Pattern Recognition 45 (2012) 4204–4213

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Attribute-restricted latent topic model for person re-identification

Xiao Liu [a], Mingli Song [a,*], Qi Zhao [b], Dacheng Tao [c], Chun Chen [a], Jiajun Bu [a]

[a] Zhejiang Provincial Key Laboratory of Service Robot, College of Computer Science, Zhejiang University, Hangzhou 310027, China
[b] Department of Electronic and Computer Engineering, National University of Singapore, Singapore
[c] Centre for Quantum Computation and Information Systems, University of Technology, Sydney, Australia

ARTICLE INFO

Article history:
Received 11 July 2011
Received in revised form
19 May 2012
Accepted 28 May 2012
Available online 7 June 2012

Keywords:
Visual attribute
Attribute-restricted latent topic model
Person re-identification
Semantic topic

ABSTRACT

Searching for specific persons from surveillance videos captured by different cameras, known as person re-identification, is a key yet under-addressed challenge. Difficulties arise from the large variations of human appearance in different poses, and from the different camera views that may be involved, making low-level descriptor representation unreliable. In this paper, we propose a novel Attribute-Restricted Latent Topic Model (ARLTM) to encode targets into semantic topics. Compared to conventional topic models such as LDA and pLSI, ARLTM performs best by imposing semantic restrictions onto the generation of human specific attributes. We use MCMC EM for model learning. Experimental results show that our method achieves state-of-the-art performance.

1. Introduction

Many surveillance-based applications rely on the ability to re-identify persons of interest across different cameras. Given a small number of instances in which the persons of interest are captured by one camera, a person re-identification system aims to find all occurrences of these persons in other cameras. Most of the time, the system has to rely solely on the visual appearance of the persons [1,2].

Since the appearance of a person varies significantly in different poses and from different camera views, low-level descriptor representations [1,3] can prove unreliable due to feature misalignment or even missing features. Conventional methods of appearance modeling [4,5] rely on face recognition technology [6] and have problems dealing with low-quality videos and irregular views. Among these is a part-based method [7] that divides the human shape into six parts – head, torso, two legs, and two arms – each of which is represented as a rectangle. This method works well in certain cases. However, not all of the parts can be properly recognized in practice. Spatial segmentation or decomposition methods are used in [2,8,9], but there are still no satisfactory invariance guarantees to changes in view and pose.

In this work, we consider how humans can ascertain human identity. To illustrate this, an example is illustrated in Fig. 1. Although the appearance of the target varies in different poses and from different viewpoints, humans recognize them all as “the target is wearing a black and gray patched jacket and dark blue uniform pants”. This semantic description is invariant with pose, view, and other such changes in appearance. Thus, we propose to use such intermediate invariable descriptions as a reliable representation for person re-identification. In the computer vision literature, the idea of using intermediate description to achieve stability has been widely explored. For example, attribute learning [10], a popular method to represent targets in such an intermediate level, utilizes human-specific high-level semantic descriptions as the intermediate representation. One of the largest difficulties of attribute-based classification, however, is the problem of heavy dependence on pre-trained classifiers for each type of attribute, which is both fussy and unreliable. On the other hand, topic models, such as LDA [11], pLSI [12], and their variants, have been applied to various tasks, including scene recognition, object recognition, action recognition [13], behavior mining [14], and human detection. In these works, characteristic intermediate “themes” of targets are learned. However, compared with attribute learning, the “themes” or topics are learned in an unsupervised manner; that is, they are not relative to any semantic concepts, but rather self-organized products. Therefore, the topics cannot benefit from human-specific information and cannot guarantee a stable representation. In this paper, we propose Attribute-Restricted Latent Topic Model (ARLTM). It clusters visual words that often occur in the same semantic attribute into one topic. The model outperforms previous topic models as it benefits from high-level human-specific information. In this work, we do not directly use human-specific attribute classifiers, such as [15,16]; instead, ARLTM considers all attributes

[*] Corresponding author. Tel.: +86 571 87951277.
E-mail address: brooksong@zju.edu.cn (M. Song).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2012.05.019

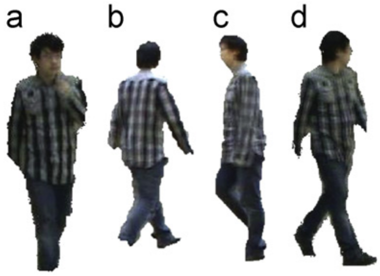

Fig. 1. The appearance of people varies significantly in different poses and views. As a result, low-level descriptor representations are unreliable. In comparison, despite the fact that the figure captured in (a), (b), (c), and (d) has different poses and views, we humans can describe each of them as "the target is wearing a black and gray patch jacket and dark blue uniform pants", and such semantic themes remain the same under different poses and viewing conditions.

as restriction priors in a principled generative process. Our model is more efficient compared with conventional attribute learning methods, as will be shown below. In summary, the major contributions of this work are as follows:

- We propose an intermediate representation that encodes high-level semantic information for human appearance modeling in multi-camera surveillance systems. (It is worth noting that the work is free of the overlapping-field-of-view assumption of the multi-camera system.)
- Our proposed new algorithm, ARLTM, bridges the gap between human-specific individual attributes and self-organized topic models, and can be easily generalized and applied to other attribute-based situations.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 gives an overview of our method and discusses the details of the ARLTM as well as model learning and inference. Section 4 demonstrates experimental results and Section 5 concludes this paper.

2. Related work

Existing work on person re-identification and appearance modeling can be roughly categorized into three categories: distance learning [17–19], local feature selection [20–22], and segmentation based matching [1,8,9].

In distance learning, a distance metric is learnt as a means of representing the similarity of individuals between camera pairs. Porikli [17] proposed a distance metric and a model function to evaluate the inter-camera radiometric properties. In the proposed method, a Brightness Transfer Function (BTF) is computed for every pair of cameras such that the BTF maps an observed color value in one camera to the corresponding observation in the other camera. Javed et al. [18] improved on Porikli's method by showing that all BTFs lie in a low dimensional subspace such that some parameters of BTF are not required for computation. Zheng et al. [19] introduced a Probabilistic Relative Distance Comparison (PRDC) model to maximize the probability of a pair of true matches having a smaller distance than that of a wrong match pair. PRDC is tolerant to both appearance changes and model over-fitting.

In local feature selection, supervised [21,22] or unsupervised [20] algorithms are designed to select the most relevant features for person re-identification. Gray et al. [22] used the AdaBoost algorithm to find a subset of optimal features for human matching by combining different types of simple features into a single similarity function. Prosser et al. [21] developed a person re-identification system based on RankSVM. In their method, the combinations of local features are learnt such that the relative ranking of the matching scores are fit to the training data.

