Cross-Domain Human Action Recognition

Wei Bian, Dacheng Tao, *Member, IEEE*, and Yong Rui, *Fellow, IEEE*

***Abstract*—Conventional human action recognition algorithms cannot work well when the amount of training videos is insufficient. We solve this problem by proposing a transfer topic model (TTM), which utilizes information extracted from videos in the auxiliary domain to assist recognition tasks in the target domain. The TTM is well characterized by two aspects: 1) it uses the bag-of-words model trained from the auxiliary domain to represent videos in the target domain; and 2) it assumes each human action is a mixture of a set of topics and uses the topics learned from the auxiliary domain to regularize the topic estimation in the target domain, wherein the regularization is the summation of Kullback–Leibler divergences between topic pairs of the two domains. The utilization of the auxiliary domain knowledge improves the generalization ability of the learned topic model. Experiments on Weizmann and KTH human action databases suggest the effectiveness of the proposed TTM for cross-domain human action recognition.**

***Index Terms*—Bag-of-words, cross-domain learning, human action recognition, topic models.**

I. Introduction

VIDEO-BASED human action recognition has received increasing attention nowadays and plays an important role in practical applications, e.g., video surveillance and abnormal detection systems. A dozen of methods for human action recognition have been proposed in the past years [4], [8], [9], [14], [19], [25], [36]. To name a few, Laptev and his colleagues [22], [23], [32] represented action videos by extracting spatial–temporal local features and recognized actions by using support vector machines (SVMs). Niebles *et al.* [27] developed an unsupervised learning method for human action recognition by exploiting topic models, e.g., probabilistic latent semantic indexing (pLSI) [17] and latent Dirichlet allocation (LDA) [5]. Recently, Wang and Mori [43] have proposed a semilatent topic model for human action recognition, which introduces supervised information to LDA for subsequent recognition. Empirical studies have shown that these conventional methods have achieved promising recognition performance when the amount of videos for model training is sufficient.

However, in many practical scenarios, the amount of available videos is insufficient to train a robust model for recognition. For example, it is impossible for a newly installed video surveillance system to collect sufficient amount of "clean" and "precisely" labeled training videos in a short period. Therefore, the aforementioned action recognition methods cannot work well. On the other hand, it is always possible to collect a large amount of human action videos elsewhere. Although these videos are generally unlabeled and may not be directly relevant to the current recognition task, it is possible to extract useful information from these videos and use them to boost the current recognition task. In this paper, this treatment is termed cross-domain human action recognition, wherein the target domain stands for the insufficient amount of training videos for the current recognition task, and the auxiliary domain stands for the sufficient amount of unlabeled videos collected elsewhere.

In this paper, we propose a transfer topic model (TTM) for cross-domain human action recognition. Although the TTM is built upon topic models similar to models used in [27] and [43], there are two important differentiae that make the TTM learn useful information from the auxiliary domain to boost the recognition task in the target domain. In particular, it first uses a cross-domain bag-of-words video representation, wherein the visual words obtained from the auxiliary domain are directly used to represent the target domain videos. Second, and more importantly, by assuming an action is a mixture of elementary movements, i.e., topics, the TTM uses the learned topics from the auxiliary domain to regularize the topic learning in the target domain. The regularization is the summation of Kullback–Leibler divergences between topic pairs of the two domains. When the training videos in the target domain are insufficient, this regularization serves as an inductive bias for the topic learning and helps improve the generalization ability of the learned topics. We summarize our contributions here.

1) We study the problem of human action recognition from a new perspective, i.e., through transfer learning. It is particularly valuable for the scenario that the target domain has limited data, whereas large relative data are available in the auxiliary domain.
2) To model the transfer human action recognition problem, we propose the TTM. In particular, the Kullback–Leibler divergence regularization is used for knowledge transfer cross domains.
3) An algorithm based on the variational method is derived for parameter estimation of the proposed TTM and predictions on new video frames.

The rest of this paper is organized as follows: In Section II, we present the TTM. In Section III, we derive an expectation–maximization (EM) algorithm for TTM parameter estimation. Section IV presents a simple toy to show the working principle of the TTM. Real data experimental results on

Manuscript received May 4, 2010; revised March 31, 2011; accepted July 25, 2011. This paper was recommended by Associate Editor Y. Fu.

W. Bian and D. Tao are with Centre for Quantum Computation and Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology, Sydney, Sydney, N.S.W. 2007, Australia (e-mail: wei.bian@student.uts.edu.au; dacheng.tao@uts.edu.au).

Y. Rui is with the Microsoft China R&D (CRD) Group, Beijing 100190, China (e-mail: yongrui@microsoft.com).

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/TSMCB.2011.2166761

Weizmann and KTH action databases are reported in Section V. Section VI concludes this paper.

II. Related Work

Video-based human action recognition has received significant attention in recent studied from both computer vision and machine learning areas. Various features were developed for action representation from videos. For instance, Cutler and Davis [11] utilized self-similarity and time–frequency techniques to present periodic motions. Efros *et al.* [14] proposed the motion descriptor based on optical-flow measurements in a spatial–temporal volume to characterize the stabilized human figures. Bobick and Davis [8] proposed to use temporal templates for motion and shape representation. Furthermore, Laptev and Lindeberg [21] developed the space–time interest points by extending the spatial interest points into the spatial–temporal domain as features for action representation. It overcomes limitations of traditional approaches, e.g., optic flow, which suffers from rapid changes of human motions [8]. Recently, specifically motivated by new techniques from machine learning area, a number of model-based methods have been developed for human action recognition. Topic models [3], [5], [6] have been most successfully applied. For instance, Niebles *et al.* [27] proposed an unsupervised learning method for human action recognition based on pLSI [17], which models human actions with intermediate topics. Most recently, Wang *et al.* [43] have also studied the problem of human action recognition by modifying topic models, namely, LDA [5] and correlated topic models (CTMs) [6], with semilatent variations. Both these studies showed encouraging potentials of topic models on human action recognition.

One fundamental assumption in traditional learning is that training and testing data are sampled from an identical distribution [29], [34], [35], [38]. However, this assumption is not always valid. For example, when the data for one learning task (called the target domain) are limited and we want to use the data from the auxiliary domain to improve the performance of the leaning task at hand, traditional learning algorithms are inapplicable because the data distributions of the target and the auxiliary domains can be different. Transfer learning emerges as a new learning strategy to deal with such knowledge transfer problem. The core problem in transfer learning is how to transfer knowledge. Concerning this key point, many strategies with distinct intuitions have been proposed, e.g., the sample selection bias correction [16], [18], [44] uses the reweighting method to obtain an approximately unbiased distribution for learning, self-taught learning approach finds new feature representations to improve target domain learning performance [12], [31], and the sharing common latent space or prior distributions idea for realizing the knowledge transfer cross domains [1], [30], [33]. A comprehensive survey on transfer learning, including its categories and algorithms, can be found in [29]. Our study integrates transfer learning with topic models in the background of human action recognition. The method we use for knowledge transfer is related to self-taught learning and sharing prior distribution approaches. However, our method is more intuitive and flexible by using the knowledge learned from the auxiliary domain as prior information for target domain learning.

III. TTM

We propose a TTM for cross-domain human action recognition. The TTM contains two aspects, namely, cross-domain bag-of-words representation and regularized topic estimation.

A. Cross-Domain Bag-of-Words Representation

The bag-of-words model is popular for action video representation [22], [27]. A codebook containing the visual words is obtained by clustering the extracted spatial–temporal local features from a collection of videos. Then, a video can be represented by quantizing the spatial–temporal local features according to the codebook. For cross-domain bag-of-words representation, we have the following procedure. First, we use the Harris3D detector and the histograms of optic flow (HOF) descriptor [21] to extract spatial–temporal local features from videos in the auxiliary domain. Afterward, k-means is used to cluster these local features into visual words, which gives an auxiliary domain a codebook $\text{CB} = \{c_1, c_2, \ldots, c_V\}$, where each c_j represents a visual word, and V is the number of words. For each video v in the target domain, the same spatial–temporal detector Harris3D and the descriptor HOF are used to extract spatial–temporal local features $F = \{f_1, f_2, \ldots, f_N\}$, wherein N is the feature number, and finally, a bag of words $\{w_1, w_2, \ldots, w_N\}$ can be obtained by quantization, i.e., $w_n = j$ if $\|f_n - c_j\| = \min_{1 \le j' \le V} \|f_n - c_{j'}\|$.

