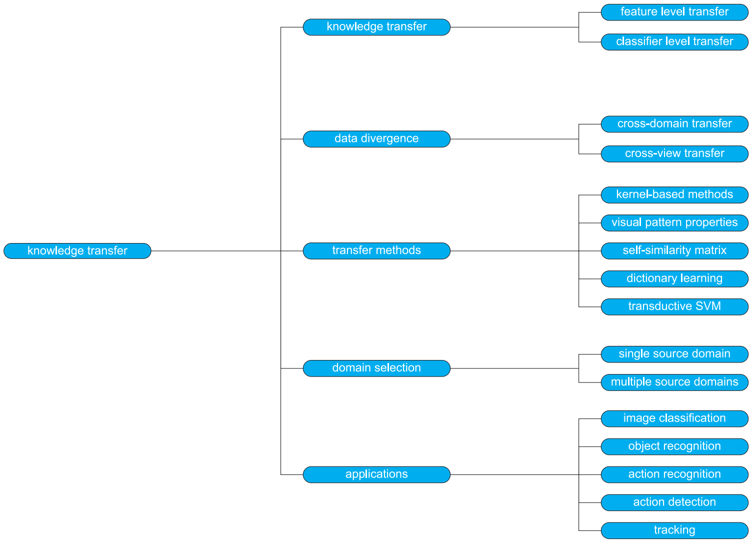


Fig. 2. Different ways of differentiating existing knowledge transfer approaches.

II. Overview

A. Developing Interests on Transfer Learning

Dating from the raising of its notion in the last century, transfer learning (also known as, cross-domain learning, domain transfer, and domain adaptation) has a long history of being studied as a particular machine learning technique. In recent years, with the information explosion on the Internet, (e.g., audio, images, and videos) and the growing demands for target tasks in terms of accuracies, data scales, and computational efficiencies, transfer learning approaches begin to attract increasing interests from all research areas in pattern recognition and machine learning. When regular machine learning techniques reach their limits, transfer learning opens the flow of a new stream that could fundamentally change the way of how we used to learn things and how we used to treat classification or regression tasks. Along with the flow, some workshops and tutorial have been held (such as the NIPS 1995 postconference workshop[2] in machine learning and data mining areas and another transfer learning survey is given in [29] for reinforcement learning). In this survey, we focus on the applications of transfer learning techniques to visual categorization, including action recognition, object recognition, and image classification.

B. Notations and Issues

Some general notations are defined as follows for later usage: let $\mathcal{D}^T = \mathcal{D}_l^T \cup \mathcal{D}_u^T$ denote the target domain data, where the partially labeled parts are denoted by $\mathcal{D}_l^T$ and the unlabeled parts are denoted by $\mathcal{D}_u^T$. In addition to the target domain data, a set of auxiliary data is seen as the source domain data, which is semilabeled or fully labeled and has the representation $\mathcal{D}^s = \{(x_i, y_i)\}_{i=1}^{a}$ in a single source case, and $\mathcal{D}_1^s, \mathcal{D}_2^s, \ldots, \mathcal{D}_M^s$ with $\mathcal{D}_k^s = \{(x_i^k, y_i^k)\}_{i=1}^{N_k^a}$ in a multiple source case. Here, $x_i \in \mathbb{R}^d$ is the ith feature vector, where d denotes the data dimension, and y_i denotes the class label of the ith sample.

According to prior proposals, common issues regarding knowledge transfer are twofold. First, the auxiliary samples are typically treated without accounting for their mutual dependency during adaptation, which may cause the adapted data to be arbitrarily distributed and the structural information beyond single data samples of the auxiliary data may become undermined. Second, during adaptation, noises, and particularly possible outliers from the auxiliary domains are blindly forced to the target domain [30].

When transferring knowledge from the auxiliary domains to the target domain, it is crucial to know the distribution similarities between the target domain data and each source domain data. So far, the most common criterion to measure the distribution similarity of two domains is a nonparametric distances metric named maximum mean discrepancy (MMD). The MMD is proposed in [31], and it compares data distributions in the reproducing kernel Hilbert space

$$\mathrm{Dist}_k(D^s, D^T) = \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \phi(x_i^s) - \frac{1}{n_T} \sum_{i=1}^{n_T} \phi(x_i^T) \right\|^2 \quad (1)$$

where $\phi(\cdot)$ is the feature space mapping function.

In the literature, transfer learning techniques are categorized according to a variety of taxonomies. In [19], considering tasks allocated to the target domain and auxiliary domains and the availability of sample labels within the target domain and auxiliary domains, transfer learning techniques are first grouped as inductive transfer learning, transductive transfer learning, and unsupervised transfer learning, upon which they are further categorized as instance-transfer, feature representation transfer, parameter transfer, and relational knowledge transfer within each initial partition. Fig. 2 shows five ways of differentiating existing knowledge transfer approaches for visual categorization. In this survey, inheriting the concepts from the computer vision community, we simply categorize transfer learning techniques into feature representation level knowledge transfer and classifier level knowledge transfer.

[2] https://nips.cc/Conferences/2005/Workshops/

III. Feature Representation Transfer

Feature representation level knowledge transfer is a popular transfer learning category that maps the target domain to the source domains exploiting a set of meticulously manufactured features. Through this type of feature representation level knowledge transfer, data divergence between the target domain and the source domains can be significantly reduced so that the performance of the task in the target domain is improved. Most existing transductive features are designed for specific domains and would not perform optimally across different data types. Thus, we review the feature level knowledge transfer techniques according to two data types: 1) cross-domain knowledge transfer and 2) cross-view knowledge transfer.

A. Cross-Domain Knowledge Transfer

In the cross-domain setting, the gap between the source domain data and the target domain data varies from images to videos and from objects to edges. According to the degree of data divergence, different approaches are proposed. In [15], knowledge transfer is made between the KTH data set [16], the TRECVID data set [17] and the Microsoft research action data set II (MSRII), where the KTH data set is seen as the target domain and both the TRECVID data set and the MSRII data set are used as the source domains. The KTH data set is limited to clean backgrounds and a single actor and each video sequence exhibits one individual action from the beginning to the end. On the other hand, the TRECVID data set and the MSRII data set are captured from realistic scenarios, with cluttered backgrounds and multiple actors in each video sequence. To take advantage of the labeled training data from both the target domain and the source domain, Daumé [32] proposed the feature replication (FR) method using augmented feature for training. Inspired by [33], which applies the Gaussian mixture model (GMM) to model the visual similarities between images or videos, the work in [15] models the spatial temporal interests points (STIPs) with the GMM and introduces a prior distribution of the GMM parameters to generate probabilistic representations of the original STIPs. Such representations can accomplish the adaptation from the source domains to the target domain. The basic setting of [34] assumes that there are labeled training data in the source domain, but no labeled training data in the target domain. Furthermore, the activities in the source domain and the target domain do not overlap, so that traditional supervised learning methods cannot be applied in this scenario. Utilizing the Web pages returned by search engines to mine similarities across the domains, the labeled data in the source domain are then interpreted by the label space of the target domain. In some extreme cases, the source domain data may not be relevant to the target domain data.

Sparseness has gained tremendous attention in various scientific fields, and computer vision is a dominant part of this trend. Sparse models can find their applications in a wide range of computer vision techniques, e.g., dictionary learning (DL) [35]–[37] and transfer learning. Raina *et al.* [38] apply 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 form a higher level representation of the training samples in the target domain, in which case the source domain data do not necessarily need to be relevant to the target domain data. Since in the regular transfer learning formalism, the source domain data have to be relevant with the target domain data, such a knowledge transfer method is named self-taught learning rather than transfer learning. Zhu and Shao [39] present a discriminative cross-domain DL (DCDDL) framework that utilizes relevant data from other visual domains as auxiliary knowledge for enhancing the learning system in the target domain. The objective function is designed to encourage similar visual patterns across different domains to possess identical representations after being encoded by a learned dictionary pair. In the part-of-speech (POS) tagging tasks, shared patterns from auxiliary categorization tasks are extracted as pivot features, which represent the frequent words emerged in the speech and are themselves indicative of their corresponding categories [40]. While the pivot features are sensitive to the POS tagging tasks, pivot visual words do not exist in typical local histogram-based low-level visual features, which indicates that no single feature dimension of the histogram bins is discriminative enough to represent the difference of the visual categories [41].

On the other hand, some works also target to identify a new lower-dimensional feature space such that the auxiliary domain and the target domain manifest some shared characteristics [42]–[44], instead of transferring the entire knowledge across the target domain and auxiliary domains making such an assumption that the smoothness property (i.e., those data points close to each other are more likely to share the same label) is satisfied in low-dimension subspaces [41].

B. Cross-View Knowledge Transfer

Cross-view knowledge transfer can be seen as a special case of cross-domain knowledge transfer, where the divergences across domains are caused by view-point changes. The task is to recognize action classes in the target view using training samples from one or more different views. Generating view-invariant features to address the cross-view visual pattern recognition problems attracts significant attention in the computer vision field, especially for cross-view action recognition. The bottom of Fig. 3 shows the cross-view knowledge transfer scenario on the multiview IXMAS [45] data set. The typical setting is to use samples captured in one view (the source view) as training data to predict the labels of samples captured from a different view (the target view). The core methodology of approaches that tackle visual categorization problems with changes in the observer's viewpoint is to discover the shared knowledge irrespective to such viewpoint changes. One common approach to attack the cross-view feature representation diversity problem is to infer 3-D scene structure for cross-view feature adaptation, where the derived features can be adapted from one view to another utilizing geometric reasoning [46]–[49]. Another family of approaches is to explore visual pattern properties, e.g., affine [50], projective [51], epipolar geometry [52]–[54], to compute such cross-view

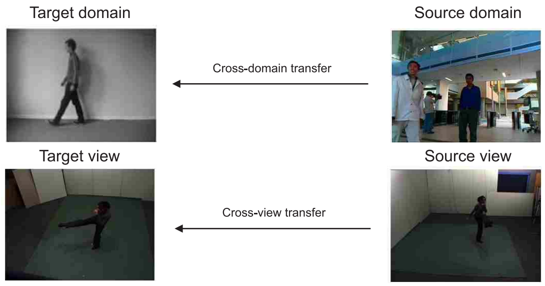


Fig. 3. Top row: cross-domain knowledge transfer scenario. In the target domain, the walking action performed by a single player in clean backgrounds comes from the KTH data set, while in the target domain, the walking action captured from much more complicated backgrounds with multiple players comes from the TRECVID data set. Bottom row: cross-view knowledge transfer scenario, where the target view data and the source view data are the same action captured from two different views of the IXMAS data set.

feature representations. On the other hand, Junejo *et al.* [55] applied a self-similarity matrix to store distances between different pairs of actions for a view-invariant representation. Spatial-temporal features of a video sequence that are insensitive to changes in view angle are studied in [12], [51], and [56]–[58].