In segmentation based matching, images of persons are first divided into small blobs and then the correspondences between these blobs are calculated. Bird et al. [1] used stripe based rigid blobs to model the appearance of individuals. They divide the image of a pedestrian into 10 equally spaced horizontal strips and the mean feature vectors of the horizontal strips are learnt in a training step. Gheissari et al. [8] proposed a spatiotemporal over-segmentation method that groups pixels that belong to the same type of fabric, after which they merge connected clusters whenever the distance between them is less than the internal variation of each of the individual clusters. The final distance between two individuals is then defined as the sum of the correspondences between these resulting segmentations. Oreifej et al. [9] extract foreground blobs in aerial images, and then they assign a weight to every blob region such that the most consistent regions are given higher weights as it is more probable that they will lead to the target's identity.

In another point of view, existing techniques can be also categorized as single-shot and multiple-shot groups based on their experimental setups. In the single-shot group, an associating pair of images, each containing one instance of the individual, are used for training and testing, respectively. The approaches [19,22,23] have to model a person by analyzing the single training image. In contrast to the single-shot group, the multiple-shot approaches [1,8] train and test the person appearance model using multiple images which are usually obtained through tracking.

3. Our approach

Given instances of a human target from one camera, our goal is to automatically train a semantic topic model, so that targets can be identified in any testing frame from other cameras. Our proposed algorithm handles challenging issues such as pose changes, view changes, and low video quality by introducing an attribute restriction.

Fig. 2 illustrates the overall flow of our approach. Background subtraction and human tracking are first carried out in order to capture the human targets in continuous video streams. For the targets captured, we then extract color and texture features such as local HSV histogram, SIFT [24], LSS [25], and SURF [26]. These local features are then clustered into sets of visual words, thus called codebooks. The values of pre-defined attributes are then manually assigned to the human target. ARLTM is then trained under the restriction of the human-specific attributes. We will show later that the topics have strong semantic information and can be a stable intermediate representation of human targets.

3.1. Human tracking

In multiple interacting target tracking [27] the joint states (such as position and scale) of the multiple targets in the videos are modeled as hidden variables, $X_{1:t}$. Given an observed video sequence $Y_{1:t}$ captured from a fixed camera, the key issue is to calculate the filtering distribution of

$$p(X_t|Y_{1:t}) = Z^{-1}p(Y_t|X_t)\cdot\int_{X_{t-1}} p(X_t|X_{t-1})p(X_{t-1}|Y_{1:t-1}) \tag{1}$$

Particle filtering (PF) approximates Eq. (1) as

$$p(X_t|Y_{1:t}) \propto Z^{-1}p(Y_t|X_t)\sum_{n=1}^{N-1} w_{t-1}^{n}p(X_t|X_{t-1}^{n}) \tag{2}$$

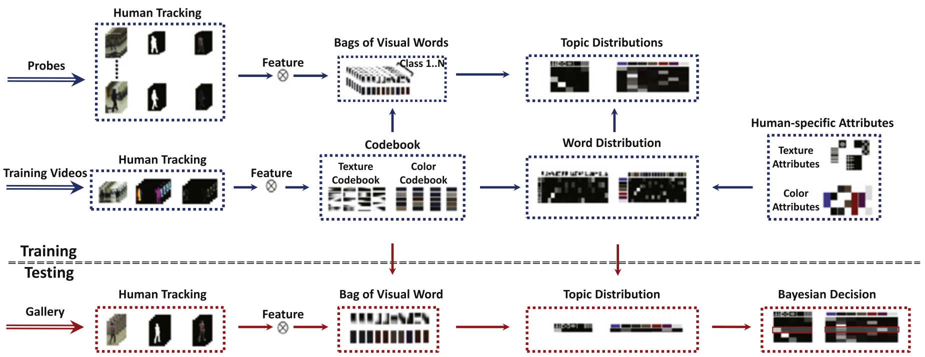

Fig. 2. Overall flow of our approach. The blue arrows show the training flows while the red arrows show the testing flows. In the training phase, color and texture features extracted from the training data are used to construct the visual words—as a result, we call them codebooks. These visual words, as well as the human-specific attributes, are combined to learn the word distributions over semantic topics. Then, for each target of interest (called a probe), topic distributions are calculated to represent the individuals. In the testing phase, given the visual word representation, each individual (called a gallery) is matched to an available probe through Bayesian decision. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

where $\{X_{t-1}^n, n=1\dots N\}$ is a set of N particles at time $t-1$, $\{w_{t-1}^n, n=1\dots N\}$ are the associated weights of the particles, and Z is a normalization factor. A pair-wised Markov Random Field (MRF) prior constraint is used to handle multiple target interactions

$$p(X_t|X_{t-1}) \propto \prod_{i \in I_t} p(X_{i,t}|X_{i,t-1}) p_0(X_t) \tag{3}$$

where $X_{i,t}$ denotes the state of the i-th target at time t and

$$p_0(X_t) = \prod_{i,j} \phi(X_{i,t}, X_{j,t}) \tag{4}$$

is the MRF prior constraint, which is expressed as a product of pair-wised interaction potentials

$$\phi(X_{i,t}, X_{j,t}) = \exp(-g(X_{i,t}, X_{j,t})) \tag{5}$$

where $g(X_{i,t}, X_{j,t})$ is a penalty function that depends on the overlap ratio of two particles.

Background subtraction technique [28] is used to estimate foreground likelihood, which, together with color likelihood, is used to calculate the observed distribution $p(Y_t|X_t)$. The tracking results are output as bounding boxes of the targets.

3.2. Feature and codebook

We use color and texture patterns as features for people. The contour descriptors are not suitable since they are not discriminative enough to distinguish between different individuals. Four types of features are extracted for describing human targets. The bounding box of each person in each frame is resized to the same size. Local histograms of HSV are densely sampled in the foreground position as the color feature. SIFT, SURF, and LSS are sampled at DoG [24] interest points as the texture features.

We then quantize each type of feature into codebooks, sized W, by using the K-means algorithm [29], and visual words are defined as the centers of the clusters. By clustering to the nearest one in the codebook, every feature can be quantized as one of the visual words in the codebooks.

3.3. Latent Dirichlet allocation

Our model is based on the Latent Dirichlet Allocation (LDA). The graphical representation of LDA is shown in Fig. 3(a). In our context, each person is a document. There are M documents in the corpus and the j-th document has N_j visual words. A visual word is the basic item in the codebook.

LDA assumes that there are K underlying latent topics, each of which is a multinomial distribution over the codebook. α and β are the Dirichlet prior parameters, while θ, Z and φ are the latent parameters. The generative process of LDA for a person is as follows:

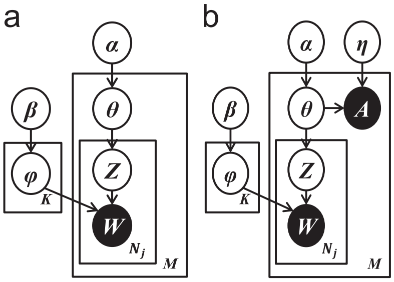

Fig. 3. Graphical representations of (a) LDA and (b) ARLTM.