Recent studies have shown that direct dense sampling from the spatial–temporal volume of videos, instead of using spatial–temporal detectors, offers promising recognition performance [40]. In this paper, however, we prefer to use detectors, e.g., Harris 3D, because the improvement by using dense sampling is limited, and detectors help extract features most related to movements, which saves computational cost. Furthermore, a codebook is generally attained via vector quantization (k-means) [23], [27] on a large set of the extracted spatial–temporal local features and is quite time consuming. The reuse of the auxiliary domain codebook can save the time of learning a codebook for the target domain video representation.

B. Transfer Topic Estimation

Recent works have shown the effectiveness of topic models for human action recognition. It is suggested that recognition in the topical space is more robust than previous recognition methods [27], [43], e.g., by using SVM [22], that are conducted in the raw feature (bag-of-words) space. However, due to statistical characterization, when the training data are insufficient, it is hard to obtain a set of reliable topics for the representation of unseen data (videos in the target domain). We address this problem by using transfer learning, which has become a hot research topic in recent years [1], [30], [37], [39]. Specifically, in the TTM, we propose a transfer topic estimation method, which utilizes the knowledge extracted from the auxiliary domain as an inductive bias for target domain learning. Fig. 1 shows the graphical representation of the proposed TTM, wherein for the auxiliary domain, standard LDA are trained to obtain auxiliary

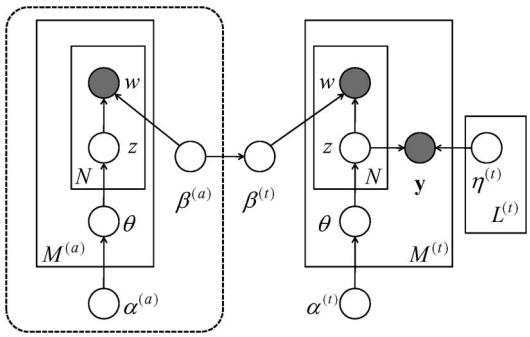

Fig. 1. Graphical representation of the transfer topic estimation: the left part marked by a dashed rectangle stands for the auxiliary domain learning, and the right part is the target domain learning.

domain topics $\beta^{(a)}$ and then use $\beta^{(a)}$ as prior information to regularize the learning of target domain topics $\beta^{(t)}$.

Human action recognition is usually deemed as a multiclass classification, and thus, we exploit the error-correcting code (ECC) method, which has been widely used for multiclass classifier design [2], [13]. Specifically, for the C action concepts in the target domain, we generate C distinct L-length binary, i.e., ECC $= \{e_1, e_2, \ldots, e_C\}$, where each binary vector $e_c \in \{0,1\}^L$ represents an action concept. It has been shown that, for adequate code bits L, randomly generated ECC can work sufficiently well for multiclass classification [2]. By using this ECC, a video v can be labeled by an L-length vector $\mathbf{y} = (y_1, y_2, \ldots, y_L)$, which is equal to e_i if video v is the ith action concept.

Suppose a K-topic model, where the K topics are denoted by $\beta = \{\beta_1, \beta_2, \ldots, \beta_K\}$, then for a video v associated with a pair $(\mathbf{w}, \mathbf{y})$, wherein $\mathbf{w} = \{w_1, w_2, \ldots, w_N\}$ is its cross-domain bag-of-words representation and $\mathbf{y} = (y_1, y_2, \ldots, y_L)$ is its coded label, we have the following generation procedures.

1) Sample topic proportions θ from Dirichlet distribution $\text{Dir}(\theta|\alpha)$;
2) For each of N words w_n
   a) sample a topic z_n from multinomial distribution $\text{Multi}(z_n|\theta)$;
   b) sample a word w_n from multinomial distribution $\text{Multi}(w_n|\beta_{z_n})$;
3) Sample each of L binary labels y_l from binomial distribution $\text{Bi}(y_l|0.5 + \eta_l^T \bar{z})$, where $\bar{z} = (1/N)\sum_{n=1}^{N} z_n$, and $-0.5 \le \eta_l \le 0.5$.

The probabilities used in the preceding generation procedure are given as

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \tag{1}$$

$$p(z_n|\theta) = \prod_{k=1}^{K} \theta_k^{z_{n,k}} \tag{2}$$

$$p(w_n|\beta_{z_n}) = \prod_{j=1}^{V} \beta_{z_n,j}^{w_{n,j}} \tag{3}$$

$$p\left(y_l|\eta_l^T \bar{z}\right) = \left(0.5 + \eta_l^T \bar{z}\right)^y \left(0.5 - \eta_l^T \bar{z}\right)^{1-y}. \tag{4}$$

The aforementioned notations are explained in the following. Dirichlet parameter α is a positive vector in R^K; $\theta \in R^K$ is a multinomial probability with nonnegative entries and a sum of 1; each topic index $z_n \in \{0,1\}^K$ is an indicator vector with only one entry of 1 and the rest of all 0; each $w_n \in \{0,1\}^V$ is also an indicator vector with only one entry of 1 and the rest of all 0; and classification parameter $\eta_l \in R^K$ used in the binomial satisfies $-0.5 \le \eta_l \le 0.5$. In addition, we use a subscript after the comma to denote the entry of a vector, e.g., $z_{n,j}$ denotes the jth entry in vector z_n, so does $w_{n,j}$ for $\beta_{z_n,j}$.

According to the aforementioned generation procedure, the joint probability density of $(\mathbf{w}, \mathbf{y}, \mathbf{z}, \theta)$ is given by

$$p(\mathbf{w}, \mathbf{y}, \mathbf{z}, \theta|\alpha, \beta, \eta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|\beta_{z_n}) \prod_{l=1}^{L} p\left(y_l|\eta_l^T \bar{z}\right). \tag{5}$$

Furthermore, by integrating over θ and summing over $\mathbf{z} = [z_1, z_2, \ldots, z_N]$, the marginal probability density of $(\mathbf{w}, \mathbf{y})$ is given by

$$p(\mathbf{w}, \mathbf{y}|\alpha, \beta, \eta) = \int_{\theta} p(\theta|\alpha) \sum_{z_1, \ldots, z_N} \prod_{n=1}^{N} p(z_n|\theta) p(w_n|\beta_{z_n}) \times \prod_{l=1}^{L} p\left(y_l|\eta_l^T \bar{z}\right) d\theta. \tag{6}$$

Suppose there are M videos $(\mathbf{w}_m, \mathbf{y}_m)$, $m = 1, 2, \ldots, M$, in target domain $D^{(t)}$, the probability of all M videos is given by

$$p\left(D^{(t)}|\alpha, \beta, \eta\right) = \prod_{m=1}^{M} p(\mathbf{w}_m, \mathbf{y}_m|\alpha, \beta, \eta) \tag{7}$$

where each of the independent terms is determined by (6).

