Neurocomputing 151 (2015) 1283–1292

Contents lists available at ScienceDirect

Neurocomputing

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

Set-label modeling and deep metric learning on person re-identification

Hao Liu [a], Bingpeng Ma [a], Lei Qin [b,*], Junbiao Pang [c], Chunjie Zhang [a], Qingming Huang [a,b]

[a] School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100190, China
[b] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[c] Beijing Key Laboratory of Multimedia and Intelligent Software Technology, College of Metropolitan Transportation, Beijing University of Technology, 100124, China

ARTICLE INFO

Article history:
Received 27 July 2014
Received in revised form
21 October 2014
Accepted 2 November 2014
Communicated by Qingshan Liu
Available online 11 November 2014

Keywords:
Person re-identification
Mutual-information
Metric learning
Deep learning
Neighborhood component analysis

ABSTRACT

Person re-identification aims at matching individuals across multiple non-overlapping adjacent cameras. By condensing multiple gallery images of a person as a whole, we propose a novel method named Set-Label Model (SLM) to improve the performance of person re-identification under the multi-shot setting. Moreover, we utilize mutual-information to measure the relevance between query image and gallery sets. To decrease the computational complexity, we apply a Naive–Bayes Nearest-Neighbor algorithm to approximate the mutual-information value. To overcome the limitations of traditional linear metric learning, we further develop a deep non-linear metric learning (DeepML) approach based on Neighborhood Component Analysis and Deep Belief Network. To evaluate the effectiveness of our proposed approaches, SLM and DeepML, we have carried out extensive experiments on two challenging datasets i-LIDS and ETHZ. The experimental results demonstrate that the proposed methods can obtain better performances compared with the state-of-the-art methods.

1. Introduction

In the recent years, the task of person re-identification (Re-Id) is becoming largely attractive in video surveillance. It aims to match people across multiple non-overlapping cameras, for example, identify people across multi-view cameras in the multi-camera network, or recognize the identical person who disappeared in one camera and appeared in another camera later. It also can be embedded in widespread applications such as tracking and target re-acquisition.

According to the experimental setting, the methods of Re-Id can be divided into two groups, single-shot and multi-shot. The former group selects only one image for each person, while the latter group describes multiple images as a signature for each person Id(class label). Re-Id is a challenging problem, since it suffers illumination changes, low-resolution, and view variations in multiple cameras. For recent best efforts from researchers, one kind of the Re-Id methods focuses on designing discriminative features [1–7]. By utilizing the supervised information, the other kind of methods aims at finding a global linear transformation to re-weight feature dimensions (e.g. learning a Mahalanobis distance) [8–11].

In this paper, we first propose a Set-Label Model named SLM approach to improve the performance of person re-identification under the multi-shot setting. There are three steps for SLM. Firstly, we define a set-based structure for each class, which contains concatenated features between the query feature and the gallery features. In the following, we use SET to replace the set-based structure for simplicity. We utilize mutual-information to measure the relationship between features w.r.t their class label. Secondly, since the features distribution of conditional probability can be hardly assumed, we apply a Naive–Bayes Nearest-Neighbor algorithm (NBNN) [12] to approximate the mutual-information value instead of directly accessing the probability form. In addition, the NBNN algorithm can also provide a significant efficiency. Finally, the mutual-information values can be ranked in a descending order and the corresponding class label of the highest value is assigned to the query.

By utilizing the labeled data, we further develop a deep non-linear metric learning method named DeepML based on Neighborhood Component Analysis (NCA) [13] and Deep Belief Network (DBN) [14]. NCA aims to maximize the expected numbers of classified sample in training data via a data transformation. By NCA, an improvement can be performed on the algorithms, which are based on computing the distance of two features (such as k-nearest-neighbor classification). To extend the data transformation

* Corresponding author.
E-mail addresses: h-liu14@mails.tsinghua.edu.cn (H. Liu), bpma@ucas.ac.cn (B. Ma), qinlei@ict.ac.cn (L. Qin), jbpang@jdl.ac.cn (J. Pang), zhangcj@ucas.ac.cn (C. Zhang), qmhuang@ict.ac.cn (Q. Huang).

http://dx.doi.org/10.1016/j.neucom.2014.11.002
0925-2312/

in NCA, we utilize DBN to learn a nonlinear feature transformation. NCA is placed on the top layer of RBM to adjust the weights of top layer. Then fine-tuning is carried out to adjust the weights of other layers. By this way, the discriminative power of features can be enhanced by the learned new metric (transformation).

There are two main contributions for this paper. (1) We model the relevance among multiple image features by mutual-information. Furthermore, we apply an approximate algorithm NBNN to value the mutual-information instead of directly accessing the probability form. As our knowledge, it is the first time that mutual-information theory is applied in the task of person re-identification. (2) By considering the labeled images, we develop a deep non-linear metric learning method to improve the discriminative power of our features. As most metric learning methods focus on learning a linear transformation, we apply deep learning architecture to provide a non-linear mapping from the origin features to new non-linear features.

We evaluate SLM and DeepML on two benchmark datasets i-LIDS and ETHZ, both of which have multiple images per person and are undergoing the changes of illumination, view angle, low-resolution and occlusion. The experimental results demonstrate that SLM can obtain 100 percent matching accuracy with simple color features (HSV) on ETHZ after rank 3, and DeepML can gain additional improvements by combining SLM with deep non-linear metric learning.

The remainder of this paper is organized as follows: related works are introduced in Section 2. The details of SLM and DeepML are described in Section 3. The experimental performance and results are presented in Section 4. Finally, we draw some conclusions and put forward future works in Section 5.

2. Related work

Recently, the task of person re-identification, aiming at matching the same individual across multiple disjoint cameras, has obtained increasing attention in video surveillance. To improve the performance of Re-Id, existing works mainly focus on two aspects, appearance feature extraction and distance metric learning.

The appearance based methods mainly rely on designing descriptive features such as low-dimensional discriminative features [1], viewpoint invariance features [2,15], accumulation of multiple features [3], combination of both local and global features [4], bio-inspired features [5,7], discriminative features by attributes [6] and Fisher vector encoded features [7].

Different from the appearance based methods, other methods in person re-identification care more about how to use the metric learning method to improve the measure of the features [8–11,16,17]. By a feature mapping, these methods project the original features into another feature space. The traditional metric learning approaches such as [16,17] aim to learn an optimal transformation to weight the features. By this transformation, the true matches are clustered closer, and the false matches are pushed farther. In this way, a budget of metric learning based methods specific for Re-Id has been proposed. Ref. [8] develops a fast and scalable method of learning metric by inference of likelihood ratio test. Ref. [11] formulates Re-Id as a distance comparison learning problem by maximizing the probability between a true match and a false match. Most of these methods are limited to learn just a linear transformation to provide a projection from the source feature space onto the target space.