In [12], a bipartite graph is built 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 generated. Beyond the bag-of-visual-words representation, which have been successfully applied to natural language processing, information retrieval, and computer vision, the proposed bag-of-bilingual-words representation discovers the shared set of common action concepts between two different views, even though the two view domains are highly independent. Similar to the work of [12], Li and Zickler [58] captured the conceptual idea of virtual views construction to represent an action descriptor continuously from one observer's viewpoint to another. Another family of approaches is proposed in [59] and [60], where a pair of over-complete dictionaries are constructed utilizing correspondence samples across two view domains. Encouraged by the learned dictionary pair, the labeled source view data and unlabeled target view data are forced to the same feature space that satisfies the smoothness assumption.

We summarize the main characteristics of the feature representation level knowledge transfer approaches according to their adaptation methods, the target domain label, the source domain label, adaptation data types and applications, and list them in Table I. Among these approaches, [38], [59], and [61] utilize the sparseness property to generate sparse representations for data adaptation.

IV. Classifier-Based Knowledge Transfer

Similar as the feature representation level knowledge transfer, classifier-based knowledge transfer is another significant part of existing visual transfer learning techniques and it has attracted much attention in recent years. However, unlike the feature representation level knowledge transfer techniques, where only the training samples themselves in the source domain are adapted to the target learning framework, classifier-based knowledge transfer methods share the common trait that the learned source domain models are utilized as prior knowledge in addition to the training samples when learning the target model. Instead of minimizing the cross-domain dissimilarity by updating instances' representations, classifier-based knowledge transfer methods aim to learn a new model that minimizes the generalization error in the target domain via provided training instances from both domains and the learned model. We structure this section according to the following categories of classifier-based knowledge transfer techniques.

A. SVM-Based

Support Vector Machine (SVM) is a supervised learning method for solving classification and regression problems, and the majority of existing work on classifier-based knowledge transfer are constructed from the original SVM classifier. As a direct application of SVM, adaptive-SVM (A-SVM) [18], and projective model transfer SVM (PMT-SVM) [63] learn from the source model w_s by regularizing the distance between the target model w_t and the learned model w_s. The A-SVM uses the following objective function:

$$L_A = \min_{w_t,b} \|w_t - \Gamma w_s\|^2 + C\sum_{i}^{N} l(x_i, y_i; w_t, b) \qquad (2)$$

where $y_i \in \{-1, 1\}$ indicates the corresponding labels, $l(x_i, y_i; w_t, b) = \max(0, 1 - y_i({w_t}^\top x_i + b))$ is the hinge loss, C controls the weight of the loss function, and Γ controls the amount of transfer regularization. By regularizing the distances between the two models, knowledge transfer for A-SVM is like a spring between Γw_s and w_t, which is equivalent to providing samples from the source classes. By expanding the regularization term

$$\|w_t - \Gamma w_s\|^2 = \|w_t\|^2 - 2\Gamma\|w_t\|\cos\theta + \Gamma^2 \qquad (3)$$

where $\|w\|^2$ provides the margin maximization as in regular SVM and the second term $-2\Gamma\|w\|\cos\theta$ induces the transfer by maximizing $\cos\theta$, i.e., by minimizing the angle θ between w_t and w_s. Instead of maximizing the term $\cos\theta$, knowledge transfer can be induced by minimizing the projection of w_t onto the separating hyperplane orthogonal to w_s for PMT-SVM using the following objective function:

$$L_{\text{PMT}} = \min_{w_t,b} \|w_t\|^2 + \Gamma\|Pw_t\|^2 + C\sum_{i}^{N} l(x_i, y_i; w_t, b)$$
$$\text{s.t.}: w_t^\top w_s \geq 0 \qquad (4)$$

where $P = I - (w_s w_s^\top)/(w_s^\top w_s)$ is the projection matrix. Compared with A-SVM, PMT-SVM can increase the amount of transfer (Γ) without penalizing margin maximization.

Opposed to the rigid transfer methods A-SVM and PMT-SVM, the deformable adaptive SVM (DA-SVM) [63] provides more flexible transfer regularization through a deformable source template, where small local deformations can be tolerated for the template fit of the source domain to the target domain. Aytar and Zisserman [63] used a simple example to explain such visual deformation in knowledge

TABLE I

MAIN CHARACTERISTICS OF LISTED FEATURE REPRESENTATION LEVEL KNOWLEDGE TRANSFER APPROACHES. AVAILABILITY OF BOTH TARGET DOMAIN LABELS, ADAPTATION TYPE, AND APPLICATIONS OF ALL STATED FEATURE REPRESENTATION LEVEL KNOWLEDGE TRANSFER METHODS ARE LISTED

Adaptation methods	**Target domain label**	**Source domain label**	**Adaptation type**	**Application**
Sparse Approximation[61]	A small set available	Unavailable	Cross-domain	Image classification
Kernel Based [41]	Unavailable/a small set available	Available	Cross-domain	Object recognition
Sparse Coding [38]	Available	Unavailable	Cross-domain	Digit recognition
Gaussian Mixture Model [15]	Unavailable	Available	Cross-domain	Action detection
Geometric Reasoning [46], [47], [48], [49]	Available	Available	Cross-view	Gesture detection, tracking, action recognition
Similarity Mining [34]	Unavailable	Available	Cross-domain	Activity recognition
Visual Pattern Properties [50], [51], [52], [53], [54]	Available	Available	Cross-view	Action recognition
Self-Similarity Matrix [55]	Available	Available	Cross-view	Action recognition
Bipartite Graph[12]	Unavailable	Available	Cross-view	Action recognition
Linear Transformation[58]	Unavailable/a small set available	Available	Cross-view	Action recognition
Dictionary Learning[59], [60]	Unavailable/a small set available	Available	Cross-view	Action recognition
Quantized Aspect[62]	Unavailable	Available	Cross-view	Action recognition

transfer that the wheel part of a motorbike template can be increased in radius and reduced in thickness when fitting to a bicycle wheel template. The DA-SVM can also be seen as the generalization form of the rigid A-SVM by replacing w_s in (2) with $\tau(w_s)$

$$L_{\text{DA}} = \min_{f,w_t,b} \|w_t - \Gamma\tau(w_s)\|^2 + C\sum_{i}^{N} l(x_i, y_i; w_t, b) + \lambda\left(\sum_{i\neq j}^{M,M} f_{i,j}^2 d_{i,j} + \sum_{i}^{M}(1-f_{ii})^2 d\right) \quad (5)$$

where $d_{i,j}$ is the spatial distance between the ith and jth cell, d is the penalization for the additional flow from the ith source cell to the ith target cell, and $\tau(w_s)_i = \sum_j^M f_{ij} w_{s_j}$ is the flow transformation, where the parameter f_{ij} denotes the amount of transfer from the jth cell in the source template to the ith cell in the transformed template. The cells are extracted from local image regions, on which local descriptors, (e.g., HOG [64] and SIFT [65]) are computed. Thus, different from other classifier-based knowledge transfer techniques, DA-SVM has such a constraint that it has to be constructed using low-level visual features that measure the geometrical information of local image parts.

Tommasi *et al.* [66] proposed a discriminative transfer learning method based on least squares support vector machine (LS-SVM) that learns the new category through adaptation. By replacing the regularization term in classical LS-SVM, the new learning objective function for knowledge transfer is formulated as

$$L_{\text{KTLS}} = \min_{w_t,b} \frac{1}{2}\|w_t - \theta w_s\|^2 + \frac{C}{2}\sum_{i=1}^{l}[y_i - w_t\phi(x_i) - b]^2 \quad (6)$$

where θ is a scaling factor in the range of (0, 1) to control the degree of transfer across the learned model w_s and the target model w_t. When being extended to multimodel knowledge transfer (multi-KT), the scaling factor θ is substituted with the vector $\Theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$, where each θ_j is the weight of a corresponding prior model. Thus, (6) can be rewritten as

$$L_{\text{Multi-KT}} = \min_{w_t,b} \left\| w_t - \sum_{j=1}^{k} \theta_j w_{s\,j} \right\|^2 + \frac{C}{2}\sum_{i=1}^{l} \zeta_i (y_i - w_t \cdot \phi(x_i) - b)^2. \quad (7)$$

The ζ_i in (7) is used for resampling the data so that training samples are balanced. Taking the advantage of LS-SVM that the leave-one-out (LOO) error, which measures the proper amount of knowledge to be transferred, can be written in a closed form [67], the best values of θ_j are those that minimize the LOO error.

Typically, the kernel functions need to be specified in advance to learning and the associated kernel parameters, (e.g., the mean and variance in the Gaussian kernel) are determined during optimization. On top of the various kernel learning methods [68]–[71], the domain transfer SVM (DT-SVM) [72] unified the cross-domain learning framework by searching for the SVM decision function $f(x) = w'\phi(x) + b$ as well as the kernel function simultaneously instead of the two-step approaches [28], [73]. In general, DT-SVM achieves cross-domain classification by reaching two objective criteria: 1) DT-SVM minimizes the data distribution mismatch between the target domain and source domains using the MMD criterion mentioned in Section II and 2) DT-SVM pursues better classification performance by minimizing the structural risk of SVM. By meeting both criteria, an effective kernel function can be learned for better separation performance in linear space over different domains, and thus samples from the source domains are infused to the target domain to improve the classification performance of the SVM classifier.