To estimate parameters (α, β, η), we use the following regularized log likelihood:

$$\left(\alpha^{(t)}, \beta^{(t)}, \eta^{(t)}\right) = \arg\max_{\alpha,\beta,\eta} \sum_{m=1}^{M} \log\, p(\mathbf{w}_m, \mathbf{y}_m|\alpha, \beta, \eta) + \lambda \sum_{k=1}^{K} \sum_{j=1}^{V} \beta_{k,j}^{(a)} \log \frac{\beta_{k,j}}{\beta_{k,j}^{(a)}} \tag{8}$$

where $\beta^{(a)}$ denotes the auxiliary domain topics, and λ is a weighting parameter between the log likelihood and the regularization. The regularization term in (8) is the negative of the summation of the relative entropy between each pair of topics in $\beta^{(a)}$ and β, i.e.,

$$\sum_{k=1}^{K} \sum_{j=1}^{V} \beta_{k,j}^{(a)} \log \frac{\beta_{k,j}}{\beta_{k,j}^{(a)}} = -\sum_{k=1}^{K} KL\left(\beta_k^{(a)} \| \beta_k\right). \tag{9}$$

Thus, it penalizes the dissimilarity of the topics between two domains. By the regularization, the topic learning for the target domain becomes possible even with a small amount of training videos.

IV. Parameter Estimation

Here, we derive an EM algorithm to solve the aforementioned regularized maximum-likelihood estimation. For an EM algorithm, generally, it needs to construct a lower bound for the log likelihood, which requires an explicit posterior distribution of the latent variables. However, for a topic model, posterior $p(\theta, \mathbf{z}|\mathbf{w}, \mathbf{y}, (\alpha, \beta, \eta))$ is hard to obtain [5], and thus, the variational method is usually exploited. According to the variational method [20], we restrict (approximate) posterior $p(\theta, \mathbf{z}|\mathbf{w}, \mathbf{y}, (\alpha, \beta, \eta))$ to a fully factorized distribution family, i.e.,

$$q(\theta, \mathbf{z}|\gamma, \Phi) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi_n) \tag{10}$$

where γ and $\Phi = [\phi_1, \phi_2, \ldots, \phi_N]$ are variational parameters, $q(\theta|\gamma)$ is a Dirichlet distribution, and $q(z_n|\phi_n)$ is a multinomial distribution. The best variational (approximate) distribution $q^*(\theta, \mathbf{z}|\gamma^*, \Phi^*)$ is given by the one within the family [see (10)] that minimizes the KL divergence between the approximation and the true posterior $p(\theta, \mathbf{z}|\mathbf{w}, \mathbf{y}, (\alpha, \beta, \eta))$, i.e.,

$$\begin{aligned} &q^*(\theta, \mathbf{z}|\gamma^*, \Phi^*; (\alpha, \beta, \eta)) \\ &\quad = \arg \min_{q(\theta, \mathbf{z}|\gamma, \Phi)} KL\left(q(\theta, \mathbf{z}|\gamma, \Phi) \| p\left(\theta, \mathbf{z}|\mathbf{w}, \mathbf{y}, (\alpha, \beta, \eta)\right)\right). \end{aligned} \tag{11}$$

The preceding minimization can be solved by (12), shown at the bottom of the page, and details to obtain them are given in Appendix B. Note that $\circ$ denotes the elementwise multiplication between vectors.

To estimate the model parameters, we need the following M-step:

$$\begin{aligned} (\alpha, \beta, \eta)^{[k+1]} = \arg \max_{\alpha, \beta, \eta} &\sum_{m=1}^{M} E_{q_m^*(\theta, \mathbf{z}|\gamma_m^*, \Phi_m^*; (\alpha, \beta, \eta)^{[k]})} \\ &\times \log p(\mathbf{w}_m, \mathbf{y}_m, \theta, \mathbf{z}|\alpha, \beta, \eta) \\ &+ \lambda \sum_{k=1}^{K} \sum_{j=1}^{V} \beta_{k,j}^{(a)} \log \frac{\beta_{k,j}}{\beta_{k,j}^{(a)}} \end{aligned} \tag{13}$$

where $(a, \beta, \eta)^{[k]}$ is the output of the kth iteration round. We derive the solution of (13) in Appendix C. Particularly, α is given by

$$\begin{aligned} \alpha^{[k+1]} = \arg \max_{\alpha} &M \log \Gamma\left(\sum_{i=1}^{K} \alpha_i\right) - M \sum_{i=1}^{K} \log \Gamma(\alpha_i) \\ &+ \sum_{m=1}^{M} \sum_{i=1}^{K} (\alpha_i - 1) \left(\Psi\left(\gamma_{m,i}^*\right) - \Psi\left(\sum_{i=1}^{K} \gamma_{m,i}^*\right)\right) \end{aligned} \tag{14}$$

which can be solved by the Newton method detailed in [5] or the fixed-point method detailed in [26]. For β, we have the closed form

$$\beta_{i,j}^{[k+1]} \propto \sum_{m=1}^{M} \sum_{n=1}^{N_m} \phi_{m,n,i}^* \delta(w_{m,n,j} = 1) + \lambda \beta_{i,j}^{(a)}. \tag{15}$$

For η, it is given by

$$\eta_l^{[k+1]} = \arg \min_{-0.5 \le \eta_l \le 0.5} 0.5 \eta_l^T A \eta_l - b^T \eta_l \tag{16}$$

where $A = \sum_{m=1}^{M} \left(1/N_m^2\right) \left(\sum_{n=1}^{N_m} \sum_{n' \neq n} \phi_{m,n} \phi_{m,n'}^T + \sum_{n=1}^{N_m} \mathrm{diag}(\phi_{m,n})\right)$, and $b = \sum_{m=1}^{M} (3y_{m,l} - 1.5)(1/N_m) \times \sum_{n=1}^{N} \phi_{m,n}$.

Based on the learned $(\alpha^{(t)}, \beta^{(t)}, \eta^{(t)})$, we can predict the corresponding label of a new action video. For a new video v^* represented by $\mathbf{w}$, we have

$$\begin{aligned} \mathbf{y}^* &= E_{\mathbf{y}}\left(\mathbf{y}|\mathbf{w}, \alpha^{(t)}, \beta^{(t)}, \eta^{(t)}\right) \\ &= E_{\mathbf{z}}\left(E_{\mathbf{y}}(\mathbf{y}|\mathbf{z})|\mathbf{w}, \alpha^{(t)}, \beta^{(t)}, \eta^{(t)}\right) \\ &= \left(\eta_1^{(t)}, \eta_2^{(t)}, \ldots, \eta_L^{(t)}\right)^T E_{\mathbf{z}}\left(\bar{z}|\mathbf{w}, \alpha^{(t)}, \beta^{(t)}, \eta^{(t)}\right). \end{aligned} \tag{17}$$

The calculation of $E_{\mathbf{z}}(\bar{z}|\mathbf{w}, \alpha^{(t)}, \beta^{(t)}, \eta^{(t)})$ requires the exact posterior probability of $\mathbf{z} = [z_1, z_2, \ldots, z_N]$, which is hard to calculate as mentioned earlier. Thus, we use variational distribution $q^*(\theta, \mathbf{z}|\gamma^*, \Phi^*)$ defined in (11) to get an approximation of $E_{\mathbf{z}}(\bar{z}|\mathbf{w}, \alpha^{(t)}, \beta^{(t)}, \eta^{(t)})$, and thus, we can obtain the prediction

$$\mathbf{y}^* \approx \left(\eta_1^{(t)}, \eta_2^{(t)}, \ldots, \eta_L^{(t)}\right)^T \frac{1}{N} \sum_{n=1}^{N} \phi_n^*. \tag{18}$$

Then, comparing $\mathbf{y}^*$ with ECC $= \{e_1, e_2, \ldots, e_C\}$, we can find the action concept of v^*.

$$\begin{cases} \gamma = \alpha + \sum\limits_{n-1}^{N} \phi_n \\ \phi_n \propto \beta_{:,w_n} \circ \exp\left\{\Psi(\gamma) - \Psi\left(\sum\limits_{i=1}^{K} \gamma_i\right) + (1/N) \sum\limits_{l=1}^{L} (3y_l - 1.5)\eta_l \right. \\ \qquad \left. - (1/N^2)\left(\sum\limits_{l=1}^{L} \sum\limits_{n' \neq n} \phi_{n'}^T \eta_l \eta_l + 0.5(\eta_l \circ \eta_l)\right)\right\}, \qquad n = 1, 2, \ldots, N \end{cases} \tag{12}$$