According to the ways of verifications, Re-Id can be grouped into two categories: single-shot and multi-shot. Generally, to validate the effectiveness of a Re-Id approach, we should randomly choose the same number of images (candidates) from each person first, and then group these candidates with their labels as a gallery set. In the single-shot setting, there is only one image for each person in the gallery set [1,2]. Since single image of the target can hardly cover the changes of multi-pose, multi-camera and illumination, the traditional approaches have obtained few improvements under the single-shot setting.

Different with the single-shot, the multi-shot setting chooses two or more images to model a person [3–5,7]. In the multi-shot case, the query image matches with the different signatures (a signature represents a person with multiple images) and then the label of the query image is assigned to the signature, which has the smallest distance with the query image. The multi-shot setting provides more information and probability matching clues to classify the query image.

The proposed method is different from the previous works in three aspects. First, unlike the feature designing in the appearance based methods, SLM designs a framework. Under this framework, the features are fed into a features-class (Set-Label) structure, which can deeply discover the information from multiple features in the multi-shot setting. Second, different from the linear projection in the traditional metric learning methods, DeepML learns a non-linear transformation by a deep network, which can enhance the discriminative power of features. Finally, SLM and DeepML are combined into a single pipeline via the similarity measurement.

3. The proposed method

In this section, we introduce our proposed person Re-Id method. Our method consists of two parts: SLM and DeepML. In Section 3.1, the feature modeling approach SLM is introduced. Specifically, we construct SET between query feature and gallery set, and utilize mutual-information to measure relationship between features and their labels. In Section 3.2, we provide a nearest neighborhood based algorithm [12] to estimate the mutual information value. In Section 3.3, to enhance the discriminative power of the pairwise features in SLM, we further develop a non-linear metric learning approach named DeepML based on NCA and DBN. More details are given below.

3.1. Set-class model

The target of Re-Id is to predict the person Id (class label) of the given query image. Focusing on the multi-shot setting, we model multiple images as a representative signature, and propose SLM. The overview of SLM is shown in Fig. 1.

Following [18], we concatenate the query feature x_q and the feature x_j^c into the pairwise feature x_{qj}^c, where x_j^c is the feature of j-th image with label c in the gallery set, j is in the range of $[1, N_c]$, N_c is the number of images with label c in the gallery set. For simplicity in this paper, N_c is set to the constant N for all the labels. Then, features x_{qj}^c constitute a SET S_q^c, $x_{qj}^c \in S_q^c$. Fig. 2 demonstrates how these sets are formed.

To model the relationships between SETs and class labels, we propose a novel method based on the mutual-information theory. Mutual-information can measure the relevance between two random variation and has gained better results in many computer vision tasks, such as action detection [19]. In the task of person Re-Id, if the mutual-information value between the person feature and the class label is maximal, we can consider that the image has a higher probability to belong to this class(person). Thus, we reformulate the person Re-Id problem as follows:

$$\hat{c} = \underset{c \in \{1,2,\ldots,|C|\}}{\arg\max}\ MI\left(S_q^c; C = c\right) \tag{1}$$

where $MI(\bullet)$ denotes the mutual-information between SET and

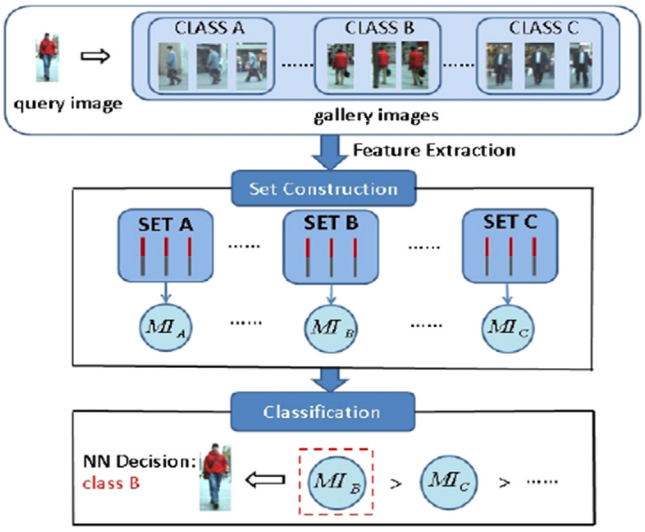


Fig. 1. The flowchart of SLM. (The strips combined with red and blue within one *SET* are pairwise features and *MI* represents the mutual-information between the *SET* and the class.) (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

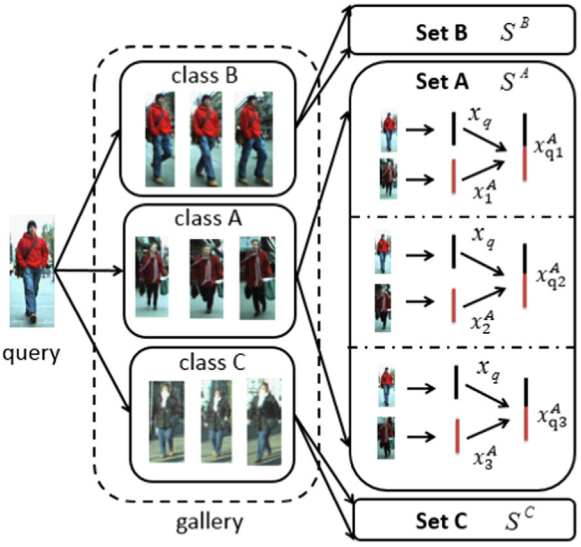


Fig. 2. The schematic diagram of *SET* construction: given a query image, it should be paired with all images in gallery to form a structure defined *SET* for each class.

class label, S_q^c denotes the SET for class label C (c is the specific value of C) and $|C|$ is the number of classes.

Assuming that the pair-wise features within one SET are i.i.d. (independently and identically distribute), we can define the mutual-information formulation in the following probability form:

$$MI\left(S_q^c; C=c\right)=\sum_{x_{qj}^c \in S_q^c} \log \frac{p\left(x_{qj}^c, C=c\right)}{p\left(x_{qj}^c\right)p(C=c)} \tag{2}$$

where $p\left(x_{qj}^c, C=c\right)$ donates the joint distribution of the pairwise feature x_{qj}^c and class label $C=c$, while $p\left(x_{qj}^c\right)$ and $p(C=c)$ are the factored distribution.

