Activity-Based Human Identification Using Discriminative Sparse Projections and Orthogonal Ensemble Metric Learning

Haibin Yan[1], Jiwen Lu[2], and Xiuzhuang Zhou[3],*

[1]National University of Singapore, Singapore
[2]Advanced Digital Sciences Center, Singapore
[3]Captial Normal University, Beijing, China

Abstract. In this paper, we propose an activity-based human identification approach using discriminative sparse projections (DSP) and orthogonal ensemble metric learning (OEML). Unlike gait recognition which recognizes person only from his/her walking activity, this study aims to identify people from more general types of human activities such as eating, drinking, running, and so on. That is because people may not always walk in the scene and gait recognition fails to work in this scenario. Given an activity video, human body mask in each frame is first extracted by background substraction. Then, we propose a DSP method to map these body masks into a low-dimensional subspace and cluster them into a number of clusters to form a dictionary, simultaneously. Subsequently, each video clip is pooled as a histogram feature for activity representation. Lastly, we propose an OEML method to learn a similarity distance metric to exploit discriminative information for recognition. Experimental results show the effectiveness of our proposed approach and better recognition rate is achieved than state-of-the-art methods.

Keywords: Human identification, activity analysis, subspace learning, sparse coding, metric learning.

1 Introduction

Over the past two decades, gait recognition has attracted much attention in computer vision [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14] because human gait provides a noninvasive way to human identification at a distance. One key shortcoming of gait recognition is that only the walking activity is exploited for human identification and these gait recognition systems are likely to fail to work when people perform other activities such as eating, drinking, and running rather than walking. In many real-world applications, people may not always walk in the scene and it is very likely that they are performing other activities besides walking in the scene. Since gait can provide enough discriminative information for human identification, a natural question arises: is it possible

[*] Corresponding author.

to identify people from different types of activities rather than gait since gait can be considered as a special case of general human activities? If so, how to effectively explore discriminative features of these activities to achieve this goal? In this paper, we provide a positive answer to these two questions.

Intuitively, the manner with which humans perform different activities can provide some distinctive information for human identification because human body information is generally distinct for different persons. Moreover, different dynamic information observed in other activities are also discriminative. Similar to gait recognition, people may perform the same activity in different manners. While gait recognition [1], [2], [3], [4], [5], [7], [12] has been extensively studied over the past decade, there has been extremely few attempts on using other activities rather than gait for human identification. In this paper, we present a new approach to activity-based human identification. For each activity video, human body mask in each frame is extracted by background substraction. Then, we project these body masks into a low-dimensional subspace and cluster them into a number of clusters, simultaneously. Subsequently, each video clip is pooled as a histogram feature for activity representation. Finally, we propose an OEML method to learn a discriminative distance metric for discriminative feature extraction. Experimental results show the effectiveness of our proposed approach.

2 Related Work

Human Activity Analysis: In computer vision, a large number of activity recognition methods have been proposed in recent years [15], [16], [17], [18], [19], [20], [21], [22]. Unlike activity recognition which aims to recognize the type of human activity from videos, activity-based human identification is a relatively new research topic, and there has been only a few seminal studies in recent years [23], [24], [25]. To our best knowledge, Gkalelis *et al.* [23] was the first attempt to formally address the problem of activity-based human identification by using fuzzy c-means (FCM) and linear discriminant analysis (LDA). Their method was further evaluated on more activity datasets and encouraging results were achieved to show the feasibility of human identification using activities [24]. More recently, Lu *et al.* [25] presented a sparse coding method for activity-based human identification. Since the the quantization error is reduced, their method achieved better performance than [23]. However, both FCM and sparse coding are not discriminative enough since they are generative methods. Moreover, these methods performed feature quantization in the original feature space, which may not be effective enough because some irrelevant and redundancy information are contained in this space. To address these shortcomings, we propose a discriminative sparse projections (DSP) method to learn a low-dimensional subspace for feature quantization, so that the irrelevant information of human body masks is discarded in the learned subspace and a discriminative codebook can be obtained for feature encoding.

Metric Learning: Metric learning has been proven to be an effective tool for visual analysis and many such algorithms have been presented over the past decade [26], [27], [28], [29], [30], [31], [32], [33], [34]. While these methods have achieved reasonably good performance in many computer vision applications, these methods usually suffer from high-dimensional feature representations. To address this, PCA is usually applied to reduce the feature dimensionality before metric learning. However, such a preprocessing may lose some discriminative information. In this paper, we propose a new OEML method to learn multiple projections from randomly sampled subsets of training samples, and orthogonalize these projections and combine them into a distance metric. Hence, no PCA preprocessing is required in our method. Moreover, the basic vectors of our learned distance metric are orthogonal to each other such that they are more compact than those of most existing metric learning methods [26], [27], [28], [29], [31].

