Attribute-based Learning for Gait Recognition using Spatio-Temporal Interest Points

Worapan Kusakunniran

Faculty of Information and Communication Technology, Mahidol University, Thailand

Abstract

This paper proposes a new method to extract a gait feature from a raw gait video directly. The Space-Time Interest Points (STIPs) are detected where there are significant movements of human body along both spatial and temporal directions in local spatio-temporal volumes of a raw gait video. Then, a histogram of STIP descriptors (HSD) is constructed as a gait feature. In the classification stage, the support vector machine (SVM) is applied to recognize gaits based on HSDs. In this study, the standard multi-class (i.e. multiple subjects) classification can often be computationally infeasible at test phase, when gait recognition is performed by using every possible classifier (i.e. SVM models) trained for all individual subjects. In this paper, the attribute-based classification is applied to reduce the number of SVM models needed for recognizing each probe gait. This process will significantly reduce the test-time computational complexity and also retain or even improve the recognition accuracy. When compared with other existing methods in the literature, the proposed method is shown to have the promising performance for the case of normal walking, and the outstanding performance for the cases of walking with variations such as walking with carrying a bag and walking with varying a type of clothes.

Keywords: Gait recognition, human identification, spatio-temporal, attribute-based learning

1. Introduction

Gait recognition has been studied broadly in a surveillance perspective because it is capable of identifying humans at a distance by inspecting their walking manners. A large number of gait recognition methods have been developed and published, which can be roughly divided into two categories: 1) model-based methods [1][2][3][4][5][6][7][8][9][10]; 2) model-free methods [11][12][13][14][15][16][17][18][19][20][21][19]. These methods require a pre-processing to extract gait appearances such as shape contours, silhouettes, skeletons, and body joints for further gait analysis. This pre-processing will consume additional computing time. It may also extract incomplete gait appearances which will negatively influence performances of gait analyses in later stages. Also, these gait appearances are sensitive to variations and partial occlusions caused by several factors such as carrying a bag and varying a type of clothes [25].

1.1. Related Works

The model-based methods generally aim to model kinematics of human joints in order to measure physical gait parameters such as trajectories, limb lengths, and angular speeds [24]. For example, Cunado et al. [3] modeled legs as an interlinked pendulum. Then, a phase-weighted Fourier magnitude spectrum was used to recognize gait features which were derived from frequency components of changes in an inclination of human thighs. Bobick and Johnson [1][10] used activity-specific static body parameters for gait recognition without directly analyzing dynamics of gait patterns. Kovac and Peer [2] proposed a skeleton model-based gait recognition method by modeling gait dynamics and eliminating influences of subjects’ appearances on recognition. Tafazzoli and Safabakhsh [8] applied active contour models and the Hough transform to
model movements of articulated parts of the human body. Bouchrika and Nixon [5] used elliptic Fourier descriptors to extract crucial features from human joints. However, methods in this category have to deal with localizations of human joints, which are not robust on a markerless motion [7]. It is also difficult to extract underlying models from gait sequences [15].

The model-free methods typically analyze gait sequences without explicit modeling of human body structure [24]. The different methods in this category have been developed from different perspectives. For example, BenAbdelkader et al. [21] proposed an eigengait method using image self-similarity plots. Chai et al. [2] introduced a Perceptual Shape Descriptor technique for recognizing gaits. Tan et al. [25] used eight kinds of projective features to describe human gait and PCA was applied for gait feature dimension reduction. Han and Bhanu [11] proposed a concept of Gait Energy Image (GEI), and combined real and synthetic templates to improve the accuracy of gait recognition. Liu et al. [12] employed a population Hidden Markov Model (pHMM) to model human gaits and generated dynamics-normalized stance-frames to recognize individuals. Recently, Wang et al. [13] developed a temporal template, named Chrono-Gait Image (CGI), by encoding gait contours using a multi-channel mapping function. Roy et al. [14] modeled a gait cycle using a chain of key poses which were then averaged to generate the gait feature, called Pose Energy Image (PEI).

1.2. The Proposed Framework

Without any pre-processing on a raw gait video, the proposed method is to extract and recognize gait features directly from the video on a spatio-temporal feature domain. In this paper, the pre-processing is referred only to a foreground-background segmentation and/or a shape contour/boundary/skeleton extraction. Figure 1 shows the framework of the proposed method for gait recognition (the detailed frameworks of its key processes will be shown in Figures 2, 3, and 4). In these figures, rectangles represent inputs/outputs, while ellipses represent processing steps.

In Figure 1, gait recognition is to find the best matched identity of a probe gait against other gaits in a gallery dataset. First, spatio-temporal interest points (STIPs) are detected from each gait video individually. STIPs are local structures in a spatio-temporal domain where image values have significant local variations in both space and time. These variations are linked to significant movements of human body in a gait video. Therefore, STIP is an interest point of a dominant walking pattern, which can be used to represent characteristics of each individual gait.

Second, a histogram of STIP descriptors (HSD) is constructed as a gait feature. Each STIP descriptor is computed by applying a histogram of oriented gradients (HOG) and a histogram of optical flow (HOF) on a 3D video patch (i.e. width × height × time) in a neighborhood of each detected STIP. It well describes walking patterns around the interest point in space and time.

Third, a support vector machine (SVM) [26] is used as a classification method. SVM has been employed to classify gaits efficiently [17, 24, 25, 29, 30]. However, SVM has been applied based on the standard one-vs.-all strategy [31] which can be computationally infeasible at test phase in a large scale multi-class classification (i.e. a large number of subjects). This is because its test-time complexity grows linearly with a number of subjects.

Recently, several tree-based methods [32, 33, 34, 35, 36, 37] have been proposed to address sublinear testing cost for large multi-class tasks. However, they are limited to several practical constraints [38] including: 1) the tree-based methods cannot be further speeded up by using parallel computing which is a very important and useful methodology to solve
a computational problem in real applications; 2) the tree-based methods consume expensive memory for storing and loading all classifiers at all nodes of the tree in test phase.