Based on Eq. (2), we can obtain a further derivation:

$$\frac{p\left(x_{qj}^c, C=c\right)}{p\left(x_{qj}^c\right)p(C=c)}=\frac{p\left(x_{qj}^c|C=c\right)}{p\left(x_{qj}^c\right)}$$

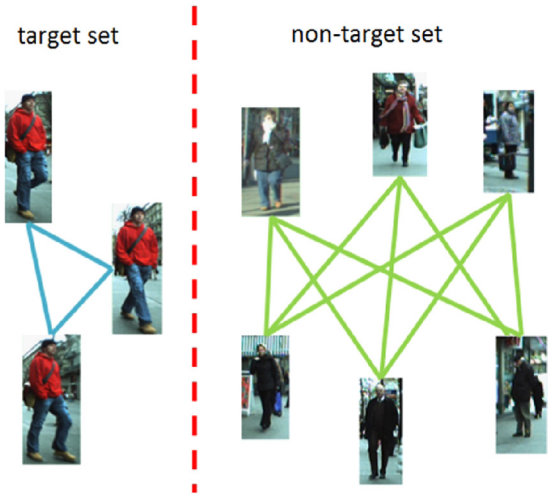


Fig. 3. The demonstration of constructing target set and non-target set.

$$=\frac{p\left(x_{qj}^c|C=c\right)}{p(x_{qj}^c|C=c)\cdot p(C=c)+p(x_{qj}^c|C\neq c)\cdot p(C\neq c)}$$

$$=\frac{1}{p(C=c)+\dfrac{p(x_{qj}^c|C\neq c)}{p(x_{qj}^c|C=c)}\cdot p(C\neq c)} \tag{3}$$

In Eq. (3), $p(C=c)$ and $p(C\neq c)$ can be directly considered as a known prior. Finally, if we can get the likelihood ratio item $p(x_{qj}^c|C\neq c)/p(x_{qj}^c|C=c)$, $MI\left(S_q^c; C=c\right)$ can be computed easily.

3.2. The approximation of the likelihood ratio

The likelihood ratio item has suffered the difficulties of direct accessing probability densities. To overcome this difficulty, we use an efficient algorithm to approximate the ratio value. Specially, we calculate the likelihood ratio by using NBNN [12].

NBNN holds a very simple form $\log p\left(x_{qj}^c|C\right)\propto -||x_{qj}^c-x_{NN}^c||^2$ to approximate the Gaussian kernel without dependence on the variance, where x_{NN}^c is the nearest neighbor of x_{qj}^c in class c. We construct a target set (c^+) and non-target set (c^-) based on the samples in the gallery set (Fig. 3). The target set consists of the positive pairwise features, while the non-target set consists of negative pairwise features of the dataset. As demonstrated in Fig. 3, the positive pairwise features for class c are the combinations of two arbitrary features within class c. On the contrary, the negative pairwise features for class c are the combinations of two features which do not belong to class c. The likelihood ratio item can be estimated as

$$\frac{p(x_{qj}^c|C\neq c)}{p(x_{qj}^c|C=c)}\propto \exp^{-\left(||x_{qj}^c-x_{NN}^{c-}||^2-||x_{qj}^c-x_{NN}^{c+}||^2\right)} \tag{4}$$

where x_{NN}^{c-} and x_{NN}^{c+} are the nearest neighbors of x_{qj}^c in the non-target c^- and target set c^+, respectively. Generally speaking, one of the drawbacks in the task of person re-identification is that there is no enough positive samples for training. By the combinations of two features, the number of positive samples is increased greatly and the information in the gallery set can be used completely. For example, the number of positive samples for class c is C_N^2 and N for our pairwise features and traditional features, respectively. N is the number of images with label c in the gallery set.

Based on (Eqs. (3) and 4), the mutual-information in Eq. (2) can be computed in another form by

$$MI(S^c; C=c) \approx \sum_{S^c} \log \frac{|C|}{1+\exp^{-\omega(x_{qj}^c)} \cdot (|C|-1)} \tag{5}$$

where $\omega(x_{qj}^c)$ denotes $||x_{qj}^c - x_{NN}^{c-}||^2 - ||x_{qj}^c - x_{NN}^{c+}||^2$, $P(C=c)$ and $P(C \neq c)$ are set to $1/|C|$ and $(1-1/|C|)$, respectively, and $|C|$ is the number of classes.

After constructing the SETs, the query features can be utilized for classification, because each query pair x_{qj}^c within one SET can provide a positive or a negative vote for class c. From Eq. (5), if the mutual-information score is positive, it is indicated that SET S_q^c votes a positive score for class c, and the query image is considered to belong to class c with a high probability. Otherwise, when the mutual-information is negative, the query image seems not belonging to class c. Finally, we make a decision by ranking all of the mutual-information scores. The query is classified to the class c, which obtains the highest score.

3.3. Deep metric learning

SLM does not consider the discriminative information of the gallery samples and can be seen as an unsupervised method. Generally speaking, supervised methods such as metric learning can improve the performances by using the discriminative information. In the following part, we introduce how to enhance the discriminative power of the SLM pairwise features by metric learning.

In Eq. (4), both $||x_{qj}^c - x_{NN}^{c-}||^2$ and $||x_{qj}^c - x_{NN}^{c+}||^2$ can be calculated by a distance metric. The distance metric utilized in SLM is Euclidean distance. There are many metric learning methods, which use a linear transformation (mapping) to project the features into a new subspace. But in the scenario of Re-Id, the appearance of persons can change violently for the large illumination changes, versatile view angles and intricate background noise. These changes cannot be well modeled by a linear transformation. We use a non-linear transformation, not the linear projection, to improve the distance metric. Specially, we address the metric learning problem by combining DBN and NCA. Through DBN, we can learn a non-linear feature representation, which has more powerful representation ability (as shown in many computer vision tasks [20–26]). Through NCA, the features sharing the same label can be closer and more discriminative than other neighbors. The flow chart of DeepML is illustrated in Fig. 4. To build the neural network, we use unsupervised training data to pre-train RBMs and stack the RBMs layer by layer into a DBN architecture according to [14]. After accomplishing the construction of the DBN network, we solve NCA by fine-tuning the parameters of deep neural network, and learn a non-linear distance metric.

Given the labeled training features $\{(x_{qj}^c, c_q)\}$, where x_{qj}^c is the pairwise feature of training sample q, c_q is the label of this sample, which is in the range of $\{1, 2, \ldots, C\}$. C is the number of classes. In the following, we use x_q to replace x_{qj}^c for simplicity. Then the DeepML objective function can be written as follows:

$$\Theta_{DeepML} = \sum_{a=1}^{N} \sum_{x_b : c_a = c_b} p(x_a, x_b) \tag{6}$$

$$p(x_a, x_b) = \frac{\exp(-d_W(x_a, x_b))}{\sum_{z \neq a} \exp(-d_W(x_a, x_z))} \tag{7}$$

$$d_W(x_a, x_b) = ||f(x_a|W) - f(x_b|W)||^2 \tag{8}$$

where x_a, x_b and x_z are training pairwise features, $x_b : c_a = c_b$ is the sample which has the same label with x_a, x_z is the neighbor of x_a. $p(x_a, x_b)$ is the likelihood of two pairwise features. $d_W(x_a, x_b)$ is the non-linear distance metric under the DBN network W, where