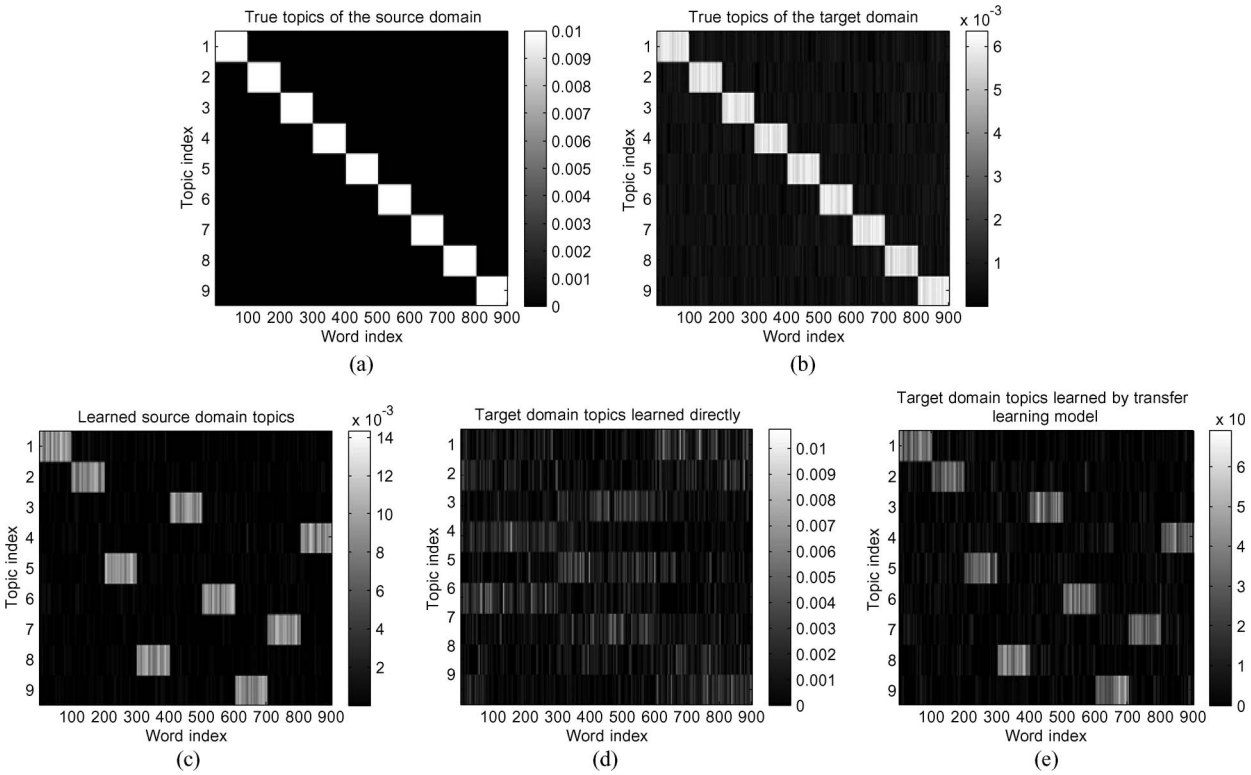

Fig. 2. Topical structure plots for the toy example. (a) True topics of the auxiliary domain. (b) True topics of the target domain. (c) Topics directly learned from the auxiliary domain data. (d) Topics directly learned from the target domain data. (e) Target domain topics learned by using the transfer learning model.

Remarks: In the aforementioned EM-type algorithm for parameter estimation, the major computational cost is in solving (11) via iterative updating formulas (12). In (12), updating of γ is ignorable, and updating of ϕ_n requires a computational cost of order $O(NL^2 + K)$, where NL^2 is for the computation within the exponent, and K is for the elementwise multiplication. Since there are ϕ_n's, $n = 1, 2, \ldots, N$, to be calculated, the computation cost for (12) is $O(N^2L^2 + NK)$. Suppose M documents have on average N words and T iterations are required for convergence, then the total computational cost for solving is $O(TMN^2L^2 + TMNK)$.

V. Toy Example

We use a toy example to illustrate the working principle of the proposed TTM. In particular, we show that the learned topics from the auxiliary domain can facilitate the topic learning in the target domain. In the example, the auxiliary domain has a topical structure of nine topics over a vocabulary of 900 words, whereas the target domain has a similar topical structure, but the topics are slight variations from the auxiliary domain ones by adding additional noise. Fig. 2(a) and (b) shows the topics of the two domains. We suppose a three-class classification problem, and for each class, we generate 200 documents for the auxiliary domain and 50 documents for the target domain. Note the amount of labeled documents in the target domain is insufficient. This design meets the scenario of transfer learning. The topics learned from the auxiliary domain are shown in Fig. 2(c). Due to the sufficient amount of labeled documents in the target domain, the learned topics are consistent with the true topics shown in Fig. 2(a). On the contrary, it is difficult to directly learn topics from data in the target domain. This is evident by comparing the learned topics in Fig. 2(d) against the true ones shown in Fig. 2(b). However, by using the proposed TTM, we relearn topics of the target domain by using the learned topics of the auxiliary domain as regularization, and Fig. 2(e) shows that the learned topics can be significantly improved.

Furthermore, the topics obtained by the TTM are more discriminative. Fig. 3 shows the embedded (by principal component analysis (PCA), from nine to two dimensions) proportions of the target domain documents under the true topics, the topics learned by using only the target domain data, and the topics learned by the TTM. One can see that the three classes are better separated by using the topics learned by the TTM.

VI. Experiments

We evaluate the proposed TTM by using a cross-domain human action recognition experiment from the Weizmann database [4] to the KTH database [32]. These two databases are frequently used for testing human action recognition methods [23], [27], [32]. In addition, the actions in the two databases share some common properties, which are particularly suitable for our cross-domain human action recognition experiments. The Weizmann human action database contains ten classes of human actions performed by nine people and has 90 videos in total. The KTH human motion data set contains six classes of human actions, including walking, jogging, running, boxing, hand waving, and hand clapping. Each action is performed by

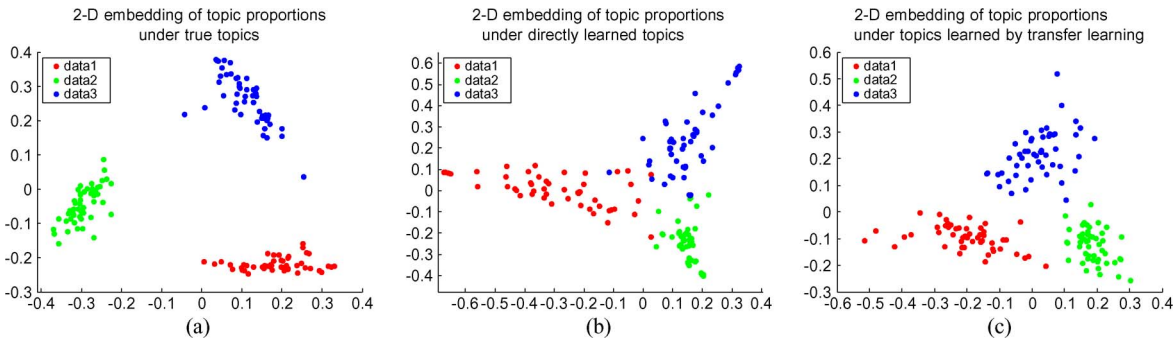

Fig. 3. Two-dimensional embedding for the topic proportions of the target domain data. (a) Topic proportions under the true topics. (b) Topic proportions under the directly learned topics. (c) Topic proportions under the topics learned by using the transfer learning model.