B. TrAdaboost

Adaptive boosting (AdaBoost) [74] is a popular boosting algorithm, which has been used in conjunction with a wide range of other machine learning algorithms to enhance their performance. At every iteration, AdaBoost increases the accuracy of the selection of the next weak classifier by carefully adjusting the weights on the training instances. Thus, more importance is given to misclassified instances since they are believed to be the most informative for the next selection. The transfer learning AdaBoost (TrAdaBoost) is introduced in [21] to extend AdaBoost for transfer learning by weighting less on the different-distribution data, which are considered as dissimilar to the same-distribution data in each boosting iteration. The goal of TrAdaBoost is to reduce the weighted training error on the different-distribution data, and meanwhile preserving the properties of AdaBoost. Since the quality of different-distribution data is not certain, the performance of TrAdaBoost cannot be always guaranteed to outperform AdaBoost.

C. Generative Models

The learning to learn concept via rich generative models has emerged as one promising research area in both computer vision and machine learning. Recently, researchers have begun developing new approaches to deal with transfer learning problems using generative models. One workshop in conjunction with NIPS 2010 was held specifically for the discussion of transfer learning via rich generative models. In general, the generative knowledge transfer methods can lead to higher-impact transfer, including more information than those discriminative approaches and they can be more adaptive to a single specific task.

Fei-Fei *et al.* [75] proposed a Bayesian-based unsupervised one-shot learning object categorization framework that learns a new object category using a single example (or just a few). Since Bayesian methods allow us to incorporate prior information about objects into a prior probability density function when observations become available, general information coming from previously learnt unrelated categories is represented with a suitable prior probability density function on the parameters of the probabilistic models. Thus, priors can be formed from unrelated object categories. For example, when learning the category motorbikes, priors can be obtained by averaging the learnt model parameters from other three categories spotted cats, faces, and airplanes, so that the hyperparameters of the priors are then estimated from the parameters of the existing category models. Yu and Aloimonos [76] applied the generative author-topic [77] model to learn the probabilistic distribution of image features-based object attributes. Since object attributes can represent common properties across different categories, they are used to transfer knowledge from source categories to target categories. Both the zero-shot learning problem and the one-shot learning problem are addressed, where in the first problem, the attribute model learned from the source domain categories is used to generate synthesized target training examples through the generative process, and in the second problem, the learned attribute model is used to reduce the uncertainty of parameters of the Dirichelt priors.

D. Fuzzy System-Based Models

Transfer learning also finds its application in fuzzy systems. Deng *et al.* [78] and [79] proposed two knowledge-leverage-based fuzzy system models, respectively. The former is based on the Takagi–Sugeno–Kang fuzzy system, and the latter is based on the reduced set density estimator-based Mamdani–Larsen-Type fuzzy system. In both works, the training set is decomposed to training data of the current scene and model parameters of reference scenes. The same knowledge leverage strategy is adopted by both works, where model parameters obtained from the reference scenes are fed to the current scene for parameter approximation. The knowledge leverage strategy is performed through a unified objective function, which emphasizes on both learning from the data of the current scene and transferring model parameters from reference scenes.

1) Discussion: The stated SVM-based knowledge transfer methods can act as a plug in to the SVM training process. A common trait shared amid these methods according to their objective functions is that they all include a regularization term that measures the similarity between the learned model and the target model. In A-SVM, PMT-SVM, and DA-SVM, Γ is the tradeoff parameter between margin maximization and knowledge transfer, so it defines the amount of transfer regularization. The DA-SVM is specialized in dealing with the transfer of visually deformable templates, while A-SVM and PMT-SVM are more likely to be generalized. The advantage of PMT-SVM over A-SVM is that it can increase the amount of transfer without penalizing margin maximization, while A-SVM encourages $\|w\|$ to be larger when increasing Γ. A large $\|w\|$ indicates small margins to the hyperplane, and thus the generalization error of the classifier fails to gain an optimal bound. In general, PMT-SVM is expected to outperform A-SVM.

Compared with SVM-based approaches, the boosting-based method, TrAdaBoost, is simpler in terms of implementation, and it does not require the parameters from the prelearned models. Like other boosting-based techniques, TrAdaBoost has a fairly strong generalization ability. However, TrAdaBoost relies heavily on the relevance of the source domain data to the target domain data, thus it is vulnerable to negative transfers. In addition, TrAdaBoost can easily overfit in the presence of noise in either domain. The generative models are more adaptive to a specific task, however, but computationally more complex.

V. Model Selection in Knowledge Transfer

In real-world applications, knowledge transfer techniques have to consider more complicated scenarios than adapting the samples or prelearned models from a single source domain

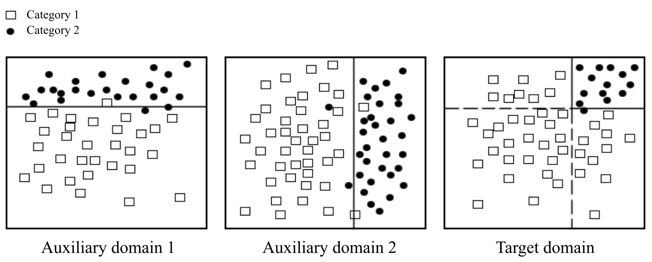


Fig. 4. Knowledge transfer from multiple auxiliary domains. (a) Auxiliary domain 1. (b) Auxiliary domain 2. (c) Target domain. The two subfigures on the left denote the two different auxiliary domain data and their corresponding decision boundaries, where auxiliary domain 1 is partitioned by a horizontal line and auxiliary domain 2 is partitioned by a vertical line. By brutally combining the decision boundaries from the two auxiliary domains, ambiguous predictions will be caused in the top-left region and the bottom-right region of the target domain.

to obtain the target learner. In the first case, more than one source domains are available yet we have no idea which source domain contains more useful information that potentially improves the target learner or whether the knowledge in a specific domain is against the smoothness property in the target domain. On the other hand, in visual categorization tasks, the shared information across the two domains can be hidden in different visual forms, e.g., appearance, local symmetry, and layout, which can be captured by different feature descriptors. A fusion strategy is required to mine the most helpful knowledge from multiple features. The third case is that some knowledge transfer techniques are constructed from prelearned models, e.g., a learned bicycle classifier or a learned bird classifier, and these models can lead to different scales of contributions to the target model. In advance to knowledge transfer, the bad prelearned models need to be filtered out so that the good models can achieve more effective transfer. All the above three cases generalize the common many-to-one adaptation situations in knowledge transfer, and they can all be deemed as the model selection problem. Fig. 4 shows a typical example of multisource binary classification. A straightforward approach to reduce such prediction ambiguity is to measure the model similarity between each auxiliary domain and the target domain, and apply the closest model for prediction in the target domain, i.e., if auxiliary domain 1 is more similar with the target domain, the decision boundary in Fig. 4(c) will inherit the decision boundary in Fig. 4(a). However, data in auxiliary domain 2, which also contain useful information for the prediction of target domain data, are abandoned.

In general, extending the existing single-source knowledge transfer techniques to the multiple-source scenario can evoke two challenges: 1) how to leverage the distribution differences among multiple source-domains to promote the prediction performance on the target domain task? and 2) how to extend the single-source knowledge transfer techniques to a distributed algorithm, while only sharing some statistical data of all source domains instead of revealing the full contents? Since most existing multiple-source knowledge transfer methods are extended from their corresponding single-source algorithms, we structure this section in a similar manner as Sections III and IV.

A. SVM-Based

In the one-to-one adaptation scenario of A-SVM [18], the new target classifier $f^T(x)$ is adapted from the existing source classifier $f^s(x)$ using the form

$$f^T(x) = f^s(x) + \triangle f(x) \tag{8}$$

where the perturbation function $\triangle f(x)$ is learned using the labeled data D_l^T from the target domain. Intuitively, when encountering with multiple source domains $\mathcal{D}_1^s, \mathcal{D}_2^s, \ldots, \mathcal{D}_M^s$, which are assumed to possess similar distributions to the primary domain D^t, the adapted classifier can be constructed using the ensemble of all the source domain classifiers $f_1^s(x), f_2^s(x), \ldots, f_M^s(x)$

$$f^T(x) = \sum_{k=1}^{M} \gamma_k f_k^s(x) + \triangle f(x) \tag{9}$$

where $\gamma_k \in (0, 1)$ is the predefined weight of each source classifier $f_k^s(x)$, which sums to one: $\sum_{k=1}^{m} \gamma_k = 1$. The MMD criterion can be applied for obtaining the value of γ_k. The perturbation function can be formulated as $\triangle f(x) = \sum_{i=1}^{n_l} \alpha_i^T y_i^T k(x_i^T, x)$, where α_i^T is the coefficient of the ith labeled pattern in the target domain and $k(\cdot, \cdot)$ is a kernel function induced from the nonlinear feature mapping $\phi(\cdot)$. When applying the same kernel function to the source classifiers, (9) can be expanded as

$$f^T(x) = \sum_{s} \gamma_s \sum_{i=1}^{n_l} \alpha_i^s y_i^s k\left(x_i^T, x\right) + \sum_{i=1}^{n_l} \alpha_i^T y_i^T k\left(x_i^T, x\right), \tag{10}$$

which is the sum of a set of weighted kernel evaluations between the test pattern x and all labeled patterns x_i^T and x_i^s, respectively, from the target domain and all the source domains. Obviously, the learning process is inefficient when being applied to large-scale data sets, which is the first disadvantage of A-SVM on the many-to-one adaptation application. The second disadvantage of A-SVM is its failure on using the unlabeled target domain data D_u^T.

Duan *et al.* [72] proposed the domain adaptation machine (DAM) to overcome the two disadvantages of A-SVM. To utilize the unlabeled target domain data D_u^T, a data-dependent regularizer is defined for the target classifier f^T

$$\Omega(f_u^T) = \frac{1}{2} \sum_{s=1}^{S} \gamma_s \sum_{i=1}^{m} \left(f_i^T - f_i^s\right)^2 \tag{11}$$

where $f_u^T = [f_{n_l+1}^T, \ldots, f_{n_T}^T]'$ and $f_u^s = [f_{n_l+1}^s, \ldots, f_{n_T}^s]'$ are defined as the decision values from the target classifier and the sth source classifier, respectively. Based on the smoothness assumption for domain adaptation, DAM minimizes the structural risk function of LS-SVM as well as the data-dependent regularizer simultaneously. DAM is formulated as

$$\min_{f^T} \Omega(f^T) + \frac{1}{2} \sum_{i=1}^{n_l} \left(f_i^T - y_i^T\right)^2 + \Omega_D\left(f_u^T\right) \tag{12}$$

where $\Omega(f^T)$ is a regularizer to control the complexity of the target classifier f^T. Since the target classifier in DAM is learned in a sparse representation, the computation inefficiency problem of A-SVM is overcome.