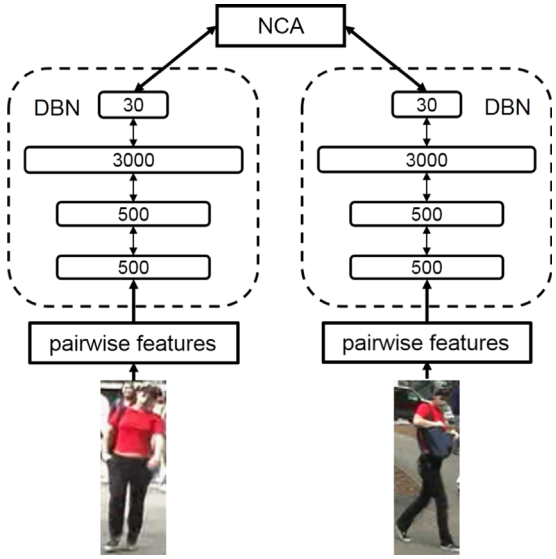


Fig. 4. The flow chart of DeepML.

$W = (W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)}, \ldots, W^{(n-1)}, b^{(n-1)})$, $W^{(l)}$ denotes the weight associated with the connection between layer l and layer $l+1$, $b^{(l)}$ is the bias associated in layer $l+1$. $f(x|W)$ is the non-linear feature transformation parameterized by W, which can be obtained by the following:

$$\begin{aligned} z^{(1)} &= x \\ z^{(2)} &= f(W^{(1)}x + b^{(1)}) \\ z^{(3)} &= f(W^{(2)}z^{(2)} + b^{(2)}) \\ &\vdots \\ f(x|W) &= z^{(n)} = f(W^{(n-1)}z^{(n-1)} + b^{(n-1)}) \end{aligned}$$

where $f(\cdot)$ can be chosen from sigmoid function, tangent function and so on. In our approach, we utilize sigmoid activation function.

The DeepML objective function can be optimized by the back propagation algorithm. Fine-tuning is carried out to adjust the weights of the whole DBN network [27–29]. We obtain the derivation of the learning objective function w.r.t parameter W (Eq. (9)):

$$\frac{\partial \Theta_{DeepML}}{\partial W} = \frac{\partial \Theta_{DeepML}}{\partial f(x|W)} \cdot \frac{\partial f(x|W)}{\partial W} \tag{9}$$

where $\partial \Theta_{DeepML} / \partial f(x|W)$ can be calculated by gradient descending algorithm [17]:

$$\begin{aligned} \frac{\partial \Theta_{DeepML}}{\partial f(x_a|W)} &= 2 \sum_{x_l : c_l = c_b} p(x_a, x_l)\, \mathrm{diff}(x_a, x_l) \\ &\quad - 2 \sum_{x_b : c_a = c_b} p(x_a, x_b)\, \mathrm{diff}(x_a, x_b) \\ &\quad + 2 \sum_{x_b : c_a = c_b} p(x_a, x_b) \left[\sum_{z \neq a} p(x_a, x_z)\, \mathrm{diff}(x_a, x_z) \right] \\ &\quad - 2 \sum_{x_l \neq x_a} \left[\sum_{q : c_l = c_q} p(x_l, x_q) \right] p(x_a, x_l)\, \mathrm{diff}(x_a, x_l) \\ \mathrm{diff}(x_a, x_b) &= f(x_a|W) - f(x_b|W) \end{aligned} \tag{10}$$

where $\partial f(x|W) / \partial W$ can be calculated by the standard back-propagation [29]. Through fine-tuning the weights of DBN, NCA and DBN are combined together to obtain a non-linear distance metric learning method. The procedure of the Fine-Tuning algorithm of DeepML is shown in Table 1.

Table 1
The Fine-Tuning algorithm of DeepML.

Input
$(x_i, c_i), i \in \{1, 2, \ldots, N\}$: the training samples; N: the number of training samples; $f(\bullet|W)$ the pre-trained DBN architecture.

Algorithm
for $i = 1 : N$ do gradient descending
1. Calculate $\frac{\partial \Theta_{DeepML}}{\partial f(x_i|W)}$ according to Eq. (10)
2. Calculate $\frac{\partial f(x_i|W)}{\partial W}$ using back-propagation
3. Obtain the derivation of the learning objective according to Eq. (9)
4. Update W using gradient descend
end

Output
$f(\bullet|W)_{FT}$ the fine-tuning DeepML network architecture

Fig. 5. Some images in the i-LIDS dataset. The images in the same column are belonging to the same person.

4. Experiments

In this section, to show the effectiveness, we test the proposed method on two public datasets: i-LIDS and ETHZ. For evaluation, we use the standard measurement named Cumulative Match Characteristic (CMC) curve, where the vertical coordinate donates the correct matching rates and the horizontal coordinate exploits top ranking k. According to the CMC curve, we can figure out the correct matches from the top suspected pedestrian images.

Dataset: In our validation, we use two standard datasets: i-LIDS and ETHZ, which are both widely used in the task of person re-identification. The i-LIDS dataset contains 476 images of 119 persons in total. All the images are captured by the multiple non-overlapping cameras at a busy airport. Each person has a set of averaging four images under the different camera views. The dataset is undergoing illumination changes, image blurring, low resolution, and occlusions. Meanwhile, as to dataset ETHZ, it consists of 8555 images of 146 persons. The images in ETHZ are captured by a moving camera in the busy streets. There are three video sequences including seq1 with 4857 images for 83 persons; seq2 with 35 persons in 1936 images; seq3 with 28 persons in 1762 images. Compared with seq3, seq1 and seq2 have suffered challenges incorporated in the large illumination changes, versatile view angles and intricate background noise. Since both of the dataset have more than two images per person, they can be used under the multi-shot Re-Id setting. The samples of i-LIDS and ETHZ are shown in Figs. 5 and 6, respectively.

Settings: In our experiments, we divide test images into two sets: gallery set and probe set. N images for each person are randomly selected to build gallery sets. Specially, for a person which has M images, we randomly permutate the M images. Then we choose the first N images to build the gallery sets. The remaining images are used as probe set. In this way, the gallery set has more than one images per person for multi-shot setting. For the discriminative learning method DeepML, we divide the images into two parts. One part is used for training and the other part for testing. Duringthe training stage, the samples of the same person are combined to form the positive pairs while the samples of the different person are randomly combined to form the negative pairs. In the testing phase, we apply DeepML on the pairwise features and execute the whole test via SLM. To make the comparison more fairly, we repeat the whole test procedure 10 times and average the recognition results as the final performance.

Features: For the features, we follow the feature extraction of Zheng in [30]. First, since the image sizes are quite different for both datasets, we normalize the images to the size of 128×64. Then, each image is divided into six horizontal stripes. For each stripe, we extract the features of the color histograms and texture information. Finally, an image is represented by a 2784 dimensional vector.