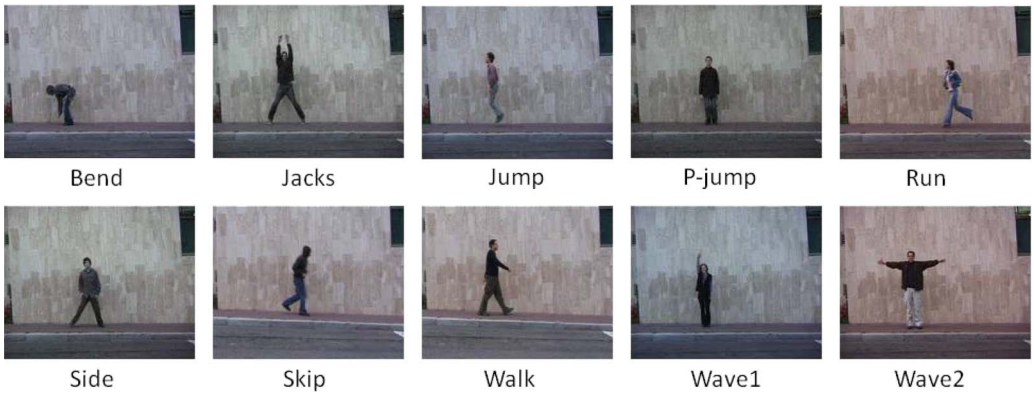

Fig. 4. Sample frames from the Weizmann human action database [4].

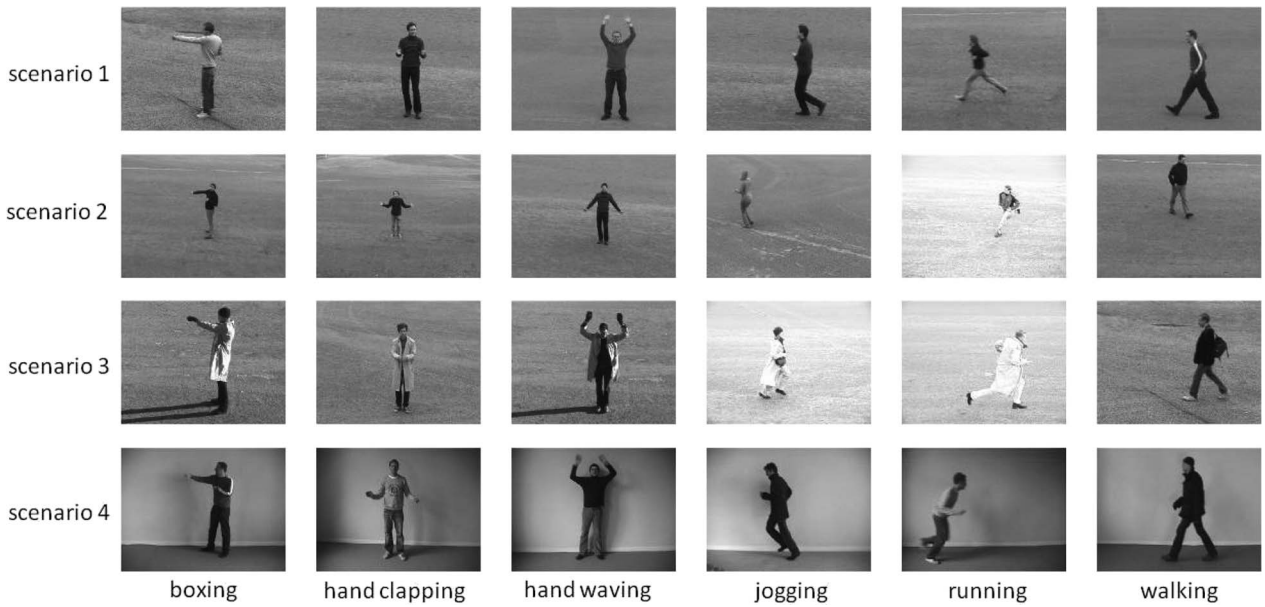

Fig. 5. Sample frames from the KTH human action database [32].

25 subjects in four scenarios. Example frames of both databases are shown in Figs. 4 and 5.

In our experiments, we use the Weizmann database as the auxiliary domain. We first learn a codebook for cross-domain bag-of-words representation by using this database. Specifically, the Harris3D detector [21] is used to detect the spatial–temporal interest points from the videos, and then, the HOF descriptor [21] is used to extract local features on the detected points. In all, 10 562 HOF features of 90 dimensions are obtained from all videos in the database. Then, k-means is used to cluster the features into 1500 visual words, which comprise the auxiliary domain codebook. Based on the bag-of-words representation of the videos, we conduct LDA to extract topics from the database. Particularly, five sets of topics are obtained by varying topic numbers from 20 to 40 with an interval of 5.

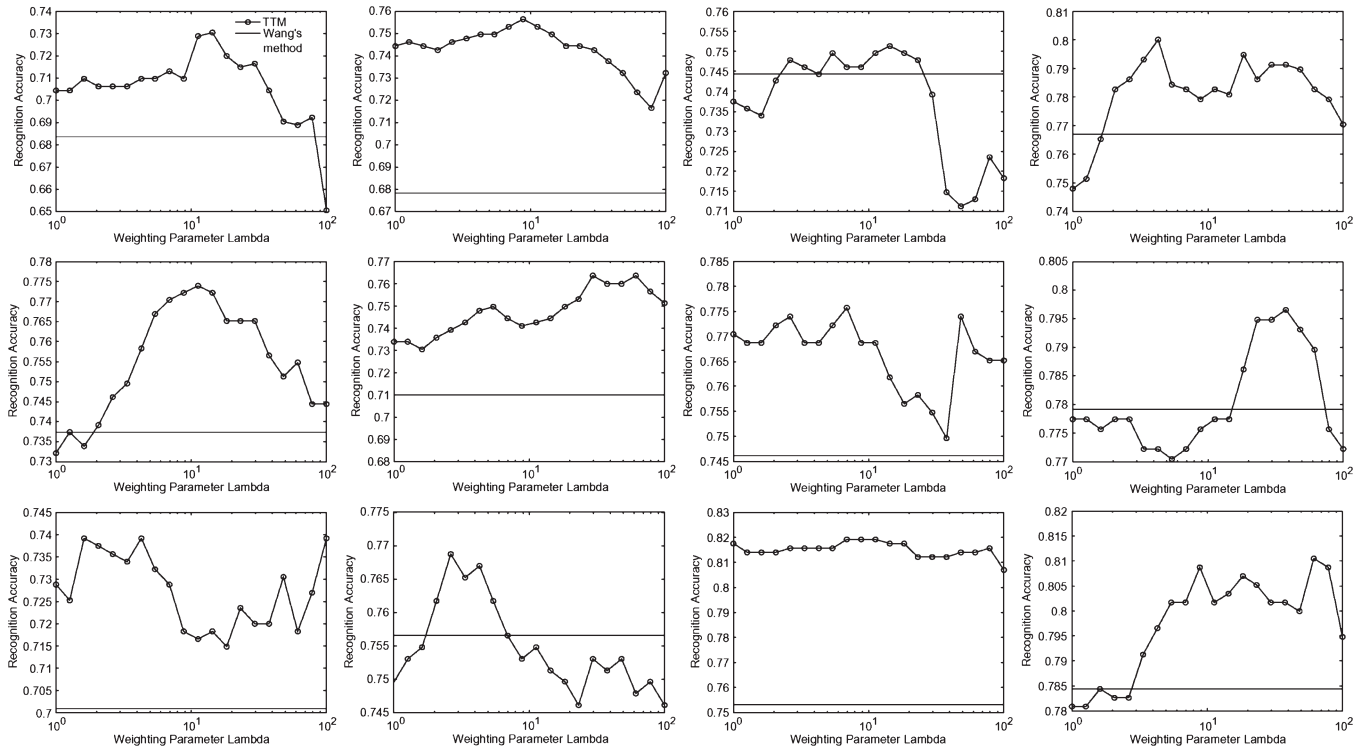

Fig. 6. Comparison of different methods: the proposed TTM, Wang's method [43], Niebles's method [27], and SVM-based recognition.

