Pattern Recognition 72 (2017) 446–457

Contents lists available at ScienceDirect

Pattern Recognition

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

Multi-modal uniform deep learning for RGB-D person re-identification

Liangliang Ren[a,b,c], Jiwen Lu[a,b,c,*], Jianjiang Feng[a,b,c], Jie Zhou[a,b,c]

[a] Department of Automation, Tsinghua University, Beijing, 100084, China
[b] State Key Lab of Intelligent Technologies and Systems, Beijing, 100084, China
[c] Tsinghua National Laboratory for Information Science and Technology (TNList), Beijing, 100084, China

ARTICLE INFO

Article history:
Received 25 January 2017
Revised 10 May 2017
Accepted 30 June 2017
Available online 4 July 2017

Keywords:
Person re-identification
Deep learning
Multi-model learning

ABSTRACT

In this paper, we propose a multi-model uniform deep learning (MMUDL) method for RGB-D person re-identification. Unlike most existing person re-identification methods which only use RGB images, our approach recognizes people from RGB-D images so that more information such as anthropometric measures and body shapes can be exploited for re-identification. In order to exploit useful information from depth images, we use the deep network to extract efficient anthropometric features from processed depth images which also have three channels. Moreover, we design a multi-modal fusion layer to combine these features extracted from both depth images and RGB images through the network with a uniform latent variable which is robust to noise, and optimize the fusion layer with two CNN networks jointly. Experimental results on two RGB-D person re-identification datasets are presented to show the efficiency of our proposed approach.

1. Introduction

Person re-identification aims at recognizing individuals across different cameras with non-overlapping areas, which is a significant problem in computer vision and has gained a lot of attention in recent years [1–3]. While a variety of methods have been proposed in the literature [4–6], it is still a challenging problem to re-identify persons in wild conditions where large intra-class variations of illumination, pose, resolution and occlusion usually occur in pedestrian images.

Most current person re-identification methods focus on matching pedestrians with appearance features, and those methods can be divided into two categories: image-based [3,7,8] and video-based [1,9]. For the first category, methods [3,8,10–12] focus on seeking effective feature descriptors which are robust to the changes of light, pose and viewing angle, and discriminative similarity metrics for person matching. Video-based methods [13,14] focus on promising video modeling and matching techniques to reduce the influences of occlusion and illumination changes. Due to the large intra-class divergence (i.e., persons in different viewpoints or different lighting conditions) and low inter-class divergence (i.e., persons with similar clothes), the accuracy of appearance-based methods is usually low in some specific situations such as schools where students wear uniform. To this end, some researchers proposed methods to combine the appearance features with other modalities, such as thermal data [15], gait [16], and anthropometric measures [15,17,18]. These modalities are robust to varying light conditions, view points, and clothes changing. Several datasets collected with RGB-D sensors and infrared cameras have been proposed recently to evaluate the performance of these methods [17–19]. Experimental results on these datasets show that those features can improve the re-identification performance.

In this paper, we focus on person re-identification from RGB-D images. While RGB-D images contain more information than RGB images, there are two main challenges: (1) how to combine those two modalities, and (2) how to extract efficient discriminative features from depth images. Unlike current anthropometric measures [17,18] which only use twenty skeletal points extracted from depth images, our approach uses a convolutional neural networks (CNN) network to extract more discriminative anthropometric feature from processed depth images. As shown in Fig. 1, given the depth images and RGB images of a pedestrian, we use another CNN which has the same framework but the different parameters to extract appearance features from RGB images, and combine the output of the two CNN networks to a uniform latent variable which has three parts: the depth specific part, the sharable part and the RGB specific part. Experimental results on two RGB-D person re-identification datasets are presented to show the efficiency of our proposed approach.

* Corresponding author at: Department of Automation, Tsinghua University, Beijing, 100084, China.
E-mail address: lujiwen@tsinghua.edu.cn (J. Lu).

http://dx.doi.org/10.1016/j.patcog.2017.06.037
0031-3203/

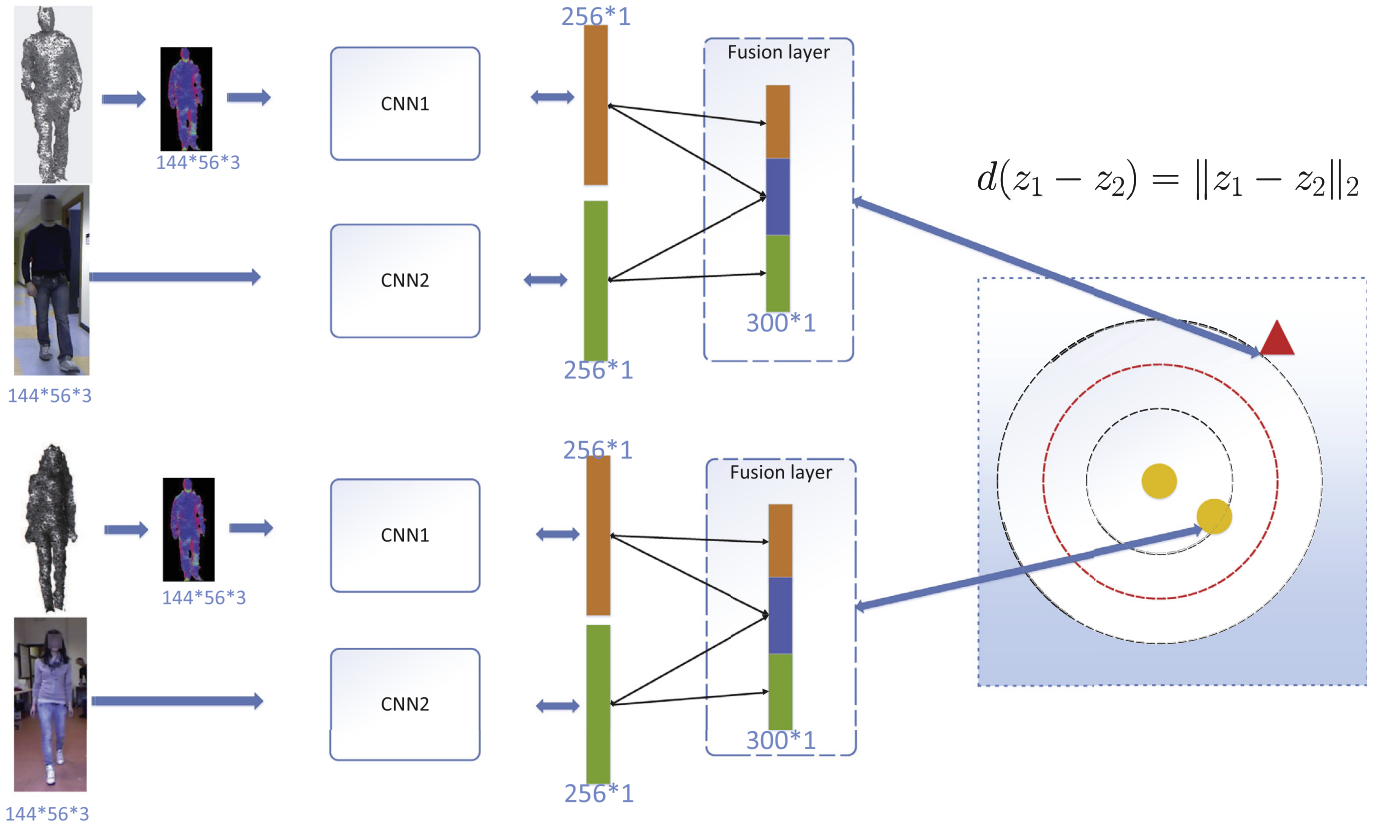

Fig. 1. The basic idea of the proposed multi-modal uniform deep learning (MMUDL) method for RGB-D person re-identification. First, we process each depth image to 3 channels, and extract an anthropometric feature vector by using a CNN network. We extract an appearance feature vector from RGB image by using another CNN networks. The anthropometric feature vector and the appearance feature vector are combined in a fusion layer with a uniform latent variable. The latent variable contains three parts: the depth specific part, the sharable part and the RGB specific part. Lastly, we compute the distance of the uniform latent variables of two persons for person re-identification and adjust the parameters of the fusion layer and CNN networks accordingly.

The contributions of this work are summarized as follows:

(1) To our best knowledge, our work is the first attempt to use deep network to deal with depth images for person re-identification. To exploit effective features from RGB-D images, we propose a multi-modal uniform deep learning (MMUDL) method to extract the anthropometric and appearance features from RGB-D images for RGB-D person re-identification.
(2) Experimental results on two RGB-D based person re-identification datasets, the Kinect-REID dataset and the RGBD-ID dataset, are presented to show the efficiency and the robustness of our proposed approach.