1. For each person j, its topic multinomial prior θ_j is sampled from the Dirichlet prior α. $\theta_j \sim Dirichlet(\alpha)$.
2. For each topic k, the multinomial prior φ_k over the visual word is sampled from the Dirichlet prior β. $\varphi_k \sim Dirichlet(\beta)$.
3. For each visual word i of person j, a topic label is discretely sampled from the multinomial prior θ_j, $z^{ji} \sim Discrete(\theta^j)$.
4. The value w^{ji} is discretely sampled from distribution z^{ji}, $w^{ji} \sim Discrete(\varphi_{z^{ji}})$.

The parameter estimation problem of LDA can be solved by the variational EM algorithm or the Gibbs sampling algorithm.

3.4. Attribute-restricted latent topic model

LDA models each document (person) as a mixture of several topics. However, as LDA is an unsupervised learning method, no human-specific high-level information is encoded in the model. In the left of Fig. 6, there is no connection between the topics found by LDA and any semantic concepts. These "LDA topics" may be noise that are not good for classification. Various tricks, e.g., utilizing SVM [11] or supervised information [30], have been applied when LDA is used for recognition and classification. However, the performance is still far from satisfactory. In this work, since the semantic attributes of given people can be manually specified, we have reason to believe that this high-level information will help to improve the model. Specifically, in this work we introduce the Attribute-Restricted Latent Topic Model (ARLTM). Like LDA, ARLTM models each document as a mixture of underlying topics and generates each visual word from a topic. However, unlike LDA, ARLTM imbues the topics with semantics by enforcing a one-to-one restriction between the topics and semantic attributes. As Fig. 3(b) illustrates, we use a node A to represent the attributes $A=\{A_1,A_2,\ldots,A_K\}$, where K is the number of attributes or the number of latent topics. Here, the number of attributes is the same as the number of topics, and each topic is relative to an attribute. Each attribute is a binary variable that can be either 1 or 0, which is manually labeled. η is a threshold. The restriction operates as a human would: a topic can be generated only when human observers can recognize its relative attribute, thus giving the attribute a value of 1, and the topic prior is larger than the observing threshold. The generative process of ARLTM is as follows:

1. For each person j, its topic multinomial prior θ_j is sampled from the Dirichlet prior α. $\theta_j \sim Dirichlet(\alpha)$.
2. For each person's topic prior θ^{jk}, it will be observed as an attribute only if the prior is larger than pre-defined threshold η. Thus, the observed attribute list A_i is sampled as $p(A^{jk}|\theta^{jk},\eta) \sim \lfloor \theta^{jk} > \eta \rfloor$.
3. For each topic k, the multinomial prior φ_k over the visual word is sampled from the Dirichlet prior β. $\varphi_k \sim Dirichlet(\beta)$.
4. For each visual word i of person j, a topic label is discretely sampled from the multinomial prior θ_j, $z^{ji} \sim Discrete(\theta^j)$.
5. The value w^{ji} is discretely sampled from distribution z^{ji}, $w^{ji} \sim Discrete(\varphi_{z^{ji}})$.

In the training step, two parameters are learnt: φ , which models the topic elements and θ, which indicates the contents of each document. They can be learnt by sampling z^{ji} through the MCMC EM procedure. Specifically, we use Gibbs Sampling EM here. The conditional distribution of z^{ji} can be obtained by multiplying two factors:

$$p(z^{ji}=k|z_{-ji},w,\alpha,\beta,A,\eta) \propto p(w^{ji}|z^{ji}=k,z_{-ji},\beta)$$

$$p(z^{ji}=k|\alpha,z_{-ji},A,\eta) \tag{6}$$

where z_{-ji} represents all the topic labels except z_{ji}. Intuitively, the first factor expresses the probability of word w^{ji} under topic k, and the second factor expresses the probability of topic k in document j. Neither $p(w^{ji}|z^{ji}=k,z_{-ji},\beta)$ nor $p(z^{ji}=k|\alpha,z_{-ji},A,\eta)$ can be directly calculated, however, by estimating the values of latent nodes φ and θ, the overall distribution can be formed as

$$p(w^{ji}|z^{ji}=k,z_{-ji},\beta) \propto \varphi^{k}_{-ji} \tag{7}$$

$$p(z^{ji}=k|\alpha,z_{-ji},A,\eta) \propto \theta^{jk}_{-ji} \tag{8}$$

where φ^{k}_{-ji} is the multinomial prior distribution of topic k over the visual words and the distribution is estimated by using the labels of visual words excluding word i of person j. θ^{jk}_{-ji} is the multinomial prior distribution of topic k with person j and the distribution is estimated by using labels of visual words excluding word i of person j.

As shown in Fig. 3(b), φ^{k}_{-ji} can be estimated using β and w as

$$\varphi^{k}_{-ji} \propto \frac{n^{k}_{-ji,w_{ji}}+\beta}{\sum_{w=1}^{W} n^{k}_{-ji,w}+\beta W} \tag{9}$$

where W is the size of the codebook, $n^{k}_{-ji,w_{ji}}$ is the number of words assigned to topic label k excluding word i in person j, and $n^{k}_{-ji,w}$ is the number of words with value w and topic label k excluding word i in person j.

Estimating θ^{j}_{-ji} involves three terms to determine the log-likelihood

$$\hat{Q}(\theta^{j}_{-ji}) = \log p(\theta^{j}_{ji}|\alpha) \cdot \sum_{l} \log p(z^{jl}_{-ji}|\theta^{j}_{-ji}) \sum_{k} \log p(A^{jk}|\theta^{jk}_{-ji},\eta) \tag{10}$$

The first two terms are for the original LDA generative process, while the third term is for the attribute restriction. To maximize the log-likelihood $\hat{Q}(\theta^{j}_{-ji})$, we estimate the value of θ^{j}_{-ji} in two steps. Specifically, we first estimate θ^{j}_{-ji} ignoring the effect of attribution restriction

$$\hat{\theta}^{jk}_{-ji} \propto \frac{n^{j}_{-ji,k}+\alpha}{\sum_{l=1}^{K} n^{j}_{-ji,l}+\alpha K} \tag{11}$$

where $n^{j}_{-ji,k}$ is the number of words in person j and labeled with topic k excluding word i in person j. We then renormalize $\hat{\theta}^{jk}_{-ji}$ so that it can satisfy the restriction of attributes

$$\theta^{jk}_{-ji} = \begin{cases} \eta : A^{jk}=1 & \text{and} \quad \hat{\theta}^{jk}_{-ji} < \eta \\ \quad A^{jk}=0 & \text{and} \quad \hat{\theta}^{jk}_{-ji} > \eta \\ \dfrac{\hat{\theta}^{jk}_{-ji}}{\Psi} : & \quad \text{else} \end{cases} \tag{12}$$

where Ψ is a normalization factor.

The entire MCMC EM algorithm is shown in Algorithm 1.