We then do action recognition on the KTH database, which is referred to as the target domain. To simulate a scenario of limited training data, we use the videos of one person for training and use the rest videos from others for test. Similarly, the Harris3D detector and the HOF descriptor are used to extract spatial–temporal local features, and the auxiliary domain codebook is used to get the bag-of-words of the target domain videos. Then, we use the proposed TTM with the auxiliary domain topics as regularization to perform action recognition.

We compare TTM against three state-of-the-art human action recognition methods. The first one is the SVM-based method [32], where a linear form is used and the parameter is tuned by leave-one-out cross validation on the training set, whereas the other two are topic model based, which are referred to as Niebles's method [27] and Wang's method. The average recognition rates (over 25 training cases) of different methods are shown in Fig. 6. It is shown that the proposed TTM and Wang's method [43] performed better than the other two, and with a suitable number of topics, the TTM can outperform Wang's method. The requirement of a suitable topic number is reasonable because too few topics cannot reliably represent human actions while too many of topics will lose the dimension reduction function of topic models.

To investigate the effectiveness of transfer learning, we take the 30-topic case as an example and show the recognition performance against weighting parameter λ in the proposed TTM. Results on 12 out of the 25 training cases are shown in Fig. 7. In most cases, recognition performance first increases and then decreases along with the increase in transfer learning parameter λ, which reflects that cross-domain knowledge is helpful to improve the performance of current recognition task while, of course, much knowledge transfer (or regularization) should not be dominated. In addition, the best parameter λ varies on different training examples due to the diversities among the videos of different people.

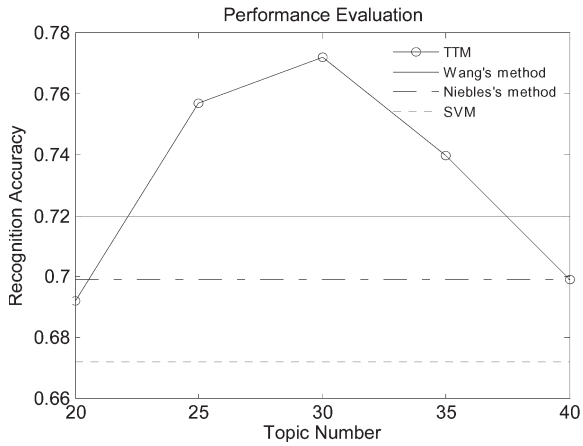

Fig. 7. Performance evaluation by varying transfer learning weighting parameter λ.

VII. Conclusion

In this paper, we have proposed a TTM for cross-domain human action recognition. The TTM consists of cross-domain bag-of-words representation and regularized target domain topic estimation. It overcomes the limitation of existing action recognition methods in which both video representation and model learning are domain specific, i.e., it saves much computational cost in obtaining a codebook for bag-of-words representation and improves the recognition performance by utilizing auxiliary domain knowledge for scenarios where only

TABLE I
VARIABLES AND NOTATIONS IN THE TTM AND PARAMETER ESTIMATION

$(\mathbf{w}, \mathbf{y})$	Bag of words representation and label of video frame
$\alpha^{(a)}$, $\alpha^{(t)}$	Dirichlet parameters for auxiliary and target domains
$\beta^{(a)}$, $\beta^{(t)}$	Topics for auxiliary and target domains
$\eta^{(t)}$	Classification parameter for the target domain
θ	Topic proportions
z	Topic index
γ	Variational parameter for θ
ϕ	Variational parameter for z
$KL(\cdot \|\| \cdot)$	Kullback-Leibler divergence
$\log \Gamma(\cdot)$	Log-Gamma function
$\Psi(\cdot)$	Digamma function
$\delta(\cdot)$	Indicator function

insufficient amount of training videos in the target domain is available. For the implementation of the TTM, we derived an efficient EM learning algorithm based on the variational approximation. Experiments on the KTH and Weizmann databases suggested the effectiveness of the TTM for cross-domain human action recognition. It has been shown that the TTM outperformed several state-of-the-art human action recognition methods by utilizing cross-domain knowledge.

APPENDIX A

The variables and notations in the TTM and parameter estimation are shown in Table I.

APPENDIX B

We show the details for solving (11). First, we expand the KL divergence in (11) as

$$\begin{aligned}
&KL(q(\theta, \mathbf{z}|\gamma, \Phi)||p(\theta, \mathbf{z}|\mathbf{w}, \mathbf{y}, (\alpha, \beta, \eta))) \\
&= \log\, p(\mathbf{w}, \mathbf{y}|(\alpha, \beta, \eta)) - E_{q(\theta|\gamma)} \log\, p(\theta|\alpha) \\
&\quad - \sum_{n=1}^{N} E_{q(\theta|\gamma)} \log\, p(z_n|\theta) - \sum_{n=1}^{N} E_{q(z_n|\phi_n)} \log\, p(w_n|\beta_{z_n}) \\
&\quad - \sum_{l=1}^{L} E_{q(\mathbf{z}|\Phi)} \log\, p\left(\mathbf{y}|\eta_l^T \bar{z}\right) + E_{q(\theta|\gamma)} \log\, q(\theta|\gamma) \\
&\quad + \sum_{n=1}^{N} E_{q(z_n|\phi_n)} \log\, q(z_n|\phi_n) \\
&\approx \log\, p(\mathbf{w}, \mathbf{y}|(\alpha, \beta, \eta)) - \log\, \Gamma\left(\sum_{i=1}^{K} \alpha_i\right) \\
&\quad + \sum_{i=1}^{K} \log\, \Gamma(\alpha_i) - \sum_{i=1}^{K} (\alpha_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\sum_{i=1}^{K} \gamma_i\right)\right) \\
&\quad - \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni}\left(\Psi(\gamma_i) - \Psi\left(\sum_{k=1}^{K} \gamma_i\right)\right) \\
&\quad - \sum_{n=1}^{N} \sum_{i=1}^{K} \sum_{j=1}^{V} \phi_{ni} \delta(w_{n,j} = 1) \log\, \beta_{ij} \\
&\quad + 0.625L - (1/N) \sum_{l=1}^{L} (3y_l - 1.5) \eta_l^T \sum_{n=1}^{N} \phi_n \\
&\quad + 0.5(1/N^2) \sum_{l=1}^{L} \eta_l^T \left(\sum_{n=1}^{N} \sum_{m \neq n} \phi_n \phi_m^T + \sum_{n=1}^{N} \mathrm{diag}(\phi_n) \right) \eta_l \\
&\quad + \log\, \Gamma\left(\sum_{i=1}^{K} \gamma_i\right) - \sum_{i=1}^{K} \log\, \Gamma(\gamma_i) \\
&\quad + \sum_{i=1}^{k} (\gamma_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\sum_{i=1}^{K} \gamma_i\right)\right) \\
&\quad + \sum_{n=1}^{N} \sum_{i=1}^{K} \phi_{ni} \log\, \phi_{ni} \qquad (19)
\end{aligned}$$

where we apply an approximation on $\log\, p(\mathbf{y}|\eta_l^T \bar{z})$ by using $\log(1 - x) \approx -x - 0.5x^2$ for any $0 > x > 1$, i.e.,

$$\begin{aligned}
&\log\, p\left(y_l|\eta_l^T \bar{z}\right) \\
&\quad = y_l \log\, (0.5 + \eta_l^T \bar{z})) + (1 - y_l) \log\left(0.5 - \eta_l^T \bar{z}\right) \\
&\quad \approx -0.625 + (3y_l - 1.5)\eta_l^T \bar{z} - 0.5\left(\eta_l^T \bar{z}\right)^2 \qquad (20) \\
&E_{q(\mathbf{z}|\Phi)} \log\, p\left(\mathbf{y}|\eta_l^{[k]T} \bar{z}\right) \\
&\quad \approx -0.625 + (3y_l - 1.5)\eta_l^T E_{q(\mathbf{z}|\Phi)} \bar{z} - 0.5 \eta_l^T E_{q(\mathbf{z}|\Phi)}(\bar{z}\bar{z}^T)\eta_l \\
&\quad = -0.625 + (3y_l - 1.5)\eta_l^T (1/N) \sum_{n=1}^{N} \phi_n \\
&\qquad - 0.5(1/N^2)\eta_l^T \left(\sum_{n=1}^{N} \sum_{n' \neq n} \phi_n \phi_{n'}^T + \sum_{n=1}^{N} \mathrm{diag}(\phi_n) \right) \eta_l. \qquad (21)
\end{aligned}$$

