Contents lists available at ScienceDirect

CAAI Transactions on Intelligence Technology

journal homepage: http://www.journals.elsevier.com/caai-transactions-on-intelligence-technology/

Original article

Online RGB-D person re-identification based on metric model update

Hong Liu [a], Liang Hu [a, *], Liqian Ma [b]

[a] The Engineering Lab on Intelligent Perception for Internet of Things (ELIP), Shenzhen Graduate School, Peking University, Shenzhen, 518055, China
[b] The VISICS, ESAT, KU Leuven, Kasteelpark Arenberg 10, Heverlee, 3001, Belgium

ARTICLE INFO

Article history:
Received 18 March 2017
Accepted 1 April 2017
Available online 20 April 2017

Keywords:
Person re-identification
Online metric model update
Face information
Skeleton information

ABSTRACT

Person re-identification (re-id) on robot platform is an important application for human-robot-interaction (HRI), which aims at making the robot recognize the around persons in varying scenes. Although many effective methods have been proposed for surveillance re-id in recent years, re-id on robot platform is still a novel unsolved problem. Most existing methods adapt the supervised metric learning offline to improve the accuracy. However, these methods can not adapt to unknown scenes. To solve this problem, an online re-id framework is proposed. Considering that robotics can afford to use high-resolution RGB-D sensors and clear human face may be captured, face information is used to update the metric model. Firstly, the metric model is pre-trained offline using labeled data. Then during the online stage, we use face information to mine incorrect body matching pairs which are collected to update the metric model online. In addition, to make full use of both appearance and skeleton information provided by RGB-D sensors, a novel feature funnel model (FFM) is proposed. Comparison studies show our approach is more effective and adaptable to varying environments.

1. Introduction

The essence of person re-id is associating a person under different cameras over time. It is a key technology for service robotics, especially for modeling human long-term activities across different scenes to provide friendly human-computer interaction.

In recent years, many re-id methods for video surveillance applications have been proposed. Generally, they can be classified into two categories [1]: feature representation and matching strategy. For feature representation, researchers mainly focus on generating robust and efficient body appearance representation using information such as color [2] and texture [3], since in surveillance environment the captured RGB images are usually in low-resolution. For matching strategy, metric learning [4] and learning to rank methods are explored a lot. And methods based on metric learning achieved good performances.

As matures of high-resolution RGB-D sensors, robotics can afford to use them to improve re-id performances. Compared with person RGB images in low resolution, RGB-D data can provide not only appearance information, but also skeleton, depth and even clear face information. Taking advantages of RGB-D data, some researchers combine the skeleton or face information with color or texture features to obtain a multi-modal system which is more robust to pose changes and complex backgrounds [5–7].

Although these multi-modal methods achieved good performances on public dataset, these methods might obtain unsatisfactory performance on handling relatively varying environments for two reasons. On one hand, most conventional approaches usually learn the similarity metric model offline, so that they cannot adapt to new scenes which are significantly different from the training data, caused by varying illuminations, camera views, backgrounds. On the other hand, blindly combining a bunch of features to calculate person similarity is prone to error accumulation, and also brings unnecessary computational cost.

To overcome the drawbacks of above methods, an effective online re-id framework is proposed in this paper. Based on observations, clear face images contain more reliable and distinguishing information but may be difficult to obtain in many situations such as people with his back to the camera. While, body images are more ambiguous but are usually easy to capture. Therefore, the face and body images are complementary.

Firstly, each person is described by appearance-based and geometric features using skeleton information following [8].

* Corresponding author.
E-mail address: lianghu@sz.pku.edu.cn (L. Hu).
Peer review under responsibility of Chongqing University of Technology.

http://dx.doi.org/10.1016/j.trit.2017.04.001

Secondly, the metric model is learned offline using labeled training data. Then, the face information is utilized to update the metric model online. Finally, the feature similarities are fused by feature funnel model which is based on the degree of feature reliability.

Our contributions are mainly threefold. First, we propose a novel online re-id framework which is robust to the changing environments. Second, a fusion strategy named feature funnel model is proposed to fuse multiple features effectively. Third, a novel publicly available RGB-D re-id dataset named RobotPKU RGBD-ID dataset is collected, which contains 180 video sequences of 90 persons collected using Kinect.

The rest of this paper is organized as follows. In section 2, the related works are reviewed. In section 3, our method is presented in its main components: features extraction, online metric model update and feature funnel model. In section 4, experiments run for assessing the performance and comparing our method to the state-of-the-art methods are discussed. Finally conclusions are drawn in Section 5.

2. Related works

2.1. Depth-based re-identification

With the appearing of cheap depth-sensing devices in the market, computer vision for robotics was revolutionized and some RGB-D based re-id approaches have also been proposed. These methods can be divided into two categories, the first type of method is appearance-based methods which integrate appearance and depth information together [9–11]. Munaro et al. propose a re-id approach based on skeletal information [11]. Feature descriptors are extracted around person skeletal joints and final person signature is obtained by concatenating these descriptors. The second type of method is based on geometric features: in Ref. [12], re-id is performed by matching body shapes in terms of whole point clouds warped to a standard pose with the described method. Matteo et al. [8] adopt the anthropometric measure method for re-id. They use the 3D location of body joints provided by skeletal tracker (see Fig. 2) to compute the geometric features, such as limb lengths and ratios.