The rest of this paper is organized as follows: Section 2 reviews the current studies of person re-identification and deep learning. Section 3 details our multi-modal uniform deep learning approach for RGB-D person re-identification. Section 4 shows the experimental results and Section 5 concludes this paper finally.

2. Related work

2.1. Person re-identification

Most existing person re-identification methods focus on matching pedestrians with appearance features [3,5–8,12,13,20–30], which can be mainly divided into two categories: image-based and video-based. Methods in the first category aim to extract static characteristics such as clothing and body shapes for person re-identification. For example, Zhao et al. [12] proposed a matching strategy based on the relationship of salience distributions between still images. Xiong et al. [3] employed LBP features and color histograms to evaluate the effects of different spatial splitting criterion. Bazzani et al. [31] proposed a symmetry-driven accumulation of local feature method by using three important static characteristics. Ma et al. [10] introduced a biologically inspired features and covariance descriptors method for person re-identification. For the second category, statics features of each pedestrian image set are extracted for person representation. For example, Bazzani et al. [21] extracted and matched features which embed global and local appearance features. Karanam et al. [13] learned a dictionary to discriminate sparse codes corresponding to the feature vectors and computed the Euclidean distance between codes for person matching. Wu et al. [32] proposed an image sequence hierarchical clustering method and used the representative samples to learn a feature subspace. Gheissari et al. [1] presented a spatiotemporal segmentation method to generate salient edges which are robust to appearance changes. Bak et al. [5] minimized the perspective distortions from the video streams by using affine transformations. Wang et al. [22] introduced a model to select discriminative video fragments to obtain spatiotemporal features. McLaughlin et al. [9] presented a recurrent feature aggregation network to address the multi-shot person re-identification problem. Ma et al. [30] proposed an unsupervised method with surveillance image-sequences for video-based person re-identification.

In order to improve the accuracy of appearance based person re-identification methods, some researchers proposed methods to combine the appearance features with other modalities, such as thermal data [15], gait [16], and anthropometric measures [15,17,18]. For example, Bialkowski [19] combined color, height and texture information with the pose and lighting conditions. Kawai [16] introduced a view-dependent score-level fusion method to

combine gait and color features. Barbosa [17] presented a new approach for person re-identification that only uses soft biometrics cues as features. Mogelmose [33] proposed a tri-modal re-identification system based on RGB, depth, and thermal descriptors. Figueira [34] proposed a semi-supervised multi-features learning framework to process the appearance-based and learning-based re-identification problem. Pala [18] investigated that anthropometric measures can improve the re-identification performance of the widely used clothing appearance cue, and combined some chosen anthropometric measures with several different clothes appearance descriptors.

2.2. Deep learning

Deep learning has gained great successes on several computer vision applications such as image classification [35–39], face recognition [40–42], and object detection [43–47]. There are also some methods which applied deep learning to person re-identification in recent years. For example, Li et al. proposed a deep filter pairing neural network [48] to jointly handle misalignments, photometric and geometric transforms, occlusions and background clutter. Yi et al. [49] presented a Siamese CNN deep architecture for person re-identification, where three S-CNNs were employed for deep feature learning. Ding et al. [27] proposed a scalable deep feature learning method for person re-identification via the maximum relative distance. Ahmed et al. proposed a cross-input neighborhood difference method [50] to extract the cross-view relationships of the features. Cheng et al. proposed a framework to deal with local features and the global features [51]. Wang et al. [52] proposed a framework which contains one shared sub-network together with two sub-networks to extract single-image and cross-image representations respectively. Yan et al. [53] proposed a recurrent feature aggregation network to generate highly discriminative sequence representations. Xiao et al. [54] proposed a domain guided drop out (DGD) method [2] to improve feature learning by selecting the neurons specific to certain domains. Varior et al. [55] proposed a long short term memory method to process image regions sequentially and enhance the discriminative capability of local feature representation by leveraging contextual information. More recently, Varior et al. also proposed a gated Siamese CNN architecture to selectively emphasize fine common local patterns by comparing the mid-level features across pairs of images.

3. Proposed approach

3.1. Multi-modal learning

Let $\mathbf{X}=\{X_1, X_2, \ldots, X_n\}$ be the training set of n samples. For each sample $X_i=\{x_i^1, x_i^2, \ldots, x_i^K\}$, feature representations from K views $x_i^k \in \mathbb{R}^{d_k}$ are extracted. We assume that those samples have a uniform but unknown variable $z \in \mathbb{R}^{d_0}$ in a latent space $\mathbf{Z}$. If z_i is known, the conditional distribution of each view x_i^k is independent to each other. The joint probability distribution can be decomposed as follows:

$$p(x_i^1, x_i^2, \ldots, x_i^K, z_i) = p(z_i)\prod_{k=1}^{K} p(x_i^k|z_i) \tag{1}$$

Since it is hard to get the value of the latent variable z_i, we estimate the distribution of z_i through the Bayesian probability theory:

$$p(z_i|x_i^1, x_i^2, \ldots, x_i^K) = \frac{p(z_i, x_i^1, x_i^2, \ldots, x_i^K)}{m(x_i^1, x_i^2, \ldots, x_i^K)} \tag{2}$$

where

$$m(x_i^1, x_i^2, \ldots, x_i^K) = p(x_i^1, x_i^2, \ldots, x_i^K) = \int p(z_i)\prod_{k=1}^{K} p(x_i^k|z_i)dz_i \tag{3}$$

We assume that the distribution of latent variable z is a standard normal distribution, and the conditional distribution of $p(x_i^k|z_i)$ is also a normal distribution with the covariance matrix Σ_k and the mean value of $\omega_k z_i$.

$$z_i \sim \mathcal{N}(0, I) \tag{4}$$

$$p(x_i^k|z_i) \sim \mathcal{N}(\omega_k z_i, \Sigma_k) \tag{5}$$

Then, the joint probability distribution is computed as:

$$\begin{aligned} p(z_i, x_i^1, x_i^2, \ldots, x_i^K) &= \frac{e^{-\frac{z_i^T z}{2}}}{\sqrt{(2\pi)^{d_0}}}\prod_{k=1}^{K}\frac{e^{-\frac{1}{2}(x_i^k-\omega_k z_i)^T\Sigma_k(x_i^k-\omega_k z_i)}}{\sqrt{(2\pi)^{d_k}|\Sigma_k|}} \\ &= \alpha e^{-\frac{1}{2}(z_i^T\beta z_i+2\gamma^T z_i+\theta)} \\ &= \alpha e^{-\frac{1}{2}(z_i^T\beta z_i+(\beta^{-1}\gamma)^T\beta z_i+z_i^T\beta(\beta^{-1}\gamma)+\theta)} \\ &= \alpha e^{-\frac{1}{2}((z_i-\beta^{-1}\gamma)^T\beta(z_i-\beta^{-1}\gamma)-\gamma^T\beta^{-1}\gamma+\theta)} \end{aligned} \tag{6}$$

where α, β, γ, θ:

$$\alpha = \frac{1}{(2\pi)^{\frac{\sum_{k=0}^{K} d_k}{2}}\prod_{k=1}^{K}|\Sigma_k|^{\frac{1}{2}}} \tag{7}$$

$$\beta = I + \sum_{k=1}^{K}\omega_k^T\Sigma^{-1}\omega_k \tag{8}$$

$$\gamma = \sum_{k=1}^{K} x_i^{kT}\Sigma_k^{-1}\omega_k \tag{9}$$

$$\theta = \sum_{k=1}^{K} x_i^{kT}\Sigma_k^{-1}x_i^k \tag{10}$$

We calculate the conditional exception of z_i as the best estimation.

$$\begin{aligned} \mathbb{E}_{\Sigma_i,\omega_i}(z_i|x_i^1, x_i^2, \ldots, x_i^K) &= \int z_i p_{\Sigma_k,\omega_k}(z|x_i^1, x_i^2, \ldots, x_i^K)dz \\ &= \beta^{-1}\gamma^T \\ &= \left(I + \sum_{k=1}^{K}\omega_k^T\Sigma^{-1}\omega_k\right)^{-1} \\ &\quad\times\left(\sum_{k=1}^{K}\omega_k^T\Sigma_k^{-1}x_i^k\right) \end{aligned} \tag{11}$$

The conditional exception of z_i contains two terms: the first term is $(I+\sum_{k=1}^{K}\omega_k^T\Sigma^{-1}\omega_k)^{-1}$, which is a normalization term, the second term is $(\sum_{k=1}^{K}\omega_k^T\Sigma_k^{-1}x_k)$, which refers to the weighted summation of each views x_i^k, and the weight of view x_i^k is determined by the variance of probability distribution of prior condition $p_{x_i^k|z}$. Fig. 2 shows an example of the weights learned in the latent space with depth images and RGB images.