To avoid these limitations, this paper employs the attribute-based classification using SVM which is shown to achieve comparable performance to the state-of-the-art regarding sublinear test-time complexity and accuracy. In this paper, the attribute-based learning is to define attributes and their associations with human subjects automatically and simultaneously. Then, a classifier is learned for each attribute instead of each subject. These trained classifiers are used to predict subjects’ identities based on their attribute representations. This can be seen that the test-time complexity grows linearly with a total number of attributes (M) instead of a total number of human subjects (N) in the gait database. Thus, M can be selected to be much less than N, in order to significantly reduce the complexity.

The rest of this paper is organized as follows. Section 2 explains the detection process of the Spatio-temporal interest points (STIPs) in a gait video. Section 3 proposes the gait feature extraction. Section 4 discusses the attribute-based learning for gait classification. Then, experimental results are shown in Section 5, and conclusions are drawn in Section 6.

2. Key Points Detection in a Gait Video

Given a raw gait video G, key spatio-temporal interest points (STIPs) \{S_n\}_{n=1}^N are detected to represent dominant walking patterns of a person in the video, where G(x, y, t) is a point at the spatio-index (x, y) of the image frame t in G, and S_n is a detected STIP. STIP is detected by considering large variations along both spatial and temporal directions in local spatio-temporal volumes of G. Figure 2 shows the proposed framework of the key points detection in a gait video.

First, a linear scale-space representation L of G is constructed by using an anisotropic Gaussian kernel \(g(x, y, t; \sigma^2, \tau^2)\) with independent spatial variance \(\sigma^2\) and temporal variance \(\tau^2\) as:

\[
L(\cdot; \sigma^2, \tau^2) = g(\cdot; \sigma^2, \tau^2) * G(\cdot) \quad (1)
\]

where * is the convolution, and \(g(\cdot; \sigma^2, \tau^2)\) is the Gaussian kernel.  

A raw gait video (G)

A linear scale-space representation (L) of G

A spatio-temporal second moment matrix (µ)

A spatio-temporal Harris corner function (H)

STIPs

Figure 2: The proposed framework of the key points detection in a gait video.

Second, a spatio-temporal second moment matrix \(\mu(\cdot; \sigma^2, \tau^2)\) within a Gaussian neighborhood of each point is constructed as:

\[
\mu(\cdot; \sigma^2, \tau^2) = g(\cdot; \sigma^2, \tau^2) * \begin{bmatrix}
L_x^2 & L_xL_y & L_xL_t \\
L_xL_y & L_y^2 & L_yL_t \\
L_xL_t & L_yL_t & L_t^2
\end{bmatrix} \quad (2)
\]

where \(L_i\) is the partial derivative of \(L\) with respect to the variable \(i\), and \(i = x, y\) or \(t\).

Third, it has been demonstrated in (11) that the interest points are located in regions of \(G\) having significant eigenvalues \(\lambda_1, \lambda_2, \lambda_3\) of \(\mu\). Thus, the spatio-temporal Harris corner function (12) is constructed by using a combination of the determinant and trace of \(\mu\) as:

\[
H = \det(\mu) - h \; \text{trace}^3(\mu) = \lambda_1\lambda_2\lambda_3 - h (\lambda_1 + \lambda_2 + \lambda_3)^3 \quad (3)
\]

Then, STIPs of G are detected at local positive maxima in \(H\) with a sufficiently large value of \(h\). Local positive maxima of \(H\) correspond to points in \(G\) with high variations in both directions of space and time (11).

The scale parameters (\(\sigma, \tau\)) in the Gaussian kernel can be automatically adjusted to match the spatio-temporal extent of underlying image structures, as shown in (11). Alternatively, STIPs can be extracted by using a set of multiple combinations of spatial (\(\sigma\)) and temporal (\(\tau\)) scales. When compared to the automatic scale selection in (11), this simplification has shown to produce similar (or
3. Gait Feature Extraction

Figure 3 shows the proposed framework of the gait feature extraction. Given a set of STIPs detected in Section 2 for each raw gait video, HOG [45] and HOF [46] are used to compute a point description for each STIP. By using each STIP as a center point, HOG and HOF are applied on a 3D volume of its neighborhood in the linear scale-space representation \( L(\sigma, \tau) \) (see Section 2). This 3D volume is partitioned into a grid with \( N_b \) spatio-temporal blocks. Then, \( a \)-bin HOG and \( b \)-bin HOF descriptors are computed for all blocks and concatenated into an \( a \times N_b \)-element HOG descriptor and \( b \times N_b \)-element HOF descriptor. Finally, the HOG and HOF descriptors are concatenated to generate the point descriptor with \( (a + b) \times N_b \) elements for each STIP.

This paper proposes a histogram of STIP descriptors (HSD) as a gait feature. It contains two main steps including: 1) a histogram quantization; and 2) a histogram computation. In the histogram quantization step, the \( k \)-means clustering [47] is applied to all STIP descriptors from the training dataset. It clusters these STIP descriptors into \( k \) groups, in which the center of each group is used to define each bin of the histogram.

In the histogram computation step, for each raw gait video, its gait feature is extracted by assigning each of its STIP descriptors to the corresponding bin (i.e. by using a chosen similarity measurement function e.g. Euclidean distance [48]) and accumulating this bin’s value by one. Then, L2-norm normalization is applied on the histogram to generate the \( k \)-bin HSD-based gait feature.

4. Attribute-based Learning for Gait Classification

In this paper, support vector machine (SVM) is used as a classification method. It has been shown to be very efficient for recognizing gaits [17][27][28][29][30]. However, these methods applied SVM based on the standard one-vs.-all strategy [31] which can often be computationally infeasible at test-phase for large scale multi-class classifications i.e. gait recognitions with a large number of individual subjects. This is because its test-time complexity grows linearly with a number of subjects, by using every possible classifiers trained for each individuals. In this paper, the test-time complexity is a complexity for determining a subject identity of a gait sequence.

The concept of attribute-based classification [38][49][40] is adopted for this study to reduce test-time complexity and also retain or even improve the recognition accuracy. Figure 4 shows the proposed framework of the attribute-based learning for gait classification.