4.1. SLM versus appearance based methods

4.1.1. SLM versus Baseline

Under the standard multi-shot setting, Re-Id methods compare all possible pairs between the query feature and gallery features, and select the one which obtained lowest distance as the convinced signature [3]. Differently, we provide a framework for feature combination before the feature matching step. To valid the effectiveness of our framework SLM, we download Zheng's features from the website http://www.eecs.qmul.ac.uk/~jason/ilids.html and test them under the traditional multi-shot setting. In the following parts, simply we name the way of the traditional multi-shot setting as *Baseline*. We compare SLM with Baseline on both i-ILDS and

Fig. 6. Some images in the ETHZ dataset. The images in the same column are belonging to the same person.

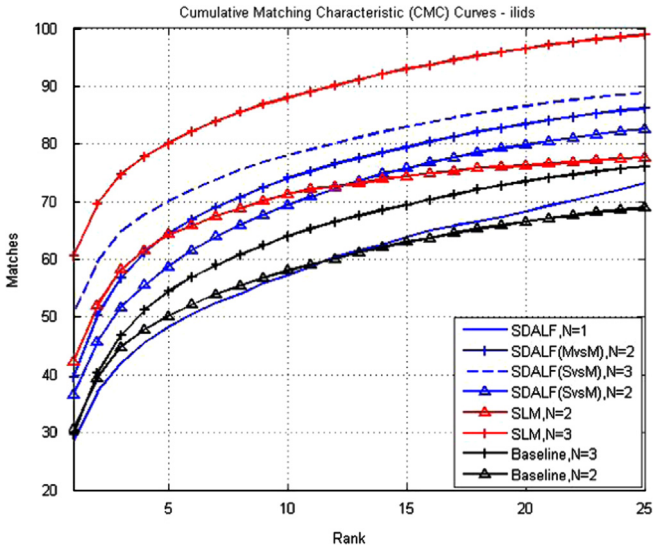


Fig. 7. CMC performance on dataset i-LIDS compared to [3].

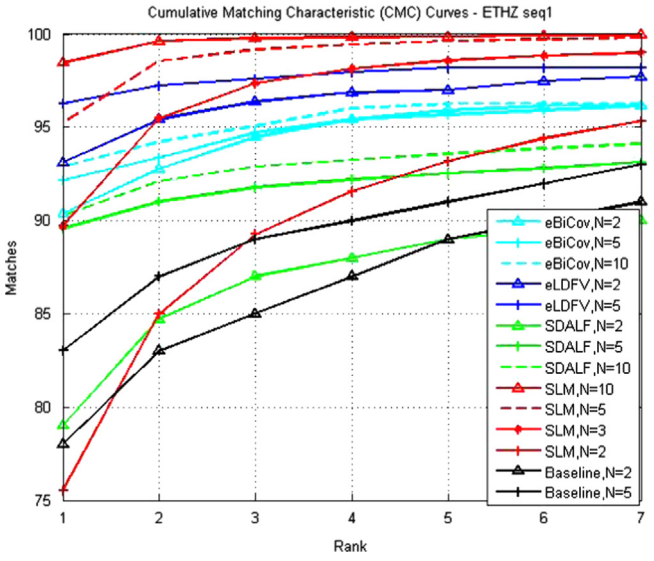


Fig. 8. CMC performance on the dataset ethz seq1 compared with the state-of-the-art methods.

ETHZ. The results are shown in Figs. 7–10. From these figures, we can know that on both the datasets, SLM can improve the performances significatively. We attribute the improvements to the proposed framework of SLM. Under the framework of SLM, the different features are combined to form the positive and negative samples. By this way, Re-Id can easily discover the similarity and dissimilarity from features of pedestrians.

4.1.2. *Performances on i-LIDS*

For the dataset i-LIDS, we compare SLM with the popular method SDALF [3]. Since the dataset has multiple images for each person, we randomly choose two or three images belonging to each person to build the gallery set. Then the positive and negative samples are selected from the gallery set to form the pairwise training set. By the training set, we compute the distances between the query feature and training samples. The distances can be used to gain the final classification results. Fig. 7 gives the final results. According to the CMC curve, it is easy to see that SLM outperforms SDALF in an obvious way. In Fig. 7, when N is changed from 2 to 3, the performance improvement of SLM is much bigger than that of other methods. For example, the accuracies of SLM, SDALF and Baseline are improved about 16%, 12% and 5% under rank 5, respectively. Thus, when using more images in the gallery set, we can obtain a further improvement in our experiment. It is obvious

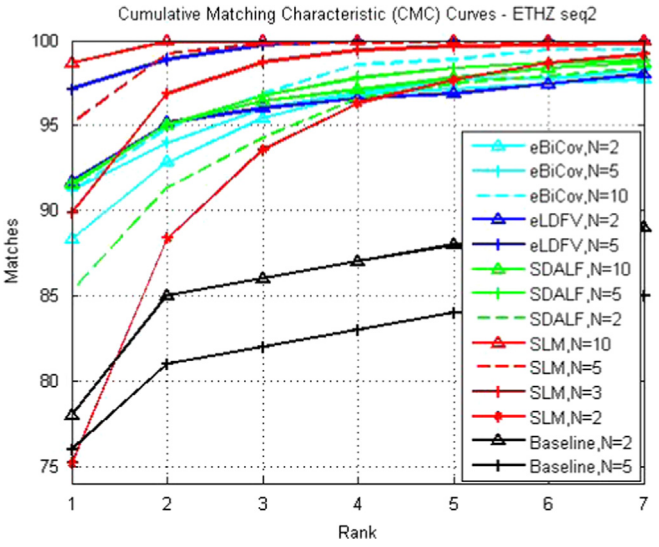


Fig. 9. CMC performance on the dataset ethz seq2 compared with the state-of-the-art methods.

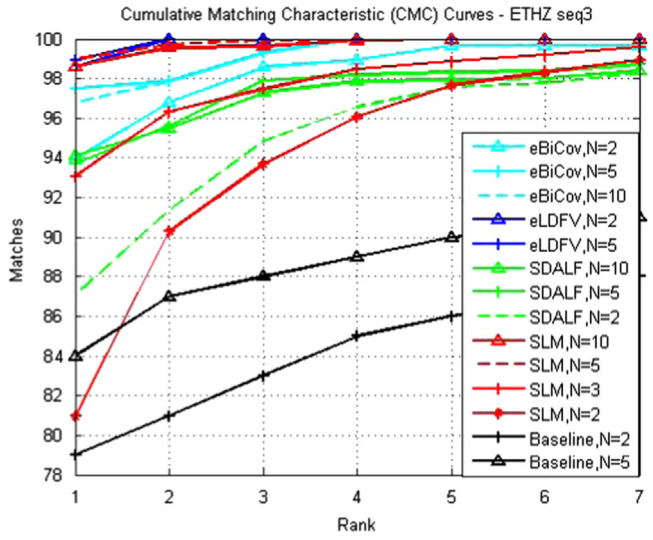


Fig. 10. CMC performance on the dataset ethz seq3 compared with the state-of-the-art methods.

to figure out that more feature combinations can contribute to higher accuracy on the Re-Id performance via our SLM framework.