3.2. MMUDL

We take the depth images and color images of the ith people as two views x_i^1, x_i^2. we consider that the latent variable z contains three terms: $z^{(1)}$, $z^{(2)}$, $z^{(1,2)}$.

$$\mathbb{E}(x_i^1|z_i) = W_1 z_i^{(1)} + W_3 z_i^{(1,2)} = \omega_1 z_i \tag{12}$$

$$\mathbb{E}(x_i^2|z_i) = W_2 z_i^{(2)} + W_4 z_i^{(1,2)} = \omega_2 z_i \tag{13}$$

where

$$z_i = [z_i^{(1)}, z_i^{(1,2)}, z_i^{(2)}], \omega_1 = \begin{bmatrix} W_1 & W_3 & 0 \end{bmatrix}, \omega_2 = \begin{bmatrix} 0 & W_4 & W_2 \end{bmatrix}$$

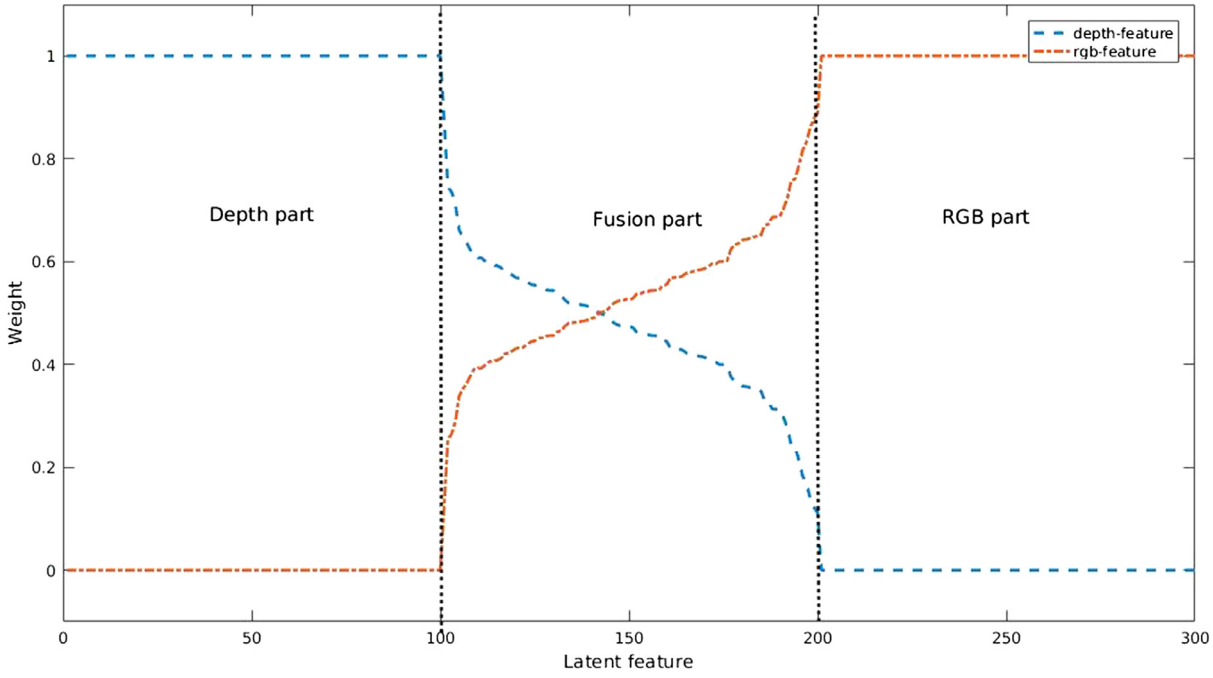

Fig. 2. The contribution weight of depth images (blue) and RGB images(red) to the latent variable on the Kinect-REID dataset. The latent variable contains three parts: the depth specific part, the sharable part and the RGB specific part. For the specific part, the latent variable value is only determined by one view, while the latent variable in the fusion part is determined by two views. As shown in this figure, for the fusion part, the weights of the depth part and the RGB part in each dimension are not always $\frac{1}{2}$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$z_i^{(1)}$ refers to the depth specific part, and $z_i^{(2)}$ refers to the RGB specific part, and $z_i^{(1,2)}$ refers to the share part.

We formulate our multi-modal uniform deep learning approach as the following optimization problem:

$$\begin{aligned}\arg\min_{w_i}\ & \sum_{i,j}^{N} g(\ell_{ij}, d(z_i, z_j)) - \lambda_1 \sum_{j=1}^{N} (p(x_i^1|z_j) + p(x_i^2|z_j)) \\ s.t.\ & \omega_i^T \omega_i = \mathbb{I}_{d_i}, 1 \le i \le K \\ & z_i = \left(I + \sum_{k=1}^{2} \omega_k^T \Sigma^{-1} \omega_k\right)^{-1} \left(\sum_{k=1}^{2} \omega_k^T \Sigma_k^{-1} x_i^k\right) \\ & \Sigma_i = diag\left(\frac{1}{N} \sum_{j=1}^{N} (x_{i,j} - \omega_i z_j)^2\right)\end{aligned} \tag{14}$$

where $p(x_i^k|z_j)$ is the posterior probability of x_i, and $g(\ell_{ij}, d(z_i, z_j))$ [56] is the generalized logistics loss function to approximate the hinge loss function $z = \max(z, 0)$. $p(x_i^k|z_j)$ and $g(\ell_{ij}, d(z_i, z_j))$ are defined as follows:

$$p(x_i^k|z_j) = \frac{1}{\sqrt{(2\pi)^{d_k}|\Sigma_k|}} \exp\{-(x_i^k - W_{k,2+k} z_i)^T \Sigma^{-1} (x_i^k - W_{k,2+k} z_i)\}$$

$$g(\ell_{ij}, d(z_i, z_j)) = \frac{1}{\theta} log(1 + \exp(\theta \ell_{ij} (\tau - d(z_i - z_j))))$$

where θ is a sharpness parameter and τ is the threshold parameter.

There are two terms in the optimization problem in (14). The first term is to minimize the distance of positive pairs and maximize the distance of negative pairs, and the second term is to maximize the posterior probability of the uniform latent variables, which also presents the weighted reconstruction error. λ_1 is parameter to balance the influence of different parts. To simplify the optimization, we add the orthogonal constraints of ω_1, ω_2 to the objective function with parameter λ_2.

$$\begin{aligned}\arg\min_{\omega_1,\omega_2} J = & \sum_{i,j}^{N} g(\ell_{ij}, d(z_i, z_j)) - \lambda_1 \sum_{j=1}^{N} (p(x_i^1|z_j) + p(x_i^2|z_j)) \\ & + \lambda_2 \sum_{k=1}^{2} \|\omega_k \omega_k^T - \mathbb{I}_{d_k}\|_F^2 \\ s.t.\ \ & z_i = \left(I + \sum_{k=1}^{2} \omega_k^T \Sigma^{-1} \omega_k\right)^{-1} \left(\sum_{k=1}^{2} \omega_k^T \Sigma_k^{-1} x_i^k\right) \\ & \Sigma_i = diag\left(\frac{1}{N} \sum_{j=1}^{N} (x_{i,j} - \omega_i z_j)^2\right)\end{aligned} \tag{15}$$

To solve the optimization problem, we use the stochastic subgradient descent algorithm to obtain the parameters ω_1, ω_2 and fine tune the CNN networks. The gradients of the objective function J with respect to the parameters ω_1, ω_2 can be computed as follows:

$$\begin{aligned}\frac{\partial J}{\partial \omega_1} = & \sum_{i,j}^{N} g'(\ell_{ij}, d(z_i, z_j))(z_i - z_j)(x_i^1 - x_j^1)^T \\ & + \lambda_1 \left(\sum_{i=1}^{N} 2p(x_i^1|z_i)(\Sigma_1^{-1}(\omega_1 z_i - x_i^1) z_i' + x_i^1 (\omega_1 z_i - x_i^1)^T \Sigma_1^{-1} \omega_1) \right. \\ & \left. + \sum_{i=1}^{N} 2p(x_i^2|z_i) x_i^1 (\omega_2 z_i - x_i^2)^T \Sigma_2^{-1} \omega_2 \right) \\ & + 2\lambda_2 (\omega_1 \omega_1^T - \mathbb{I}_{d_1}) \omega_1\end{aligned} \tag{16}$$

Algorithm 1: MMUDL.