3 Proposed Approach

Our key objective of this work is to learn discriminative identity information from activities for person recognition. Such information can be exploited at two levels: the single frame level and the whole video level. To extract discriminative information at the single frame level, we propose simultaneously learning a low-dimensional subspace and a discriminative dictionary, so that the irrelevant and redundancy information of body masks are discarded in the learned subspace and discriminative information can be exploited in the learned dictionary. To extract discriminative information at the whole video level, we propose OEML to learn a discriminative distance metric to enhance their separability. We will detail the proposed approach in the following subsections.

3.1 Body Mask Extraction

For each activity video, we first extract human body silhouette in each frame by background subtraction by using the method in [4]. Then, we align each body mask into 64×48 in each frame to make all body masks in different frames are of the same size. Fig. 1 shows several extracted body masks from different types of activities.

3.2 Discriminative Sparse Projections

Let $Y = [y_1, y_2, \cdots, y_N] \in R^{d \times N}$ be a training set of binary masks, where $y_i \in R^d$ is the ith sample, d is the feature dimension of each y_i, and N is the number of training samples. The aim of DSP is to learn a low-dimensional subspace $P \in R^{l \times d}$ and a codebook $U \in R^{l \times K}$, under which each sample y_i is encoded as $v_i \in R^K$ so that

1. Each sample y_i is sparsely reconstructed by v_i over U;
2. The intraclass and interclass variations of each y_i are minimized and maximized, simultaneously.

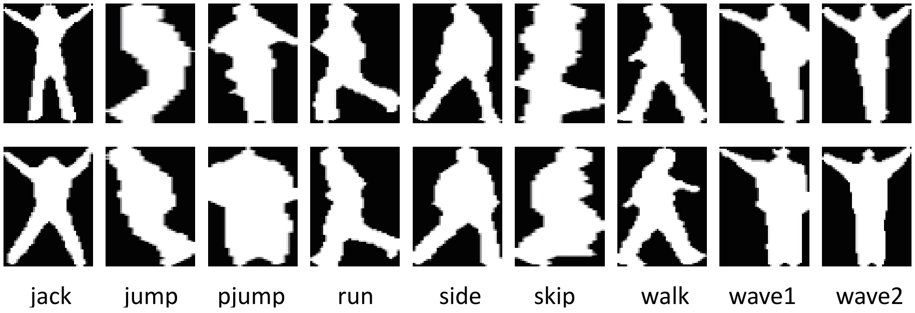

Fig. 1. Extracted and aligned body masks from different activities in the Weizmann dataset.

We present the following optimization objective function to achieve the above goals:

$$\begin{aligned} \min_{P,U,V} \quad & \|PY - UV\|_F^2 + \alpha\|Y - P^T PY\|_F^2 \\ & +\beta(\sum_{ij}^{N} \|v_i - v_j\|^2 W_{ij}^c - \sum_{ij}^{N} \|v_i - v_j\|^2 W_{ij}^p) \\ \text{subject to} \quad & PP^T = I, \|v_i\|_0 \leq T_0, \text{and} \|u_i\|_F^2 \leq 1, \forall i. \end{aligned} \tag{1}$$

where $I \in R^{l \times l}$ is the identity matrix, α and β are non-negative constants and they were empirically set as 1.0 and 1.0 in our experiments, P is the learned low-dimensional subspace, and rows of P are enforced to be orthogonal and normalized to unit norm. U is the dictionary learned in the low-dimensional subspace, $\|u_i\|_F^2 \leq 1$ is to constrain the scale of u_i, V is the sparse representation of Y over U, and T_0 is the sparsity level, W^c and W^p are two affinity matrices to characterize the geometrical structure of the samples in the training set, which are defined as [35]:

$$W_{ij}^c = \begin{cases} 1 & \text{if } x_i \in N_{k_1}^+(x_j) \text{ or } x_j \in N_{k_1}^+(x_i) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

and

$$W_{ij}^p = \begin{cases} 1 & \text{if } x_i \in N_{k_2}^-(x_j) \text{ or } x_j \in N_{k_2}^-(x_i) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $N_{k_1}^+(x)$ and $N_{k_1}^-(x)$ denote the k_1-intra-class and k_2-inter-class nearest neighbors of x, respectively, and k_1 and k_2 are two empirically pre-specified parameters to define the sizes of the local neighborhoods. With some algebraic deduction, the third term of Eq. (1) can be simplified as

$$\begin{aligned} &\sum_{ij}^{N} \|v_i - v_j\|^2 W_{ij}^c - \sum_{ij}^{N} \|v_i - v_j\|^2 W_{ij}^p \\ = \ & tr(V^T L^c V) - tr(V^T L^P V) \end{aligned} \tag{4}$$

where $L^c = D^c - W^c$ and $L^p = D^p - W^p$ are two Laplacian matrices, $D_{ii}^c = \sum_j W_{ij}^c$ and $D_{ii}^p = \sum_j W_{ij}^p$ are two diagonal matrices to reflect the degree of the ith sample, respectively.