By arguing that it is more beneficial to transfer from a few relevant source domains rather than using all the source domains as in A-SVM and DAM, Duan *et al.* [80] further design a new data-dependant regularizer in domain selection machine (DSM) for source domain selection

$$\Omega(f) = \frac{1}{2}\sum_{s=1}^{S} d_s \sum_{i=1}^{m} \left(f_i^T - f_i^s\right)^2. \tag{13}$$

Similar as γ_s in (11), which is a predefined weight measuring the relevance between the sth source domain and the target domain, $d_s \in \{0, 1\}$ in (13) is a domain selection indicator for the sth source domain. When the objective function is optimized, the value of d_s is 1 if the sth source domain is relevant to the target domain, and the value of d_s is 0 otherwise. Another advantage of DSM over most existing transfer learning methods is its ability to work when the source domains and the target domain are represented by different types of features, e.g., using static 2-D SIFT features to represent the source domain data and 3-D spatio-temporal (ST) features to represent the target domain data. The learning function of DSM can be formulated as

$$f(x) = f_2D(x) + f_3D(x) = \sum_{s=1}^{S} d_s\beta_s f^s(x) + w'\varphi(x) + b \tag{14}$$

where $f_2D(x) = \sum_{s=1}^{S} d_s\beta_s f^s(x)$ is a weighted combination of source classifiers based on SIFT features, β_s is a real-valued weight for the sth source domain, $f_{3D}(x) = w'\varphi(x) + b$ is the adaptation error function of space-time features, $\varphi(\cdot)$ is a feature mapping function that maps x into $\varphi(x)$, w is a weight vector, and b is a bias term.

B. Boosting-Based

As discussed in Section IV-B, TrAdaBoost relies only on one source domain, which makes it intrinsically vulnerable to negative samples in the source domain. To avoid such a problem, Yao and Doretto [81] proposed two boosting approaches multisource-TrAdaBoost and task-TrAdaBoost for knowledge transfer with multiple source domains.

Multisource-TrAdaBoost is an extension of TrAdaBoost to multiple source domains. Instead of searching for a weak classifier by leveraging a single source domain, a mechanism is introduced to apply all the weak classifiers in the selected source domain that appears to be the most relevant to the target domain at the current iteration. Specifically, the training data of each source domain are combined with the training data in the target domain to generate a candidate weak classifier at each iteration, while all the source domains are considered independent from each other. Thus, the multisource-TrAdaBoost approach significantly reduces the effects of negative transfer caused by the imposition to knowledge transfer from a single source domain, which is potentially not relevant to the target domain.

On the other hand, task-TrAdaBoost is a parameter-transfer approach, that tries to identify which parameters that come from various source domains can be used. Task-TrAdaBoost is constituted of two separate phases. In phase-I, traditional AdaBoost is employed to extract suitable weak classifiers from each source domain, respectively, under the assumption that some parameters are shared between the source domain and the target domain. Thus, the source domain is described explicitly rather than implicitly with only the labeled source domain data. Phase-II runs the AdaBoost loop again over the target training data using the collection of all the candidate weak classifiers obtained from phase-I. At each iteration, the weak classifier with the lowest classification error on the target training data is picked out to ensure the knowledge being transferred is more relevant to the target task. In addition, the update of the weights on the target training data drives the search of the most helpful candidate classifiers in the next round for boosting the target classifier.

C. Multikernel Learning

There are many types of hidden knowledge that can be transferred across different visual domains, for example, the appearance or shape of an object part, (e.g., the shape of a wheel), local symmetries between parts, (e.g., the symmetry between front- and back-legs for quadrupeds), and the partially shared layout, (e.g., the layout of torso and limbs of a human). When employing knowledge transfer between the visual domains, though the shared knowledge exists among the target data and the source data, the exact type of knowledge that needs to be transferred is uncertain. Alternately, since these different types of knowledge can be represented by different features or different prior models, all types of knowledge can be considered by fusing these features or prior models when constructing the target model. Instead of using predefined weights for all the features or prior models, multikernel learning provides a more appropriate solution by learning the linear combination of coefficients of the prelearned classifiers to assure the minimization of domain mismatches.

Motivated by A-SVM, Duan *et al.* [82] proposed an adaptive multiple kernel learning (A-MKL) method to cope with the considerable variation in feature distributions between videos from two domains. As described above, in A-SVM, the target classifier is adapted from an existing classifier trained with the source domain data. When A-SVM employs multiple source classifiers, those classifiers are fused with fixed weights. Different from A-SVM, A-MKL learns the optimal combination of coefficients corresponding to each prelearned classifier to minimize the mismatch between the data distributions of two domains under the MMD criterion.

The multimodel knowledge transfer (multi-KT) [66] method modifies the l_2-norm regularizer in the LS-SVM objective function and constrains the new hyperplane w to be close to hyperplanes of F prior models. The regularization term is given as $\|w - \sum_{j=1}^{F}\beta^j\mu^j\|$, where μ^j is the hyperplane of the jth model, and β^j determines the amount of transfer from each model, while subjecting to the constraint that $\|\beta\|_2 \leq 1$.

For a sample x, the decision function is given by

$$s(x) = w \cdot \phi(x) + \sum_{j=1}^{F} \beta^j \mu^j \cdot \phi(x). \quad (15)$$

While the solution to multi-KT is through two separate optimization problems, Jie *et al.* [83] proposed a multiple kernel transfer learning (MKTL) method that learns the best hyperplanes and corresponding weights assigned to each prior model in a unified optimization process. The MKTL utilizes the prior knowledge as experts evaluating the new query instances and addresses such a knowledge transfer problem with a multikernel learning solver. In addition to the training sample x_i, the prediction score $s_p(x_i, z), z = 1, \ldots, F$ (F is the total number of classes), predicted by the prior models are considered when learning the new model. The intuition behind such an idea is that if prior knowledge of a bicycle gives a high prediction score to images of a motorbike, this information may also be useful for the new model of motorbikes, since certain visual parts, (e.g., the wheels) are shared between the two categories. Priors are built over multiple features instead of only one, and meanwhile, different learning methods are considered.

D. Cross-View Multiple Source Adaptation

For the cross-view action recognition problem, some shared visual patterns (either spatial or ST) can exist in actions captured from more than one view-points, thus transferring knowledge from multiple source views to the target view is more beneficial rather than transferring from a single view.

Liu *et al.* [12] apply the locally weighted ensemble (LWE) approach introduced in [45] to fuse the multiple classification models. Specifically, for a set of prelearned models $f_1, f_2, \ldots, f_k$, the general Bayesian model averaging approach computes the posterior distribution of y as $P(y|x) = \sum_{i=1}^{k} P(y|x, D, f_i)P(f_i|D)$, where $P(y|x, D, f_i) = P(y|x, f_i)$ is the prediction made by each model and $P(f_i|D)$ is the posterior of model f_i after observing the training set D. Considering the data distribution mismatch across the target domain and the source domains, the model prior for $P(f_i|T)$ is incorporated, where T is the test set. By replacing $P(f_i|D)$ with $P(f_i|T)$, the difference between the target and the source domains are considered during learning

$$P(y|x) = \sum_{i=1}^{k} w_{f_i,x} P(y|x, f_i) \quad (16)$$

where $w_{f_i,x} = P(f_i|x)$ is the true model weight that is locally adjusted for x representing the model's effectiveness on the target data.

Li and Zickler [58] achieve multiview fusion by aggregating the response values from the w MKL-SVM [69] classifiers on their corresponding cross-view features $\hat{x}$, beyond which a binary decision is made. Similar as the idea in MKTL [83], MKL-SVM solves a standard SVM optimization problem, where the kernel is defined as a linear combination of multiple kernels.

1) Discussion: The multiple source A-SVM is an intuitive extension of A-SVM that it assembles all the source domain classifiers by allocating a weight γ_k to each source classifier. The DAM and DSM are proposed to overcome the disadvantages of multiple source A-SVM in both inefficiency and the failure of using unlabeled target domain data, where DSM precedes over DAM by filtering out those less relevant source domain data.

By introducing multiple source domains rather than one in both multisource-TrAdaBoost and task-TrAdaBoost, the first imperfection of TrAdaBoost has been compensated. The convergence properties of multisource-TrAdaBoost can be inherited directly from TrAdaBoost [21], whereas for task-TrAdaBoost they can be inherited directly from AdaBoost [74]. It has been proved in [81] that since the convergence rate of task-TrAdaBoost has a reduced upper bound compared with multisource-TrAdaBoost, it requires fewer iterations to converge.

Compared with A-SVM, the unlabeled data in the target domain are used in the MMD criterion of A-MKL, and the weights in the target classifier are learned automatically together with the optimal kernel combination. Calling the theorem in [84], for the binary-class classification of multi-KT, multi-KT is equivalent to multiple source A-SVM based on the Mahalanobis distance measure [85]. Since the relationship between A-SVM and PMT-SVM is demonstrated in (2)–(4), the connection between multi-KT and PMT-SVM can be naturally discovered.

VI. Evaluation, Analysis, and Discussion

In general, there are three types of benefits that transfer learning can provide for performance improvements [66], [86], including: 1) higher start—improved performance at the initial points; 2) higher slope—more rapid growth of performance; and 3) higher asymptote—leading to improved final performance. In the following, several simple experiments are conducted with some selected representative knowledge transfer techniques discussed above to make a comparison between these methods and to see whether they can meet the stated criteria.

A. Feature-Level Knowledge Transfer Methods

Comparison between different feature representation cross-view transfer learning methods is given in Tables II and III, where experiments are conducted on every possible pairwise view combination of the IXMAS data set (i.e., twenty combinations in total) and columns demonstrate the results of target views, while rows demonstrate the results of auxiliary training views. According to previous cross-view action recognition works, there are two different experimental settings, which are the correspondence mode and the partially labeled mode. In the correspondence mode, the leave-one-action-class-out scheme is applied, where one action class is considered as the orphan action in the target view, while all action videos of the selected class are excluded when establishing the correspondences. Approximately 30% of the nonorphan samples are randomly selected to serve as the correspondences, and

TABLE II
COMPARISON BETWEEN DIFFERENT FEATURE REPRESENTATION CROSS-VIEW TRANSFER LEARNING METHODS IN THE CORRESPONDENCE MODE. RESULTS ARE REPORTED ON EVERY POSSIBLE PAIRWISE VIEW COMBINATION OF THE IXMAS DATA SET, WHERE COLUMNS CORRESPOND TO THE TARGET VIEWS AND ROWS CORRESPOND TO SOURCE VIEWS

%		Camera 0	Camera 1	Camera 2	Camera 3	Camera 4	Average
Camera 0	WO	-	16.1	10.3	11.2	8.8	11.6
	BW	-	81.2	79.6	73.0	82.0	79.0
	QA	-	69.0	62.0	63.0	51.0	61.0
	SS	-	77.3	66.1	69.4	39.1	63.0
	CV	-	72.0	71.0	75.0	80.0	74.0
	VV	-	87.5	85.3	82.1	78.8	83.4
	DL	-	97.3	92.1	97.0	83.0	92.4
Camera 1	WO	14.4	-	11.8	8.6	8.5	10.8
	BW	79.9	-	76.6	74.1	68.3	74.7
	QA	72.0	-	67.0	72.0	55.0	67.0
	SS	77.6	-	70.6	70.0	38.8	64.3
	CV	79.0	-	82.0	75.0	73.0	77.0
	VV	81.8	-	82.6	81.5	73.8	79.9
	DL	96.7	-	89.7	94.2	70.6	87.8
Camera 2	WO	10.7	11.1	-	10.0	9.2	10.3
	BW	76.8	75.8	-	74.4	74.0	75.2
	QA	61.0	64.0	-	68.0	51.0	61.0
	SS	69.4	73.9	-	63.0	51.8	64.5
	CV	79.0	74.0	-	79.0	73.0	76.0
	VV	88.1	82.0	-	80.2	77.7	82.0
	DL	97.9	96.4	-	96.7	89.7	95.1
Camera 3	WO	10.6	7.4	12.9	-	10	10.2
	BW	76.8	78.0	79.8	-	71.1	76.4
	QA	62.0	68.0	67.0	-	53.0	63.0
	SS	70.3	67.3	63.6	-	34.2	58.9
	CV	68.0	70.0	76.0	-	79.0	73.0
	VV	87.5	92.3	82.6	-	78.7	85.3
	DL	97.6	89.7	94.9	-	83.7	91.2
Camera 4	WO	19.1	9.2	8.1	9.3	-	11.4
	BW	74.8	70.4	72.8	66.9	-	71.2
	QA	30.0	41.0	43.0	44.0	-	40.0
	SS	44.8	43.9	53.6	44.2	-	46.6
	CV	76.0	66.0	72.0	76.0	-	72.0
	VV	81.4	74.2	76.5	70.0	-	75.5
	DL	84.9	81.2	89.1	83.9	-	84.8

TABLE III
COMPARISON BETWEEN CROSS-VIEW KNOWLEDGE TRANSFER METHODS IN THE PARTIALLY LABELED MODE. RESULTS ARE REPORTED ON EVERY POSSIBLE PAIRWISE VIEW COMBINATION OF THE IXMAS DATA SET, WHERE COLUMNS CORRESPOND TO THE TARGET VIEWS AND ROWS CORRESPOND TO SOURCE VIEWS

%		Camera 0	Camera 1	Camera 2	Camera 3	Camera 4	Average
Camera 0	SVMSUT	-	35.7	36.1	31.6	24.7	32.0
	AUGSVM	-	44.1	53.7	46.3	37.0	45.3
	MIXSVM	-	39.4	49.1	39.3	40.3	50.0
	VV	-	61.0	63.2	64.2	50.0	59.6
	DL	-	98.8	99.4	98.2	85.8	95.5
Camera 1	SVMSUT	39.8	-	42.0	30.3	27.0	34.8
	AUGSVM	42.8	-	50.5	42.5	35.0	42.7
	MIXSVM	36.8	-	49.4	42.5	42.5	42.8
	VV	63.6	-	62.4	71.0	59.7	64.2
	DL	98.8	-	96.4	97.6	81.5	93.6
Camera 2	SVMSUT	42.1	42.0	-	36.0	36.7	39.2
	AUGSVM	45.2	43.5	-	48.8	44.4	45.4
	MIXSVM	46.8	51.8	-	51.2	40.4	47.5
	VV	60.6	62.1	-	64.3	60.7	61.9
	DL	99.1	99.7	-	99.7	93.3	98.0
Camera 3	SVMSUT	41.6	28.5	43.0	-	31.1	36.1
	AUGSVM	47.2	47.1	53.5	-	37.2	46.2
	MIXSVM	42.7	45.8	45.0	-	40.7	43.5
	VV	61.2	65.1	71.7	-	61.1	64.8
	DL	99.4	92.7	97.3	-	83.9	93.3
Camera 4	SVMSUT	28.8	25.1	30.4	28.7	-	28.3
	AUGSVM	30.5	43.6	39.1	37.5	-	37.6
	MIXSVM	36.7	40.2	46.9	38.9	-	40.7
	VV	52.6	54.2	58.2	56.6	-	55.4
	DL	92.7	90.6	95.5	90.0	-	92.4

none of these correspondences are labeled. On the other hand, there are a small set of samples labeled in the partially labeled mode. We list the performance comparison of the above mentioned methods of the correspondence mode in Table II and of the partially labeled mode in Table III, respectively. Seven pairwise view scenarios are shown in Table II: 1) without (WO) transfer learning techniques [12]; 2) using the method in [12] with bilingual-words (BW); 3) using the method in [62] with quantized aspect (QA); 4) using the method in [55] with self-similarity metrics (SS); 5) using the method in [87] with continuous model of aspect (CV); 6) using the method in [58] with discriminative virtual views (VV); and 7) using the transferable dictionary pair in [59] constructed by DL. According to Table II, DL significantly outperforms the other methods and its most significant improvement over WO is 87% when treating Camera 0 as the source view and Camera 3 as the target view. Loosening the experimental restrictions by abandoning the correspondence instances from both views, while adding a small set of labeled training instances in the target view, comparisons between SVMSUT, AUGSVM, MIXSVM [88], VV, and DL are given in Table III, where DL still achieves the best results with the most significant improvement of 88.7% over WO when treating Camera 0 as the source view and Camera 3 as the target view. In general, Camera 4 has relatively weak performance. The reason is that Camera 4 is set above the actors, so that actions are captured in a totally different view. On the other hand, the performance involving Camera 4 can effectively demonstrate the capability of a transfer learning system. The BW, VV, and DL significantly outperform QA, SS, and CV. However, one limitation for the former three lies in that they implicitly assume that the target view is known for a query sequence.

B. Classifier-Level Knowledge Transfer Methods

We conduct experiments on both image classification and action recognition tasks, where the PASCAL VOC 2007 data set [89] is used for image classification and the UCF YouTube and HMDB51 data set [90] are used for action recognition. The PASCAL VOC 2007 data set contains 20 object classes, including bird, bicycle, motorbike, and so on, among which we choose samples from the bicycle class and the motorbike class as positive samples of the target domain and the source domain, respectively, and samples from the remaining classes as negative testing samples in the target domain. The histogram of oriented gradients (HOG) features are extracted from each image by dividing each image into eight cells. The task is to learn a bicycle classifier to achieve a binary decision over whether the test sample belongs to the bicycle category or a different category. The target classifier is learned by transferring information from a motorbike classifier via the guidance of a few bicycle samples. We compare the methods of nontransfer SVM, A-SVM, PMT-SVM, DA-SVM, and MKTL in Table IV with different numbers of training examples that vary from 1 to 25 with the interval of 3. Among these methods, DA-SVM achieves the best performance in terms of higher start and higher slope, while the PMT-SVM achieves the best final performance.

The UCF YouTube action data set is a realistic data set that contains camera shaking, cluttered background, variations in actors' scale, variations in illumination, and view point changes. There are 11 actions contained in the UCF YouTube data set, including biking, diving, golf swinging, and so on.

TABLE IV

PERFORMANCE COMPARISON ON THE IMAGE CLASSIFICATION TASK BETWEEN SVM, A-SVM, PMT-SVM, DA-SVM, AND MKTL. MODELS ARE LEARNED WITH DIFFERENT NUMBERS OF TRAINING EXAMPLES OF THE BICYCLE CLASS AND THE MOTORBIKE CLASS AS THE SOURCE DOMAIN. FIRST ROW INDICATES THE NUMBER OF TRAINING SAMPLES USED IN THE SOURCE DOMAIN

Methods	1	4	7	10	13	16	19	22	25
Non-transfer SVM	50.00%	50.00%	51.66%	46.00%	51.66%	49.33%	47.66%	50.33%	44.33%
A-SVM	97.95%	99.10%	98.49%	98.42%	98.32%	98.35%	98.59%	99.11%	99.31%
PMT-SVM	96.78%	98.84%	98.49%	98.61%	98.55%	98.63%	98.93%	99.24%	99.44%
DA-SVM	98.00%	99.11%	98.48%	98.41%	98.28%	98.35%	98.59%	99.11%	99.31%
MKTL	84.00%	87.66%	86.00%	86.00%	87.66%	89.66%	88.66%	90.00%	86.33%

TABLE V

PERFORMANCE COMPARISON ON THE ACTION RECOGNITION TASK BETWEEN SVM, A-SVM, PMT-SVM, AND MKTL MODELS ARE LEARNED WITH DIFFERENT NUMBERS OF TRAINING EXAMPLES OF THE BIKING CLASS AND THE DIVING CLASS AS THE SOURCE DOMAIN ON THE UCF YOUTUBE DATA SET. FIRST ROW INDICATES THE NUMBER OF TRAINING SAMPLES USED IN THE SOURCE DOMAIN

Methods	1	4	7	10	13	16	19	22	25
Non-transfer SVM	49.66%	46.66%	50.00%	47.00%	46.33%	49.33%	53.66%	48.00%	47.33%
A-SVM	59.77%	56.61%	60.30%	76.98%	81.48%	88.00%	92.26%	93.92%	94.11%
PMT-SVM	54.70%	56.25%	60.24%	77.49%	81.06%	85.94%	90.36%	93.14%	93.13%
MKTL	51.00%	50.33%	53.67%	54.33%	54.47%	60.00%	53.67%	59.00%	55.67%

TABLE VI

RECOGNITION RESULTS ON THE UCF YOUTUBE DATA SET WHEN USING THE HMDB51 DATA SET AS THE SOURCE DOMAIN

Algorithm	LLC [94]	LLC [94]	K-SVD [95]	K-SVD [95]	FR [32]	A-SVM [18]	DCDDL [39]
Supervision	N/A	N/A	UN	SU	SU	SU	SU
Source data	No	Yes	No	Yes	Yes	Yes	Yes
24 actors	86.67%	86.67%	82.22%	77.78%	83.74%	82.51%	88.89%
20 actors	75.42%	70.21%	68.75%	72.08%	74.88%	79.05%	77.50%
16 actors	70.88%	70.17%	63.96%	67.54%	71.56%	72.46%	73.03%
09 actors	61.41%	61.80%	55.70%	59.15%	62.77%	61.65%	66.31%
05 actors	54.10%	53.35%	50.05%	48.88%	54.09%	51.54%	56.66%

"UN" denotes "unsupervised" and "SU" denotes "supervised".