Algorithm 1. Given W and A, estimate θ and φ.

```
1. Initialize all $z^{ji}$, and set iteration index $c=0$
2. Iterate over $c$ for the sampling step:
    • For each $z^{ji}$ , use EM to estimate $\theta^{j}_{-ji}$
    • E-step:
      $\hat{Q}(\theta^{j}_{-ji}) = \log p(\theta^{j}_{ji}|\alpha) \cdot \sum_{l} \log p(z^{jl}_{-ji}|\theta^{j}_{-ji}) \cdot \sum_{k} \log p(A^{jk}|\theta^{jk}_{-ji},\eta)$
    • M-step: Maximize ($\hat{Q}$)
      $\hat{\theta}^{jk}_{-ji} \propto \frac{n^{j}_{-ji,k}+\alpha}{\sum_{l=1}^{K} n^{j}_{-ji,l}+\alpha K}$
      $\theta^{jk}_{-ji} = \begin{cases} \eta : A^{jk}=1 & \text{and} \quad \hat{\theta}^{jk}_{-ji} < \eta \\ \quad A^{jk}=0 & \text{and} \quad \hat{\theta}^{jk}_{-ji} > \eta \\ \frac{\hat{\theta}^{jk}_{-ji}}{\Psi} : & \quad \text{else} \end{cases}$
    • Sample $z^{ji}$
      $p(z^{ji}=k|z_{-ji},w,\alpha,\beta,A,\eta)$
      $\propto \frac{n^{k}_{-ji,w_{ji}}+\beta}{\sum_{w=1}^{W} n^{k}_{-ji,w}+\beta W} \theta^{jk}_{-ji}$
    • $c \leftarrow c+1$, and go to 2 until $c \geq$ Max Iterations
3. Estimate parameters:
      $\theta^{jk} = \frac{n^{j}_{k}+\alpha}{\sum_{l=1}^{K} n^{j}_{l}+K\alpha}$
      $\varphi^{kw} = \frac{n^{k}_{w}+\beta}{\sum_{l=1}^{W} n^{k}_{l}+W\beta}$
```

Fig. 4 illustrates two examples to show the discriminative ability of ARLTM. The right side of Fig. 4(a) and (b) shows the topic probability distributions learned for the target persons shown on the left side. The most likely topics of the target in Fig. 4(a) are: "black" and "uniform" for the head; "dark blue", "black", and "stripe" for the upper body; and "light blue" and "uniform" for the lower body. Thus, ARLTM describes the target as "black uniform head, wearing a dark blue and black stripe jacket and light blue uniform pants". Similarly, Fig. 4(b) demonstrates the most likely topics of another target: "black" and "uniform" for

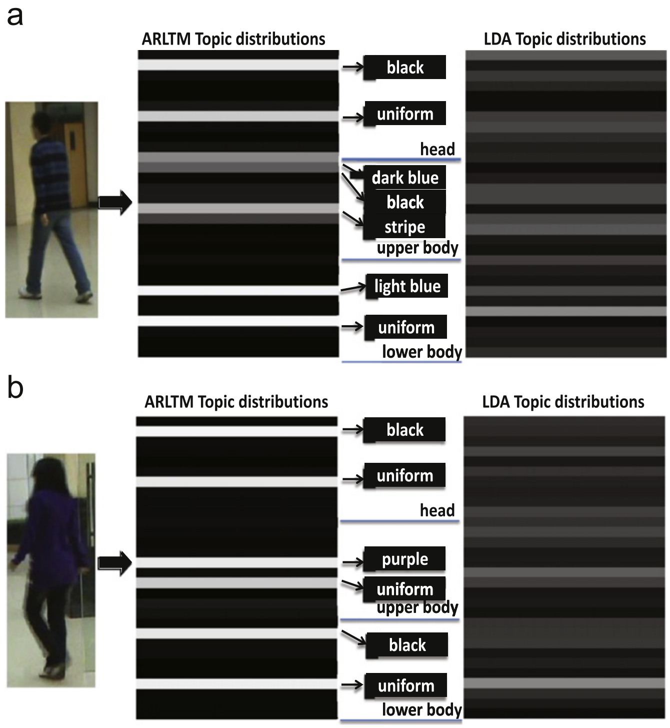

Fig. 4. Examples of the topic probability distributions of different targets. A lighter stripe signifies a higher probability distribution. Compared with LDA, the topics of ARLTM are consistent with attributes and are sparser. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

the head; "purple" and "uniform" for the upper body; and "black" and "uniform" for the lower body. One big advantage of ARLTM is that the learnt topics encode the method used by humans to recognize targets. For example, without human labeling, LDA may classify "dark blue" and "black" into one topic; while with attribute restriction, ARLTM learns that "dark blue" and "black" are two different topics. Such descriptions are closer to how humans identify targets compared with low-level features; hence, they are more powerful in handling appearance changes. Another advantage comes from the sparsity of topics induced from attributes supervision. This sparsity, as experimental results show, will significantly improve recognition performance.

3.5. Bayesian decision

We utilize human tracking and then represent each person at each frame as a bag of visual words. Given these visual words w and the parameters we learnt, the probability that a person X appearing in the new video is matched to the target person j can be calculated as

$$p(w_x|\theta^j,\varphi)=\prod_{n=1}^{N}\sum_{z_n}p(z_n|\theta^j)p(w_{xn}|z_n,\varphi)=\prod_{n=1}^{N}\sum_{k=1}^{K}\theta^{jk}\varphi^{kw_{xn}} \qquad (13)$$

where N is the number of visual words in the new frame and K is the number of topics.

The attribute values of the newly appearing person are unknown, so we omit the effect of attribute restriction in the decision step, but use the original topic decision model to make the decision. The matched target is the one with the highest matching probability to the newly appearing target, where the matching identity is defined as

$$f(X,M)=\arg\max_{j} p(w_x|\theta^j_M,\varphi_M) \qquad (14)$$

3.6. Attributes

For each type of attribute, we make an attribute list. The attribute list for color includes 12 different values inclusive of colors such as "blue", "black", and "red". The attribute list for texture includes eight different values inclusive of textures, such as "uniform", "patch", "stripe", and "spot". Prior to the training step, the model has no knowledge of what "red" or "stripe" is. It only knows that one person is "blue", "white" and "spot"; and the other is "black", "red" and "stripe". However, by means of the training process, ARLTM begins to learn the correlation between attributes and semantic labels. Fig. 5 is the visualization of a section of topic distribution over visual words. It can be seen that the visual words organize semantic topics as we expected.

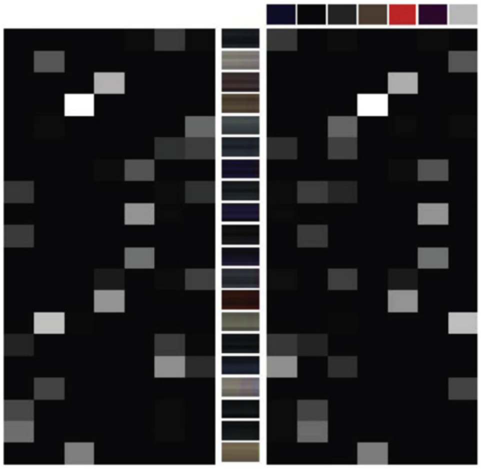

Fig. 5. Visualization of a part of a color topic distribution over visual words. Each row represents a color topic, and each column represents a color visual word. The left is for LDA and the right is for ARLTM. A lighter square signifies the closer relation with the visual word that the topic has. The attribute restrictions for ARLTM from left to right are: "blue", "black", "gray", "flesh color", "red", "purple" and "white". (Best viewed in color.) (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

Fig. 6. Examples of targets captured on video.