To minimize (19) with respect to γ, we isolate related terms and get

$$\begin{aligned}
O_{[\gamma]} = &- \sum_{i=1}^{K} (\alpha_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\sum_{i=1}^{K} \gamma_i\right)\right) \\
&- \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni}\left(\Psi(\gamma_i) - \Psi\left(\sum_{k=1}^{K} \gamma_i\right)\right) \\
&+ \log\, \Gamma\left(\sum_{i=1}^{K} \gamma_i\right) - \sum_{i=1}^{K} \log\, \Gamma(\gamma_i) \\
&+ \sum_{i=1}^{k} (\gamma_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\sum_{i=1}^{K} \gamma_i\right)\right) \qquad (22)
\end{aligned}$$

and by setting $\partial O_{[\gamma]}/\partial \gamma = 0$, we get the updating equation in (12). To minimize (19) with respect to ϕ_n, we isolate related

terms and use the Lagrangian multiplier method, i.e.,

$$\begin{aligned} O_{[\phi_n]} = &- \sum_{i=1}^{K} \phi_{ni} \left(\Psi(\gamma_i) - \Psi\left(\sum_{k=1}^{K} \gamma_i \right) \right) \\ &- \sum_{i=1}^{K} \sum_{j=1}^{V} \phi_{ni} \delta(w_n = j) \log \beta_{ij} \\ &- (1/N) \sum_{l=1}^{L} (3y_l - 1.5) \eta_l^T \phi_n + 0.5(1/N^2) \sum_{l=1}^{L} \eta_l^T \\ &\times \left(\sum_{n=1}^{N} \sum_{n' \neq n} \phi_n \phi_{n'}^T + \sum_{n=1}^{N} \mathrm{diag}(\phi_n) \right) \eta_l \\ &+ \sum_{i=1}^{K} \phi_{ni} \log \phi_{ni} + \lambda \left(\sum_{i=1}^{K} \phi_{n,i} - 1 \right) \end{aligned} \tag{23}$$

where λ is the multiplier. By setting $\partial O_{[\phi_n]} / \partial \phi_n = 0$, we get the updating equation in (12).

APPENDIX C

We show the details of solving (13). First, similar to (19), we have

$$\begin{aligned} &\sum_{m=1}^{M} E_{q_m^*\left(\theta, \mathbf{z} | \gamma_m^*, \Phi_m^*; (\alpha, \beta, \eta)^{[k]}\right)} \\ &\quad \times \log p\left(\mathbf{w}_m^{(s)}, \mathbf{y}_m^{(s)}, \theta, \mathbf{z} | \alpha, \beta, \eta \right) + \lambda \sum_{k=1}^{K} \beta_k^{(a)} \log \frac{\beta_k}{\beta_k^{(a)}} \\ &\approx M \log \Gamma\left(\sum_{i=1}^{K} \alpha_i \right) - M \sum_{i=1}^{K} \log \Gamma(\alpha_i) \\ &\quad + \sum_{m=1}^{M} \sum_{i=1}^{K} (\alpha_i - 1) \left(\Psi(\gamma_{m,i}) - \Psi\left(\sum_{i=1}^{K} \gamma_{m,i} \right) \right) \\ &\quad + \sum_{m=1}^{M} \sum_{n=1}^{N_m} \sum_{i=1}^{k} \phi_{m,n,i} \left(\Psi(\gamma_{m,i}) - \Psi\left(\sum_{k=1}^{K} \gamma_{m,i} \right) \right) \\ &\quad + \sum_{m=1}^{M} \sum_{n=1}^{N_m} \sum_{i=1}^{K} \sum_{j=1}^{V} \phi_{m,n,i} \delta(w_{m,n,j} = 1) \log \beta_{i,j} \\ &\quad - 0.625LM + \sum_{m=1}^{M} (1/N_m) \sum_{l=1}^{L} (3y_{m,l} - 1.5) \eta_l^T \sum_{n=1}^{N} \phi_{m,n} \\ &\quad - 0.5 \sum_{m=1}^{M} (1/N_m^2) \sum_{l=1}^{L} \eta_l^T \\ &\quad \times \left(\sum_{n=1}^{N_m} \sum_{n' \neq n} \phi_{m,n} \phi_{m,n'}^T + \sum_{n=1}^{N_m} \mathrm{diag}(\phi_{m,n}) \right) \eta_l \\ &\quad + \lambda \sum_{k=1}^{K} \beta_k^{(a)} \log \frac{\beta_k}{\beta_k^{(a)}}. \end{aligned} \tag{24}$$

To maximize (24) with respect to α, we isolate related terms and get

$$\begin{aligned} O_{[\alpha]} = &M \log \Gamma\left(\sum_{i=1}^{K} \alpha_i \right) - M \sum_{i=1}^{K} \log \Gamma(\alpha_i) \\ &+ \sum_{m=1}^{M} \sum_{i=1}^{K} (\alpha_i - 1) \left(\Psi(\gamma_{m,i}) - \Psi\left(\sum_{i=1}^{K} \gamma_{m,i} \right) \right). \end{aligned} \tag{25}$$

To maximize (24) with respect to $\beta_{i,j}$, we isolate related terms and use the Lagrangian multiplier method for $\sum_{j=1}^{V} \beta_{i,j} = 1$, i.e.,

$$\begin{aligned} O_{[\beta]} = &\sum_{m=1}^{M} \sum_{n=1}^{N_m} \sum_{i=1}^{K} \sum_{j=1}^{V} \phi_{m,n,i} \delta(w_{m,n,j} = 1) \log \beta_{i,j} \\ &+ \sum_{i=1}^{K} \lambda_i \left(\sum_{j=1}^{V} \beta_{i,j} - 1 \right) + \lambda \sum_{k=1}^{K} \sum_{j=1}^{V} \beta_{k,j}^{(a)} \log \frac{\beta_{k,j}}{\beta_{k,j}^{(a)}} \end{aligned} \tag{26}$$

where λ is the multiplier. By setting $\partial O_{[\beta]} / \partial \beta_{i,j} = 0$, we get optimal solution (15). To maximize (24) with respect to η_l, we isolate related terms and get

$$\begin{aligned} O_{[\eta_l]} = &\eta_l^T \sum_{m=1}^{M} (3y_{m,l} - 1.5)(1/N_m) \sum_{n=1}^{N} \phi_{m,n} \\ &- 0.5 \eta_l^T \sum_{m=1}^{M} (1/N_m^2) \left(\sum_{n=1}^{N_m} \sum_{n' \neq n} \phi_{m,n} \phi_{m,n'}^T + \sum_{n=1}^{N_m} \mathrm{diag}(\phi_{m,n}) \right) \eta_l \end{aligned} \tag{27}$$

which leads to problem (16) with constraint $-0.5 \leq \eta_l \leq 0.5$.

REFERENCES

[1] J. Baxter, "A model of inductive bias learning," *J. Artif. Intell. Res.*, vol. 12, no. 1, pp. 149–198, Feb. 2000.

[2] A. Berger, "Error-correcting output coding for text classification," in *Proc. IJCAI: Workshop Mach. Learn. Inf. Filter.*, Stockholm, Sweden, 1999.

[3] W. Bian and D. Tao, "Dirichlet mixture allocation," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2009, pp. 711–715.