2.2. Multimodal person re-identification

Appearance-based methods are easily impacted by varying environments such as illuminations, and geometric method has low inter-class variability. Thus, re-identifying persons only relying on a single source of biometric information can actually be difficult. For this reason, multimodal biometric systems are adopted to make re-id more reliable.

Many multimodal systems have been proposed which can be classified into two categories [13]: One approach is to fuse information at feature level [14,15] by concatenating the feature vectors as final feature. However, this method often overlooks the reliability differences between different features. The other is to fuse information at score level [6,7], i.e. combining the scores of different sub-systems, and our feature funnel model belongs to this type.

2.3. Online learning

Most re-id methods are corresponding learning the similarity metric model offline [4], but these methods can not adapt to unknown scenes. Recently, the semi-supervised learning methods become more and more popular in many computer vision applications [16], since they can utilize both labeled and unlabeled data in classifier training [17] to improve the performance using unlabeled data. P-N learning [18] is an effective semi-supervised learning method which is guided by positive (P) and negative (N) constraints restricting the labeling procedure of the unlabeled set. In this paper, we adopt P-N learning in our online re-id framework to utilize the clear face images to improve the metric model online. Through face information, the measured-error examples are screened out to retaining the metric model and make the metric model adaptive to new environment.

3. Framework description

In this section, an overview of the system framework is provided. As depicted in Fig. 3, our framework contains three parts: feature extraction, online metric model update and feature funnel model.

Primarily, each person is described by appearance and geometric features utilizing skeleton information. Then metric models are updated for each feature modality. The concrete steps of online metric learning are follows: (I) Train the initial metric model offline using labeled data; (II) Measure the similarities of unlabeled data pairs obtained online with metric model; (III) Label the unlabeled data pairs whose measure results are inconsistent with reliable face verification results; (IV) Extend the training set with the new labeled data; (V) Retrain the metric model. Finally, the feature similarity is obtained by our feature funnel model which will be explained in 3.3.

3.1. Feature extraction

The first step of our framework is to extract features which are robust to varying illumination and pose. Inspired by Ref. [11], we extract features around person skeletal joints to overcome pose variation. In order to overcome illumination varying, we adopt noise-insensitive appearance-based features and illumination-invariant geometric features.

3.1.1. Appearance-based features

Person appearance-based feature is an intuitive and effectual expression for re-id. In order to reduce the impact of pose variation, the features extracted from skeleton joints are used to describe person appearance [11]. As shown in Fig. 2, 20 keypoints are obtained by the Microsoft Kinect SDK for each person. Through these keypoints, the real joints of person body can be located and the local information can be described accurately. To describe local appearance information, color and texture features are extracted.

The feature evaluated from skeleton joints is represented as ft $\mathbf{x}_A^{p,i}$, where (p,i) denotes the i-th skeleton joint on the p-th person image. A denotes the appearance-based method. The feature of the p-th person image is defined as:

$$\mathbf{X}_A^p = \left[\mathbf{x}_A^{p,1}, \mathbf{x}_A^{p,2}, \cdots, \mathbf{x}_A^{p,20}\right] \tag{1}$$

In order to extract features which are robust to illumination variation, the color feature employs 8 × 8 × 8 bins HSV histograms. To describe texture features, the Scale Invariant Local Ternary Pattern (SILTP) [19] histograms are extracted from each joint. SILTP is an improved operator over the Local Binary Pattern (LBP) [20]. LBP is a gray-scale invariant texture feature, but is susceptible to the noises. In order to overcome this drawback, SILTP introduces a scale invariant local comparison tolerance, achieving invariance to intensity scale changes and insensitivity to noises.

3.1.2. Geometric feature

In addition to appearance-based description, we also apply the anthropometric measure method for geometric description. Although the distinguishable power of geometric feature is not as good as appearance-based feature, it has a good property that insensitive to varying illumination and dim lighting conditions, as shown in Fig. 1(b). When appearance information is ambiguous, geometric feature can provide supplemental information. In order to describe body geometric, following anthropometric measures from Ref. [8] are chosen:

a) head height,
b) neck height,
c) distance between neck and left shoulder,
d) distance between neck and right shoulder,
e) distance between torso and right shoulder,
f) the length of right arm,
g) the length of left arm,
h) the length of right upper leg,
i) the length of left upper leg,
j) the length of torso,
k) distance between right hip and left hip,
l) ratio between torso and right upper leg length (j/h),
m) ratio between torso and left upper leg (j/i).

These distances are shown in Fig. 2. Our geometric features are composed of these distances and ratios, which is defined as:

$$\mathbf{X}_G^p = \left[\mathbf{x}_G^{p,a}, \mathbf{x}_G^{p,b}, \cdots, \mathbf{x}_G^{p,m}\right] \tag{2}$$

3.2. Online metric model update

In order to evaluate the similarity of features (X^p, Y^q) between probe person p and gallery person q, Mahalanobis distance is adopted to measure the similarity distance $S(\bullet)$ of features:

$$S^2(\mathbf{X^p}, \mathbf{Y^q}) = (\mathbf{X^p} - \mathbf{Y^q})^T \mathbf{M} (\mathbf{X^p} - \mathbf{Y^q}) \tag{3}$$

where **M** is the Mahalanobis distance matrix.

First, in the initialization phase, the **M** is learned offline with some labeled data.

However, due to the varying changes of the environment, the metric model trained offline may fail in real scenes. Considering that in the human-robot-interaction scenario, distance between robot and human is always close, so sometimes we can obtain clear face images which are very helpful to re-id. In our online re-id framework, we use reliable face information to label person image pairs online which are used to update our metric model.

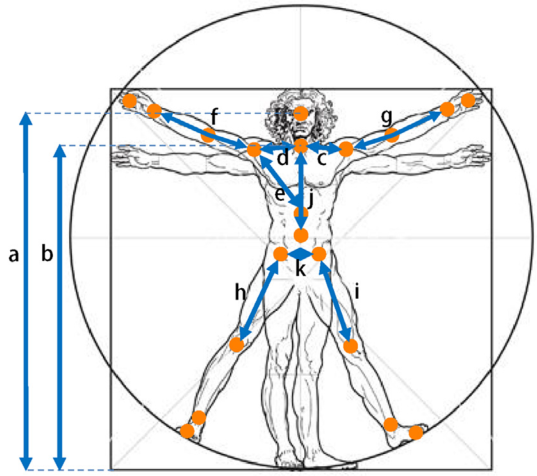

Fig. 2. Twenty body joints tracked by the Microsoft Kinect SDK.

3.2.1. Metric Learning

An effective and efficient metric learning method, XQDA [21], is applied to learn the Mahalanobis distance matrix **M**. Unlike other methods which reduce feature dimension and learn distance matrix separately, XQDA further learns a discriminant subspace together with a matrix. From a statistical inference point of view, XQDA defines the Mahalanobis distance matrix **M** by:

$$\mathbf{M} = \mathbf{W}\left({\mathbf{\Sigma_I}'}^{-1} - {\mathbf{\Sigma_E}'}^{-1}\right)\mathbf{W}^T \tag{4}$$

where ${\mathbf{\Sigma_I}'}^{-1} = \mathbf{W}^T\mathbf{\Sigma_I}\mathbf{W}$ and ${\mathbf{\Sigma_E}'}^{-1} = \mathbf{W}^T\mathbf{\Sigma_E}\mathbf{W}$. Here, $\mathbf{\Sigma_I}$ and $\mathbf{\Sigma_E}$ denote the covariance matrices for similar pairs I and dissimilar pairs E, respectively. $\mathbf{W} = (w_1, w_2, \cdots, w_r)$ is a discriminant subspace which is learned through calculating the top r largest eigenvalues of $\mathbf{\Sigma_I}^{-1}\mathbf{\Sigma_E}$. Similar pairs are composed of $(\mathbf{X}^p, \mathbf{Y}^p)$, and dissimilar pairs are composed of $(\mathbf{X}^p, \mathbf{Y}^q)(p \neq q)$. In addition, the distance matrix **M** can be learned fast with XQDA which is based on statistics without optimization procedure.

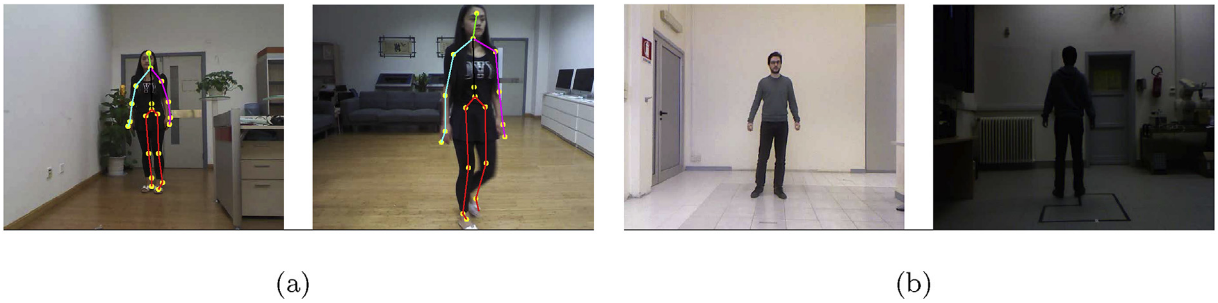

Fig. 1. Some example image pairs from datasets. (a) Example image pairs from the new proposed RobotPKU RGBD-ID dataset. (b) Example image pairs from the proposed dataset with strong illumination changes.

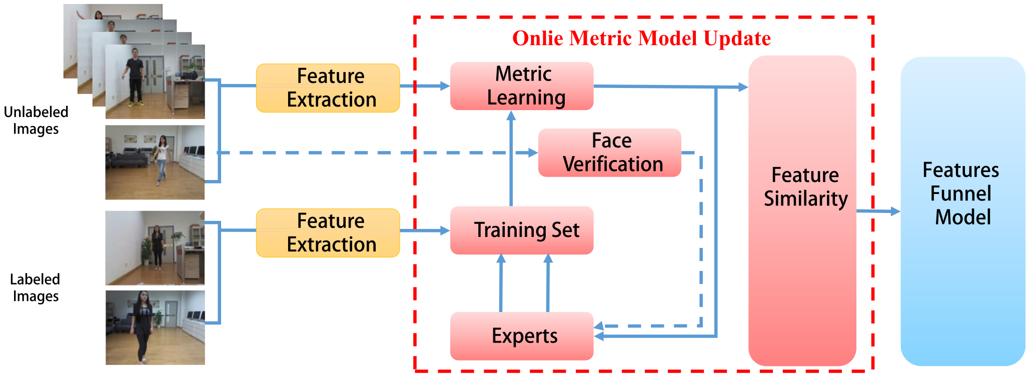

Fig. 3. The proposed framework contains three major modules in three different colors. The features of each person are extracted from labeled and unlabeled images. Then, metric models are updated for each feature. Finally the similarity is obtained by feature funnel model. Best viewed in color.

3.2.2. Online metric model update strategy

Face contains more powerful information for person identification than body appearance. Besides, face verification research has achieved highly reliable performance on large unconstrained LFW dataset [22]. Additionally, towards re-id on robot platform, captured face images are usually in high resolution compared to surveillance environment. Thus, we use face information to update the metric model online.

After initializing the metric model with labeled data, the metric model is then updated using unlabeled data iteratively in online stage.

Firstly, the similarities $S(X^p, Y^q)(q = 1...M)$ between probe image feature X^p and gallery image feature Y^q are calculated.

Secondly, the matching pairs $S(X^p, Y^q)(q = 1...M)$ are ranked according to their similarities. Let $R(X^p, Y^q)$ denotes the ranking result.

Thirdly, ripe face detection and recognition algorithm are utilized to obtain the face similarity score θ $(0 \le \theta \le 1)$. Since the research of face detection and recognition is very mature and is not the focus of this paper, we simply adopt the Face++ SDK[1] to detect face and calculate the similarity score θ between two faces. In Fig. 5, the face similarity score θ distributions of positive samples and negative samples pairs are shown. The confidence of face pair is defined as:

$$F(p,q) = \begin{cases} \theta & \text{if both faces are detected} \\ -1 & \text{if one face is not detected} \end{cases} \tag{5}$$

Then, we verify that whether the similarities $S(X^p, Y^q)(q = 1...M)$ calculated by the metric are in line with the face verification result $F(p,q)$. Two types of error samples are be selected:

- **Error positive samples**: Positive samples refer to the samples from the same person. When their reliable face images are captured, the similarity score θ between the positive samples is usually high, as shown in Fig. 5. Therefore, threshold θ_1 is selected to choose the positive samples $F(p,q) \ge \theta_1$, and the error positive samples are selected if the ranking result $R(X^p, Y^q) > E$, where the E is a threshold to determine ranking result;
- **Error negative samples**: Negative samples refer to the samples from the different person. When their reliable face images are captured, the similarity score θ between the negative samples is usually low, as shown in Fig. 5. Therefore, threshold θ_2 is selected to choose the negative samples $F(p,q) < \theta_2$, and the error negative samples are selected if the $R(X^p, Y^q) < E$;

Finally, the two types of error samples are placed into the training set to retrain the matrix **M** by Eq. (4).

The procedure is shown in Algorithm 1.

Algorithm 1 Online Metric Model Update Strategy

```
Inputs: probe person $p$, gallery persons $\{q\}_{q=1:m}$, matrix $M$
Outputs: updated matrix $M$
initial threshold: $\theta_1$, $\theta_2$ and $E$
do
    Calculate the similarities $S(X^p, Y^q)(q = 1...M)$ between prob feature $X^p_{prob}$ and gallery features $\{X^q_{gal}\}_{q=1:M}$
    Rank similarity $R(X^p, Y^q)$
    Find error positive samples which meet the conditions: $F(p,q) \ge \theta_1$ and $R(X^p, Y^q) < E$
    Find error negative samples which meet the conditions: $R(X^p, Y^q) > E$ and $0 \le F(p,q) < \theta_2$
    Place the two types of error samples into the training set stack.
    Use the new training set to retrain matrix M by Eq. (4).
while $(q > m)$
```

3.3. Feature funnel model

In the feature fusion stage, fusing multiple types of features blindly sometimes will not increase discrimination power, since different features have different reliability. In order to obtain an effective fusion strategy, feature funnel model is proposed.