4.1.3. Performances on ETHZ

For the dataset ETHZ, we compare SLM with the popular method SDALF [3] and the other two state of the art methods eBiCov [5] and eLDFV [7]. Different from the i-LIDS dataset, in the ETHZ dataset, we randomly select more images from the gallery set to build the training set. The results when N is set to {2,3,5,10} are reported in Figs. 8–10 for each sequence. When setting N to 3, SLM obtains better performance than the Baseline method. When N is assigned to 5, the experimental results show that SLM can obtain a big improvement. While N reaches 10, we perform the highest performance compared with the state-of-the-art. Especially, our CMC curve is close to 100% after rank 3 with $N=10$. As to seq3, we obtain a comparable performance with eLDFV, and better performance than the other methods. As our known, eLDFV [7] exhibits the best performance on ETHZ dataset in the literatures because it utilizes the powerful Fisher vector representation. Compared with eLDFV, SLM can get the better performances even with the simple features (Zheng's features).

4.2. DeepML versus metric learning based methods

4.2.1. Parameters setting

In the following experiments, we test the effectiveness of our DeepML integrated with SLM. N is set to 2 and 5 on i-LIDS and ETHZ, respectively. datasetOne affected parameter in deep learning is the number of the layers during the pre-training phase [14]. Table 2 demonstrates that the results are influenced by the number of layers in the deep architecture. In this experiment, we give the results of two layers RBM. Each layer has 1000 nodes. According to Table 2, two layers RBM obtains better performance than single layer RBM. This falls in with the common sense in deep leaning. But deeper the architecture is, more expensive it is in computing. In the following experiments, we choose a four layers RBM network [14]. The number of nodes of four layers are 500, 500, 3000, 30. The number of layers and nodes are chosen by the experimental results. The DeepML feature is composed by concatenating the responses of all layers. And further more, we concatenate the DeepML features with the original Zhang's feature. The combined feature is named as DeepML-hybrid.

Table 2
CMC performance of the pre-training phase on i-LIDS.

Layer	Matching rates								
	$R=1$	$R=2$	$R=3$	$R=4$	$R=5$	$R=6$	$R=7$	$R=8$	$R=9$
1	35.9	49.7	57.8	62.7	66.4	69.9	73.3	75.7	77.9
2	41.5	51.3	58.1	64.5	69.3	72.3	75.4	77.3	80.1

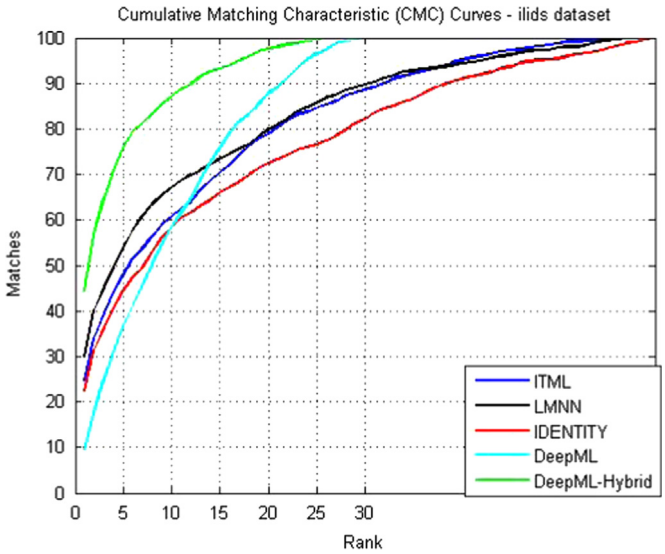


Fig. 11. CMC performance on the dataset i-LIDS based on metric learning approaches compared with the state-of-the-art methods.

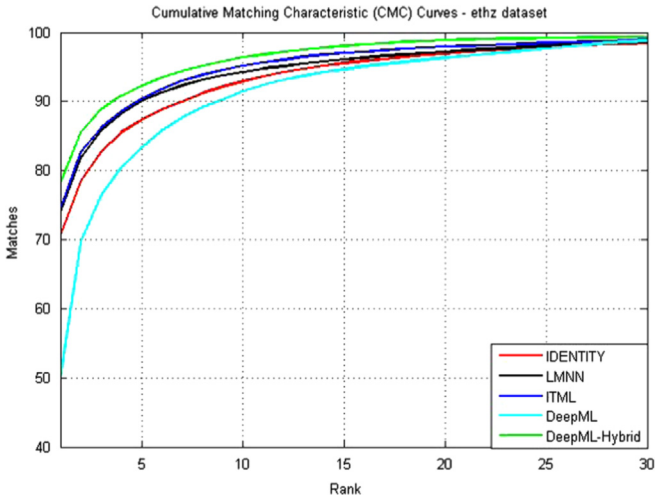


Fig. 12. CMC performance on the dataset ETHZ based on metric learning approaches compared with the state-of-the-art methods.

4.2.2. Performances

We compare the effectiveness of DeepML with many traditional metric learning methods, i.e. LMNN [17] and ITML [16]. The results are demonstrated in Figs. 11 and 12, respectively. According to the results, the improvements can be observed when the different methods apply the metric learning in re-weighting the features. From Fig. 11, we also can know that on the i-LIDS dataset, the matching rate of DeepML at rank 10 is lower than traditional metric learning methods. However, the combination feature DeepML-hybrid can gain a big higher performance compared with the deep features only. The same phenomenon can be observed in

Table 3
Matching performance in dataset i-LIDS: p is the number of persons used for testing (approach with ∗ donates the multi-shot experimental performance).

Approaches	$p=30$			$p=50$			$p=80$		
	$r=1$	$r=5$	$r=10$	$r=1$	$r=5$	$r=10$	$r=1$	$r=5$	$r=10$
RDC	44.05	72.74	84.69	37.83	63.70	75.09	32.60	54.55	65.89
Adaboost	35.58	66.43	79.88	29.62	55.15	68.14	22.79	44.41	57.16
LMNN	33.68	63.88	78.17	27.97	53.75	66.14	23.70	45.42	57.32
LMNN*	37.55	69.85	81.89	33.85	61.54	78.97	28.64	54.01	66.42
ITML	36.37	67.99	83.11	28.96	53.99	70.50	21.67	41.80	55.12
ITML*	37.72	68.87	85.09	31.28	63.85	81.54	23.46	48.33	64.63
MCC	40.24	73.64	85.87	31.28	59.30	75.62	12.00	33.66	47.96
Xing's	31.80	62.62	77.29	27.04	52.28	65.35	23.18	45.24	56.90
PLS	25.76	57.36	73.57	22.10	46.04	59.95	18.32	38.23	49.68
L1-norm	35.31	64.62	77.37	30.72	54.95	67.99	26.73	49.04	60.32
Bhat.	31.77	61.43	74.19	28.42	51.06	64.32	24.76	45.35	56.12
DeepML-Hybrid	**53.95**	**84.34**	**97.63**	**42.99**	**73.30**	**88.14**	**37.69**	**66.81**	**79.13**

The best results are marked with bold.