TABLE VII

MEAN AVERAGE PRECISIONS (MAPS) OF SVM, DASVM, DAM, DSM$_{\text{SIM}}$, AND DSM METHODS ON KODAK, YOUTUBE, AND CCV DATA SETS

Dataset	SVM	DASVM	DAM	DSM$_{sim}$	DSM
Kodak	27.95%	25.68%	27.66%	33.67%	35.46%
YouTube	31.17%	29.40%	32.58%	33.75%	35.26%
CCV	17.14%	18.38%	17.01%	17.80%	21.76%

The binary action recognition task aims at distinguishing actions between the biking class and the diving class with corresponding source domain actions from the HMDB51 data set, which is an even more challenging data set. Dense trajectories [91] are extracted from raw action video sequences with eight spatial scales spaced by a factor of $1/\sqrt{2}$, and feature points are sampled on a grid spaced by five 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. The HOG-HOF [92] and MBH [93] are computed within a $32 \times 32 \times 15$ volume along the dense trajectories, where each volume is subdivided into a ST grid of size $2 \times 2 \times 3$ to impose more structural information in the representation. The LLC coding scheme [94] is applied to the low-level local dense trajectory features. We compare the methods of nontransfer SVM, A-SVM, PMT-SVM, and MKTL in Table V. Obviously, the overall performance when transferring knowledge from the motorbike class to the bicycle class on the PASCAL VOC 2007 data set significantly outperforms the performance for transferring knowledge from the biking class to diving class on the UCF YouTube data set. This is due to that the relevance between motorbike and bicycle is much higher than the relevance between the actions biking and diving. In addition, the shared visual commons in video sequences are more difficult to capture than those in images. Compared with the results demonstrated in the image classification task, adding more training samples in the action recognition task leads to more significant improvements. As discussed in Section III, PMT-SVM is expected to outperform A-SVM in general. As shown in Tables IV and V, A-SVM outperforms PMT-SVM when a single or a few training instances are available, while PMT-SVM outperforms A-SVM in most cases when sufficient training instances are available. This can be explained as that PMT-SVM is relatively more sensitive to bad training samples.

We additionally conduct experiments on the action recognition task to compare the performance between the feature-level knowledge transfer techniques (FR and DCDDL), the classifier-level knowledge transfer technique (A-SVM) and nonknowledge transfer techniques (LLC and K-SVD). The experiments are conducted using the same setting as described above on the UCF YouTube data set and the HMDB51 data set. The results are demonstrated in Table VI.

TABLE VIII
MEANS AND STANDARD DEVIATIONS OF MAPS OVER SIX EVENTS FOR METHODS IN THREE CASES: 1) CLASSIFIERS LEARNED BASED ON SIFT FEATURES; 2) CLASSIFIERS LEARNED BASED ON ST FEATURES; AND 3) CLASSIFIERS LEARNED BASED ON BOTH SIFT AND ST FEATURES

Dataset	SVM-AT	SVM-A	A-SVM	MKL	DTSVM	A-MKL
MAP-(a)	$42.32 \pm 5.50\%$	$53.93 \pm 5.58\%$	$38.42 \pm 7.93\%$	$47.19 \pm 2.59\%$	$52.36 \pm 1.88\%$	$57.14 \pm 2.34\%$
MAP-(b)	$32.56 \pm 2.08\%$	$24.73 \pm 2.22\%$	$24.95 \pm 1.25\%$	$35.34 \pm 1.55\%$	$31.07 \pm 2.60\%$	$37.24 \pm 1.58\%$
MAP-(c)	$42.00 \pm 4.94\%$	$36.23 \pm 3.37\%$	$32.40 \pm 4.99\%$	$46.92 \pm 2.53\%$	$53.78 \pm 2.99\%$	$58.20 \pm 1.87\%$

By comparing the results of nonknowledge transfer techniques LLC and K-SVD by brutally using the source domain data to the same techniques without the source domain data, we can conclude that brutal forcing the source domain data into the target task could degrade the performance of original learning systems. Among the listed techniques, the recently proposed cross-domain DL method DCDDL achieves the best performance.

C. Knowledge Transfer From Multiple Source Domains

To demonstrate the performance comparisons between multisource knowledge transfer methods, we quote the experimental results in [80] and [82] for cross-domain multisource knowledge transfer, and the results in [12] and [58] for cross-view multisource knowledge transfer.

Duan *et al.* [80] chose the two large-scale image data sets to construct multiple source domains, where the first data set is the NUS-WIDE data set [96], which consists of 269, 648 images downloaded from the Flickr, and the second data set is collected from the photo forum called photosig.com, which contains about 1.3 million images. Three real-world consumer video data sets, the Kodak data set [82], the YouTube data set [80], and the CCV data set [97], are used as the target domains for performance evaluation, where the former contains 195 videos from six event classes, (e.g., birthday, picnic, parade, show, sports, and wedding), the middle is collected from YouTube using the same event classes as in the Kodak data set, and the latter contains 2726 videos for the same event classes. In the source domain, one hundred thousand training images are randomly selected from the two image sources and SIFT features are extracted from each image. After that, five source domains are constructed by randomly sampling 100 relevant images and 100 irrelevant images for clustering. In the target domain, both static SIFT features and space-time features are extracted from each video sequence in all the three data sets, where space-time interest point (STIP) feature and the Mel-frequency cepstral coefficients audio feature are extracted from the CCV data set, and three types of space-time features, HOG, HOF, and MBH, are extracted from Kodak and YouTube data sets. Since the standard SVM and DASVM cannot handle the domain adaptation problem when the data from the source and the target domain are with different feature types, only static SIFT features are used to learn classifiers in the target domain. The MAPs of SVM, DASVM, DAM, DSM$_{\text{sim}}$, and DSM methods on the three data sets are show in Table VII, where DSM$_{\text{sim}}$, as a simplified version of DSM, only considers the ST features in the target domain. Compared with the standard SVM, DASVM, and DAM achieve worse or equivalent performance on all three data sets, which indicate that the source domain data are not successfully used by these two methods. Based on the observation that the DSM$_{\text{sim}}$ consistently outperforms the related DAM method, it clearly demonstrates the benefits of employing the selected relevant source domains rather than using all the source domains. The DSM method achieves the best performance on all three data sets, which further demonstrates the effectiveness of integrating static SIFT features and ST features. Experiments are also conducted in [82] using the Kodak data set and videos downloaded from YouTube using keywords-based search to evaluate the performance of A-SVM, DTSVM, and A-MKL by transferring knowledge from the source image domains to the target video domain. Table VIII reports the means and the standard deviations of MAPs over all six events for the standard SVM, MKL, DTSVM, and A-MKL methods in the three cases, which are: 1) classifiers learned based on SIFT features; 2) classifiers learned based on ST features; and 3) classifiers learned based on both SIFT and ST features. Two forms of the standard SVM method, SVM-AT and SVM-T, are evaluated, where SVM-AT is learned based on samples from both the target domain and the source domain and SVM-T is learned based on samples only from the target domain. SVM-T outperforms SVM-AT in both cases 2) and 3), which indicates that directly, including the source domain knowledge may degrade the event recognition performances in the target domain. For all methods, the MAPs based on SIFT features are better than those based on ST features. This is consistent with our evaluation for classifier level knowledge transfer methods,

TABLE IX
CROSS-VIEW TRANSFER LEARNING ACTION RECOGNITION WITH MULTIPLE AUXILIARY VIEWS UNDER BOTH CORRESPONDENCE MODE AND PARTIALLY LABELED MODE

%	Partially labeled			
Target view	SVMSUT	AGUSVM	MIXSVM	MKL-SVM
Camera 0	38.5	54.2	46.4	62.0
Camera 1	43.4	50.8	44.2	65.5
Camera 2	50.3	58.1	52.3	64.5
Camera 3	51.0	49.5	47.7	69.5
Camera 4	35.1	46.9	44.7	57.9
	Correspondence			
	LWE		MKL-SVM	
Camera 0	86.6		85.1	
Camera 1	81.1		82.1	
Camera 2	80.1		82.2	
Camera 3	83.6		85.7	
Camera 4	82.8		77.6	

which indicates that the shared commons are more difficult to be captured in ST features than in static SIFT features. The effectiveness of fusing average classifiers and multiple base kernels is proved in A-MKL by providing the best performances for all cases.

The LWE fusing approach [12] and MKL-SVM approach [58] are compared with the SVMSUT, AUGSVM, MIXSVM methods on both the correspondence mode and the partially labeled mode in Table IX for cross-view multisource knowledge transfer. The overall results in the correspondence mode significantly outperforms the results in the partially labeled mode. In the correspondence mode, LWE and MKL-SVM achieve equivalent performance, while in the partially labeled mode, MKL-SVM consistently leads to the best performance.