Firstly, a feature space F is constructed which contains different types of features $\{f_k\}$ for describing person. Based on the feature space F, the similarity between probe p and gallery q is defined as follows:

$$Sim(p,q|F) = \prod_{f_k \in F} S\left(X^p_{f_k}, Y^q_{f_k}\right) \tag{6}$$

Secondly, F is rebuilt into K levels according to the following rules:

[1] http://www.faceplusplus.com.

```
Algorithm 2 Feature Funnel Model
Inputs: probe $p$, gallery set $G$
Outputs: $Sim(p,q|F_K)$
$G_1 = G$, $F_0 = \phi$
for i=1 to $K-1$ do
    $F_i = F_{i-1} \bigcup f_i$, $f_i$ is the $i$-th reliable feature.
    Based on the feature space $F_i$, calculate the similarity $Sim(p,q|F_i)$ between probe $p$ and gallery set $G_i$.
    Find the minimum similarity $Min_k$.
    Find the $Sim(p,q|F_i) < \alpha_k \cdot Min_k$ gallery set $G_{i+1}$.
end for
calculate the similarity $Sim(p,q|F_K)$ between probe $p$ and gallery set $G_K$.
```

1) The 1st step of the feature space F_1 includes the most reliable feature f_1 in F.
2) The k-th step of the feature space F_k includes all features in F_{k-1} and the most reliable feature f_k in $F - F_{k-1}$.

Then, we use these feature spaces rebuilt feature space $\{F_1, F_2, \ldots, F_k\}$ to filter the gallery set G for probe p. Based on the first feature space F_1, we calculate the similarity $Sim(p,q|F_1)$ between probe p and gallery set G to find the minimum similarity Min_1.

In the k-th level $(k = 1, 2, \cdots, K-1)$, the gallery image q will be selected from the gallery set G_k to G_{k+1} in the $(k+1)$-th level if $Sim(p,q|F_k) < \alpha_k \bullet Min_k$. And the rest distracters are then abandoned.

Finally, the similarity $Sim(p,q|F_K)$ is used to re-identify the probe p based on the last feature level F_K.

4. Experiments and analysis

In this section, some re-id experiments are presented to demonstrate the effectiveness of the method.

4.1. Datasets and evaluation protocol

Our approach is evaluated on three RGB-D re-id datasets: the BIWI RGBD-ID, IAS-Lab RGBD-ID and the new proposed RobotPKU RGBD-ID dataset. Results are reported in the form of average Cumulated Matching Characteristic (CMC) [1] which is commonly used in re-id problem.

4.1.1. BIWI RGBD-ID dataset [8]

This dataset[2] is collected with Kinect sensors using the Microsoft Kinect SDK. It contains a training set and two testing sets. The training set includes video sequences of 45 persons. The testing set contains 56 sequences and each person has both one still sequence and one walking sequence. People wear different clothes in the training video with respect to their two testing sequences, and people are wearing the same clothes in still and walking sequences. This dataset includes RGB images, depth images, persons' segmentation maps and skeletal data.

4.1.2. IAS-Lab RGBD-ID dataset

This dataset[3] is acquired using the OpenNI SDK and the NST tracker. It contains 33 sequences of 11 people. Unlike BIWI RGBD-ID, the Training and TestingB sequences of this dataset have strong illumination varying because of the different auto-exposure level of the Kinect in the two rooms.

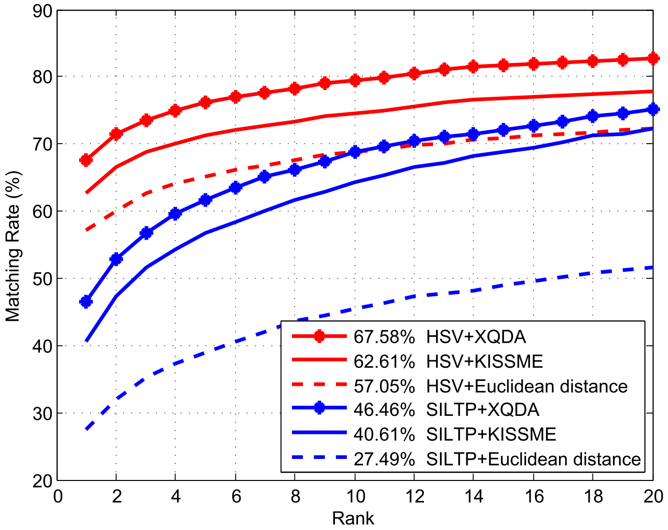

Fig. 4. The performances of metric learning methods with different features on the RobotPKU RGBD-ID Dataset.

4.1.3. RobotPKU RGBD-ID dataset

To perform more extensive experiments on a larger amount of data we collected our own RGB-D dataset called RobotPKU RGBD-ID Dataset[4]. This dataset is collected with Kinect sensors using the Microsoft Kinect SDK. This dataset contains 180 video sequences of 90 person, and for each one the Still and Walking sequences were collected in two different rooms. This dataset includes RGB images, depth images, persons' segmentation maps and skeletal data.