Input: Training set X, parameters: λ_1, λ_2, learning rate ρ, total iterative number Γ, and convergence error ε.
Output: Parameters: W_1, W_2, W_3, W_4.
Initialize W_1, W_2, W_3, W_4 according to (21–22)
Estimation Σ_1 and Σ_2 according to (14)
for $t = 1, 2, \cdots, \Gamma$ **do**
  Randomly select a batch of **X**.
  for $X_i \in$ **X do**
    Extract features $\{x_i^1\}$ and $\{x_i^2\}$ for **X** using neural network CNN1 and CNN2
  end
  Calculate the latent variable $\{z_i\}$
  Calculate the gradient $\frac{\partial J}{\omega_1}, \frac{\partial J}{\omega_2}$ according to (16 - 17)
  Update for parameter set W_1, W_2, W_3, W_4.
  for $X_i \in$ **X do**
    Calculate the gradient $\frac{\partial J}{x_i^1}, \frac{\partial J}{x_i^2}$ according to (16 - 17)
    Back propagation $\frac{\partial J}{x_i^1}$ to adjust CNN1
    Back propagation $\frac{\partial J}{x_i^2}$ to adjust CNN2
  end
  Estimation Σ_1 and Σ_2 according to (14)
  Calculate J_t using (15).
  If $t > 1$ and $|J_t - J_{t-1}| < \varepsilon$, go to **Return**.
end
Return: W_1, W_2, W_3, W_4.

$$\begin{aligned}\frac{\partial J}{\partial \omega_2} &= \sum_{i,j}^{N} g'(\ell_{ij}, d(z_i, z_j))(z_i - z_j)(x_i^2 - x_j^2)^T \\ &+ \lambda_1 \left(\sum_{i=1}^{N} 2p(x_i^2|z_i)(\Sigma_2^{-1}(\omega_2 z_i - x_i^2)z_i' + x_i^2(\omega_1 z_i - x_i^2)^T \Sigma_2^{-1}\omega_2) \right. \\ &\left. + \sum_{i=1}^{N} 2p(x_i^1|z_i)x_i^2(\omega_1 z_i - x_i^1)^T \Sigma_1^{-1}\omega_1 \right) \\ &+ 2\lambda_2(\omega_2\omega_2^T - \mathbb{I}_{d_2})\omega_2 \end{aligned} \tag{17}$$

Then, ω_1, ω_2 can be updated as follows:

$$\omega_k = \omega_k - \rho \frac{\partial J}{\omega_k} \tag{18}$$

where ρ is the learning rate.

The gradients of the objective function J with the output of CNN networks x_i^1 and x_i^2 can be computed as follows:

$$\begin{aligned}\frac{\partial J}{\partial x_i^1} &= \sum_{j}^{N} g'(\ell_{ij}, d(z_i, z_j))\omega_1(z_i - z_j) \\ &+ 2\lambda_1(p(x_i^1|z_i)(\omega_1\omega_1^T - I)\Sigma_1^{-1}(\omega_1 z_i - x_i^1)) \\ &+ p(x_i^2|z_i)\omega_1\omega_2^T\Sigma_2^{-1}(\omega_2 z_i - x_i^2)) \end{aligned} \tag{19}$$

$$\begin{aligned}\frac{\partial J}{\partial x_i^2} &= \sum_{j}^{N} g'(\ell_{ij}, d(z_i, z_j))\omega_2(z_i - z_j) \\ &+ 2\lambda_1(p(x_i^2|z_i)(\omega_2\omega_2^T - I)\Sigma_2^{-1}(\omega_2 z_i - x_i^2)) \\ &+ p(x_i^1|z_i)\omega_2\omega_1^T\Sigma_1^{-1}(\omega_1 z_i - x_i^1)) \end{aligned} \tag{20}$$

Algorithm 1 summarizes the procedure of our proposed MMUDL.

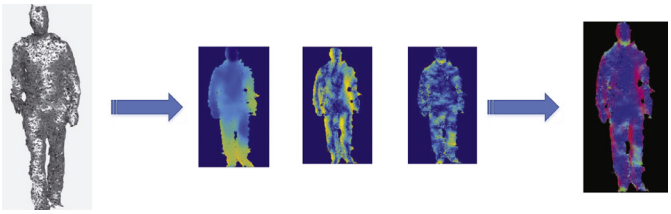

Fig. 3. The depth image representation of RGBD-ID dataset. The left image is a 3D model of pedestrian, and the middle three images refer to depth value, height value and the angle with horizontal direction. The right image is the three channel combination in RGB space.

3.3. Implementation details

Due to the limitation of the size of the training set, it is difficult to learn W_1, W_2, W_3, W_4 from a random initialized value. So we initialize these parameters using principle component analysis [57] as follows:

$$[W_3, W_4] = PCA([X^{(1)}, X^{(2)}]), Z = W_3'X^{(1)} + W_4'X^{(2)} \tag{21}$$

$$W_1 = PCA(X^{(1)} - W_3 Z), W_2 = PCA(X^{(2)} - W_4 Z) \tag{22}$$

where $X^{(1)}$ and $X^{(2)}$ are extracted from a depth image and a RGB image in the training set using pre-trained CNN networks, respectively.

More specifically, we chose the pre-trained model provided in [2]. The CNN starts with 4 concatenated convolution layers followed by a pooling layer, which is shown in Table 1. The next is a series of 6 inception units. At the final fully connected layer, the CNN produces a 256-dimensional feature. Unlike other pre-trained model trained on ImageNet, it takes input of size 144 × 56 which is much more suitable to the person body image, and it was pre-trained with several person re-identification datasets. As shown in Table 2, even the GoogLeNet is actually deeper than the selected model, results show that selected model from [2] outperforms GoogLeNet on CUHK03 [48].

The depth image of person contains the anthropometric information, such as height and body shape, which help us to identify person. Unlike most existing methods which use the skeleton information only and usually ignore the information of body shape, we aim to exploit more useful information to identify person from depth images. Deep learning can extract efficient feature of RGB images in various computer vision tasks, and we expect to take advantage of deep learning to extract discriminative feature from depth images.

In order to better use pre-trained deep network better, we process the single channel depth images to learn rich features using the deep network [43]. For the RGBD-ID dataset, the 3D points cloud was provided for each frame. We computed the angle with horizontal direction of the each point as the value of the second channel, and computed the height of each point as the third channel, and the value of depth as the first channel. Fig. 3 shows an example of a person in the RGBD-ID dataset, and the body structure information is clear in the last 3 channel image.

For the KinectREID dataset, the depth images were provided, so we directly use the depth information as the first channel and calculate the height of each point as the third channel. It is hard to estimate the gravity direction, and the obtained result has larger error due to the low quality of depth images captured by Kinect V1 sensor. We use the width information of each point as the second channel. As shown in Figs. 3 and 4, the processed images contain more intuitive information than original depth images and 3D points cloud.

Table 1
The detailed parameter settings of the CNN network used in our approach.

Name	patch size /stride	output size	#1 × 1	#3 × 3 reduce	#3 × 3	double #3 × 3 reduce	double #3 × 3	pool + proj
input		3 × 144 × 56						
conv1 - conv3	3 × 3/2	32 × 144 × 56						
pool3	2 × 2/2	32 × 72 × 28						
inception (4a)		256 × 72 × 28	32	32	32	32	32	avg + 32
inception (4b)	stride 2	384 × 72 × 28	32	32	32	32	32	max+ pass through
inception (5a)		512 × 36 × 14	64	64	64	64	64	avg + 64
inception (5b)	stride 2	768 × 36 × 14	64	64	64	64	64	max+ pass through
inception (6a)		1024 × 36 × 14	128	128	128	128	128	avg + 128
inception (6b)	stride 2	1536 × 36 × 14	128	128	128	128	128	max+ pass through
fc7		256						

Table 2
Comparison of the 2 models.

Model	Rank 1	Rank 5	Rank 10	final loss
GoogLeNet	40.8%	76.1%	89.1%	0.027
Deep model [2]	63.6%	92.0%	95.9%	0.0078

Fig. 4. The depth image representation of Kinect-REID dataset. The left image is a depth image of pedestrians, and the middle three images which refer to depth value, height value and width value. The right image is the combination of these channels in the RGB space.

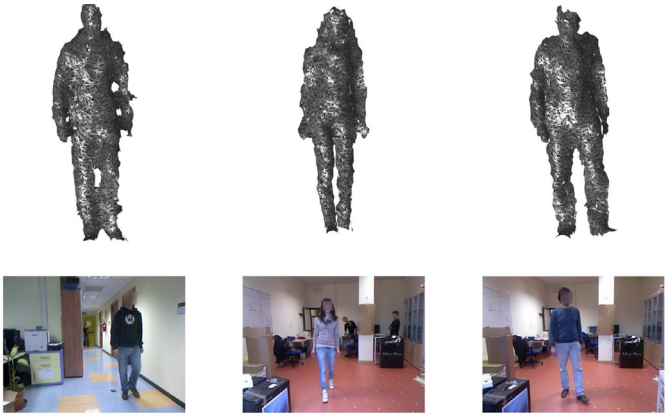

Fig. 5. Samples from RGBD-ID dataset. The images in first row are the 3D points cloud of pedestrians and the images in second row are RGB images of pedestrians.