Table 4
Matching performance in dataset ETHZ: p is the number of persons used for testing (approach with ∗ donates the multi-shot experimental performance).

Approaches	$p=40$			$p=120$		
	$r=1$	$r=5$	$r=10$	$r=1$	$r=5$	$r=10$
RDC	72.65	90.08	95.59	61.58	79.70	86.65
Adaboost	69.21	87.76	93.54	60.73	78.82	85.66
LMNN	64.88	84.23	92.04	47.87	67.90	76.96
LMNN*	84.46	90.78	92.99	70.96	79.09	82.76
ITML	65.38	86.81	94.06	43.09	65.95	76.55
ITML*	**86.79**	91.49	93.49	72.35	80.20	84.03
MCC	71.92	90.96	**95.96**	31.08	59.40	73.19
Xing's	60.78	80.28	87.37	47.09	66.68	76.04
PLS	54.55	75.09	83.30	43.12	63.00	71.77
L1-norm	60.71	80.85	87.90	51.30	70.49	78.20
Bhat.	60.97	80.91	87.79	51.60	70.49	78.45
DeepML-Hybrid	85.13	**92.68**	95.09	**80.63**	**87.04**	**89.89**

The best results are marked with bold.

Fig. 12. The hybrid features outperform both LMNN and ITML. These results show that the DeepML feature and the original feature are mutually complementary. The combined features can carry more information about the samples.

To supplement more comparison experiments, we have compared our DeepML approach with the implemented results of the state-of-the-art in [31] (Tables 3 and 4). l_1-norm distance and Bhattacharyya distance are two nonlearning distances, which are commonly utilized in person Re-Identification method. Relative Distance Comparison (RDC) is formulated in [31], the Adaboost algorithm is proposed in [2] and the partial least squares (PLS) approach is introduced in [1]. All these three methods are learning-based person Re-Identification methods. Xing's method [32], LMNN [17], ITML [16], and MCC [33] are four popular distance metric learning methods.

In the both tables (Tables 3 and 4), p donates the number of testing persons, which means that the remaining persons are used in the training stage, r represents the top rank and the numbers filled in the table are the recognition rates. Upon the results of the i-LIDS dataset in Table 3, it is distinct to discover that our deep learning method has gained the highest performance over other metric learning methods. At the same time, according to Table 4, when p is assigned to 40, we have obtained a comparable results. Moreover, when we set p to 120, which means only a small fraction of images is put into training, the proposed method even makes the better results than other methods.

From the above results based on metric learning approaches, we can figure out three aspects of our conclusions. First, our non-linear deep metric learning method has gained significant improvements when combining the deep features and the original ones. Second, integrated with the SLM, DeepML can work well for the task of person Re-Id by using only small scale of training samples. Finally, the multi-shot outperforms the single-shot while using the traditional metric learning methods such as LMNN and ITML, because the multi-shot setting utilizes more information to obtain more stable features against vibrations.

4.3. Computational efficiency analysis

In our method, the number of positive samples for class c is C_N^2, where N is the number of randomly selected images for each person to build gallery sets. The number of negative samples for class c is $C_{(C-1)N}^2$, where C is the number of persons. The computational efficiency of Eq. (4) is $O(C^2N^2)$. N is usually less than 5 in our method. So the computational efficiency of our method is most influenced by the amount of persons.

5. Conclusion and future work

In this paper, we re-formulate Re-Id as a set-based classification problem from the perspective of information theory. Specifically, we define the set-based structure between the query image and the gallery images, and signify the relationship between the set and the class label by mutual-information. Further, we propose a non-linear metric learning method (DeepML), which is based on NCA and DBM. The proposed DeepML can both introduce the supervised information into SLM, and improve the generalization ability of SLM. The experimental results on the popular datasets demonstrate the effectiveness of our proposed methods.

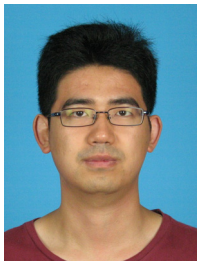
As a matter of fact, our approaches encounter high computational cost problem since we have to compute all positive and negative pairwise neighbors for each class. In the future, we will optimize the algorithm and extend our method with a speed-up algorithm.

Acknowledgments

This work was supported in part by National Basic Research Program of China (973 Program): 2012CB316400, in part by National Natural Science Foundation of China: 61025011, 61133003, 61332016, 61390510, 61303154.