4.2. Evaluation of online metric update

The effectiveness of XQDA is first evaluated on RobotPKU RGBD-ID dataset, comparing with some state-of-the-art metric learning methods as shown in Fig. 4. The dataset is split into two parts, both consisting of 50 individuals, one for training and the other for testing. It is indicated that the metric learning methods (XQDA and KISSME [23]) can learn more information and perform better than conventional Euclidean distance [11]. Especially, XQDA achieves an improvement of 4.97% for HSV and 5.85% for SILTP respectively compared with KISSME.

Since face is an important information in update stage, we analyze the face reliability calculation procedure. According to Table 1, and we can see that 28.26% of positive pairs and 28.36% of negative pairs can detect both two effective face images, respectively. Fig. 5 shows the similarity score θ distributions of positive samples and negative samples pairs respectively, we can see that they are independent and identical distributions. The average (standard deviation) of positive samples and negative samples are 0.7397 (0.0997) and 0.5540 (0.0924). So when $\theta_1 > 0.8$, the two face images are considered to be a positive pair, i.e. face images are from same person. On the contrary, when $\theta_2 < 0.5$, the two face images are considered to be a negative pair.

To validate the performances of online metric model update

[2] http://robotics.dei.unipd.it/reid.

[3] http://robotics.dei.unipd.it/reid.

[4] https://github.com/lianghu56/RobotPKU-RGBD-ID-dataset.

Table 1
The result of face detection and verification.

Results		Positive pairs $(p = q)$	Negative pairs $(p \neq q)$
Percent of the Detected Face Pairs		28.26%	28.36%
Similarity Score θ	$F(X^p, Y^q) \geq 0.8$	26.42%	0.20%
	$0.8 > F(X^p, Y^q) \geq 0.5$	71.51%	66.84%
	$F(X^p, Y^q) < 0.5$	2.07%	32.96%
	Average	0.7397	0.5540
	Standard deviation	0.0997	0.0924

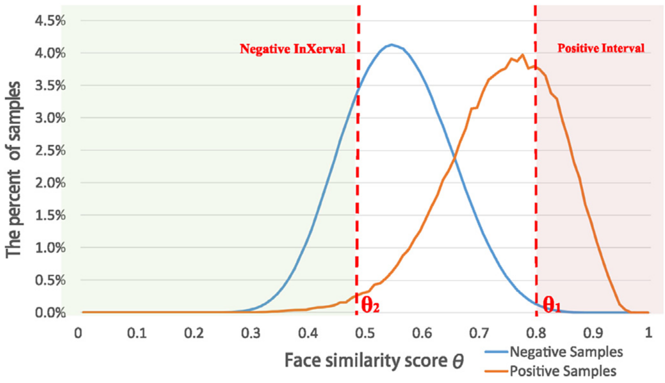

Fig. 5. The face similarity score θ distributions of positive samples and negative samples pairs.

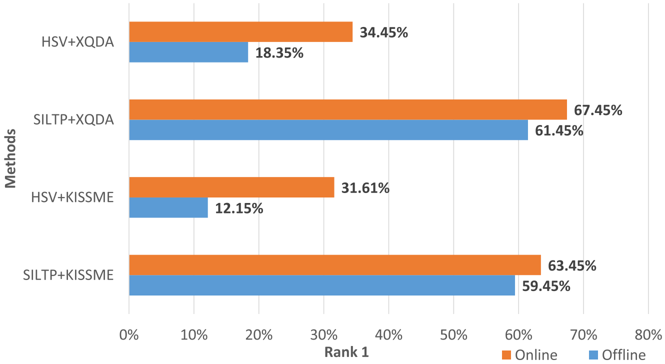

Fig. 7. The results of different features and metric methods on the IAS-Lab RGBD-ID Dataset.

strategy, we compare the online update strategy with offline strategy using different features and metric learning methods. As shown in Fig. 6, there are four sets of comparative experiments. The result shows that rank1 of online update strategy is higher than offline but the performance is limited. The reasonable interpretation is that online update strategy can correct its mistakes through adding new labeled datas which are error positive samples and error negative samples coming from the new scenes. However, the initial training data and test data are drawn from the same environment, where the online learning strategy help little on the performance.

When we initially train the metric model using the RobotPKU RBGD-ID dataset and test it on IAS-Lab RGBD-ID dataset, the rank 1 of offline HSV+XQDA method and offline HSV+KISSME method are 18.35% and 12.15%, respectively (see Fig. 7). The offline SILTP+XQDA method and the offline SILTP+KISSME method achieve 61.45% and 59.45%. The performance of HSV is poorer than that of SILTP. The reason is that IAS-Lab RGBD-ID dataset has strong illumination changes which has a great influence on the color feature (see Fig. 1(a) and (b)). Therefore, the performance of HSV is not good, but varying illumination has less performance on SILTP. Comparing online methods with offline methods, online update strategy can increase the performance obviously particularly HSV+XQDA methods with an improvement of 16.10%. The performance improvement on the IAS-Lab is more obvious than on the RobotPKU RGB-D dataset. This is due to the obvious scenery variation. However, it is not obvious on SILTP owing to SILTP is insensitivity to varying illumination. In general, the online update strategy can adapt to the new environment than offline strategy.