4. Experimental results

4.1. Identity recognition in videos

To test our proposed technique, video sequences containing 12 target persons (Fig. 6) were captured using four cameras. Videos from one camera were used for training and those from the others for testing. The videos were 720 × 480 px in resolution at a frame rate of 25 (FPS). The orientation of the cameras and sample frames are shown in Figs. 7 and 8, respectively. It should be noted that only appearance information was used to recognize and identify each person. When a person in the training video entered specified areas, 10 frames of that person were used to train an ARLTM, while the interval between each frame used was 0.5 s (so as to obtain sufficient variations in the training data). Consequently, a video 5 s long was needed for each target. In the testing of the appearance model, the videos used more than 1000 frames per individual.

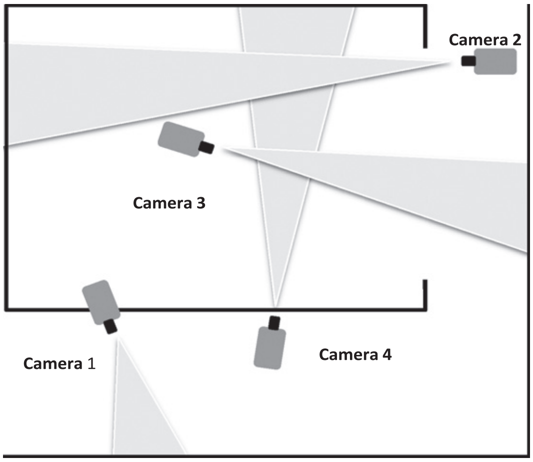

Fig. 7. The arrangement of the four cameras. Cameras 2 and 3 as well as cameras 3 and 4 have small overlapping FOVs, while camera 1 has a non-overlapping FOV.

The confusion matrix among the 12 targets is depicted in Fig. 9. The horizontal axis represents the trained models. The vertical axis represents ground truth query persons for testing. The results show that there are clear separations among the different targets. Targets 3, 7, 8 and 9 have all been flawlessly identified.

We chose the work of Bird et al. [2] for comparison, as its experimental setup is similar to our work. In addition to the complete model with both color and texture features, we evaluated the model with either feature. We also implemented LDA for comparison, as the human tracking and visual word computation steps in the LDA are the same as those in the ARLTM, with only the topic sampling step being different. Fig. 10 shows the comparison results and Table 1 summarizes the average

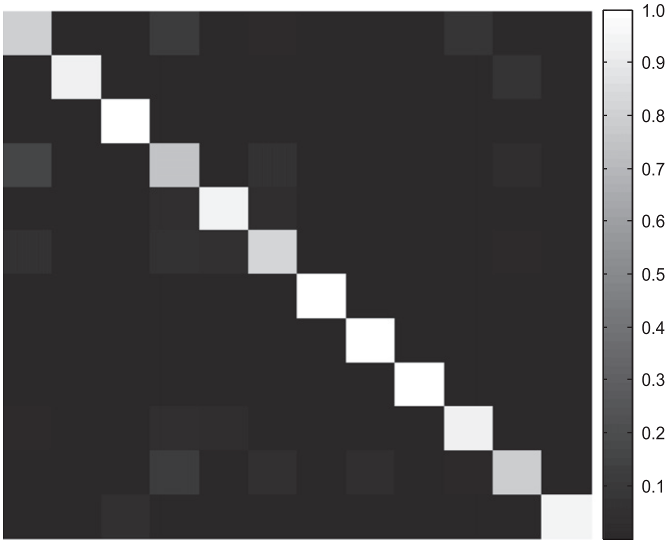

Fig. 9. Confusion matrix among the 12 targets with each using 10 frames for training and more than 1000 frames for testing.

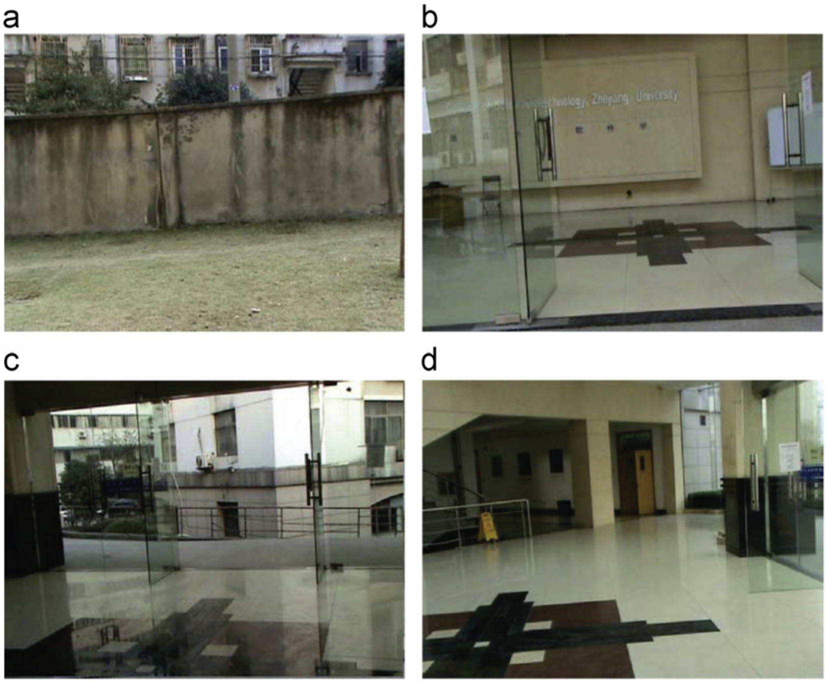

Fig. 8. Sample frames from the four cameras used in the experiments. Frame (d) was taken from the video captured by the training camera, while frames (a), (b), and (c) were taken from the testing videos captured by the three other cameras.

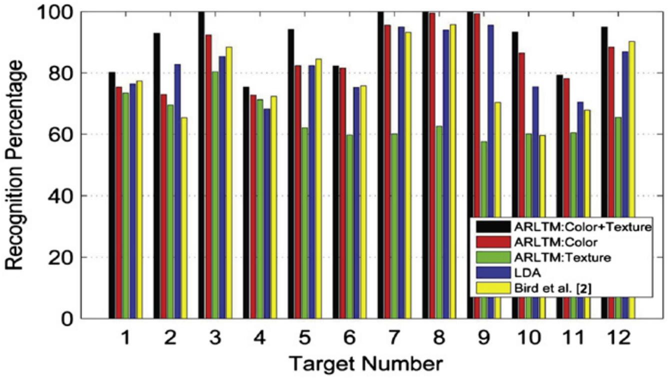

Fig. 10. Recognition precisions for each of the 12 targets.