VII. Conclusion

In this survey, we have reviewed transfer learning techniques on visual categorization tasks. There are three types of knowledge that are useful for knowledge transfer: 1) source domain features; 2) source domain features and the corresponding labels; and 3) parameters of the prelearned source domain models, which indicate instance-based transfer learning, inductive transfer learning and parameter-based transfer learning, respectively. Through the performance comparisons between knowledge transfer techniques and nonknowledge transfer techniques, we can conclude that brutal forcing the source domain data for learning can degrade the performance of the original learning system, which demonstrates the significance of knowledge transfer. To transfer the source domain knowledge to the target domain, methods are designed from either the feature representation level or the classifier level. In general, the feature representation level knowledge transfer aims to unify the mismatched data in different visual domains to the same feature space and the classifier level knowledge transfer aims to learn a target classifier based on the parameters of prelearned source domain models, while considering the data smoothness in the target domain. Thus, the feature representation level knowledge transfer techniques belong to either instance-based transfer or inductive transfer, while most classifier level knowledge transfer techniques belong to the parameter-based transfer. To avoid transferring the negative knowledge and deal with the many-to-one adaptation problem, many strategies are proposed to learn a set of weights for each source domain to achieve multiple source domain knowledge fusion.

Transfer learning is a tool for improving the performance of the target domain model only in the case that the target domain labeled data are not sufficient, otherwise the knowledge transfer is meaningless. So far, most research on transfer learning only focuses on small scale data, which cannot well reflect the potential advantage of transfer learning over regular machine learning techniques. The future challenges of transfer learning should lie in two aspects: 1) how to mine the information that would be helpful for the target domain from highly noisy source domain data and 2) how to extend the existing transfer learning methods to deal with large-scale source domain data.

References

[1] L. Shao, L. Liu, and X. Li, "Feature learning for image classification via multiobjective genetic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 7, pp. 1359–1371, Jul. 2014.

[2] L. Liu, L. Shao, and P. Rockett, "Boosted key-frame selection and correlated pyramidal motion-feature representation for human action recognition," *Pattern Recognit.*, vol. 46, no. 7, pp. 1810–1818, 2013.

[3] L. Shao, D. Wu, and X. Li, "Learning deep and wide: A spectral method for learning deep networks," *IEEE Trans. Neural Netw. Learn. Syst.*, doi: 10.1109/TNNLS.2014.2308519.

[4] L. Zhang, X. Zhen, and L. Shao, "Learning object-to-class kernels for scene classification," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3241–3253, Aug. 2014.

[5] L. Shao, X. Zhen, D. Tao, and X. Li, "Spatio-temporal Laplacian pyramid coding for action recognition," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 817–827, Jun. 2014.

[6] L. Shao, S. Jones, and X. Li, "Efficient search and localization of human actions in video databases," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 3, pp. 504–512, Mar. 2014.

[7] F. Zhu, L. Shao, and M. Lin, "Multi-view action recognition using local similarity random forests and sensor fusion," *Pattern Recognit. Lett.*, vol. 34, no. 1, pp. 20–24, 2013.

[8] X. Cao, Z. Wang, P. Yan, and X. Li, "Transfer learning for pedestrian detection," *Neurocomputing*, vol. 100, no. 1, pp. 51–57, 2013.

[9] X. Gao, X. Wang, X. Li, and D. Tao, "Transfer latent variable model based on divergence analysis," *Pattern Recognit.*, vol. 44, nos. 10–11, pp. 2358–2366, 2011.

[10] C. Orrite, M. Rodríguez, and M. Montañés, "One-sequence learning of human actions," in *Proc. 2nd Int. Workshop Human Behavior Understand.*, Amsterdam, The Netherlands, Nov. 2011, pp. 40–51.

[11] D. Wu, F. Zhu, and L. Shao, "One shot learning gesture recognition from RGBD images," in *Proc. 25th IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Providence, RI, USA, Jun. 2012, pp. 7–12.

[12] J. Liu, M. Shah, B. Kuipers, and S. Savarese, "Cross-view action recognition via view knowledge transfer," in *Proc. 24th IEEE Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, Jun. 2011, pp. 3209–3216.

[13] L. Fei-Fei, "Knowledge transfer in learning to recognize visual objects classes," in *Proc. 5th Int. Conf. Develop. Learn.*, Bloomington, IN, USA, Jun. 2006.

[14] I. Biederman, "Recognition-by-components: A theory of human image understanding," *Psychol. Rev.*, vol. 94, no. 2, pp. 115–147, 1987.

[15] L. Cao, Z. Liu, and T. S. Huang, "Cross-dataset action detection," in *Proc. 23rd IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 1998–2005.

[16] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. 17th Int. Conf. Pattern Recognit.*, Cambridge, U.K., Aug. 2004, pp. 32–36.

[17] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *Proc. 8th ACM Int. Workshop Multimedia Inform. Retrieval*, Santa Barbara, CA, USA, Oct. 2006, pp. 321–330.

[18] J. Yang, R. Yan, and A. G. Hauptmann, "Cross-domain video concept detection using adaptive SVMs," in *Proc. 15th ACM Int. Conf. Multimedia*, Augsburg, Germany, Sep. 2007.

[19] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[20] G.-J. Qi, C. Aggarwal, Y. Rui, Q. Tian, S. Chang, and T. Huang, "Towards cross-category knowledge propagation for learning visual concepts," in *Proc. 24th IEEE Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, Jun. 2011, pp. 897–904.

[21] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for transfer learning," in *Proc. 24th Int. Conf. Mach. Learn.*, Corvallis, OR, USA, Jun. 2007, pp. 193–200.

[22] Y. Wang and G. Mori, "Max-margin hidden conditional random fields for human action recognition," in *Proc. 22nd IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 872–879.

[23] A. Yao, J. Gall, and L. Van Gool, "A Hough transform-based voting framework for action recognition," in *Proc. 23rd IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 2061–2068.

[24] T. Xia, D. Tao, T. Mei, and Y. Zhang, "Multiview spectral embedding," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 6, pp. 1438–1446, Dec. 2010.

[25] H. Wang, F. Nie, H. Huang, and C. Ding, "Dyadic transfer learning for cross-domain image classification," in *Proc. 13th IEEE Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 551–556.

[26] T. Li and C. Ding, "The relationships among various nonnegative matrix factorization methods for clustering," in *Proc. 6th IEEE Int. Conf. Data Mining*, Hong Kong, Dec. 2006, pp. 362–371.

[27] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix t-factorizations for clustering," in *Proc. 12th ACM Int. Conf. Knowl. Discovery Data Mining*, Philadelphia, PA, USA, Aug. 2006, pp. 126–135.

[28] S. J. Pan, J. T. Kwok, and Q. Yang, "Transfer learning via dimensionality reduction," in *Proc. 23rd Nat. Conf. Artif. Intell. (AAAI)*, Chicago, IL, USA, Jul. 2008, pp. 677–682.

[29] M. E. Taylor and P. Stone, "Transfer learning for reinforcement learning domains: A survey," *J. Mach. Learn. Res.*, vol. 10, no. 1, pp. 1633–1685, 2009.

[30] I. Jhuo, D. Liu, D. Lee, and S. Chang, "Robust visual domain adaptation with low-rank reconstruction," in *Proc. 25th IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 2168–2175.

[31] K. Borgwardt, A. Gretton, M. Rasch, H. Kriegel, B. Schölkopf, and A. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

[32] H. Daumé, "Frustratingly easy domain adaptation," in *Proc. 45th Meeting Assoc. Comput. Linguist.*, Prague, Czech Republic, Jun. 2007.

[33] X. Zhou, X. Zhuang, S. Yan, S.-F. Chang, M. Hasegawa-Johnson, and T. S. Huang, "SIFT-bag kernel for video event analysis," in *Proc. 16th ACM Int. Conf. Multimedia*, Vancouver, BC, Canada, Oct. 2008, pp. 229–238.

[34] V. W. Zheng, D. H. Hu, and Q. Yang, "Cross-domain activity recognition," in *Proc. 11th Int. Conf. Ubiquitous Comput.*, Orlando, FL, USA, Jun. 2009, pp. 61–70.

[35] R. Yan, L. Shao, and Y. Liu, "Nonlocal hierarchical dictionary learning using wavelets for image denoising," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4689–4698, Dec. 2013.

[36] L. Shao, R. Yan, X. Li, and Y. Liu, "From heuristic optimization to dictionary learning: A review and comprehensive comparison of image denoising algorithms," *IEEE Trans. Cybern.*, vol. 44, no. 7, pp. 1001–1013, Jul. 2014.

[37] J. Tang, L. Shao, and X. Li, "Efficient dictionary learning for visual categorization," *Comput. Vis. Image Understand.*, vol. 124, no. 1, pp. 91–98, 2014.

[38] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data," in *Proc. 24th Int. Conf. Mach. Learn.*, Corvallis, OR, USA, Jun. 2007, pp. 759–766.

[39] F. Zhu and L. Shao, "Weakly-supervised cross-domain dictionary learning for visual recognition," *Int. J. Comput. Vis.*, vol. 109, nos. 1–2, pp. 42–59, 2014.

[40] J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Sydney, Australia, Jul. 2006, pp. 120–128.

[41] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. 25th IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 2066–2073.

[42] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. 25th IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 1794–1801.

[43] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 2, no. 1, pp. 37–52, 1987.

[44] P. O. Hoyer, "Non-negative sparse coding," in *Proc. 12th IEEE Workshop Neural Netw. Signal Process.*, Miami, FL, USA, Jun. 2002, pp. 557–565.

[45] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Comput. Vis. Image Understand.*, vol. 104, nos. 2–3, pp. 249–257, 2006.

[46] T. J. Darrell, I. A. Essa, and A. P. Pentland, "Task-specific gesture analysis in real-time using interpolated views," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 12, pp. 1236–1242, Dec. 1996.

[47] D. M. Gavrila and L. S. Davis, "3-D model-based tracking of humans in action: A multi-view approach," in *Proc. 9th IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 1996, pp. 73–80.

[48] F. Lv and R. Nevatia, "Single view human action recognition using key pose matching and Viterbi path searching," in *Proc. 20th IEEE Conf. Comput. Vis. Pattern Recognit.*, Minneapolis, MN, USA, Jun. 2007, pp. 1–8.