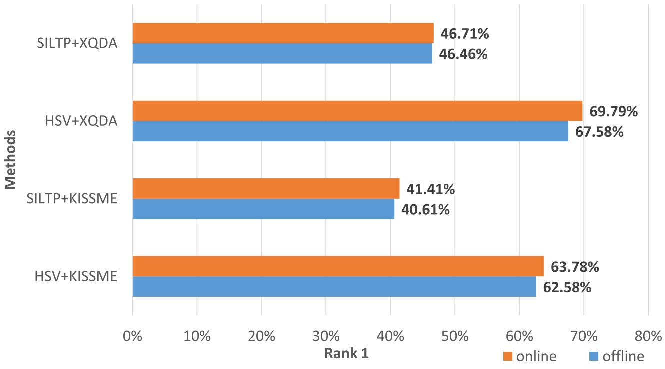

Fig. 6. The results of different features and metric methods on the BIWI RGBD-ID Dataset.

4.3. Evaluation of feature funnel model

Based on the discussion in the 4.2, the two face images are considered to be positive pair when $\theta_1 > 0.8$ and the two face images are considered to be negative pair when $\theta_2 < 0.5$. Therefore, $\theta_1 = 0.8$, $\theta_2 = 0.5$, $E = 10$ is set for this dataset.

Fig. 8 shows the results of using different feature fusion methods. In the experiments, concatenation algorithm denotes the method that these features are concatenated to obtain the final person feature. Score-level fusion denotes the method summing the scores of the different features. In addition, the single feature extraction methods which are used in feature fusion methods are also shown. As we can see, the feature funnel model achieves 77.94% accuracy rate under rank-1 identification, which is better than all other single feature extraction methods and has an improvement of 3.01% over score-level fusion method. This is because different feature methods have different reliability, concatenating directly or summing feature similarity scores brings unnecessary error. Using stable feature to filter out some special samples, it achieves better performance than all other feature fusion methods.

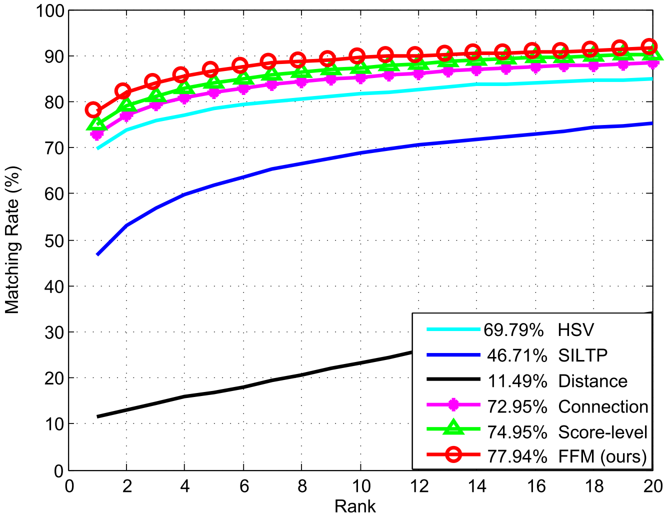

Fig. 8. The results of different features and fusion methods on the RobotPKU RGBD-ID Dataset.

Table 2
BIWI RGBD-ID dataset: comparison of state-of-the-art rank1 matching rates.

Method	Rank1
Online+FFM (Ours)	**91.6%**
Offline+FFM	86.2%
Point Cloud Matching [12]	22.4%
Skeleton(NN) [12]	21.1%
PCM+Skeleton [12]	27.4%
Face(SVM) [12]	36.7%
Face+Skeleton(SVM) [12]	43.9%
SIFT+Skeleton Keypoint [11]	65.7%

4.4. Evaluation of cross-dataset system

The proposed framework is also compared with the state-of-the-arts on the BIWI RGBD-ID dataset. In according to most re-id methods, we assume that the same person wears the same clothes, therefore we just use still sequences and walking sequences. Besides, the frames with missing skeleton joints are also discarded. In our approach, the metric model is trained initially by the RobotPKU RGBD-ID dataset, and tested on BIWI RGBD-ID dataset. The experiments of other methods are made on BIWI RGBD-ID dataset directly. The rank-1 identification rates of various algorithms are shown in Table 2. As expected, combining with extra information, like face and skeleton, will bring a certain promotion to accuracy (row 6 and row 7). Therefore, our methods (row 1 and row 2) which combine appearance-based and body geometric information also achieve good performances. As we can see, our proposed method achieves 91.6% accuracy rate under rank1 identification, with an improvement of 5.4% over the offline strategy. This was due to the online update strategy can quickly adapt to new environment by the increase of typical training data.

5. Conclusions and future work

In this paper, we present an online re-id learning framework. To overcome the drawback of offline training, that metric model can not well adapt to the changing environment, face information is utilized to update metric model online. In particular, the face information is used to find measured incorrect examples, then these examples are added to training set to update metric model. In addition, the feature funnel model is proposed to fuse the similarity scores from different feature expressions.

In order to validate the efficiency of the method, experiments are conducted on two public dataset, BIWI and IAS-Lab. In addition, a bigger dataset named RobotPKU RGBD-ID is collected to perform more extensive experiments. The results demonstrate that our method makes metric model adapt to environmental changes and has powerful transplantable ability, and the feature funnel model makes full use of features information to improve the recognition rate. Therefore, the proposed method is ideal for the use in robotic applications dealing with human-computer interactions robot.

In future work, we plan to extend the work to solve the occupation problem which may lead to the skeleton missing problem and poor re-id performance.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (NSFC, nos. 61340046), the National High Technology Research and Development Programme of China (863 Programme, no. 2006AA04Z247), the Scientific and Technical Innovation Commission of Shenzhen Municipality (nos. JCYJ20130331144631730), and the Specialized Research Fund for the Doctoral Programme of Higher Education (SRFDP, no. 20130001110011).

References

[1] S. Gong, M. Cristani, S. Yan, C.C. Loy, Person Re-identification, 2014, pp. 1–20.
[2] Y. Yang, Salient color names for person re-identification, in: European Conference on Computer Vision, 2014.
[3] B. Ma, Y. Su, F. Jurie, Local descriptors encoded by fisher vectors for person re-identification, in: International Conference on Computer Vision, 2012, pp. 413–422.
[4] X. Xu, Distance metric learning using privileged information for face verification and person re-identification, Trans. Neural Netw. Learn. Syst. (2015) 1.
[5] A. Mogelmose, T.B. Moeslund, K. Nasrollahi, Multimodal person re-identification using rgb-d sensors and a transient identification database, in: International Workshop on Biometrics and Forensics, 2013, pp. 1–4.
[6] A. Mogelmose, C. Bahnsen, T.B. Moeslund, A. Clapes, S. Escalera, Tri-modal person re-identification with rgb, depth and thermal features, in: Computer Vision and Pattern Recognition Workshops, 2013, pp. 301–307.
[7] 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: International Conference on Pattern Recognition, 2012, pp. 2694–2697.
[8] M. Munaro, A. Fossati, A. Basso, E. Menegatti, L.V. Gool, One-Shot Person Re-identification with a Consumer Depth Camera, 2014.
[9] D. Baltieri, R. Vezzani, R. Cucchiara, A. Utasi, Multi-view people surveillance using 3d information, in: International Conference on Computer Vision Workshops, 2011, pp. 1817–1824.
[10] J. Oliver, A. Albiol, A. Albiol, 3d descriptor for people re-identification, in: International Conference on Pattern Recognition, 2011, pp. 1395–1398.
[11] M. Munaro, S. Ghidoni, D.T. Dizmen, E. Menegatti, A feature-based approach to people re-identification using skeleton keypoints, in: International Conference on Robotics and Automation, 2014, pp. 5644–5651.
[12] M. Munaro, A. Basso, A. Fossati, L. Van Gool, E. Menegatti, 3d Reconstruction of Freely Moving Persons for Re-identification with a Depth Sensor, 2014, pp. 4512–4519.
[13] A. Ross, A. Jain, Information fusion in biometrics, Pattern Recognit. Lett. (2010) 2115–2125.
[14] F. Pala, R. Satta, G. Fumera, F. Roli, Multi-modal person re-identification using rgb-d cameras, Trans. Circuits Syst. Video Technol. (2015), 1–1.
[15] R. Satta, G. Fumera, F. Roli, Fast person re-identification based on dissimilarity representations, Pattern Recognit. Lett. (2012) 1838–1848.
[16] N. Noceti, F. Odone, Semi-supervised learning of sparse representations to recognize people spatial orientation, in: International Conference on Image Processing, 2014, pp. 3382–3386.
[17] C. Chen, J. Odobez, We are not contortionists: coupled adaptive learning for head and body orientation estimation in surveillance video, in: Computer Vision and Pattern Recognition, 2012, pp. 1544–1551.
[18] Z. Kalal, J. Matas, K. Mikolajczyk, P-n learning: bootstrapping binary classifiers by structural constraints, in: Computer Vision and Pattern Recognition, 2010, pp. 49–56.
[19] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikainen, Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes, in: Computer Vision and Pattern Recognition, 2010, pp. 1301–1306.
[20] T. Ojala, M. Pietikäinen, D. Harwood, A comparative study of texture measures

with classification based on featured distributions, Pattern Recognit. (1996) 51—59.

[21] S. Liao, Y. Hu, X. Zhu, S.Z. Li, Person re-identification by local maximal occurrence representation and metric learning, in: Computer Vision and Pattern Recognition, 2015, pp. 2197—2206.

[22] Y. Sun, X. Wang, X. Tang, Deep learning face representation by joint identification-verification, Adv. Neural Inf. Process. Syst. (2014) 1988—1996.

[23] P.M. Roth, P. Wohlhart, M. Hirzer, M. Kostinger, H. Bischof, Large scale metric learning from equivalence constraints, in: Computer Vision and Pattern Recognition, 2012, pp. 2288—2295.