Table 1
Comparison of ARLTM with other methods.

Algorithm	Average precision
ARLTM	91.0
ARLTM with color only	85.7
ARLTM with texture only	65.2
LDA	82.3
Bird et al. [2]	78.4

precision, which is defined as the ratio of correctly matched frame number to the total testing frame number. Our proposed method greatly outperformed both the LDA and Bird et al.'s [2], largely due to the attributes restriction that helped to raise the most discriminative topics – as we illustrated in the previous section – and the combination of color and texture features.

4.2. Identity recognition in still images

We also tested the performance of our approach on still images using the public VIPeR database [31], which contains two views of 632 individuals seen from widely differing viewpoints. The images in one view were used as the probes, while the images in the other views were used as the galleries. VIPeR [31] is currently the most challenging database for evaluating appearance models for recognition. We split the dataset into a training and a test set. Individuals in the training dataset were labeled with attributes and semantic topics of ARLTM were learnt from the training set. Each probe individual in the test set was represented as the topic distribution and each gallery was matched over probes through Bayesian decision. To evaluate the effectiveness of ARLTM, we compared our results with three state-of-the-art methods: the Ensemble of Localized Features (ELFs) [1], the Weighted Region Matching (WRM) [9] and the Symmetry-Driven Accumulation of Local Features (SDALFs) [32]. As human tracking is not needed in still images, we followed the set-up steps used by SDALF [32], thus separating background and foreground by inferring over the STEL generative model [33], and then extracted features in the foreground area.

Fig. 11 shows examples of the query results of our proposed method on the VIPeR database. Probe images from one view are shown in the left-most column and the top 18 query results are sorted from left to right. The correct matches are indicated by the red boxes. The right-most column shows the true matches. The bottom row shows a failed attempt. The correct match failed to show up in the top 18 queries. This is because the appearance of this individual was radically altered in the different views. The recognition accuracies of our proposed ARLTM compared with the state-of-the-art works on VIPeR with different number of searching classes are summarized in Table 2.

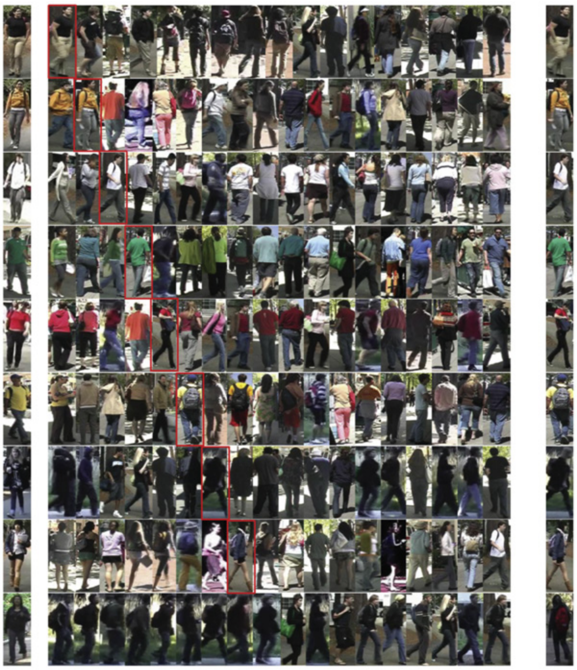

Fig. 11. Example of query results using ARLTM on the VIPeR database [31]. Probe images are shown in the left-most column. The top 18 query results are sorted from left to right. The correct matches are indicated by the red boxes. The right-most column also shows the true matches. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

More detailed results are graphically depicted by the CMC curves in Fig. 12 and the SDR/SRR rate in Fig. 13. From these figures, it is clear that ARLTM gives the best results.

4.3. ARLTM and supervised topic models

ARLTM is a supervised variation of LDA. The supervision is induced from the use of attributes. The attributes provide a richer description than that which can be achieved using only class labels. Most existing supervised topic models, however, are not able to utilize attributes information. For example, Author Topic Model [30] induces authorship information, but it is not suitable to model attributes as authors. DiscLDA [34] associates each generated topic with a class label, but attributes can be more powerful supervised information than mere class labels. Labeled LDA [35] may be used for this task by generating the binary attributes as a multiple label list. However, it uses a hard constraint when selecting topics, which possesses less tolerance for noise than the soft threshold constraint in ARLTM. We also tested labeled LDA and ARLTM on VIPeR. Figs. 14 and 15 graphically depict the results of our test, which show that ARLTM is more suitable for human identity recognition than existing supervised topic models.

4.4. Parameters

To train an ARLTM, we used 18 attribute restrictions for color and 8 for texture. As a result, the total number of topics in ARLTM was 26 for each human part. The size of the codebook was 200 in our experiments. We used symmetric Dirichlet priors with $\alpha = 1$ and $\beta = 0.01$ as a Gibbs learning process usually does.

The parameter η has a significant effect on the performance of the model. When η approaches zero, distributions of topics

Table 2
Top ranked matching rate on VIPeR. *P* is the number of searching classes in the testing set. *R* is the rank.

Methods	$P=300$				$P=400$				$P=500$			
	$R=1$	$R=5$	$R=10$	$R=20$	$R=1$	$R=5$	$R=10$	$R=20$	$R=1$	$R=5$	$R=10$	$R=20$
ARLTM	21.2	38.7	52.9	67.5	18.4	33.5	45.1	58.5	12.9	26.4	36.9	48.0
ELF	15.1	28.5	35.0	61.4	11.2	29.8	40.0	53.8	8.7	18.2	28.5	34.8
WRM	8.2	17.0	38.6	48.4	6.7	14.8	22.5	34.9	4.8	13.7	21.5	30.9
SDALF	19.9	39.1	49.3	66.1	17.5	32.9	43.9	57.0	13.5	25.7	34.5	46.7

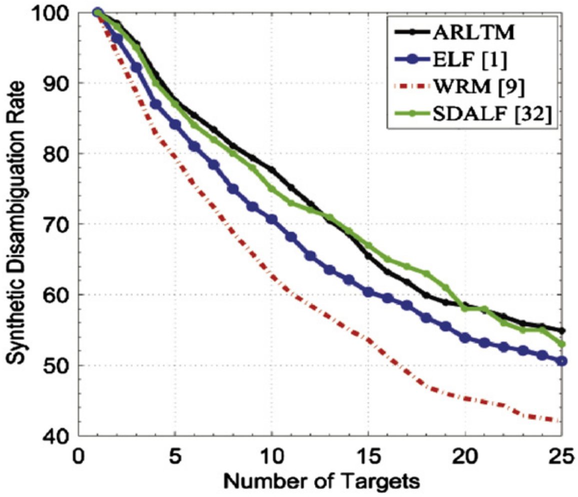

Fig. 13. SDR/SRR rates for ARLTM and the state-of-the-art works with different numbers of targets using the VIPeR database [31].