In Eq. (1), the first term aims to seek sparse signals in the low-dimensional subspace, the second term preserves the energy of the samples in the learned subspace as much as possible, the third term aims to maximize the between-class margin in a local neighborhood.

While the objective function in Eq. (1) is not convex over P, U and V, it is convex to one of them when the other two are fixed. Following the work [36], we iteratively optimize P, U and V using the following three-stage method:

Step 1: Solve P with fixed U and V: when U and V are fixed, Eq. (1) can be rewritten as

$$\begin{aligned} &\min_{P} \ \|PY - UV\|_F^2 + \alpha \|Y - P^T PY\|_F^2 \\ &\text{subject to } \ PP^T = I. \end{aligned} \tag{5}$$

Let $Q = UVY^{-1}$. Eq. (5) can be formulated as

$$\begin{aligned} &\min_{P} \ \|P - Q\|_F^2 + \alpha \|I - P^T P\|_F^2 \\ &\text{subject to } \ PP^T = I. \end{aligned} \tag{6}$$

We construct a Lagrange function as follows

$$\mathcal{L}(P, \mu) = \|P - Q\|_F^2 + \alpha \|I - P^T P\|_F^2 - \mu(PP^T - I) \tag{7}$$

Let $\frac{\partial \mathcal{L}(P,\mu)}{\partial P} = 0$ and $\frac{\partial \mathcal{L}(P,\mu)}{\partial \mu} = 0$, we have

$$\frac{\partial \mathcal{L}(P, \mu)}{\partial P} = (1 - \alpha - \mu)P - 2Q = 0 \tag{8}$$

$$\frac{\partial \mathcal{L}(P, \mu)}{\partial \mu} = PP^T - I = 0 \tag{9}$$

According to Eqs. (8) and (9), P can be obtained as

$$P = \frac{UVY^{-1}}{2\|UVY^{-1}\|_F^2} \tag{10}$$

Step 2: Solve U with fixed P and V: when P and V are fixed, Eq. (1) can be rewritten as

$$\begin{aligned} &\min_{U} \ \|PY - UV\|_F^2 \\ &\text{subject to } \ \|u_i\|_F^2 \le 1, \forall i. \end{aligned} \tag{11}$$

```
Input: Training set $Y = [y_1, y_2, \cdots, y_N] \in R^{d \times N}$, affinity matrices $W^c$ and $W^p$,
        parameters $\alpha$, $\beta$, $T_0$, iteration number $R$, convergence error $\epsilon$.
Output: Projection matrix $P$, dictionary $U$, and sparse coefficient matrix $V$.
Step 1 (Initialization):
        Compute the initiations: $P^0$, $U^0$ and $V^0$.
Step 2 (Local optimization):
        For $r = 1, 2, \cdots, R$, repeat
        2.1. Solve $P$ with fixed $U$ and $V$ via Eq. (10).
        2.2. Solve $U$ with fixed $P$ and $V$ via Eq. (11).
        2.3. Solve $V$ with fixed $P$ and $U$ via Eq. (13).
        2.3. If $r > 2$ and $|U^r - U^{r-1}| < \epsilon$, go to Step 3.
Step 3 (Output):
        Output $P^r$, $U^r$, and $V^r$.
```

Algorithm 1: DSP

Eq. (11) is a least square problem with quadratic constraints. There are many possible methods to solve this problem. Following [36], we use the conjugate gradient decent method to learn the dictionary U.

Step 3: Solve V with fixed P and U: when P and U are fixed, Eq. (1) can be rewritten as

$$\begin{aligned} \min_{V} \quad & \|PY - UV\|_F^2 + \beta(tr(V^T L^c V) - tr(V^T L^P V)) \\ \text{subject to} \quad & \|v_i\|_0 \le T_0, \forall i. \end{aligned} \tag{12}$$

Following the work in [36], we optimize each v_i individually by fixing other coefficients v_j $(j \neq i)$. We rewrite Eq. (12) as

$$\begin{aligned} \min_{v_i} \quad & \|PY - Uv_i\|_F^2 + \beta G(v_i) \\ \text{subject to} \quad & \|v_i\|_0 \le T_0, \forall i. \end{aligned} \tag{13}$$

where

$$G(v_i) = (v_i V L_i^c + (V L_i^c)^T v_i - v_i L_{ii}^c v_i) - (v_i V L_i^p + (V L_i^p)^T v_i - v_i L_{ii}^p v_i) \tag{14}$$

We apply the feature sign search algorithm [36] to solve each v_i.

Now, we discuss how to set the initiations of our proposed DSP method. According to the second term of Eq. (1), the objective of P is to preserve the energy of the samples in the learned subspace as much as possible. Hence, we first learn a PCA subspace on Y as the initiation of P^0. Then, we apply P_0 to map Y into a low-dimensional subspace Y_1. Lastly, we employ the conventional sparse coding method [37] on Y_1 to learn U^0 and V^0 as the initiations of U and V. The proposed DSP method is summarized in **Algorithm 1**.

Having obtained $V = [v_1, v_2, \cdots, v_M]$ for a set of human body masks extracted from one activity video clip, we represent it as $S = [s_1, s_2, \cdots, s_K]$ by a pre-defined pooling function:

$$s = \mathcal{F}(V) \tag{15}$$

where

$$s_j = \max\{|v_{1j}|, |v_{2j}|, \cdots, |v_{Mj}|\} \tag{16}$$

s_j is the jth element of s, $j = 1, \cdots, K$, K is the size of the codebook U, which is empirically set as 200 in our implementations.

3.3 Orthogonal Ensemble Metric Learning

Let $S = [s_1, s_2, \cdots, s_n]$ be the training set of C different persons, where $s_i \in R^K$ is the feature of the ith sample and n is the number of activity video clips, $L = [l_1, l_2, \cdots, l_n]$ be the labels of the training samples, where $l_i \in [1, 2, \cdots, C]$. OEML aims to seek a distance metric M which pushes s_i and s_j $(l_i = l_j)$ as close as possible, and pull s_i and s_j $(l_i \neq l_j)$ as far as possible, simultaneously, where

$$d_M(s_i, s_j) = \sqrt{(s_i - s_j)^T M (s_i - s_j)} \tag{17}$$

where M is a $K \times K$ square matrix, and $1 \leq i, j \leq n$. Since M is a distance metric, it should be symmetric and positive semi-definite. hence, we can seek a non-square matrix Q of size $K \times K'$, where $K' \leq K$, such that

$$M = QQ^T \tag{18}$$

Then, Eq. (17) can be rewritten as

$$\begin{aligned} d_M(s_i, s_j) &= \sqrt{(s_i - s_j)^T M (s_i - s_j)} \\ &= \sqrt{(s_i - s_j)^T Q Q^T (s_i - s_j)} \\ &= \sqrt{(t_i - t_j)^T (t_i - t_j)} \end{aligned} \tag{19}$$

where $t_i = Q^T s_i$ and $t_j = Q^T s_j$.

Different from most existing distance metric learning methods [26], [27], [28], [29] which learn the distance metric over the whole training samples, we randomly sample two groups of samples from the training set and consider them as positive and negative samples for SVM learning. Assume there are C persons in the training set, we generate one group by randomly sampling F $(F \leq \frac{C}{2})$ classes from the whole training samples as positive samples. Then, we generate another group by randomly sampling F classes from the remaining training samples as negative samples. Hence, these two groups don't share any the same-class sample because we need to learn a projection vector to distinguish them.

Then, we learn a linear SVM on these two groups of samples and seek a projection vector $p_i = (w_i - b_i)^T$ to maximize the margin of these two groups of samples, where w_i and b_i are the normal vector and bias of the SVM model. We randomly iterate this procedure K' times and have multiple projection vectors

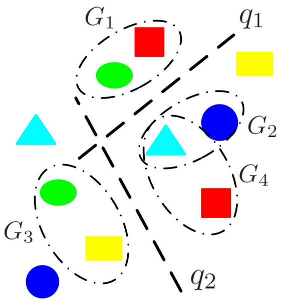

Fig. 2. Learning different projection vectors by SVM from different subsets of the training samples, where q_1 are learned from G_1 and G_2, and q_2 are learned from G_3 and G_4, respectively.

```
Input: Training set: $S = [s_1, s_2, \cdots, s_n]$, label vector $L = [l_1, l_2, \cdots, l_n]$,
       parameter $K'$.
Output: Projection matrix $Q$.
Step 1 (Learning projection vectors with SVM):
    For $k = 1, 2, \cdots, K'$, repeat
        1.1. Sampling two groups of samples from S.
        1.2. Obtain $z_k$ with SVM.
Step 2 (Orthogonazation):
    Orthogonalize $Z$ to obtain $Q$.
Step 3 (Output projection matrix):
    Output projection matrix Q.
```

Algorithm 2: OEML

$Z = [z_1, z_2, \cdots, z_{K'}]$. Fig 2 illustrates the basic idea of the learning procedure. In our experiments, we empirically set K' as 200.

Since the projection vectors are learned from the randomly sampled samples, they are not orthogonal. To reduce the redundancy of these projection vectors, we orthogonalize them to make more succinct feature extraction as follows.

Assume $Q = [q_1, q_2, \cdots, q_{K'}]$ be the orthogonal basis vectors of Z. Let $q_1 = p_1$. The ith projection vector q_i can be computed as follows:

$$q_i = z_i - \sum_{j=1}^{i-1} \frac{(q_j)^T z_i}{(q_j)^T q_j} q_j \tag{20}$$

Algorithm 2 summarizes the proposed OEML method.

4 Experimental Results

In this section, we conduct experiments on five different activity databases including the Weizmann [38], AIIA-MOBISERV [24], KTH [39], MSR [12] and TUM [40] databases to evaluate the performance of our proposed approach.