The proposed framework contains three main steps in the training process, as shown in the first three rectangles of Figure 4. The first step is to compute a similarity matrix (\( S \)) based on one-vs.-all
classification using a training dataset. The matrix $S$ represents correlations between gaits of different subject classes. Then, the second step is to create relations between subject classes and attributes which are virtually defined based on $S$. Subject classes are represented by different sets of attributes and each attribute may belong to several related subject classes. The third step is to learn a SVM-based classifier for each attribute instead of each subject class. In this way, subject classes are associated with the trained classifiers by using their attribute representations.

Later at the test phase, as shown in the last rectangle of Figure 4, the per-attribute classifiers are applied to generate predictions of attribute values for each probe gait. The final prediction score of the probe gait against each subject class is calculated from the relevant attribute values. Based on this proposed method, it can be seen that the test-time complexity grows linearly with the total number of attributes ($M$) instead of the total number of subject classes ($C$). Thus, $M$ can be selected to be much less than $C$, in order to significantly reduce the computational complexity.

### 4.1. Attribute-based Gait Classification

Assume that there are total $M$ attributes $a_1, a_2, ..., a_M$, each subject class $c$ ($z_c$) is represented by a set of attributes $A_c = \{a_1^c, a_2^c, ..., a_M^c\}$ where $a_m^c (1 \leq m \leq M)$ is a binary number. $a_m^c = 1$ means $z_c$ is associated with $a_m$, otherwise $z_c$ is not associated with $A_m$. Attributes are used as in between layer to decouple images from layer of class labels.

In this study, the probabilistic formulation of the direct attribute prediction (DAP) in [39] is adopted and revised for our attribute-based gait classification, as shown in Figure 5. SVM [26] is applied to learn a probabilistic classifier $\beta_m$ for each attribute $a_m$. $\beta_m$ is trained by using HSDs of gaits from all subject classes $c$ for which $a_m^c = 1$ as positive training samples and the rest as negative training samples. In this way, $\beta_m$ can provide a posterior probability $p(a_m|HSD)$ of that $a_m$ being present in HSD of a gait video.

$\beta_m$ models the relation between a HSD and an attribute $a_m$. Next is to consider the relation between an attribute $a_m$ and a subject class $z_c$, based on Bayes rule as:

$$ p(z_c|a_m) = \frac{p(a_m|z_c)p(z_c)}{p(a_m)} = \frac{a_m^c}{p(a_m)C} \quad (4) $$

![Figure 5: The direct attribute prediction (DAP) for attribute-based gait classification. $z_1, z_2, ..., z_C$ are $C$ subject classes. $a_1, a_2, ..., a_M$ are $M$ attributes. $a_m^c$ is an association between $z_c$ and $a_m$. $p(a_m|HSD)$ is a posterior probability of that $a_m$ being present in HSD of a gait video, which is estimated from a classifier for $a_m$.]

where $p(z_c) = 1/C$ by assuming identical class priors, and $p(a_m|z_c) = a_m^c$ which represents the association between an attribute $a_m$ and a subject class $z_c$. Then, the posterior probability of a subject class $c$ given a HSD can be calculated as:

$$ p(z_c|HSD) = \frac{\sum_{m=1}^{M} p(z_c|a_m)p(a_m|HSD)}{p(a_m)} \quad (5) $$

where $p(a_m) = 1/C$ by empirical means over all subject classes. Otherwise, $p(a_m)$ can be set to 0.5 [39].

Thus, the best output subject class label ($\hat{c}$) for HSD of a gait video is predicted as:

$$ \hat{c} = \arg\max_{1 \leq c \leq C} p(z_c|HSD) \quad (6) $$

where $p(z_c|HSD)$ is obtained from the equation (5).

The remaining challenge is to construct the associations ($a_m^c$) between subject classes and attributes. This will be explained in the following Section 4.2.

### 4.2. Attribute Representation

In this paper, attributes ($a_m$) are learned based on correlations between different subject classes [35]. A matrix $A$ is defined to represent the associations ($a_m^c$) between $M$ attributes and $C$ subject classes, as:

$$ A = [A_1, A_2, ..., A_C]^T = [B_1, B_2, ..., B_M] \quad (7) $$

where $A_c$ is a set of attributes representing a subject class $z_c$ such that $A_c = [a_1^c, a_2^c, ..., a_M^c] \in \{0, 1\}^{1 \times M}$,
Algorithm 1 Construction of the associations between $C$ subject classes and $M$ attributes

Input: $Q, S$
Output: The between-class attribute associations $A = \{a_m^c\}_{C \times M} \in \{0, 1\}^{C \times M}, 1 \leq c \leq C, 1 \leq m \leq M$

1: Initialize all elements in $A$ to be 0
2: while (max number of loops is not reached AND there is a change in $A$)
3:  
4:     for (each $z_c, 1 \leq c \leq C$)
5:         
6:             Set $a_m^c = 0, 1 \leq m \leq M$
7:             for ($i = 1; i \leq Q; i = i + 1$)
8:                 
9:                     $m = \text{maxsim}(A, S, c)$
10:                     $a_m^c = 1$
11:                     
12:     }
13: 
14: return $A$

Algorithm 2 $\text{maxsim}(\ldots)$

Input: $A, S, c$
Output: The index $m$ of the attribute $a_m$ such that $z_c$ is most correlated to other $z_1, 1 \leq c \leq C$ where $a_m^c = 1$

1: $\text{max.sim} = -1$
2: for (each $a_m$ such that $a_m^c = 0$)
3: 
4:     $s_m^c = \text{similarity}(A, S, c, m)$
5:     if (set $a_m^c = 1$ then $C1, C2$ are held AND $s_m^c > \text{max.sim}$)
6:         
7:             $\text{max.sim} = s_m^c, \text{max.attr} = m, \text{max.class} = c$
8:     
9: 
10: return $\text{max.attr}$

and $B_m$ is a set of classes representing an attribute $a_m$ such that $B_m = \{a_m^1, a_m^2, \ldots, a_m^C\} \in \{0, 1\}^{C \times 1}$.