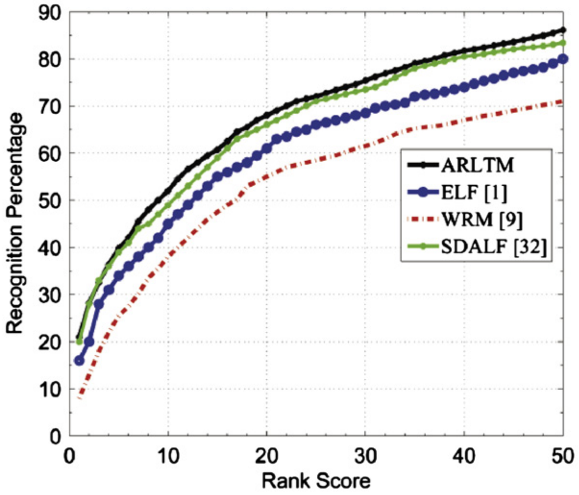

Fig. 12. CMC curves for ARLTM and the state-of-the-art works on the VIPeR database [31]. Only the first 50 ranking positions are displayed.

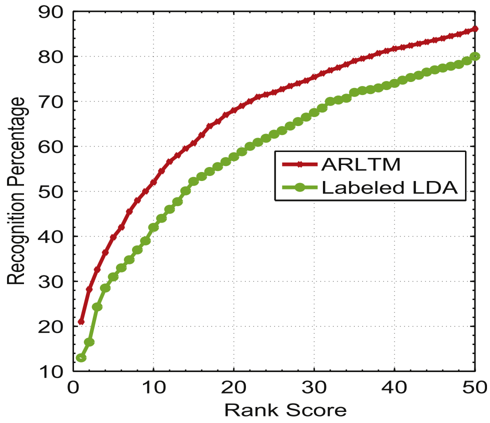

Fig. 14. CMC curves for ARLTM and labeled LDA on the VIPeR database.

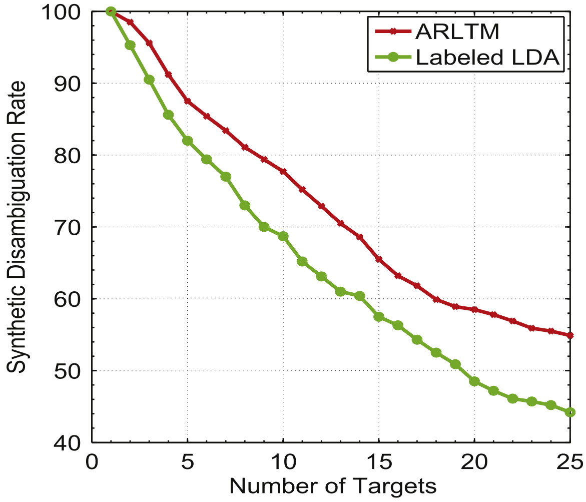

Fig. 15. SDR/SRR rates for ARLTM and labeled LDA with different numbers of targets using the VIPeR database.

related to unobserved attributes are forced to zero, and ARLTM will be so sparse that it loses details. When η approaches one, one target can only generate one topic, which tends to reduce accuracy. In fact, if the product of η and the number of observed attributes is larger than 1, the EM process of Formula (10) will collapse. This is in congruence with the real situation, e.g., we cannot observe more than five colors in one person if each color's proportion is greater than 0.2. The average precision for various values of η is shown in Fig. 16. In practice, we set η to be 0.01.

In order to estimate an appropriate number for sampling iteration, we evaluated the average precision with the model learnt in different sampling iterations. The result is depicted in Fig. 17. It can be seen that after 120 iterations, the performance is quite stable.

5. Conclusion

This paper presented a novel ARLTM for human identity recognition. The proposed model utilizes human-specific attributes as restriction priors in a principled generative process and the resulting topics learnt by the ARLTM encode strong semantic information. The model works well for general settings, including non-overlapping cameras that proved a challenge for many existing works. Quantitative and comparative results show that ARLTM achieves state-of-the-art performance for human identity recognition.

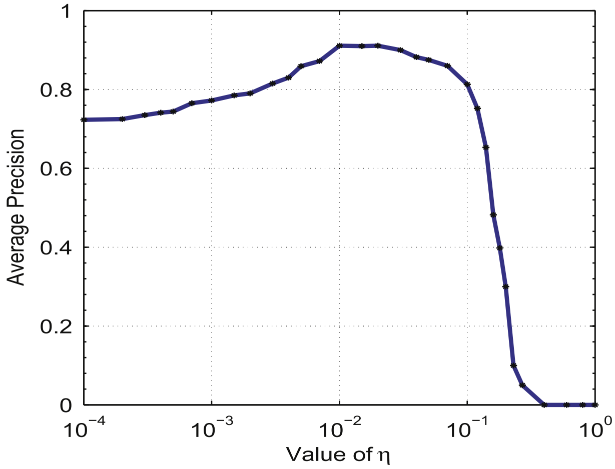

Fig. 16. Average precision for various values of η.

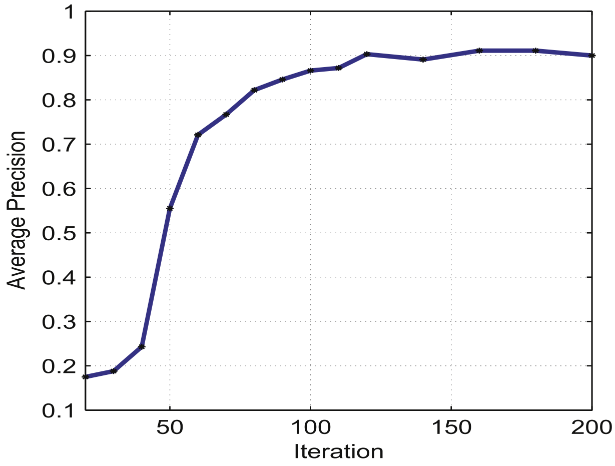

Fig. 17. Average precision with various iterations.

Acknowledgements

This work is supported by National Natural Science Foundation of China (61170142), National Key Technology R&D Program (2011BAG05B04), the Zhejiang Province Key S&T Innovation Group Project (2009R50009), and the Fundamental Research Funds for the Central Universities (2012FZA5017).

References

[1] D. Gray, H. Tao, Viewpoint invariant pedestrian recognition with an ensemble of localized features, in: Proceedings of the ECCV, 2008.

[2] N. Bird, O. Masoud, N. Papanikolopoulos, A. Isaacs, Detection of loitering individuals in public transportation areas, IEEE Transactions on Intelligent Transportation Systems 6 (2005) 167–177.

[3] K. Morioka, X. Mao, H. Hashimoto, Global color model based object matching in the multi-camera environment, in: Proceedings of the IROS, 2006, pp. 2644–2649.

[4] J. Sivic, C. Zitnick, R. Szeliski, Finding people in repeated shots of the same scene, in: Proceedings of the BMVC, 2006.

[5] Y. Song, T. Leung, Context-aided human recognition—clustering, in: Proceedings of the ECCV, 2006.

[6] S.J. McKenna, S. Gong, Y. Raja, Modelling facial colour and identity with Gaussian mixtures, Pattern Recognition 31 (1998) 1883–1892.