4.1 Datasets and Settings

The Weizmann dataset [38] contains 9 persons and each person performed 10 different activities including bending, jumping-jack, jumping-forward-on-two-legs, jumping in place-on-two-legs, running, galloping-sideways, skipping, walking, waving-one-hand, and waving-two-hands, respectively. There are 93 video clips in this database. Since some videos contain two or more cycles of a specific action performed by some subjects, we break up these videos into several single period activity videos. Hence, we obtain a database of 216 videos in total. For each person, we randomly selected 5 activities for training and the remaining 5 activities were used for testing.

The AIIA-MOBISERV dataset [24] was specifically designed for the activity-based human identification task. It contains 12 persons and each person performed eating and drinking activities with two different clothing in four different days. There are totally 96 videos in this database. Since some videos contain two or more cycles of a specific activity performed by some subjects, these sequences were segmented into several single-period activities. Following the settings in [24], we consider drinking with a cup and eating with a fork for human identification, where 776 video clips in total were selected. We use the eating activity for training and the drinking activity for testing.

The KTH dataset [39] contains 25 persons, and each person performed 6 different activities, including boxing, handclapping, handwaving, jogging, running, and walking, respectively. For each activity, it is captured at 4 different scenarios such as outdoor, indoor, outdoor with a scale variation, and outdoor with different clothes, respectively. In our experiments, we randomly chose 3 activities as training examples for each scenario and the remaining 3 activities as testing examples.

The MSR dataset [12] was captured by a Kinect device. There are 10 subjects in this dataset. For each subject, there are 16 activities: drinking, eating, reading a book, calling a cellphone, writing on a paper, using a laptop, using a vacuum cleaner, cheering up, sitting still, tossing paper, playing game, lying down on sofa, walking, playing guitar, standing up, and sitting, respectively. Each subject performed each activity twice: one in standing position and the other one in sitting position. For each person, both color and depth videos are captured. Hence, there are 320 videos in total. In our experiments, we only use the color videos to evaluate the performance of our approach. We randomly selected 8 activities for each person as training examples and the remaining 8 activities as testing examples.

The TUM dataset [40] is a collection of activity sequences recorded in a kitchen environment equipped with multiple complementary sensors. The recorded data consists of 4 subjects who naturally performed manipulation tasks in a kitchen environment with different manners. Different from previous activity datasets, this dataset offers more natural activities for evaluating activity recognition and motion tracking. There are multiple sensors used to capture human activities such as web camera, RFID and Magnetic (reed) sensors. In our experiments, we only use the video data for human identification. For each person, we selected 4 video sequences captured from 4 synchronized cameras which were installed at 4 different viewpoints. We randomly selected videos from two viewpoints as training examples and the remaining two viewpoints as testing examples.

We also construct a hybrid dataset which combines the Weizmann, AIIA-MOBISERV, KTH, MSR, and TUM databases into a larger dataset to evaluate the performance of our approach. Intuitively, this hybrid dataset is more challenging because there are 50 persons and different persons may perform different types of activities in the hybrid dataset. We followed the above experimental protocol for different datasets to construct the training and test datasets. Specifically, all training sets from each dataset which used in the above experiments were used for training and the remaining videos were used for testing.

We conducted experiments 10 times with different randomly selected training and testing samples, and the final result was shown as the mean of the correct identification rate[1]. In our experiments, the nearest neighbor classifier is used for classification. Since the advantage of our proposed approach results from two different stages: DSP feature encoding and OEML metric learning, we evaluate the performance where only one is applied to reveal their respective effects, respectively.

4.2 Results and Analysis

Comparison with Existing Feature Encoding Methods: We compare our proposed DSP method with different feature encoding methods including the K-means (KM), FCM, sparse coding (SC) [37], Laplacian sparse coding (LSC) [41] on the activity-based human identification task. For the SC and LSC methods, the maximal pooling was also used. The codebook size was set as 300 and the nearest neighbor (NN) classifier with the Euclidian distance was used for identification. Table 1 shows the rank-one identification rate of different feature encoding methods. We can see that our DSP performs better than the other four compared methods. This is because the other compared feature encoding methods are unsupervised and our DSP method is supervised, such that more discriminative information can be exploited in our method. Moreover, our DSP method performs feature encoding in the low-dimensional subspace, which can remove the noisy and irrelevant information in the learned codebook.

Comparison with Existing Metric Learning Methods: To investigate the effectiveness of the proposed OEML method in the activity-based human identification task, we compare it with five state-of-the-art metric learning methods including large margin nearest neighbor (LMNN) [27], neighborhood component analysis (NCA) [26], information theoretic metric learning [28], cosine similarity metric learning (CSML) [29], and neighborhood repulsed metric learning (NRML) [31]. For the first four compared methods, we empirically set the number of the nearest neighbors as 5. For the NRML method, two neighborhood sizes

[1] The AIIA-MOBISERV dataset was not repeated 10 times because the training and testing sets are fixed in this dataset.