To reduce the test-time complexity, $M$ must be less than $C$. $M$ can be selected empirically by trading-off between the test-time complexity and the accuracy.

Basically, subject classes that have highly related walking patterns should share some common attributes. Thus, the matrix $A$ will be determined based on the similarities between subject classes using the greedy strategy [50] under certain conditions. This process will be discussed below regarding three issues: 1) measurement of such similarities in Section 4.2.1; 2) conditions for the matrix $A$ in Section 4.2.2; 3) construction of the matrix $A$ in Section 4.2.3.

4.2.1. Correlations between different subject classes
A matrix $S$ is defined to represent similarities between subject classes. $S(i, j), 1 \leq i, j \leq C$, is the similarity of class $z_j$ to class $z_i$. It is measured from probability scores of that HSDs of subject class $z_j$ being predicted as belonging to subject class $z_i$.

The matrix $S$ will be obtained based on one-vs.-all classification in the training process, which is usually done offline beforehand. Given a training dataset, a classifier $\alpha_c$ is trained for each subject class $z_c$. That is, $\alpha_c$ can provide a posterior probability $p(z_i|HSD)$ of that HSD of a gait video being predicted as belonging to subject class $z_c$. $S(i, j)$ is computed as:

$$S(i, j) = \frac{\sum_{n=1}^{N} p(z_i|HSD_{n, j})}{N}$$

where $HSD_{n, j}$ is HSD of a gait video $n$ of $z_j$ on the training dataset.

4.2.2. Conditions for attribute representation
The attribute representation must be constructed under the following two conditions.

$C1$: No identical sets of attributes representing different subject classes (i.e., $A_i$ and $A_j$ are not identical for any $1 \leq i, j \leq C$ and $i \neq j$). In order to differentiate all subject classes, their attribute representations must be different.

$C2$: No identical sets of classes describing different attributes (i.e., $B_m$ and $B_n$ are not identical for any $1 \leq m, n \leq M$ and $m \neq n$). Otherwise, it will lead to duplicated classifiers which will increase test-time complexity without additional benefit.

4.2.3. Associations between subject classes and attributes
In this section, the associations between subject classes and attributes ($a_m^c$) are constructed using the similarity matrix $S$ based on the greedy strategy under the certain conditions (i.e., $C1, C2$), as shown in algorithms [1] [2] and [3]. Given $C$ subject classes, the number of attributes $M$ can be selected.
by trading-off between accuracy and efficiency (i.e. test-time complexity). The larger $M$ will lead to higher accuracy but lower efficiency. This will be verified in our experiments.

Given $C$ subject classes and $M$ attributes, the test-time speed-up is approximately $\frac{C}{M}$, and each subject class must be associated to at least $Q$ attributes where $(M, Q) = \frac{M}{\sqrt{(M-Q)^2}} \geq C$. For example, if $C = 124$ subjects and $M = 42$ attributes then the test-time speed-up $= \frac{124}{42} = 3$ and $Q \geq 2$.

Algorithm 1 presents the main function. It performs iteration till the maximum number of loops is reached or there is no more change in $A$. First, all elements in the matrix $A$ are initialized to be 0. That is, each subject class is initially not associated with any attributes. In each while loop, for each subject class $z_c$, the $Q$ attributes that are most associated to $z_c$ are determined and linked to $z_c$.

Algorithm 2 presents the function maxsim(...) to determine the strongest association between $z_c$ and any attributes $a_m$ under the conditions $C1, C2$. In the function, max.sim represents relative strength of the strongest association between attribute max.attr and subject class max.class. Among attributes $a_m$ which are currently not associated with subject class $z_c$, the association $a^c_m$ that has the maximum relative strength is determined.

Algorithm 3 presents the function similarity(...) to estimate the relative strength ($s^c_m$) of the association between subject class $z_c$ and attribute $a_m$. In this study, attribute $a_m$ is presented by the set of subject classes. Thus, $s^c_m$ can be equivalently calculated from correlations between class $z_c$ and all other classes that share the same attribute $a_m$.

5. Experiments

The CASIA gait database B [23] is used to evaluate the performance of the proposed method in Sections 5.1, 5.2 and 5.3. It contains 124 subjects from 11 views, namely 0°, 18°, 36°, 54°, 72°, 90°, 108°, 126°, 144°, 162°, and 180°. Under each view, ten gait sequences are captured for each person including six sequences in normal walking ($nm$), two sequences in walking when carrying a bag ($bg$), and two sequences in walking when wearing a coat ($cl$).

In our experiments, as mentioned, STIPs are detected at multiple scales based on a regular sampling of the scale parameters [13] [11]. The standard parameters are set as $h = 0.0005$, $\sigma^2 = \{4, 8\}$.

Figure 6: Examples of STIPs detected on a gait video from the CASIA gait database B [23].

$r^2 = \{2, 4\}$. In addition, for the gait feature extraction, the number of bins is coarsely set as $k = 100$. This is because each gait contains approximately 100 STIPs, based on our investigations. Also, the 4-bin HOG and 5-bin HOF are applied to a 3D video patch (i.e. with size $18\sigma \times 18\sigma \times 18\tau$) in the neighborhood of each detected STIP. The 3D patch is partitioned into a grid with $3 \times 3 \times 2$ spatio-temporal blocks.

Figure 6 shows examples of STIPs detected on a gait video from the CASIA gait database B. These STIPs are located where there are large variations along both spatial and temporal directions. Thus, they can be referred to human walking motions particularly on lower body parts i.e. legs.

The experiments are divided into two main scenarios. The first scenario is the gait recognition with no walking variations. That is, probe and gallery gait are recorded under the same walking conditions. The leave-one-out cross-validation is applied in this evaluation. The second scenario is the gait recognition with walking variations. That is, probe and gallery gaits are recorded under different walking conditions. For example, probe gait are walking when carrying a bag ($bg$) and gallery gaits are normal walking without carrying a bag ($nm$). In this evaluation, gaits under one walking condition are used as probe gaits and gaits under another walking condition are used as gallery gaits.