[49] D. Weinland, E. Boyer, and R. Ronfard, "Action recognition from arbitrary views using 3D exemplars," in *Proc. 11th Int. Conf. Comput. Vis.*, Rio de Janeiro, Brazil, Oct. 2007, pp. 1–8.

[50] C. Rao, A. Yilmaz, and M. Shah, "View-invariant representation and recognition of actions," *Int. J. Comput. Vis.*, vol. 50, no. 2, pp. 203–226, 2002.

[51] V. Parameswaran and R. Chellappa, "View invariance for human action recognition," *Int. J. Comput. Vis.*, vol. 66, no. 1, pp. 83–101, 2006.

[52] T. Syeda-Mahmood, A. Vasilescu, and S. Sethi, "Recognizing action events from multiple viewpoints," in *Proc. Workshop Detect. Recognit. Events Video*, Vancouver, BC, Canada, May 2001, pp. 64–72.

[53] A. Yilmaz and M. Shah, "Actions sketch: A novel action representation," in *Proc. 18th IEEE Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, Jun. 2005, pp. 984–989.

[54] A. Gritai, Y. Sheikh, and M. Shah, "On the use of anthropometry in the invariant analysis of human actions," in *Proc. 17th Int. Conf. Pattern Recognit.*, Cambridge, U.K., Aug. 2004, pp. 923–926.

[55] I. Junejo, E. Dexter, I. Laptev, and P. Pérez, "Cross-view action recognition from temporal self-similarities," in *Proc. 10th Eur. Conf. Comput. Vis.*, Marseille, France, Oct. 2008, pp. 293–306.

[56] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proc. 10th IEEE Int. Conf. Comput. Vis.*, Beijing, China, Oct. 2005, pp. 1395–1402.

[57] S. Seitz and C. Dyer, "View-invariant analysis of cyclic motion," *Int. J. Comput. Vis.*, vol. 25, no. 3, pp. 231–251, 1997.

[58] R. Li and T. Zickler, "Discriminative virtual views for cross-view action recognition," in *Proc. 25th IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 2855–2862.

[59] J. Zheng, Z. Jiang, P. Phillips, and R. Chellappa, "Cross-view action recognition via a transferable dictionary pair," in *Proc. 23rd British Mach. Vis. Conf.*, Surrey, U.K., Sep. 2012.

[60] F. Zhu and L. Shao, "Correspondence-free dictionary learning for cross-view action recognition," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Stockholm, Sweden, Aug. 2014.

[61] A. Quattoni, M. Collins, and T. Darrell, "Transfer learning for image classification with sparse prototype representations," in *Proc. 21st IEEE Conf. Comput. Vis. Pattern Recognit.*, Anchorage, AK, USA, Jun. 2008, pp. 1–8.

[62] A. Farhadi and M. Tabrizi, "Learning to recognize activities from the wrong view point," in *Proc. 10th Eur. Conf. Comput. Vis.*, Marseille, France, Oct. 2008, pp. 154–166.

[63] Y. Aytar and A. Zisserman, "Tabula rasa: Model transfer for object category detection," in *Proc. 13th IEEE Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 2252–2259.

[64] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. 21st IEEE Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, Jun. 2005, pp. 886–893.

[65] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[66] T. Tommasi, F. Orabona, and B. Caputo, "Safety in numbers: Learning categories from few examples with multi model knowledge transfer," in *Proc. 23rd IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 3081–3088.

[67] C. Cawley, "Leave-one-out cross-validation based model selection criteria for weighted LS-SVMs," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, San Diego, CA, USA, Jul. 2006, pp. 1661–1668.

[68] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *J. Mach. Learn. Res.*, vol. 5, pp. 27–72, Dec. 2004.

[69] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2491–2521, 2008.

[70] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf, "Large scale multiple kernel learning," *J. Mach. Learn. Res.*, vol. 7, no. 6, pp. 1531–1565, 2006.

[71] L. Duan, I. Tsang, and D. Xu, "Domain transfer multiple kernel learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 465–479, Mar. 2012.

[72] L. Duan, I. Tsang, D. Xu, and S. Maybank, "Domain transfer SVM for video concept detection," in *Proc. 22nd IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 1375–1381.

[73] B. Schölkopf *et al.*, "Correcting sample selection bias by unlabeled data," in *Proc. 20th Conf. Neural Inform. Process. Syst.*, Dec. 2006, pp. 601–608.

[74] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997.

[75] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 594–611, Apr. 2006.

[76] X. Yu and Y. Aloimonos, "Attribute-based transfer learning for object categorization with zero/one training example," in *Proc. 11th Eur. Conf. Comput. Vis. (ECCV)*, Hersonissos, Greece, Sep. 2010, pp. 127–140.

[77] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, and M. Steyvers, "Learning author-topic models from text corpora," *ACM Trans. Inform. Syst.*, vol. 28, no. 1, pp. 1–38, 2010.

[78] Z. Deng, Y. Jiang, K.-S. Choi, F.-L. Chung, and S. Wang, "Knowledge-leverage-based TSK fuzzy system modeling," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 8, pp. 1200–1212, Aug. 2013.

[79] Z. Deng, Y. Jiang, F.-L. Chung, H. Ishibuchi, and S. Wang, "Knowledge-leverage-based fuzzy system and its modeling," *IEEE Trans. Fuzzy Syst.*, vol. 21, no. 4, pp. 597–609, Aug. 2013.

[80] L. Duan, D. Xu, and S.-F. Chang, "Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach," in *Proc. 25th IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 1338–1345.

[81] Y. Yao and G. Doretto, "Boosting for transfer learning with multiple sources," in *Proc. 23rd IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 1855–1862.

[82] L. Duan, D. Xu, I. Tsang, and J. Luo, "Visual event recognition in videos by learning from web data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1667–1680, Sep. 2012.

[83] L. Jie, T. Tommasi, and B. Caputo, "Multiclass transfer learning from unconstrained priors," in *Proc. 13th IEEE Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 1863–1870.

[84] J. Ye and T. Xiong, "SVM versus least squares SVM," in *Proc. 7th Int. Conf. Artif. Intell. Stat.*, Scottsdale, AZ, USA, Apr. 2007, pp. 644–651.

[85] H. Trevor, T. Robert, and H. Friedman, *The Elements of Statistical Learning*. New York, NY, USA: Springer-Verlag, 2001.

[86] E. Olivas, M. Guerrero, M. B. M. Sober, and S. Lopez, *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods and Techniques*, vol. 2. Hershey, PA, USA: Information Science IGI Publishing, 2009.

[87] A. Farhadi, M. Tabrizi, I. Endres, and D. Forsyth, "A latent model of discriminative aspect," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Kyoto, Japan, Sep. 2009, pp. 948–955.

[88] A. Bergamo and L. Torresani, "Exploiting weakly-labeled web images to improve object classification: A domain adaptation approach," in *Proc. 24th Conf. Neural Inform. Process. Syst.*, Trento, Italy, Apr. 2010, pp. 29–37.

[89] (2007). *The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results* [Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html

[90] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. 13th IEEE Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 2556–2563.

[91] H. Wang, A. Klaser, C. Schmid, and C. L. Liu, "Action recognition by dense trajectories," in *Proc. 24th IEEE Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, Jun. 2011, pp. 3169–3176.

[92] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. 21st IEEE Conf. Comput. Vis. Pattern Recognit.*, Anchorage, AK, USA, Jun. 2008, pp. 1–8.

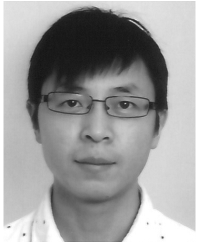
[93] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Proc. 9th Eur. Conf. Comput. Vis.*, Graz, Austria, May 2006, pp. 428–441.

[94] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. 23rd IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 3360–3367.

[95] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.

[96] T. Chua, J. Tang, R. Hong, H. Li, L. Zhiping, and Y. Zheng, "NUS-WIDE: A real-world web image database from National University of Singapore," in *Proc. ACM Int. Conf. Image Video Retrieval*, Santorini, Greece, Jul. 2009, pp. 48–56.

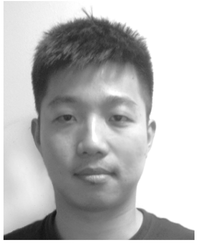
[97] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. C. Loui, "Consumer video understanding: A benchmark database and an evaluation of human and machine performance," in *Proc. 1st ACM Int. Conf. Multimedia Retrieval*, Trento, Italy, Apr. 2011, pp. 29–37.



Ling Shao (M'09–SM'10) received the B.Eng. degree from the University of Science and Technology of China, Hefei, China, and the M.Sc. and Ph.D. degrees from the University of Oxford, Oxford, U.K.

He is a Senior Lecturer (Associate Professor) with the Department of Electronic and Electrical Engineering, University of Sheffield, Sheffield, U.K., and a Guest Professor with the College of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing, China. He has authored and co-authored over 120 papers in well-known journals/conferences such as *International Journal of Computer Vision*, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON CYBERNETICS, *Pattern Recognition*, *Computer Vision and Image Understanding*, the IEEE Conference on Computer Vision and Pattern Recognition, the International Joint Conference on Artificial Intelligence, *ACM Multimedia,* and the British Machine Vision Conference, and holds more than 10 European/U.S. patents. His current research interests include computer vision, image/video processing, pattern recognition, and machine learning.

Dr. Shao is an Associate Editor of the IEEE TRANSACTIONS ON CYBERNETICS, *Information Sciences*, and several other journals. He is a fellow of the British Computer Society.



Fan Zhu (S'12) received the B.S. degree from the Wuhan Institute of Technology, Wuhan, China, in 2010, and the M.Sc. (Hons.) degree from the University of Sheffield, Sheffield, U.K., in 2012, where he is currently pursuing the Ph.D. degree with the Department of Electronic and Electrical Engineering.

His current research interests include submodular optimization for computer vision, sparse coding, and dictionary learning and transfer learning.

Xuelong Li (M'02–SM'07–F'12) is currently a Full Professor with the Center for OPTical IMagery Analysis and Learning, State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China.