Table 1. Rank-1 identification rate (%) of different feature encoding methods on different datasets.

Method	Weizmann	AIIA-MOBISERV	KTH	MSR	TUM	Hybird
KM	64.5	55.4	20.5	24.7	41.7	40.0
FCM	68.3	57.6	24.5	28.6	50.0	43.0
SC	72.1	59.3	27.5	30.6	50.0	45.5
LSC	73.4	61.3	30.4	32.5	58.3	48.8
DSP	**78.5**	**64.5**	**32.7**	**35.6**	**66.7**	**51.3**

Table 2. Rank-1 identification rate (%) of different metric learning methods on different datasets.

Method	Weizmann	AIIA-MOBISERV	KTH	MSR	TUM	Hybird
LMNN	75.5	59.5	22.5	27.8	58.3	45.0
NCA	74.3	58.3	21.8	26.9	50.0	44.5
ITML	74.6	58.0	21.6	27.3	50.0	44.0
CSML	76.3	60.5	25.7	30.4	66.7	47.5
NRML	77.5	61.7	28.6	33.5	66.7	49.0
OEML	**80.2**	**65.1**	**32.5**	**36.2**	**75.0**	**52.5**

were set as 5 and 20, respectively. We also applied principal component analysis (PCA) to reduce each encoded histogram feature learned into 100 dimensions for these five metric learning methods. For the proposed OEML method, we learned the distance metric directly from the original feature space. The FCM method was used for feature encoding. Table 2 compares the rank-1 identification rate of different metric learning methods. We can clearly see from this table that our OEML performs better than the other five compared metric learning. The reason is that the other compared metric learning methods learn the distance metric in the PCA reduced subspace and some discriminative information may be removed in the subspace because the objectives of PCA and these metric learning methods are usually not consistent. However, our OEML method learns the distance metric in the original high-dimensional feature space, which can exploit more discriminative information in the high-dimensional feature space directly.

Comparison with State-of-the-Art Activity-Based human identification Methods: we compare our approach with the state-of-the-art activity-

Table 3. Rank-1 identification rate (%) of different activity-based human identification methods on different datasets.

Method	Weizmann	AIIA-MOBISERV	KTH	MSR	TUM	Hybird
Method in [23]	70.4	58.6	25.8	29.4	50.0	44.9
Method in [24]	74.3	60.3	27.6	32.3	50.0	48.5
Method in [25]	75.4	62.5	31.4	35.7	66.7	50.2
Ours	**83.3**	**67.5**	**35.8**	**40.3**	**83.3**	**54.9**

Table 4. Rank-1 identification rate (%) of different combinations of feature encoding and metric learning methods on different datasets.

Method	Weizmann	AIIA-MOBISERV	KTH	MSR	TUM	Hybird
Baseline	76.1	61.3	24.9	31.2	66.7	50.6
Baseline+DSP	78.5	64.5	32.7	35.6	66.7	52.2
Baseline+OEML	81.3	65.8	33.9	37.5	75.0	52.6
DSP+OEML	**83.3**	**67.5**	**35.8**	**40.3**	**83.3**	**54.9**

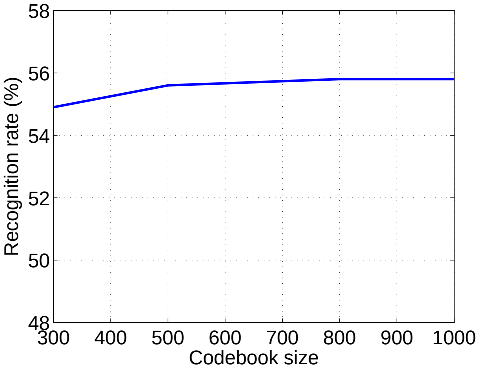

Fig. 3. Rank-1 identification rate (%) of our approach versus different codebook sizes on the hybrid dataset.

based human identification methods in [23], [24] and [25]. We implemented the three compared methods [23], [24], [25] ourselves. For a fair comparison, the number of clusters is set as 300 in our implementations for all methods. Table 3 compares the rank-1 identification rate of different methods. As can be seen from this table, our approach significantly outperforms the compared activity-based human identification methods because our approach adopts supervised feature encoding and high-dimensional metric learning, such that more discriminative information can be extracted for recognition.

Performance Analysis of Different Stages in Our Approach: We conduct experiments to analyze our approach when different modules are used. We create the baseline method which performs dictionary learning in the original feature space and uses NN for recognition without metric learning. Then, we include different modules in our approach. Table 4 compares the rank-1 identification rate of different combinations of feature encoding and metric learning methods. We see that all modules including low-dimensional subspace, discriminative dictionary learning, and discriminative metric learning contribute the final recognition rate of our approach.