In addition, five more gait databases (namely the CASIA gait database A, the CASIA gait database C, the CMU MoBo database, the OU-ISIR gait database: treadmill dataset A and the OU-ISIR gait database: treadmill dataset B) are used to further validate the proposed method. The corresponding experimental results and comprehensive comparisons are shown in Sections 5.4, 5.5, 5.6, 5.7 and 5.8.

5.1. Comparisons with the Baseline Method under Three Variations on the CASIA Gait Database B

The proposed method (with SVM based on the standard one-vs.-all classification) is first compared
with the baseline method [23] which uses the most well known gait feature i.e. gait energy image (GEI) [11], under three variations including 1) viewing angle changes; 2) clothing changes; and 3) carrying condition changes.

As set up in [23], three sets of experiments (A, B, C) are designed to evaluate gait recognitions under three variations. Experiment Set A is for investigating the effects of views on the gait recognition performance and the algorithms’ robustness to view variation. Thus, nm under one view are used as probe gait and nm under another view are used as gallery gaits. Experiment Set B is for investigating the effects of clothes on the gait recognition performance and the algorithms’ robustness to clothing change. Thus, cl under one view are used as probe gait and nm under another view are used as gallery gaits.

From Table 1 and 2, many results have been observed. For the case of no variations i.e. walking under the same views without a bag and a coat, the proposed method performs comparable to the GEI-based method [23] for most views but slightly worse for frontal views. This is because the method in [23] uses the global shape information which is visible more clearly than the local motion information used in the proposed method, under frontal views. However, the proposed method is shown to be more robust to the variations on carrying conditions and clothing types because it does not rely on shape information which is significantly altered by these variations especially under side walks.

For the case of view variations, the proposed method performs better than the GEI-based method [23] for most views but comparable for frontal views. This is because view changes significantly alter available visual features for both the global shape information used in [23] and the local motion information used in the proposed method. For the cases of clothing and carrying variations, the proposed method significantly outperforms the GEI-based method [23] because GEI relies heavily on shape information which is significantly altered.

Table 1: Comparisons with the baseline method [23] on the Experiment Set A.

<table>
<thead>
<tr>
<th>Gallery</th>
<th>Probe</th>
<th>23 (%)</th>
<th>The proposed method (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0°</td>
<td>180°</td>
<td>37.9</td>
<td>57.5</td>
</tr>
<tr>
<td>0°</td>
<td>18°</td>
<td>99.2</td>
<td>97.6</td>
</tr>
<tr>
<td>0°</td>
<td>18°</td>
<td>18</td>
<td>31.9</td>
</tr>
<tr>
<td>18°</td>
<td>18°</td>
<td>8.9</td>
<td>20.3</td>
</tr>
<tr>
<td>18°</td>
<td>36°</td>
<td>22.3</td>
<td>86.1</td>
</tr>
<tr>
<td>36°</td>
<td>148°</td>
<td>79.6</td>
<td>90.7</td>
</tr>
<tr>
<td>54°</td>
<td>148°</td>
<td>79.6</td>
<td>90.7</td>
</tr>
<tr>
<td>72°</td>
<td>148°</td>
<td>79.6</td>
<td>90.7</td>
</tr>
<tr>
<td>126°</td>
<td>148°</td>
<td>79.6</td>
<td>90.7</td>
</tr>
<tr>
<td>144°</td>
<td>148°</td>
<td>79.6</td>
<td>90.7</td>
</tr>
<tr>
<td>162°</td>
<td>148°</td>
<td>79.6</td>
<td>90.7</td>
</tr>
<tr>
<td>180°</td>
<td>148°</td>
<td>79.6</td>
<td>90.7</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>39.1</td>
<td>64.1</td>
</tr>
</tbody>
</table>

Table 2: Comparisons with the baseline method [23] on the Experiment Set B.

<table>
<thead>
<tr>
<th>Gallery</th>
<th>Probe</th>
<th>23 (%)</th>
<th>The proposed method (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0°</td>
<td>180°</td>
<td>8.1</td>
<td>24.6</td>
</tr>
<tr>
<td>0°</td>
<td>18°</td>
<td>24.6</td>
<td>38.7</td>
</tr>
<tr>
<td>18°</td>
<td>18°</td>
<td>4.0</td>
<td>144</td>
</tr>
<tr>
<td>36°</td>
<td>18°</td>
<td>4.0</td>
<td>144</td>
</tr>
<tr>
<td>36°</td>
<td>36°</td>
<td>6.5</td>
<td>19.2</td>
</tr>
<tr>
<td>54°</td>
<td>36°</td>
<td>27.0</td>
<td>38.3</td>
</tr>
<tr>
<td>72°</td>
<td>36°</td>
<td>18.5</td>
<td>15.7</td>
</tr>
<tr>
<td>90°</td>
<td>36°</td>
<td>30.2</td>
<td>39.9</td>
</tr>
<tr>
<td>108°</td>
<td>36°</td>
<td>16.5</td>
<td>20.2</td>
</tr>
<tr>
<td>126°</td>
<td>36°</td>
<td>16.5</td>
<td>20.2</td>
</tr>
<tr>
<td>144°</td>
<td>36°</td>
<td>16.5</td>
<td>20.2</td>
</tr>
<tr>
<td>162°</td>
<td>36°</td>
<td>16.5</td>
<td>20.2</td>
</tr>
<tr>
<td>180°</td>
<td>36°</td>
<td>16.5</td>
<td>20.2</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>17.5</td>
<td>24.7</td>
</tr>
</tbody>
</table>
Table 3: Comparisons with the baseline method \[23\] on the Experiment Set C.