4. Experiments

4.1. Datasets and settings

We conducted person re-identification experiments on two widely used pedestrian video datasets with the RGB-D person re-identification dataset (RGBD-ID) and the KinectREID Dataset (Fig. 5).

The RGBD-ID [17] dataset contains the RGB and depth images of 80 individuals, and each individual has four acquisitions, one rear and three frontal poses (walking1, walking2, backwards, collaborative). Four or five RGB and 3D model frames are provided for each individual in each acquisition. Some individuals wore different clothes in different acquisitions, and 43 individuals wear the same red T-shirt in walking2 and backwards acquisitions (Fig. 6).

The KinectREID [18] dataset was acquired with the Kinect V1 sensors and the official Microsoft SDK. It contains 483 videos taken at a lecture hall for 71 individuals. All of these 71 individuals walked normally along a predefined path in three scenes with different lighting conditions, and for each scene and each individual, three videos were taken in near-frontal, near-rear and lateral views. For each individual in each view point, the RGB images, the masks and skeletons about 10 key frames are provided. In order to extract features from depth images, we extracted 20 depth images per individual from the original Kinect recode files with the office Microsoft SDK.[1]

For both datasets, we randomly selected 20 individuals and used all frames of them as the training set. For each of the remaining people, one video sequence was chosen as the element in the gallery set, and others are used as the probe set. We used the cumulative matching characteristic (CMC) cure to evaluate the performance of our method. We repeated the experiments 10 times and used the average accuracy as our results.

4.2. Experiments with single depth information

The depth images contain the anthropometric information of individuals. We first conducted experiments with single depth images to evaluate the discriminative capability of the depth information. We compared the performance of the original depth images, where three channels are all depth information. As shown in Fig. 7, the processed depth images obtain a higher recognition accuracy than the original depth images, because the processed depth images can take the advantage of the deep learning pre-trained by RGB images.

We also compared our features extracted from depth images with anthropometric measures [18] extracted from skeleton:

(1) d_1: the distance between the floor and head;
(2) d_2: the ratio between the torso and legs;
(3) d_3: the height (distance between the highest body silhouette point and the floor plane);
(4) d_4: the distance between the floor and neck;
(5) d_5: the distance between the neck and shoulder;
(6) d_6: the distance between the torso center and shoulder;
(7) d_7: the distance between the torso center and hip;
(8) d_8: the arm length (sum of the distances between the shoulder and elbow, and between the elbow and wrist);
9) d_9: the leg length (sum of the distances between the hip and knee, and between the knee and ankle).

$$s = \sum_{i=1}^{9} \omega_i (d_i' - d_i'')^2 \tag{23}$$

[1] Microsoft Kinect SDK, http://www.microsoft.com/en-us/kinectforwindows/.

Fig. 6. Samples from the Kinect-REID dataset. The left images are RGB images, and the middle images are depth images, and the right images are 3D points cloud.

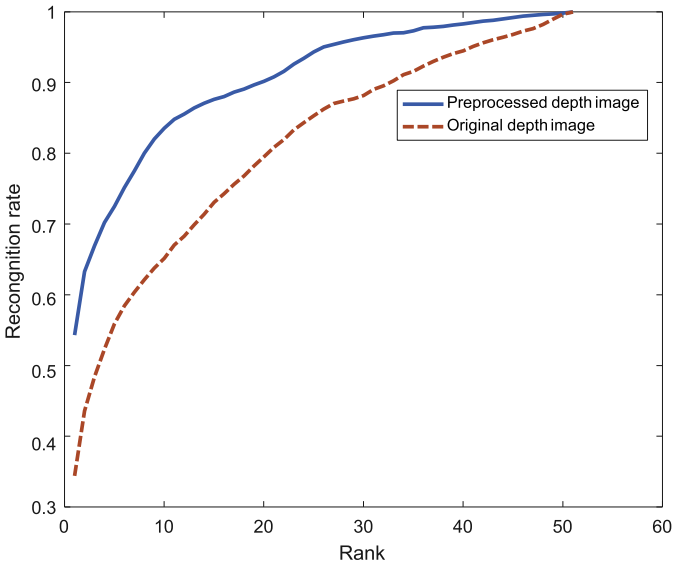

Fig. 7. The CMC curve of processed depth images versus original depth images in the Kinect-REID dataset.

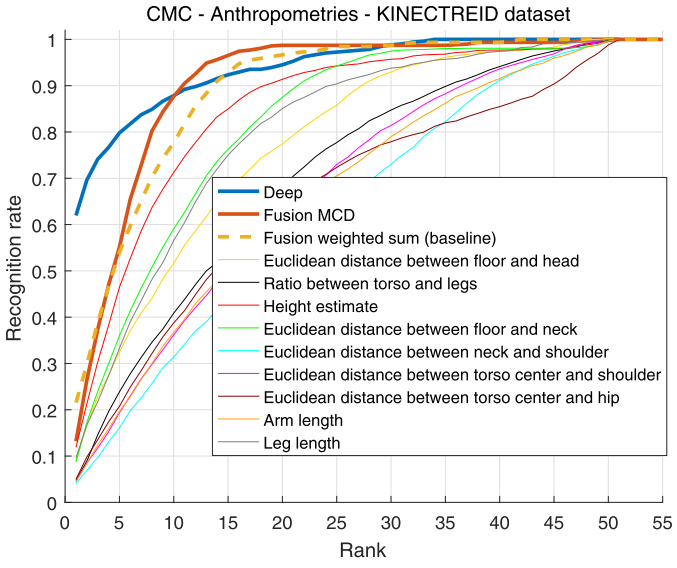

Fig. 8. The CMC curve of different anthropometric measures and our approach with depth images in the Kinect-REID dataset.

where $\omega_2 = 0.2, \omega_3 = 0.5, \omega_4 = 0.05, \omega_8 = 0.05, \omega_9 = 0.05,$ and others are equal to 0 in the Kinect-REID dataset, $\omega_1 = 0.4, \omega_3 = 0.6, \omega_8 = 0.05,$ and others are equal to 0 [18].

As shown in Fig. 8, the $\mathbf{AUC}_{20\%}$ of our approach is 72.64% versus 57% on the kinectREID dataset, and 76.5% versus 60% on the RGBD-ID dataset. The rank 1 matching rate is 54.31% versus 21.5% on the kinectREID dataset, and 46.53% versus 18.1% on the RGBD-ID dataset, respectively.

We also find that the rank-10 and rank-20 recognition rates of our approach are lower than those of the anthropometric measures. The reason is that the gallery set and probe set contain the near-rear and near-frontal views, and the views changes may influence the recognition result of our approach, while the anthropometric measures only use the skeleton information which is independent with the view point. We also conducted experiments under the same view point on the KinectREID dataset. As shown in Table 3, the view point has significant influence on the result of our approach (Fig. 9).

Table 3
The performance in specific view points using depth images in KinectREID dataset.

view point	Rank =1	Rank = 5	Rank =10	Rank = 20
front-front	87.3%	98.0%	100%	100%
rear-rear	51.6%	85.4%	94.16%	99.6%
rear-front	41.2%	60.8%	72.55%	84.3%
front-rear	35.3%	56.86%	68.63%	82.4%

4.3. Experiments with both appearance and depth information

We combined the depth information and RGB appearance information using the proposed multi-model method and conducted experiments on the KinectREID dataset. As shown in Fig. 10, the rank-1 recognition rate of our appearance feature is 82% versus 43%, and the rank-1 rate of our multi-model is 97% versus 51% compared with some anthropometric measures with several different clothes appearance descriptors(SDALF [20], eBiCov [58], MCMimpl [59]) [18].

The same individual on the kinectREID dataset wore the same clothes in all views and cameras, while many individuals wore to-

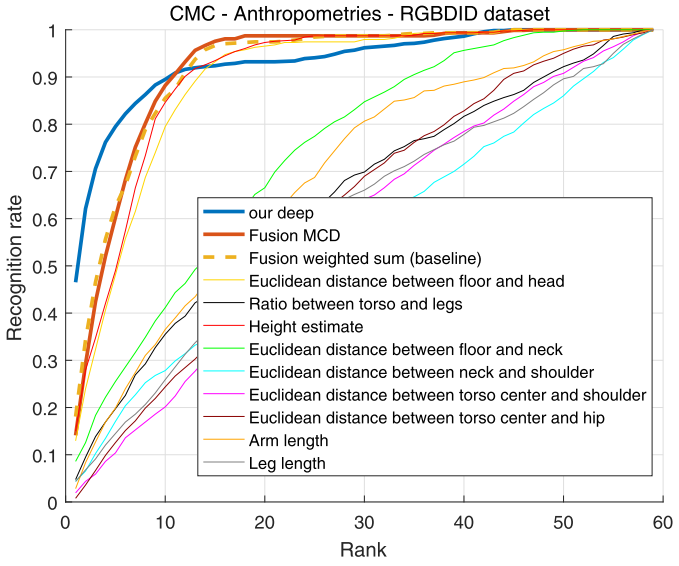

Fig. 9. The CMC curve of different anthropometric measures and our approach with depth images on the RGBD-ID dataset.