[7] N. Thome, S. Miguet, A robust appearance model for tracking human motions, in: Proceedings of the AVSS, 2005.

[8] N. Gheissari, T. Sebastian, R. Hartley, Person reidentification using spatio-temporal appearance, in: Proceedings of the CVPR, 2006, pp. 1528–1535.

[9] O. Oreifej, R. Mehran, M. Shah, Human identity recognition in aerial images, in: Proceedings of the CVPR, 2010, pp. 709–716.

[10] N. Kumar, A. Berg, P. Belhumeur, S. Nayar, Attribute and simile classifiers for face verification, in: Proceedings of the ICCV, 2009, pp. 365–372.

[11] D. Blei, A. Ng, M. Jordan, Latent Dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993–1022.

[12] T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the SIGIR, 1999, pp. 50–57.

[13] M. Bregonzio, J. Li, S. Gong, T. Xiang, Discriminative topics modelling for action feature selection and recognition, in: Proceedings of the BMVC, 2010.

[14] T. Hospedales, S. Gong, T. Xiang, Video behaviour mining using a dynamic topic model, International Journal of Computer Vision (2011).

[15] H. Christoph, N. Hannes, H. Stefan, Learning to detect unseen object classes by between-class attribute transfer, in: Proceedings of the CVPR, 2009, pp. 951–958.

[16] F. Ali, E. Ian, H. Derek, Attribute-centric recognition for cross-category generalization, in: Proceedings of the CVPR, 2010, pp. 2352–2359.

[17] F. Porikli, Inter-camera color calibration using cross-correlation model function, in: Proceedings of the ICIP, vol. 3, 2003, pp. 133–136.

[18] O. Javed, K. Shafique, Z. Rasheed, M. Shah, Modeling inter-camera space–time and appearance relationships for tracking across non-overlapping views, Computer Vision and Image Understanding 109 (2008) 146–162.

[19] W. Zheng, S. Gong, T. Xiang, Person re-identification by probabilistic relative distance comparison, in: Proceedings of the CVPR, 2011, pp. 649–656.

[20] K. Bashir, T. Xiang, S. Gong, Feature selection on gait energy image for human identification, in: Proceedings of the ICASSP, 2008, pp. 985–988.

[21] B. Prosser, W. Zheng, S. Gong, T. Xiang, Person re-identification by support vector ranking, in: Proceedings of the BMVC, 2010.

[22] B. Olsen, A. Hoover, Calibrating a camera network using a domino grid, Pattern Recognition 34 (2001) 1105–1117.

[23] L. Bazzani, M. Cristani, A. Perina, V. Murino, Multiple-shot person re-identification by chromatic and epitomic analyses, Pattern Recognition Letters 33 (2012) 898–903.

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

[25] E. Shechtman, M. Irani, Matching local self-similarities across images and videos, in: Proceedings of the CVPR, vol. 11, 2007, pp. 1–8.

[26] H. Bay, A. Ess, T. Tuytelaars, L. Gool, Speeded-up robust features (surf), Computer Vision and Image Understanding 110 (2008) 346–359.

[27] K. Smith, D. Gatica-Perez, J. Odobez, Using particles to track varying numbers of interacting people, in: Proceedings of the CVPR, vol. 1, 2005, pp. 962–969.

[28] J. Yao, J. Odobez, Multi-layer background subtraction based on color and texture, in: Proceedings of the CVPR, 2007, pp. 1–8.

[29] A. Likas, N. Vlassis, J.J. Verbeek, The global k-means clustering algorithm, Pattern Recognition 36 (2003) 451–461.

[30] M. Rosen-Zvi, T. Griffiths, M. Steyvers, P. Smyth, The author-topic model for authors and documents, in: Proceedings of the UAI, 2004.

[31] D. Gray, S. Brennan, H. Tao, Evaluating appearance models for recognition, reacquisition, and tracking, in: PETS, 2007.

[32] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person re-identification by symmetry-driven accumulation of local features, in: Proceedings of the CVPR, 2010, pp. 2360–2367.

[33] N. Jojic, A. Perina, M. Cristani, V. Murino, B. Frey, STEL component analysis: modeling spatial correlations in image class structure, in: Proceedings of the CVPR, 2009, pp. 2044–2051.

[34] S. Lacoste-Julien, F. Sha, M. Jordan, Disclda: discriminative learning for dimensionality reduction and classification, in: Proceedings of the NIPS, vol. 21, 2008.

[35] D. Ramage, D. Hall, R. Nallapati, C. Manning, Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora, in: Proceedings of the ACL, vol. 1, 2009.

Xiao Liu is a Ph.D. student in College of Computer Science, Zhejiang University. His research interests include statistic model of human appearance and visual feature fusion.

Mingli Song is an Associate Professor in College of Computer Science, Zhejiang University. He received the Ph.D. degree in Computer Science from Zhejiang University, China, in 2006. He was awarded Microsoft Research Fellowship in 2004. His research interests include face modeling and facial expression analysis.

Qi Zhao received the B.E. degree in Computer Science from Zhejiang University, China, in 2004 and the M.S. and Ph.D. degrees in Computer Engineering from the University of California, Santa Cruz, in 2007 and 2009, respectively. In 2007 and 2008, she worked as a research intern at Microsoft Research Redmond and Google Research New York.

From 2009 to 2011, she was a postdoctoral researcher in the Computation and Neural Systems, and Division of Biology at the California Institute of Technology. Since June 2011, she has been an assistant professor in the Department of Electrical and Computer Engineering at the National University of Singapore. Her current research interests include computational visual cognition, neuromorphic visual models and systems, computer vision, machine learning, and computational neuroscience. She is a member of the IEEE and currently serves as an associate editor for the International Journal of Image and Graphics.

Dacheng Tao received the B.Eng. degree from the University of Science and Technology of China (USTC), the M.Phil. degree from the Chinese University of Hong Kong (CUHK), and the Ph.D. degree from the University of London (UoL). Currently, he is an assistant professor at the department of Computing in the Hong Kong Polytechnic University. His research interests include artificial intelligence, computer vision, data mining, machine learning, multimedia, and visual surveillance. He published extensively at IEEE TPAMI, TKDE, TIP, TMM, TCSVT, TSMC, CVPR, ICDM, ACM Multimedia, KDD, etc., with best paper award and nominations. Previously he gained several Meritorious Awards from the International Interdisciplinary Contest in Modeling, organized by COMAP. He is an associate editor of Neurocomputing (Elsevier), an editor of two books, and a guest co-editor of six special issues, including CVIU, PR, PRL and Neurocomputing. He served as a publicity chair for ICMLC2008, IMAP2008, and CW2008. He cochaired a Special Session in ICMLC2007 and a Workshop in ICDM2007.

Chun Chen is a Professor in the College of Computer Science, Zhejiang University. His research interests include computer vision, computer graphics and embedded technology.

Jiajun Bu is a Professor in the College of Computer Science, Zhejiang University. His research interests include information retrieval, computer vision and embedded system.