References

[1] W. Schwartz, L. Davis, Learning discriminative appearance-based models using partial least squares, in: Computer Graphics and Image Processing, 2009.
[2] D. Gray, H. Tao, Viewpoint invariant pedestrian recognition with an ensemble of localized features, in: European Conference on Computer Vision, 2008.
[3] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person re-identification by symmetry-driven accumulation of local features, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010.
[4] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, V. Murino, Multiple-shot person re-identification by HPE signature, in: International Conference on Pattern Recognition, 2010.
[5] B. Ma, Y. Su, F. Jurie, Bicov: a novel image representation for person re-identification and face verification, in: British Machine Vision Conference, 2012.
[6] R. Layne, T. Hospedales, S. Gong, Person re-identification by attributes, in: British Machine Vision Conference, 2012.
[7] B. Ma, Y. Su, F. Jurie, Local descriptors encoded by Fisher vectors for person re-identification, in: International Workshop Conjunction with European Conference on Computer Vision, 2012, pp. 413–422.
[8] M. Kostinger, M. Hirzer, P. Wohlhart, P. Roth, H. Bischof, Large scale metric learning from equivalence constraints, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012.
[9] M. Hirzer, P. Roth, M. Kstinger, H. Bischof, Relaxed pairwise learned metric for person re-identification, in: European Conference on Computer Vision, 2012.
[10] W. Zheng, S. Gong, T. Xiang, Transfer re-identification: from person to set-based verification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012.
[11] W. Zheng, S. Gong, T. Xiang, Person re-identification by probabilistic relative distance comparison, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011.
[12] O. Boiman, E. Shechtman, M. Irani, In defense of nearest-neighbor based image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008.
[13] R. Salakhutdinov, G. Hinton, Learning a nonlinear embedding by preserving class neighbourhood structure, in: International Conference on Artificial Intelligence and Statistics, 2004.
[14] G. Hinton, R. Salakhutdinov, Reducing the dimensionality of data with neural networks, vol. 313, 2006, pp. 504–507.
[15] L. Qin, W. Gao, Image matching based on a local invariant descriptor, in: IEEE International Conference on Image Processing, vol. 3, 2005, pp. 77–80.
[16] J. Davis, B. Kulis, P. Jain, S. Sra, D. Suvrit, S. Inderjit, Information-theoretic metric learning, in: International Conference on Machine Learning, 2007.
[17] K. Weinberger, J. Blitzer, L. Saul, Distance Metric Learning for Large Margin Nearest Neighbor Classification, vol. 18, 2006, p. 1473.
[18] T. Avraham, I. Gurvich, M. Lindenbaum, S. Markovitch, Learning implicit transfer for person re-identification, in: International Workshop Conjunction with European Conference on Computer Vision, 2012, pp. 381–390.
[19] J. Yuan, Z. Liu, Y. Wu, Discriminative subvolume search for efficient action detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009.
[20] S.H. Khan, M. Bennamoun, F. Sohel, R. Togneri, Automatic feature learning for robust shadow detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014.
[21] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, Decaf: a deep convolutional activation feature for generic visual recognition, in: International Conference on Machine learning, 2014.
[22] R.B. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014.
[23] M. Oquab, L. Bottou, I. Laptev, J. Sivic, Learning and transferring mid-level image representations using convolutional neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014.
[24] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, Overfeat: integrated recognition, localization and detection using convolutional networks, in: International Conference on Learning Representations, 2014.
[25] A.S. Razavian, H. Azizpour, J. Sullivan, S. Carlsson, Cnn features off-the-shelf: an astounding baseline for recognition, in: IEEE Conference on Computer Vision and Pattern Recognition: DeepVision Workshop, 2014.
[26] Y. Liu, L. Qin, Z. Cheng, Y. Zhang, W. Zhang, Q. Huang, Da-ccd: a novel action representation by deep architecture of local depth feature, in: IEEE International Conference on Image Processing, 2014.
[27] G. Hinton, S. Osindero, Y. Teh, A fast learning algorithm for deep belief nets, in: Neural Computation, vol. 18, 2006, pp. 1527–1554.
[28] R. Salakhutdinov, Learning deep Boltzmann machines using adaptive MCMC, in: International Conference on Machine Learning, 2010.
[29] R. Salakhutdinov, G. Hinton, Deep Boltzmann machines, in: International Conference on Artificial Intelligence and Statistics, 2009.
[30] B. Prosser, W. Zheng, S. Gong, T. Xiang, Q. Mary, Person re-identification by support vector ranking, in: British Machive Vision Conference, 2010.
[31] W. Zheng, S. Gong, T. Xiang, Re-identification by Relative Distance Comparison, vol. 35, 2013, pp. 653–668.
[32] E. Xing, A. Ng, M. Jordan, S. Russell, Distance metric learning with application to clustering with side-information, in: Advances in Neural Information Processing Systems, 2002.
[33] A. Globerson, S. Roweis, Metric learning by collapsing classes, in: Advances in Neural Information Processing Systems, 2005.

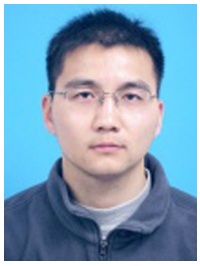


Hao Liu received the B.S. degree in software engineering from Sichuan University, China, in 2011 and the Engineering Master degree in computer technology from University of Chinese Academy of Sciences, China, in 2014. He is currently pursuing the Ph.D. degree at the department of automation, Tsinghua University. His research interests include medical imaging, computer vision, and deep learning.

Bingpeng Ma received the B.S. degree in mechanics, in 1998 and the M.S. degree in mathematics, in 2003 from Huazhong University of Science and Technology. He received Ph.D. degree in computer science at the Institute of Computing Technology, Chinese Academy of Sciences, P.R. China, in 2009. He was a post-doctorial researcher in University of Caen, France, from 2011 to 2012. He joined the School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, in March 2013 and now he is an assistant professor. His research interests cover image analysis, pattern recognition, and computer vision.

Lei Qin received the B.S. and M.S. degrees in mathematics from the Dalian University of Technology, Dalian, China, in 1999 and 2002, respectively, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2008. He is currently an associate professor with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His research interests include image/video processing, computer vision, and pattern recognition. He has authored or coauthored over 40 technical papers in the area of computer vision. Dr. Qin is a reviewer for IEEE Transactions on Multimedia, IEEE Transactions on Circuits and Systems for Video Technology, and IEEE Transactions on Cybernetics. He has served as TPC member for various conferences, including ECCV, ICPR, ICME, PSIVT, ICIMCS, PCM.

Junbiao Pang received the B.S. and the M.S. degrees in computational fluid dynamics and computer science from the Harbin Institute of Technology, Harbin, China, in 2002 and 2004, respectively, and the Ph.D. degree at the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, in 2011. He is currently a faculty member with the College of Metropolitan Transportation, Beijing University of Technology, Beijing, China. His research areas include computer vision, multi-media and machine learning, and he has authored or coauthored approximately ten technical papers.



Chunjie Zhang received his Ph.D. degree in Pattern Recognition and Intelligent Systems from Institute of Automation, Chinese Academy of Sciences, China in 2011. He received his B.E. degree from Nanjing University of Posts and Telecommunications, China, 2006. He worked as an engineer in the Henan Electric Power Research Institute during 2011–2012. He is currently working as a postdoc at Graduate University of Chinese Academy of Sciences, Beijing, China. Dr. Zhang's current research interests include image processing, machine learning, cross media content analysis, pattern recognition and computer vision.

Qingming Huang (SM'08) received the B.S. degree in computer science and Ph.D. degree in Computer Engineering from Harbin Institute of Technology, Harbin, China, in 1988 and 1994, respectively. He is currently a professor with the University of the Chinese Academy of Sciences (CAS), China, and an Adjunct Research Professor with the Institute of Computing Technology, CAS. His research areas include multimedia computing, image processing, computer vision, pattern recognition and machine learning. He has published more than 200 academic papers in prestigious international journals including IEEE Transactions on Multimedia, IEEE Transactions on CSVT and IEEE Transactions on Image Processing, and top-level conferences such as ACM Multimedia, ICCV, CVPR and ECCV. He is the associate editor of Acta Automatica Sinica, and the reviewer of various international journals including IEEE Transactions on Multimedia, IEEE Transactions on CSVT, and IEEE Transactions on Image Processing. He has served as a program chair, track chair and TPC member for various conferences, including ACM Multimedia, CVPR, ICCV, ICME and PSIVT.