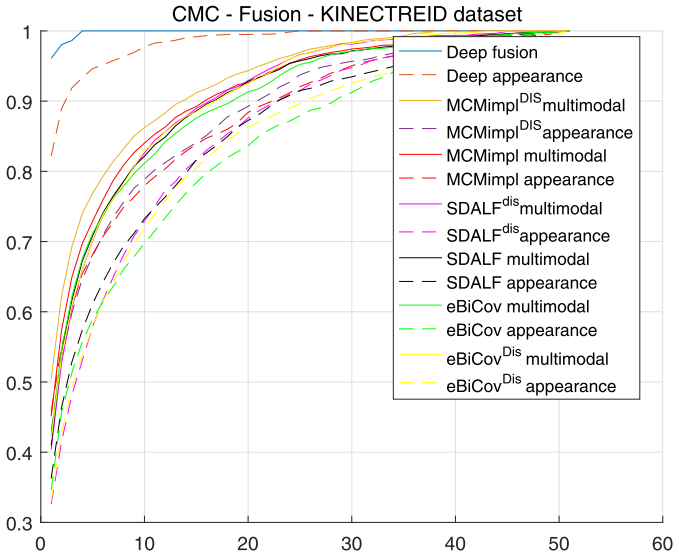

Fig. 10. The CMC curve of different appearance features in the Kinect-REID dataset.

Table 4
The performance of different methods on the complete RGBD-ID dataset.

method	Rank =1	Rank = 5	Rank =10	Rank = 20
depth images	50.1%	79.1%	89%	92.4%
RGB images	31.7%	62.03%	81.7%	92.4%
MMUDL	76.7%	87.5%	96.1%	98.0%

tally different clothes. Federico [18] removed the individuals with different clothes in RGBD-ID dataset when conducting experiments with the appearance information, and some individuals only appeared once in the test set. Federico achieved a very high matching rate in the reduced RGBD-ID due to the decent of individuals. Our appearance features achieve 100% in the reduced RGBD-ID dataset as shown in Fig. 11.

We also conducted experiments on the whole RGBD-ID dataset (Table 4). As shown in Fig. 12, the clothes of some individuals changed a lot, and several individuals wore the same T-shirt, so it is hard to recognize person with the clothes information. As shown

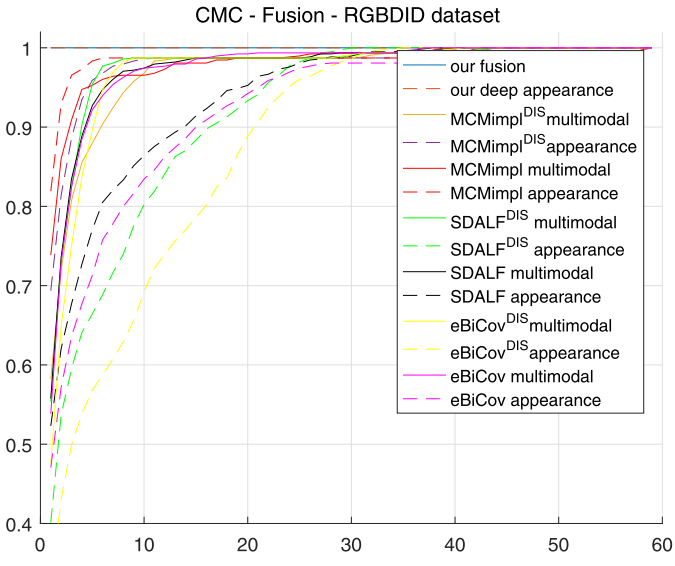

Fig. 11. The CMC curve of different appearance features on the RGBD-REID dataset.

Table 5
Comparison of matching rate (%) with state-of-the-art person re-identification methods on the KinectREID dataset.

Method	Rank =1	Rank = 5	Rank =10
SGTrP3+score level [60]	66.08	*	91.35
MCMimpl fusion [18]	50.4	87.9	86.25
depth images - our	62.0	79.8	87.8
RGB images -our	82.1	94.3	97.7
fusion directly - our	91.0	99.0	99.6
MMUDL	**97.0**	**100**	**100**

Table 6
Comparison of matching rate (%) with state-of-the-art person re-identification methods on the complete RGBD-ID dataset.

Method	Rank =1	Rank = 5	Rank =10
SGTrP3+score level [60]	76.6	*	**99.4**
DVCov+SKL [61]	71.7	**88.4**	*
depth images	50.1	79.1	89
RGB images	31.7	62.03	81.7
fusion directly	52.8	82.1	92.3
MMUDL	**76.7**	87.5	96.1

in Table 6, the performance of appearance feature is lower than depth images due to the clothes changing. However, the performance of the combined RGB images and depth images is reasonably well. Experimental results show that the depth information is more reliable than appearance information in some specific situations.

As shown in Tables 5 and 6, the performance of our method is better than the direct fusion (stitch two features together). Our approach achieves the best performance in both two datasets compared with state-of-the-art methods: 97.0% versus 66.08% on the Kinect-REID dataset, and 76.7% versus 76.6% on the RGBD-REID dataset.

4.4. Parameters analysis

Fig. 13 shows the objective function value with different number of iterations in the Kinect-REID dataset. The loss of both the training and testing sets reach the minimization at the iteration 500.

Fig. 14 shows the rank-1 matching rate of direct fusion and multi-model uniform fusion with different noises on the Kinect-

Fig. 12. Examples of individuals with clothes changing on the RGBD-ID dataset.

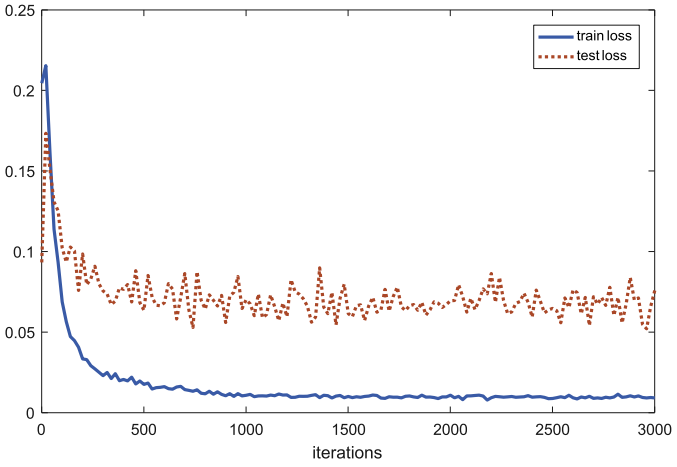

Fig. 13. The objective function value of our approach versus different number of iterations on the KinectREID dataset.

REID dataset. We included normal noises to the output vectors of CNN networks, and our proposed method can overcome the influence of strong noise compared with the directly fusion [62].

Fig. 15 shows the rank-1 matching rates versus different learning rate and batch sizes of our approach. We see that our approach achieves the best performance when learning rate equals to 0.001 and batch size equals to 9.

5. Conclusions and future work

In this paper, we have proposed a multi-modal uniform deep learning method to extract the anthropometric and appearance features from RGB-D images for person re-identification. Our approach extracted features from RGB-D images by using two CNN networks and a uniform fusion layer, which is robust to the noise. Experiments results have shown the efficiency of our proposed approach. How to combine our approach with skeleton information which is robust to different view points is an interesting future work.