Parameter Analysis: We first evaluate the effect of the codebook sizes of our approach on the hybrid dataset. Fig. 3 shows the rank-1 identification rate of our approach versus different codebook sizes on the hybrid dataset. We see

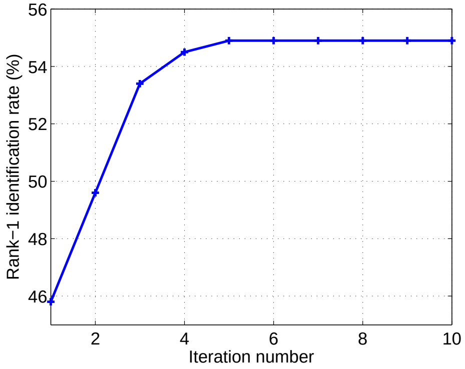

Fig. 4. Rank-1 identification rate versus different number of iterations of DSP on the hybrid dataset.

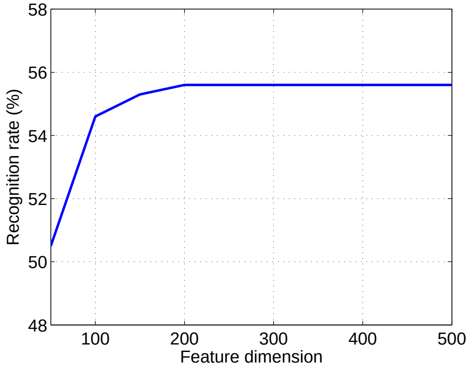

Fig. 5. Rank-1 identification rate (%) of our approach versus different number of feature dimensions on the hybrid dataset.

that the performance of our approach continues to increase as the increasing of the codebook size. However, the improvement is marginal, which indicates that the performance of our approach is not sensitive to the codebook size.

Fig. 4 shows the rank-1 identification rate versus different number of iterations on the hybrid database. We see that the recognition performance of our proposed DSP method can converge to a local optimal peak in a few iterations.

Lastly, we investigated the effect of the parameter K' in OEML. Fig. 5 shows the rank-1 identification rate versus different number of feature dimensions on the hybrid database. We see that our OEML can reach stable performance when the number of K' is above 100.

5 Conclusion

This paper presented a new activity-based human identification approach by using discriminative sparse projections and orthogonal ensemble metric learning (OEML). Experimental results demonstrate the effectiveness of the proposed approach. How to apply our proposed approach to other visual recognition applications such as face identification, object recognition, and visual tracking to further demonstrate its effectiveness seems an interesting future work.

Acknowledgement

This study is partially supported by the research grant for the Human Cyber Security Systems (HCSS) Program at the Advanced Digital Sciences Center (ADSC) from the Agency for Science, Technology and Research (A*STAR) of Singapore, and the research grant from the National Natural Science Foundation of China under Grant 61373090.

References

1. Lee, L., Grimson, W.: Gait analysis for recognition and classification. In: IEEE International Conference on Automatic Face and Gesture Recognition. (2002) 148–155
2. Wang, L., Tan, T., Ning, H., Hu, W.: Silhouette analysis-based gait recognition for human identification. IEEE Transactions on Pattern Analysis and Machine Intelligence **25**(12) (2003) 1505–1518
3. Kale, A., Sundaresan, A., Rajagopalan, A., Cuntoor, N., Roy-Chowdhury, A., Kruger, V., Chellappa, R.: Identification of humans using gait. IEEE Transactions on Image Processing **13**(9) (2004) 1163–1173
4. Sarkar, S., Phillips, P., Liu, Z., Vega, I., Grother, P., Bowyer, K.: The humanid gait challenge problem: Data sets, performance, and analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(2) (2005) 162–177
5. Boulgouris, N., Hatzinakos, D., Plataniotis, K.: Gait recognition: a challenging signal processing technology for biometric identification. IEEE Signal Processing Magazine **22**(6) (2005) 78–90
6. Lu, J., Zhang, E.: Gait recognition for human identification based on ica and fuzzy svm through multiple views fusion. Pattern Recognition Letters **28**(16) (2007) 2401–2411
7. Tao, D., Li, X., Wu, X., Maybank, S.: General tensor discriminant analysis and gabor features for gait recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(10) (2007) 1700–1715
8. Li, X., Maybank, S., Yan, S., Tao, D., Xu, D.: Gait components and their application to gender recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews **38**(2) (2008) 145–155
9. Lu, J., Tan, Y.: Uncorrelated discriminant simplex analysis for view-invariant gait signal computing. Pattern Recognition Letters **31**(5) (2010) 382–393
10. Lu, J., Tan, Y.P.: Gait-based human age estimation. IEEE Transactions on Information Forensics and Security **5**(4) (2010) 761–770