<table>
<thead>
<tr>
<th>Gallery</th>
<th>Probe</th>
<th>[23] (%)</th>
<th>The proposed method (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0(^\circ)</td>
<td>180(^\circ)</td>
<td>25.8</td>
<td>25.0</td>
</tr>
<tr>
<td>0(^\circ)</td>
<td>0(^\circ)</td>
<td>80.2</td>
<td>85.0</td>
</tr>
<tr>
<td>18(^\circ)</td>
<td>18(^\circ)</td>
<td>20.3</td>
<td>18.1</td>
</tr>
<tr>
<td>18(^\circ)</td>
<td>36(^\circ)</td>
<td>78.2</td>
<td>77.0</td>
</tr>
<tr>
<td>36(^\circ)</td>
<td>36(^\circ)</td>
<td>36.7</td>
<td>38.7</td>
</tr>
<tr>
<td>36(^\circ)</td>
<td>18(^\circ)</td>
<td>74.6</td>
<td>77.0</td>
</tr>
<tr>
<td>54(^\circ)</td>
<td>54(^\circ)</td>
<td>24.2</td>
<td>25.8</td>
</tr>
<tr>
<td>54(^\circ)</td>
<td>36(^\circ)</td>
<td>24.4</td>
<td>33.0</td>
</tr>
<tr>
<td>72(^\circ)</td>
<td>54(^\circ)</td>
<td>60.5</td>
<td>73.8</td>
</tr>
<tr>
<td>72(^\circ)</td>
<td>72(^\circ)</td>
<td>19.9</td>
<td>25.8</td>
</tr>
<tr>
<td>90(^\circ)</td>
<td>72(^\circ)</td>
<td>19.0</td>
<td>25.8</td>
</tr>
<tr>
<td>90(^\circ)</td>
<td>90(^\circ)</td>
<td>31.0</td>
<td>41.9</td>
</tr>
<tr>
<td>108(^\circ)</td>
<td>90(^\circ)</td>
<td>31.9</td>
<td>36.3</td>
</tr>
<tr>
<td>108(^\circ)</td>
<td>108(^\circ)</td>
<td>44.0</td>
<td>43.1</td>
</tr>
<tr>
<td>126(^\circ)</td>
<td>108(^\circ)</td>
<td>57.3</td>
<td>73.9</td>
</tr>
<tr>
<td>126(^\circ)</td>
<td>126(^\circ)</td>
<td>25.4</td>
<td>27.8</td>
</tr>
<tr>
<td>144(^\circ)</td>
<td>126(^\circ)</td>
<td>27.8</td>
<td>36.3</td>
</tr>
<tr>
<td>144(^\circ)</td>
<td>144(^\circ)</td>
<td>65.7</td>
<td>75.4</td>
</tr>
<tr>
<td>162(^\circ)</td>
<td>144(^\circ)</td>
<td>14.5</td>
<td>24.6</td>
</tr>
<tr>
<td>162(^\circ)</td>
<td>162(^\circ)</td>
<td>13.9</td>
<td>34.7</td>
</tr>
<tr>
<td>180(^\circ)</td>
<td>162(^\circ)</td>
<td>64.1</td>
<td>77.8</td>
</tr>
<tr>
<td>180(^\circ)</td>
<td>180(^\circ)</td>
<td>2.4</td>
<td>4.8</td>
</tr>
<tr>
<td>0(^\circ)</td>
<td>0(^\circ)</td>
<td>30.2</td>
<td>35.9</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>36.6</td>
<td>43.4</td>
</tr>
</tbody>
</table>

by changes of clothing types and carrying conditions.

5.2. Trade-off between Accuracy and Efficiency on the CASIA Gait Database B

In this section, the trade-off between accuracy and efficiency of the proposed method (with SVM based on attribute-based classification) is evaluated. The experiments have been carried out based on 5 different numbers of attributes (i.e. \( M = 124, 62, 42, 31, 25 \)) with 5 different speed-up of test-time respectively (i.e. speed-up = 1, 2, 3, 4, 5). Table 4 illustrates the trade-off between the test-time and the accuracy for six cases of (gallery, probe): \((nm, nm), (bg, bg), (cl, cl), (nm, bg), (nm, cl), \) and \((cl, bg)\), under side walk (90\(^\circ\)).

When \( M = 62 \), the highest accuracy is achieved and the test-time is reduced by approximate 2 times when compared to the standard one-vs.-all classification. The test-time can be further decreased but the accuracy will be also decreased accordingly.

Table 4: Performances (%) of the proposed method with SVM based on attribute-based classification.

<table>
<thead>
<tr>
<th>Gallery, Probe</th>
<th>M=124</th>
<th>M=62</th>
<th>M=42</th>
<th>M=31</th>
<th>M=25</th>
</tr>
</thead>
<tbody>
<tr>
<td>((nm, nm))</td>
<td>97.3</td>
<td>94.5</td>
<td>93.8</td>
<td>91.0</td>
<td>88.2</td>
</tr>
<tr>
<td>((bg, bg))</td>
<td>75.8</td>
<td>81.5</td>
<td>79.0</td>
<td>70.6</td>
<td>67.7</td>
</tr>
<tr>
<td>((cl, cl))</td>
<td>73.4</td>
<td>82.3</td>
<td>79.8</td>
<td>71.0</td>
<td>62.9</td>
</tr>
<tr>
<td>((nm, bg))</td>
<td>62.9</td>
<td>60.9</td>
<td>57.7</td>
<td>55.6</td>
<td>49.6</td>
</tr>
<tr>
<td>((nm, cl))</td>
<td>58.1</td>
<td>58.5</td>
<td>54.0</td>
<td>50.8</td>
<td>46.4</td>
</tr>
<tr>
<td>((bg, cl))</td>
<td>31.5</td>
<td>30.2</td>
<td>24.2</td>
<td>25.8</td>
<td>23.4</td>
</tr>
<tr>
<td>Average</td>
<td>66.5</td>
<td>68.0</td>
<td>65.8</td>
<td>60.8</td>
<td>56.4</td>
</tr>
</tbody>
</table>

5.3. Comparisons with Other Existing Methods under Changes of Clothing and Carrying Condition on the CASIA Gait Database B

The proposed method (with SVM based on attribute-based classification, \( M = 62 \)) is further compared with other existing methods in the literature, under changes of clothing and carrying condition. Table 5 shows the comparison results. The last column shows the average performances of the methods which reported all six results. It can be seen that the proposed method significantly outperforms the other existing methods.