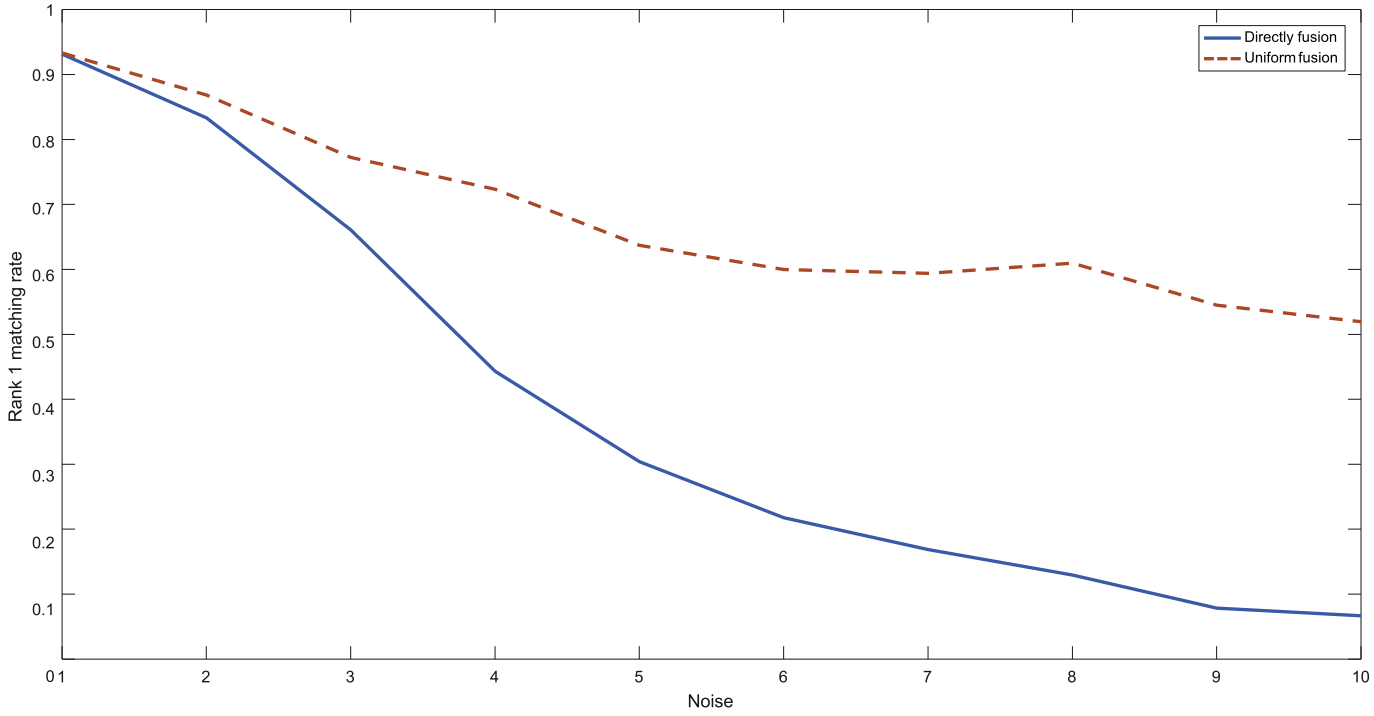

Fig. 14. The rank-1 matching rate of directly fusion (red line) and uniform fusion (red line) versus the mod of noise. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

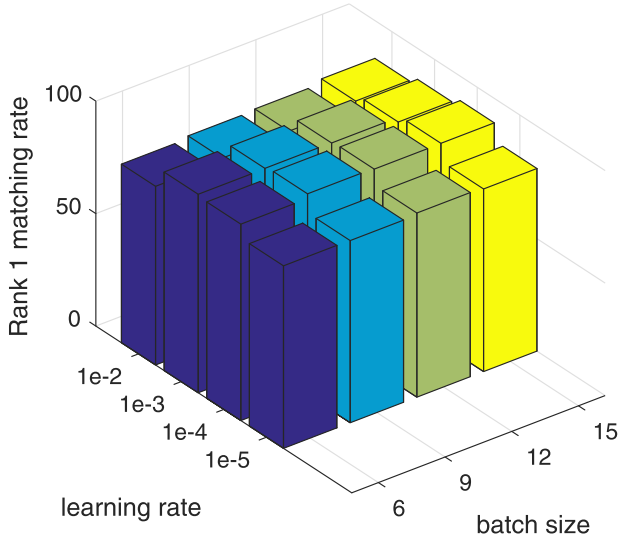

Fig. 15. The rank-1 matching rate versus different learning rate and batch size of our approach.

Acknowledgment

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1001004, in part by the National Natural Science Foundation of China under Grant 61672306, Grant 61572271, Grant 61527808, Grant 61373074, and Grant 61373090, in part by the National 1000 Young Talents Plan Program, in part by the National Basic Research Program of China under Grant 2014CB349304, in part by the Shenzhen Fundamental Research Fund (Subject Arrangement) under Grant JCYJ20170412170602564, in part by the Ministry of Education of China under Grant 20120002110033, and in part by the Tsinghua University Initiative Scientific Research Program.

References

[1] N. Gheissari, T.B. Sebastian, R. Hartley, Person reidentification using spatiotemporal appearance, in: CVPR, 2006, pp. 1528–1535.
[2] T. Xiao, H. Li, W. Ouyang, X. Wang, Learning deep feature representations with domain guided dropout for person re-identification, arXiv preprint arXiv:1604.07528 2016.
[3] F. Xiong, M. Gou, O. Camps, M. Sznaier, Person re-identification using kernel-based metric learning methods, in: ECCV, Springer, 2014, pp. 1–16.
[4] K. Bernardin, R. Stiefelhagen, Evaluating multiple object tracking performance: the clear MOT metrics, IJIVP (2008) 1.
[5] S. Bak, S. Zaidenberg, B. Boulay, F. Bremond, Improving person re-identification by viewpoint cues, in: AVSS, 2014, pp. 175–180.
[6] S. Bak, E. Corvee, F. Brémond, M. Thonnat, Person re-identification using spatial covariance regions of human body parts, in: AVSS, 2010, pp. 435–440.
[7] R. Zhao, W. Ouyang, X. Wang, Learning mid-level filters for person re-identification, in: CVPR, 2014, pp. 144–151.
[8] S. Liao, Y. Hu, X. Zhu, S.Z. Li, Person re-identification by local maximal occurrence representation and metric learning, in: CVPR, 2015, pp. 2197–2206.
[9] N. McLaughlin, D.R.J. Martine, P. Miller, Recurrent convolutional network for video-based person re-identification, in: CVPR, 2016, pp. 1325–1334.
[10] B. Ma, Y. Su, F. Jurie, Bicov: a novel image representation for person re-identification and face verification, in: BMVC, 2012, pp. 57.1–57.11.
[11] R.R. Varior, G. Wang, J. Lu, Learning invariant color features for person re-identification, arXiv preprint arXiv:1410.1035 2014.
[12] R. Zhao, W. Ouyang, X. Wang, Person re-identification by salience matching, in: ICCV, 2013, pp. 2528–2535.
[13] S. Karanam, Y. Li, R.J. Radke, Person re-identification with discriminatively trained viewpoint invariant dictionaries, in: ICCV, 2015, pp. 4516–4524.
[14] J. Lu, G. Wang, W. Deng, P. Moulin, J. Zhou, Multi-manifold deep metric learning for image set classification, in: CVPR, 2015, pp. 1137–1145.
[15] A. Mogelmose, C. Bahnsen, T. Moeslund, A. Clapés, S. Escalera, Tri-modal person re-identification with RGB, depth and thermal features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2013, pp. 301–307.
[16] R. Kawai, Y. Makihara, C. Hua, H. Iwama, Y. Yagi, Person re-identification using view-dependent score-level fusion of gait and color features, in: Pattern Recognition (ICPR), 2012 21st International Conference on, IEEE, 2012, pp. 2694–2697.
[17] I.B. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, V. Murino, Re-identification with RGB-D sensors, in: European Conference on Computer Vision, Springer, 2012, pp. 433–442.
[18] F. Pala, R. Satta, G. Fumera, F. Roli, Multimodal person reidentification using RGB-Dcameras, CSVT 26 (4) (2016) 788–799.
[19] A. Bialkowski, S. Denman, S. Sridharan, C. Fookes, P. Lucey, A database for person re-identification in multi-camera surveillance networks, in: DICTA, IEEE, 2012, pp. 1–8.
[20] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person re-identification by symmetry-driven accumulation of local features, in: CVPR, 2010, pp. 2360–2367.
[21] L. Bazzani, M. Cristani, A. Perina, V. Murino, Multiple-shot person re-identification by chromatic and epitomic analyses, Pattern Recognit. Lett. 33 (7) (2012) 898–903.
[22] T. Wang, S. Gong, X. Zhu, S. Wang, Person re-identification by video ranking, in: ECCV, 2014, pp. 688–703.
[23] A. Bedagkar-Gala, S.K. Shah, A survey of approaches and trends in person re-identification, Image Vis Comput 32 (4) (2014) 270–286.
[24] M. Gou, X. Zhang, A. Rates-Borras, S. Asghari-Esfeden, M. Sznaier, O. Camps, Person re-identification in appearance impaired scenarios, arXiv2016.
[25] W. Zheng, S. Gong, T. Xiang, Person re-identification by probabilistic relative distance comparison, in: CVPR, 2011, pp. 649–656.
[26] M. Hirzer, P.M. Roth, M. Köstinger, H. Bischof, Relaxed pairwise learned metric for person re-identification, in: ECCV, Springer, 2012, pp. 780–793.
[27] S. Ding, L. Lin, G. Wang, H. Chao, Deep feature learning with relative distance comparison for person re-identification, Pattern Recognit. 48 (10) (2015) 2993–3003.
[28] Z. Wu, Y. Li, R.J. Radke, Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features, PAMI 37 (5) (2015) 1095–1108.
[29] R. Mazzon, S.F. Tahir, A. Cavallaro, Person re-identification in crowd, Pattern Recognit. Lett. 33 (14) (2012) 1828–1837.
[30] X. Ma, X. Zhu, S. Gong, X. Xie, J. Hu, K.-M. Lam, Y. Zhong, Person re-identification by unsupervised video matching, Pattern Recognit. 65 (2017) 197–210.
[31] L. Bazzani, M. Cristani, V. Murino, Sdalf: modeling human appearance with symmetry-driven accumulation of local features, in: Person Re-Identification, Springer, 2014, pp. 43–69.
[32] M. Yang, P. Zhu, G.L. Van, L. Zhang, Face recognition based on regularized nearest points between image sets, in: FG, 2013, pp. 1–7.
[33] A. Mogelmose, T.B. Moeslund, K. Nasrollahi, Multimodal person re-identification using RGB-Dsensors and a transient identification database, in: IWBF, IEEE, 2013, pp. 1–4.
[34] D. Figueira, L. Bazzani, H.Q. Minh, M. Cristani, A. Bernardino, V. Murino, Semi–supervised multi-feature learning for person re-identification, in: AVSS, IEEE, 2013, pp. 111–116.
[35] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: NIPS, 2012, pp. 1097–1105.
[36] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 2014.
[37] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: CVPR, 2015, pp. 1–9.
[38] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, arXiv preprint arXiv:1512.03385 2015.
[39] Z. Zuo, G. Wang, B. Shuai, L. Zhao, Q. Yang, Exemplar based deep discriminative and shareable feature learning for scene image classification, Pattern Recognit. 48 (10) (2015) 3004–3015.
[40] Y. Taigman, M. Yang, M.A. Ranzato, L. Wolf, Deepface: closing the gap to human-level performance in face verification, in: CVPR, 2014, pp. 1701–1708.
[41] Y. Sun, D. Liang, X. Wang, X. Tang, Deepid3: face recognition with very deep neural networks, arXiv preprint arXiv:1502.00873 (2015).
[42] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: a unified embedding for face recognition and clustering, in: CVPR, 2015, pp. 815–823.
[43] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: CVPR, 2014, pp. 580–587.
[44] R. Girshick, Fast R-CNN, in: ICCV, 2015, pp. 1440–1448.
[45] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, in: NIPS, 2015, pp. 91–99.
[46] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, arXiv preprint arXiv:1506.02640 2015.
[47] S.M. Erfani, S. Rajasegarar, S. Karunasekera, C. Leckie, High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning, Pattern Recognit 58 (2016) 121–134.
[48] W. Li, R. Zhao, T. Xiao, X. Wang, Deepreid: deep filter pairing neural network for person re-identification, in: CVPR, 2014, pp. 152–159.
[49] D. Yi, Z. Lei, S. Liao, S.Z. Li, et al., Deep metric learning for person re-identification., in: ICPR, 2014, 2014, pp. 34–39.
[50] E. Ahmed, M. Jones, T.K. Marks, An improved deep learning architecture for person re-identification, in: CVPR, 2015, pp. 3908–3916.
[51] D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng, Person re-identification by multi-channel parts-based CNN with improved triplet loss function, in: CVPR, 2016, pp. 1335–1344.
[52] F. Wang, W. Zuo, L. Lin, D. Zhang, L. Zhang, Joint learning of single-image and cross-image representations for person re-identification, CVPR (2016) 1288–1296.

[53] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, X. Yang, Person re-identification via recurrent feature aggregation, in: ECCV, Springer, 2016, pp. 701–716.
[54] R.R. Varior, B. Shuai, J. Lu, D. Xu, G. Wang, A siamese long short-term memory architecture for human re-identification, in: ECCV, 2016, pp. 135–153.
[55] R.R. Varior, M. Haloi, G. Wang, Gated siamese convolutional neural network architecture for human re-identification, in: ECCV, 2016, pp. 791–808.
[56] J. Hu, J. Lu, Y. Tan, Discriminative deep metric learning for face verification in the wild, in: CVPR, 2014, pp. 1875–1882.
[57] I. Jolliffe, Principal Component Analysis, Wiley Online Library, 2002.
[58] B. Ma, Y. Su, F. Jurie, Covariance descriptor based on bio-inspired features for person re-identification and face verification, Image Vis. Comput. 32 (6) (2014) 379–390.
[59] R. Satta, G. Fumera, F. Roli, M. Cristani, V. Murino, A multiple component matching framework for person re-identification, in: International Conference on Image Analysis and Processing, Springer, 2011, pp. 140–149.
[60] Z. Imani, H. Soltanizadeh, Person reidentification using local pattern descriptors and anthropometric measures from videos of kinect sensor, IEEE Sens. J. 16 (16) (2016) 6227–6238.
[61] A. Wu, W.-S. Zheng, J. Lai, Robust depth-based person re-identification, TIP (2017).
[62] A. Wang, J. Lu, J. Cai, T.-J. Cham, G. Wang, Large-margin multi-modal deep learning for rgb-d object recognition, IEEE Trans.Multimedia 17 (11) (2015) 1887–1898.

Liangliang Ren received the B.S. degree in the department of automation in Tsinghua University, China, in 2016. He is currently pursuing the Ph.D. degree at the Department of Automation, Tsinghua University. His research interests include person re-identification, object tracking, camera networks and deep learning.

Jiwen Lu received the B.Eng. degree in mechanical engineering and the M.Eng. degree in electrical engineering from the Xi'an University of Technology, Xi'an, China, and the Ph.D. degree in electrical engineering from the Nanyang Technological University, Singapore, in 2003, 2006, and 2012, respectively. He is currently an Associate Professor with the Department of Automation, Tsinghua University, Beijing, China. From March 2011 to November 2015, he was a Research Scientist with the Advanced Digital Sciences Center, Singapore. His current research interests include computer vision, pattern recognition, and machine learning. He has authored/co-authored over 130 scientific papers in these areas, where more than 50 papers are published in the IEEE Transactions journals and top-tier computer vision conferences. He serves/has served as an Associate Editor of Pattern Recognition Letters, Neurocomputing, and the IEEE Access, a Guest Editor of Pattern Recognition, Computer Vision and Image Understanding, Image and Vision Computing and Neurocomputing, and an elected member of the Information Forensics and Security Technical Committee of the IEEE Signal Processing Society. He is/was a Workshop Chair/Special Session Chair/Area Chair for more than 10 international conferences. He has given tutorials at several international conferences including ACCV16, CVPR'15, FG'15, ACCV'14, ICME'14, and IJCB'14. He was a recipient of the First-Prize National Scholarship and the National Outstanding Student Award from the Ministry of Education of China in 2002 and 2003, the Best Student Paper Award from Pattern Recognition and Machine Intelligence Association of Singapore in 2012, the Top 10% Best Paper Award from IEEE International Workshop on Multimedia Signal Processing in 2014, and the National 1000 Young Talents Plan Program in 2015, respectively. He is a senior member of the IEEE.

Jianjiang Feng is an associate professor in the Department of Automation at Tsinghua University, Beijing. He received the B.S. and Ph.D. degrees from the School of Telecommunication Engineering, Beijing University of Posts and Telecommunications, China, in 2000 and 2007. From 2008 to 2009, he was a Post Doctoral researcher in the PRIP lab at Michigan State University. He is an Associate Editor of Image and Vision Computing. His research interests include fingerprint recognition and computer vision.

Jie Zhou received the B.S. and M.S. degrees both from the Department of Mathematics, Nankai University, Tianjin, China, in 1990 and 1992, respectively, and the PhD degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology (HUST), Wuhan, China, in 1995.From then to 1997, he served as a postdoctoral fellow in the Department of Automation, Tsinghua University, Beijing, China. Since 2003, he has been a full professor in the Department of Automation, Tsinghua University. His research interests include computer vision, pattern recognition, and image processing. In recent years, he has authored more than 100 papers in peer-reviewed journals and conferences. Among them, more than 30 papers have been published in top journals and conferences such as the IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing, and CVPR. He is an associate editor for the International Journal of Robotics and Automation and two other journals. He received the National Outstanding Youth Foundation of China Award. He is a senior member of the IEEE.