11. Liu, N., Lu, J., Tan, Y.P.: Joint subspace learning for view-invariant gait recognition. IEEE Signal Processing Letters, **18**(7) (2011) 431–434
12. Wang, C., Zhang, J., Wang, L., Pu, J., Yuan, X.: Human identification using temporal information preserving gait template. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(11) (2012) 2164–2176
13. Lu, J., Tan, Y.P.: Ordinary preserving manifold analysis for human age and head pose estimation. IEEE Transactions on Human-Machine Systems **43**(2) (2013) 249–258
14. Lu, J., Wang, G., Moulin, P.: Human identity and gender recognition from gait sequences with arbitrary walking directions. IEEE Transactions on Information Forensics and Security **9**(1) (2014) 51–61
15. Wu, J., Osuntogun, A., Choudhury, T., Philipose, M., Rehg, J.: A scalable approach to activity recognition based on object use. In: IEEE Interntaional Conference on Computer Vision. (2007) 1–8
16. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: IEEE International Conference on Computer Vision and Pattern Recognition. (2009) 2929–2936
17. Turaga, P., Chellappa, R., Subrahmanian, V.S., Udrea, O.: Machine recognition of human activities: A survey. IEEE Transactions on Circuits and Systems for Video Technology **18**(11) (2008) 1473–1488
18. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing **28**(6) (2010) 976–990
19. Tran, D., Sorokin, A.: Human activity recognition with metric learning. In: European Conference on Computer Vision. (2008) 548–561
20. Mikolajczyk, K., Uemura, H.: Action recognition with motion-appearance vocabulary forest. In: IEEE International Conference on Computer Vision and Pattern Recognition. (2008) 1–8
21. Junejo, I., Dexter, E., Laptev, I., Pérez, P.: View-independent action recognition from temporal self-similarities. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(1) (2011) 172–185
22. Seo, H., Milanfar, P.: Action recognition from one example. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(5) (2011) 867–882
23. Gkalelis, N., Tefas, A., Pitas, I.: Human identification from human movements. In: IEEE International Conference on Image Processing. (2009) 2585–2588
24. Iosifidis, A., Tefas, A., Pitas, I.: Activity-based person identification using fuzzy representation and discriminant learning. IEEE Transactions on Information Forensics and Security **7**(2) (2012) 530–542
25. Lu, J., Hu, J., Zhou, X., Shang, Y.: Activity-based person identification using s-parse coding and discriminative metric learning. In: ACM International Conference on Multimedia. (2012) 1061–1064
26. Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R.: Neighborhood component analysis. In: Advances in Neural Information Processing Systems. (2004) 2539–2544
27. Weinberger, K., Blitzer, J., Saul, L.: Distance metric learning for large margin nearest neighbor classification. In: Advances in Neural Information Processing Systems. (2005)
28. Davis, J., Kulis, B., Jain, P., Sra, S., Dhillon, I.: Information-theoretic metric learning. In: International Conference on Machine Learning. (2007) 209–216
29. Nguyen, H., Bai, L.: Cosine similarity metric learning for face verification. Asian Conference on Computer Vision (2011) 709–720

30. Lu, J., Tan, Y.P.: Regularized locality preserving projections and its extensions for face recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics **40**(3) (2010) 958–963
31. Lu, J., Hu, J., Zhou, X., Shang, Y., Tan, Y.P., Wang, G.: Neighborhood repulsed metric learning for kinship verification. In: IEEE International Conference on Computer Vision and Pattern Recognition. (2012) 2594–2601
32. Lu, J., Wang, G., Moulin, P.: Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning. In: IEEE International Conference on Computer Vision. (2013) 329–336
33. Lu, J., Tan, Y.P., Wang, G.: Discriminative multimanifold analysis for face recognition from a single training sample per person. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(1) (2013) 39–51
34. Yan, H., Lu, J., Deng, W., Zhou, X.: Discriminative multimetric learning for kinship verification. IEEE Transactions on Information Forensics and Security **9**(7) (2014) 1169–1178
35. Yan, S., Xu, D., Zhang, B., Zhang, H.J., Yang, Q., Lin, S.: Graph embedding and extensions: a general framework for dimensionality reduction. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(1) (2007) 40–51
36. Lee, H., Battle, A., Raina, R., Ng, A.: Efficient sparse coding algorithms. In: Advances in Neural Information Processing Systems. Volume 19. (2006) 801
37. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: IEEE International Conference on Computer Vision and Pattern Recognition. (2009) 1794–1801
38. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(12) (2007) 2247–2253
39. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: IEEE International Conference on Pattern Recognition. Volume 3. (2004) 32–36
40. Tenorth, M., Bandouch, J., Beetz, M.: The tum kitchen data set of everyday manipulation activities for motion tracking and action recognition. In: IEEE International Conference on Computer Vision Workshops. (2009) 1089–1096
41. Gao, S., Tsang, I.W., Chia, L.T., Zhao, P.: Local features are not lonely–laplacian sparse coding for image classification. In: IEEE International Conference on Computer Vision and Pattern Recognition. (2010) 3555–3561