For the case of no variation \((nm, nm)\), some other shape/appearance-based methods can perform better than the proposed method. This is because the global shape/appearance information contains more discriminant features than the local motion information used in the proposed method. For the case of small variations \((bg, bg)\) and \((cl, cl)\), the proposed method significantly performs better than all other existing methods in the literature. Also, for the case of large variations \((bg, nm), (cl, nm)\) and \((bg, cl)\), the proposed method can achieve the highest performance.

5.4. Comparisons with Other Existing Methods on the CASIA Gait Database A

The CASIA gait database A \[57\] includes 20 subjects from three views, namely frontal (0\(^\circ\)), oblique (45\(^\circ\)) and lateral (90\(^\circ\)) views. Each subject walks along a straight-line path back and forth twice. In total, there are 240 gait image sequences. The experimental results and comparisons are illustrated in Table 6. The last column shows the average performances of the methods which reported all three results. It can be seen that the proposed method outperforms the other existing methods in average performance.
Table 5: Comparisons with other existing methods under changes of clothing and carrying condition on the CASIA gait database B.

<table>
<thead>
<tr>
<th>Probe-Gallery</th>
<th>n:m-nm</th>
<th>lp-Ag</th>
<th>cl-cl</th>
<th>lp-nm</th>
<th>cl-nm</th>
<th>lp-cl</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>GE+2DLP [2]</td>
<td>95.2</td>
<td>55.7</td>
<td>44.4</td>
<td>n/a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GE1+2DLP [4]</td>
<td>95.7</td>
<td>48.8</td>
<td>48.4</td>
<td>n/a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The baseline method [2]</td>
<td>97.6</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GE1+SDA [2]</td>
<td>99.4</td>
<td>69.2</td>
<td>22.0</td>
<td>n/a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RF+PSS [2]</td>
<td>91.9</td>
<td>37.9</td>
<td>42.3</td>
<td>n/a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RF+PSS+SDA [2]</td>
<td>100.0</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RF+PSS+MDA [2]</td>
<td>99.6</td>
<td>46.0</td>
<td>53.0</td>
<td>n/a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M+ACDA [2]</td>
<td>99.6</td>
<td>54.9</td>
<td>33.0</td>
<td>n/a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LF1+AVG [2]</td>
<td>71.4</td>
<td>65.1</td>
<td>80.7</td>
<td>13.1</td>
<td>20.2</td>
<td>11.9</td>
<td>40.1</td>
</tr>
<tr>
<td>LF1+DTW [2]</td>
<td>61.9</td>
<td>17.9</td>
<td>8.9</td>
<td>21.4</td>
<td>25.0</td>
<td>22.6</td>
<td>24.8</td>
</tr>
<tr>
<td>LF1+HMM [2]</td>
<td>63.8</td>
<td>31.8</td>
<td>21.8</td>
<td>19.7</td>
<td>22.9</td>
<td>9.1</td>
<td>28.1</td>
</tr>
<tr>
<td>LF1+HMM [2]</td>
<td>94.0</td>
<td>64.2</td>
<td>57.1</td>
<td>45.2</td>
<td>42.9</td>
<td>22.6</td>
<td>54.3</td>
</tr>
<tr>
<td>GE1+PCA+LDA [5]</td>
<td>90.5</td>
<td>3.6</td>
<td>3.6</td>
<td>44.1</td>
<td>22.6</td>
<td>17.9</td>
<td>30.4</td>
</tr>
<tr>
<td>GPPW [5]</td>
<td>93.4</td>
<td>62.2</td>
<td>55.1</td>
<td>56.1</td>
<td>22.4</td>
<td>17.9</td>
<td>51.2</td>
</tr>
<tr>
<td>GRAd [5]</td>
<td>92.5</td>
<td>65.3</td>
<td>55.1</td>
<td>56.1</td>
<td>26.5</td>
<td>18.9</td>
<td>52.4</td>
</tr>
<tr>
<td>The proposed method</td>
<td>94.5</td>
<td>81.5</td>
<td>82.3</td>
<td>60.9</td>
<td>58.5</td>
<td>30.2</td>
<td>68.0</td>
</tr>
</tbody>
</table>

Table 6: Comparisons with other existing methods on the CASIA gait database A.

<table>
<thead>
<tr>
<th>Views</th>
<th>0°</th>
<th>45°</th>
<th>90°</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wavelet+TSVM [2]</td>
<td>82.9</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Curvelet+NN [2]</td>
<td>89.3</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Shape context [2]</td>
<td>91.2</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>FSA [2]</td>
<td>81.3</td>
<td>72.5</td>
<td>71.3</td>
<td>75.8</td>
</tr>
<tr>
<td>GE1 [1]</td>
<td>95.0</td>
<td>96.3</td>
<td>98.8</td>
<td>96.7</td>
</tr>
<tr>
<td>GE1+Face+Distance [2]</td>
<td>90.0</td>
<td>92.0</td>
<td>90.0</td>
<td>92.7</td>
</tr>
<tr>
<td>GE1+Face+Dynamic [2]</td>
<td>83.0</td>
<td>90.0</td>
<td>90.0</td>
<td>87.7</td>
</tr>
<tr>
<td>3D deformation [2]</td>
<td>91.0</td>
<td>97.5</td>
<td>100.0</td>
<td>98.2</td>
</tr>
<tr>
<td>2D polar-plane [2]</td>
<td>65.0</td>
<td>95.0</td>
<td>92.5</td>
<td>84.2</td>
</tr>
<tr>
<td>Neural network [2]</td>
<td>87.5</td>
<td>81.3</td>
<td>91.3</td>
<td>90.8</td>
</tr>
<tr>
<td>PSC-FSA [2]</td>
<td>93.6</td>
<td>96.3</td>
<td>97.5</td>
<td>96.3</td>
</tr>
<tr>
<td>Partial silhouette [2]</td>
<td>65.0</td>
<td>75.0</td>
<td>55.0</td>
<td>71.3</td>
</tr>
<tr>
<td>The proposed method</td>
<td>100.0</td>
<td>100.0</td>
<td>97.5</td>
<td>99.2</td>
</tr>
</tbody>
</table>

5.5 Comparisons with Other Existing Methods on the CASIA Gait Database C

The CASIA gait database C [68] contains 153 subjects. It was recorded under four walking conditions, namely normal walking (fn), slow walking (fs), fast walking (fq) and normal walking with a bag (fb). Ten gait videos were recorded for each subject (4 videos for fn, 2 videos for fs, 2 videos for fq, 2 videos for fb). The videos were all captured at night by infrared (thermal) cameras. The experimental results and comparisons are shown in Table 7. The last column shows the average performance of the methods by using four results of fn-fn, fs-fn, fq-fn and fb-fn.

In Table 7 it can be seen that the proposed method outperforms the other existing methods in average performance. For the cases of walking with speed changes (fs-fn, fq-fn), only HOG is used to compute a point description for each STIP. HOF is not used to compute such point description because it is sensitive to speed changes. More importantly, the proposed method achieves the outstanding performance under the change of a carrying condition (fb-fn). This is because carrying a small bag can alter available visual shape appearances but cannot significantly affect human walking patterns. The proposed method does not rely on shape appearances but on motion patterns. Thus, this can be a main reason why the proposed method can perform better than the appearance-based methods as shown in Table 7.

5.6. Comparisons with Other Existing Methods on the CMU MoBo Database

The CMU MoBo database [81] contains 25 subjects. Three walking conditions, namely slow walking i.e. 3.3 km/h (fs), fast walking i.e. 4.5 km/h (fq) and slow walking with a ball (fb) are used in this experiment. The experimental results and comparisons are illustrated in Table 8. In each row, the last column shows two values: 1) the average performance of the methods by using four results of fs-fs, fs-fq, fb-fb, fs-fq; 2) the average performance of the methods by using six results of fs-fs, fs-fq, fb-fb, fs-fq, fb-fq, fb-fs. As mentioned, for the cases of walking with speed changes (fs-fq, fb-fq), only HOG is used to compute a point description for each STIP. In Table 8 it can be seen that the proposed method outperforms the other existing methods in average performances. However, the proposed method is only comparable to the method in [16]. These two methods can achieve the 100 % accuracy for the cases of no walking variations (fs-fs, fq-fq, fb-fb). The proposed method performs better than the method in [16] for the cases of changes in carrying conditions (fb-fq, fb-fs), while the method in [16] performs...
better than the proposed method for the case of changes in walking speeds ($fs-fq$).

### 5.7. Comparisons with Other Existing Methods on the OU-ISIR Gait Database: Treadmill Dataset A (Speed Variation)

The OU-ISIR gait database: treadmill dataset A [82] contains six different walking speeds from 2 to 7 km/h with 1 km/h interval. A total of 34 subjects are used in this experiment. Two videos were recorded for each subject from each walking speed. To investigate and compare the recognition performance with other existing methods for the case of speed variations, the experimental setup as in [74] is used in this experiment. The walking with 5 km/h is used as the gallery and the walking with 4 km/h, 5 km/h, 6 km/h and 7 km/h are used as the probe. The experimental results and comparisons are shown in Table 9. It can be seen that the proposed method performs better than the other existing methods in the literature for the case of walking speed changes.

### 5.8. Comparisons with Other Existing Methods on the OU-ISIR Gait Database: Treadmill Dataset B (Clothes Variation)

The OU-ISIR gait database: treadmill dataset B [82] contains 68 subjects from side view with clothes variation up to 32 combinations. The probe and gallery datasets are pre-defined in the database. The experimental results and comparisons are illustrated in Table 10. It can be seen that the proposed method outperforms the other methods in the literature for the case of clothes variation.

### 6. Conclusion

This paper has proposed a new method of gait recognition. It extracts a new gait feature directly from a raw gait video without a foreground-background segmentation and/or a shape contour/boundary/skeleton extraction. The proposed...
gait feature extraction process is performed in the spatio-temporal domain. Then, SVM is applied for gait classification. The attribute-based learning is further applied to reduce the number of SVM models needed for recognizing each probe gait. It can significantly reduce the test-time computational complexity and also improve the recognition accuracy. When compared with other existing methods in the literature, the proposed method is shown to have the outstanding performances, especially for the cases of changes in clothing types and carrying conditions.

References

[27] B. Ye, Y. Wen, Gait recognition based on dct and svm, Int. Conf. on Wavelet Analysis and Pattern Recognition, China, November 2007, pp. 1382–1387.


M. Jeevan, N. Jain, H. Madasu, G. Chetty, Gait recognition based on gait pal and pal entropy image, IEEE Int. Conf. on Image Processing, Australia, September 2013, pp. 4195–4199.

K. Bashir, T. Xiang, S. Gong, Gait recognition using gait entropy image, Int. Conf. on Crime Detection and Prevention, United Kingdom, December 2009, pp. 1–6.


H. Lee, S. Hong, E. Kim, An efficient gait recognition based on a selective neural network ensemble, Int. Jour-

[66] W. Kusakunniran, Q. Wu, J. Zhang, H. Li, Pairwise shape configuration-based psa for gait recognition under small viewing angle change, IEEE Int. Conf. on Advanced Video and Signal-Based Surveillance, Austria, August 2011, pp. 17–22.


[68] D. Tan, K. Huang, S. Yu, T. Tan, Efficient night gait recognition based on template matching, Int. Conf. on Pattern Recognition, Hong Kong, August 2006, pp. 1000–1003.

[69] D. Tan, K. Huang, S. Yu, T. Tan, Recognizing night walkers based on one pseudoshape representation of gait, IEEE Int. Conf. on Computer Vision and Pattern Recognition, United States of America, June 2007, pp. 1–8.


[85] T. Whytock, A. Belyaev, N. Robertson, Gei + hog for action recognition, UK Computer Vision Student Workshop (BMVC 2012 Student Workshop), United Kingdom, September 2012.