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

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[6] D. M. Blei and J. D. Lafferty, "Correlated topic models," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2006.

[7] D. M. Blei and J. D. McAuliffe, "Supervised topic models," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2007.

[8] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

[9] A. Bissacco, M.-H. Yang, and S. Soatto, "Detecting humans via their pose," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2007, pp. 169–176.

[10] L. Cao, Z. Liu, and T. S. Huang, "Cross-dataset action recognition," in *Proc. IEEE CVPR*, 2010, pp. 1998–2005.

[11] R. Cutler and L. S. Davis, "Robust real-time periodic motion detection, analysis, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 781–796, Aug. 2000.

[12] W. Dai, G. Xue, Q. Yang, and Y. Yu, "Co-clustering based classification for out-of-domain documents," in *Proc. 13th ACM SIGKDD*, Aug. 2007, pp. 210–219.
[13] T. G. Dietterich and G. Bakiri, "Error-correcting output codes: A general method for improving multiclass inductive learning programs," in *Proc. 9th AAAI Nat. Conf. Artif. Intell.*, T. L. Dean and K. Mckeown, Eds., 1991, pp. 572–577.
[14] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 726–733.
[15] J. Gu, X. Ding, S. Wang, and Y. Wu, "Action and gait recognition from recovered 3D human joints," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 4, pp. 1021–1033, Aug. 2010.
[16] W. Fan, I. Davidson, B. Zadrozny, and P. S. Yu, "An improved categorization of classifier's sensitivity on sample selection bias," in *Proc. 5th IEEE Int. Conf. Data Mining*, 2005, pp. 605–608.
[17] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. 22nd Annu. Int. SIGIR Conf.*, 1999, pp. 50–57.
[18] J. Huang, A. Smola, A. Gretton, K. M. Borgwardt, and B. Scholkopf, "Correcting sample selection bias by unlabeled data," in *Proc. Neural Inf. Process. Syst.*, 2007, pp. 601–608.
[19] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[20] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, "An introduction to variational methods for graphical models," UC Berkeley, Berkeley, CA, Tech. Rep. CSD-98-980, 1998.
[21] I. Laptev and T. Lindeberg, "Space-time interest points," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 432–439.
[22] I. Laptev and P. Pérez, "Retrieving actions in movies," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[23] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE CS Conf. Comput. Vis. Pattern Recog.*, 2008, pp. 1–8.
[24] Y. Liang, S. Shih, A. C. Shih, H. M. Liao, and C. Lin, "Learning atomic human actions using variable-length Markov models," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 1, pp. 268–280, Feb. 2009.
[25] J. L. Little and J. E. Boyd, "Recognizing people by their gait: The shape of motion," *Videre*, vol. 1, no. 2, pp. 1–32, Winter 1998.
[26] T. Minka, "Estimating a Dirichlet distribution," MIT, Cambridge, MA, 2000.
[27] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial–temporal words," *Int. J. Comput. Vis.*, vol. 79, no. 3, pp. 299–318, Sep. 2008.
[28] A. Oikonomopoulos, I. Patras, and M. Pantic, "Spatiotemporal salient points for visual recognition of human actions," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 36, no. 3, pp. 710–719, Jun. 2005.
[29] 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.
[30] R. Raina, A. Y. Ng, and D. Koller, "Constructing informative priors using transfer learning," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 713–720.
[31] 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.*, Jun. 2007, pp. 759–766.
[32] C. Schuldt, L. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. IEEE Int. Conf. Pattern Recog.*, 2004, vol. 3, pp. 32–36.
[33] A. Schwaighofer, V. Tresp, and K. Yu, "Learning Gaussian process kernels via hierarchical Bayes," in *Proc. 17th Annu. Conf. Neural Inf. Process. Syst.*, 2005, pp. 1209–1216.
[34] S. Si, D. Tao, and B. Geng, "Bregman divergence based regularization for transfer subspace learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 7, pp. 929–942, Jul. 2010.
[35] S. Si, D. Tao, and K. P. Chan, "Evolutionary cross-domain discriminative Hessian eigenmaps," *IEEE Trans. Image Process.*, vol. 19, no. 4, pp. 1075–1086, Apr. 2010.
[36] J. Sullivan and S. Carlsson, "Recognizing and tracking human action," in *Proc. Eur. Conf. Comput. Vis.*, 2002, vol. 1, pp. 629–644.
[37] S. Thrun, in "Is learning the n-th thing any easier than learning the first?" in *Proc. Adv. Neural Inf. Process. Syst.*, 1996, vol. 8, pp. 640–646.
[38] X. Tian, D. Tao, and Y. Rui, "Sparse transfer learning for interactive video search reranking," *ACM Trans. Multimedia Comput., Commun. Appl.*, 2011.
[39] A. Torralba, K. P. Murphy, and W. T. Freeman, "Sharing features: Efficient boosting procedures for multi-class object detection," in *Proc. Conf. Comput. Vis. Pattern Recog.*, 2004, vol. 2, pp. 762–769.
[40] H. Wang, M. Muneeb Ullah, A. Kläser, and I. Laptev, "Evaluation of local spatio-temporal features for action recognition," in *Proc. British Mach. Vis. Conf.*, London, U.K., 2009.
[41] M. Wang, X.-S. Hua, J. Tang, and R. Hong, "Beyond distance measurement: Constructing neighborhood similarity for video annotation," *IEEE Trans. Multimedia*, vol. 11, no. 3, pp. 465–476, Apr. 2009.
[42] M. Wang, X.-S. Hua, R. Hong, J. Tang, G.-J. Qi, and Y. Song, "Unified video annotation via multi-graph learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 5, pp. 733–746, May 2009.
[43] Y. Wang, P. Sabzmeydani, and G. Mori, "Semi-latent Dirichlet allocation: A hierarchical model for human action recognition," in *Proc. 2nd Workshop Human Motion Underst., Model., Capture Animation*, 2007, pp. 240–254.
[44] B. Zadrozny, "Learning and evaluating classifiers under sample selection bias," in *Proc. 21st Int. Conf. Mach. Learn.*, Jul. 2004, p. 114.

Wei Bian received the B.Eng. degree in electronic engineering and the B.Sc. degree in applied mathematics in 2005 and the M.Eng. degree in electronic engineering in 2007 from the Harbin Institute of Technology, Harbin, China. He is currently working toward the Ph.D. degree at the University of Technology, Sydney, Sydney, Australia.

His research interests include topics in pattern recognition and machine learning, such as dimension reduction and feature selection.

Dacheng Tao (M'07) is Professor of Computer Science with the Centre for Quantum Computation and Information Systems and the Faculty of Engineering and Information Technology in the University of Technology, Sydney. He mainly applies statistics and mathematics for data analysis problems in data mining, computer vision, machine learning, multimedia, and video surveillance. He has authored and co-authored more than 100 scientific articles at top venues including IEEE T-PAMI, T-KDE, T-IP, NIPS, ICML, UAI, AISTATS, ICDM, IJCAI, AAAI, CVPR, ECCV; ACM T-KDD, Multimedia and KDD, with the best theory/algorithm paper runner up award in IEEE ICDM'07.

Yong Rui (F'10) received the B.S. degree from Southeast University, Nanjing, China, the M.S. degree from Tsinghua University, Beijing, China, and the Ph.D. degree from the University of Illinois at Urbana-Champaign.

He currently serves as the Director of Strategy with the Microsoft China R&D (CRD) Group, Beijing.

Dr. Rui is a Senior Member of the Association for Computing Machinery. He is an Associate Editor of the *ACM Transactions on Multimedia Computing, Communication and Applications*, the IEEE TRANSACTIONS ON MULTIMEDIA, and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He was an Editor of the *ACM/Springer Multimedia Systems Journal* and the *International Journal of Multimedia Tools and Applications* from 2005 to 2007. He also serves on the Advisory Board of the IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING.