Recognizing Gaits on Spatio-Temporal Feature Domain

Worapan Kusakunniran, Member, IEEE

Abstract—Gait has been known as an effective biometric feature to identify a person at a distance, e.g. in video surveillance applications. Many methods have been proposed for gait recognitions from various different perspectives. It is found that these methods rely on appearance (e.g. shape contour, silhouette)-based analyses which require pre-processing of foreground/background segmentation (FG/BG). This process not only causes additional time complexity, but also adversely influences performances of gait analyses due to imperfections of existing FG/BG methods. Besides, appearance-based gait recognitions are sensitive to several variations and partial occlusions, e.g. caused by carrying a bag and varying a cloth type. To avoid these limitations, this paper proposes a new framework to construct a new gait feature directly from a raw video. The proposed gait feature extraction process is performed in the spatio-temporal domain. The Space-Time Interest Points (STIPs) are detected by considering large variations along both spatial and temporal directions in local spatio-temporal volumes of a raw gait video sequence. Thus, STIPs are allocated, where there are significant movements of human body in both space and time. A histogram of oriented gradients (HOG) and a histogram of optical flow (HOF) are computed on a 3D video patch in a neighborhood of each detected STIP, as a STIP descriptor. Then, the bag-of-words (BoW) model is applied on each set of STIP descriptors to construct a gait feature for representing and recognizing an individual gait. When compared with other existing methods in the literature, it has been shown that the performance of the proposed method is promising for the case of normal walking, and is outstanding for the case of partial occlusion caused by walking with carrying a bag and walking with varying a cloth type.

Index Terms—Gait recognition, human identification, spatio-temporal, HOG, HOF, STIP, BoW

1 INTRODUCTION

Human gait analysis has been applied to many areas, including biometrics, clinical trials, computer animations, and robotics [1]. From a surveillance perspective, gait recognition is an attractive modality. This is because it is capable of identifying humans at a distance by inspecting their walking manners. It can also be performed surreptitiously in an unconstrained environment. Gait is one of the few biometric features that can be measured remotely without physical contact and proximal sensing, which makes it useful in surveillance applications. Such applications play a decisive role in monitoring high security areas including banks, airports, military bases, and railway stations.

1.1 Related Works

A large number of gait recognition methods have been published recently, which can be roughly divided into two categories, model-based methods [2][3][4][5][6][7][8][9][10][11] and appearance-based methods [12][13][14][15][16][17][18][19][20][21][22][23][24][25][26][27][28]. These methods require a pre-processing of foreground/background segmentation (FG/BG) on a gait video, in order to extract shape contours, silhouettes, skeletons, or body joints for further gait analysis.

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. For example, Cunado et al. [3] considered legs as an interlinked pendulum. Then, a phase-weighted Fourier magnitude spectrum was used to recognize gait signatures which were derived from frequency components of the variations in the inclination of human thigh. Johnson and Bobick [4] used activity-specific static body parameters for gait recognition without directly analyzing dynamics of gait patterns.

Lee and Elgammal [2] introduced a framework for simultaneous gait tracking and recognition using person-dependent global shape deformations which were modeled using a non-linear generative model with kinematic manifold embedding and kernel mapping. Then, the generative model of global shape deformation was used to estimate shape style, geometric transformation, and body pose within the Bayesian framework. 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 appearance-based methods typically analyze gait sequences without explicit modeling of human body structure. The different methods in this category
have been developed from different perspectives. For example, BenAbdelkader et al. [26][27] proposed an eigengait method using image self-similarity plots. Chai et al. [28] introduced a Perceptual Shape Descriptor technique for recognizing gaits. Tan et al. [12] used eight kinds of projective features to describe human gait and PCA was applied for gait feature dimension reduction.

Han and Bhanu [13] proposed a concept of Gait Energy Image (GEI), and combined real and synthetic templates to improve the accuracy of gait recognition. Liu et al. [14] employed a population Hidden Markov Model (pHMM) to model human gaits and generated dynamics-normalized stance-frames to recognize individuals. Recently, Wang et al. [15] developed a temporal template, named Chrono-Gait Image (CGI), by encoding gait contours using a multi-channel mapping function. Roy et al. [16] 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 Framework of the Proposed Solution

As mentioned above, the existing methods of gait recognition require FG/BG which can lead to many drawbacks. Firstly, it causes additional time complexity. Secondly, gait recognition performance can be decreased due to low-quality/incomplete silhouettes caused by the imperfection of existing FG/BG methods on low-quality videos, dynamic backgrounds, etc. Thirdly, the appearance-based gait recognition methods are sensitive to walking variations and partial occlusions. For example, GEI [13], the most well-known gait feature, has been shown to be very efficient for recognizing gaits under normal walking conditions [24][29][30]. However, it has been reported to be very sensitive to variations caused by walking with carrying a bag and walking with varying a cloth type [24].

This paper proposes a new method to extract and recognize gait feature from a raw video sequence on a spatio-temporal feature domain, without any pre-processing on the video. Fig. 1 shows the framework of the proposed solution for gait recognition. In this figure, rectangles represent inputs/outputs, while ellipses represent processing steps.

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

Second, HOG and HOF are used to compute a descriptor of each STIP. They are applied on a 3D video patch (i.e. width × height × time) in a neighborhood of each detected STIP. A concatenation of HOG and HOF features are then used as a STIP descriptor. It well describes walking patterns around the interest point in space and time. Third, BoW is used to extract a gait feature by applying on the detected STIP descriptors in each gait video. Then, the simple but widely adopted Euclidean distance is used to measure the dissimilarity between any two gait features, and nearest neighbor (NN) is used as a classification method.

It can be seen that this proposed method does not require any pre-processing on a raw video. That is, it also does not rely on any foreground/background segmentation. This makes the proposed method more robust to partial occlusions caused by many real-world factors such as carrying a bag and varying a cloth type.

The rest of this paper is organized as follows.
Spatio-temporal interest points (STIPs) detection in a gait video is explained in section 2, followed by STIP descriptor describing a walking pattern in section 3. Gait feature extraction is proposed in section 4. Experimental results are shown in section 5, and conclusions are drawn in section 6.

### 2 SPATIO-TEMPORAL INTEREST POINTS (STIPs) DETECTION IN A GAIT VIDEO

In this section, STIPs \( \{S_n\}_{n=1}^N \) are detected from each gait video sequence \( G(x, y, t) \), where \( S_n \) is a detected STIP and \( G(x, y, t) \) is a gait video sequence with spatio-index \( x, y \) and temporal-index \( t \). STIP is detected by considering large variations along both spatial and temporal directions in local spatio-temporal volumes of a raw gait video sequence [32], in order to allocate significant motion patterns of human body in both space and time.

To detect local spatio-temporal features in \( G(x, y, t) \), its linear scale-space representation \( L(x, y, t; \sigma^2, \tau^2) \) 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:

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

where \( \ast \) is the convolution, and the Gaussian kernel is defined as:

\[
g(\cdot; \sigma^2, \tau^2) = \frac{1}{\sqrt{(2\pi)^3 \sigma^2 \tau^2}} e^{-\frac{x^2+y^2}{2\sigma^2} - \frac{t^2}{2\tau^2}}
\]

Then, we consider a spatio-temporal second moment matrix \( \mu(x, y, t; \sigma^2, \tau^2) \) within a Gaussian neighborhood of each point.

\[
\mu(\cdot; \sigma^2, \tau^2) = g(\cdot; s^2 \sigma^2, s^2 \tau^2) \ast \begin{bmatrix} L_x & L_y & L_t \\ L_x & L_y & L_t \\ L_x & L_y & L_t \end{bmatrix}
\]

where the first order derivatives \( (L_x, L_y, L_t) \) are defined as:

\[
L_x(\cdot; \sigma^2, \tau^2) = \frac{\partial}{\partial x} (g(\cdot; \sigma^2, \tau^2) \ast G(\cdot))
\]

\[
L_y(\cdot; \sigma^2, \tau^2) = \frac{\partial}{\partial y} (g(\cdot; \sigma^2, \tau^2) \ast G(\cdot))
\]

\[
L_t(\cdot; \sigma^2, \tau^2) = \frac{\partial}{\partial t} (g(\cdot; \sigma^2, \tau^2) \ast G(\cdot))
\]

The interest points are located in regions of \( G(\cdot) \) having significant eigenvalues \( \lambda_1, \lambda_2, \lambda_3 \) of \( \mu \) [32]. To detect STIPs, the Harris corner function [31] is extended from the spatial domain into the spatio-temporal domain. This is done by considering a combination of the determinant and trace of \( \mu \) as:

\[
H = det(\mu) - h \text{trace}^3(\mu)
\]

\[
= \lambda_1 \lambda_2 \lambda_3 - h (\lambda_1 + \lambda_2 + \lambda_3)^3
\]

For sufficiently large values of \( h \), positive local maxima of \( H \) correspond to points in \( G(\cdot) \) with high variations in both directions of space and time [32]. That is, STIPs of \( G(\cdot) \) are detected at local positive maxima in \( H \).

It can be seen that the spatio-temporal neighborhood in the interest point detection is defined by the scale parameters \( (\sigma, \tau) \) in the Gaussian kernel. They can be automatically adjusted to match the spatio-temporal extent of underlying image structures, as shown in [32]. 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 [32], this simplification has shown to produce similar (or better) results in applications (e.g. action recognition) while resulting in a considerable speed-up and close-to-video-rate run time [33][34].

Fig. 2 shows examples of interest points detected on a gait video from the CASIA gait database B [24] by using two different methods: 1) a spatio-temporal interest point (STIP) detection (i.e. used in this paper) where \( \sigma^2 = 4 \) and \( \tau^2 = 2 \); 2) a spatial interest point detection [31] where \( \sigma^2 = 4 \). It can be seen that the spatio-temporal interest points are significantly better than the spatial interest points to represent human motion.
walking patterns. This is because human gait is a periodic dynamic action [35] which contains large variations along both spatial and temporal directions. In contrast, the Harris detector [31] (i.e. in the bottom row of Fig. 2) selects both moving (i.e. gait) and stationary (i.e. background) points in the gait image sequence because it considers local variations in space domain only. Thus, without taking the temporal domain into account, there is a highly chance to faultily cover variations in background patterns.

3 STIP DESCRIPTOR DESCRIBING A WALKING PATTERN

This section explains how to compute a point descriptor for each STIP (Sn) detected in section 2. By using Sn(x, y, t) as a center point, HOG and HOF are applied on a 3D volume of its neighborhood in the linear scale-space representation L(x, y, t; σ^2, σ^2) (i.e. see section 2). This 3D volume is partitioned into a grid with N_x × N_y × N_t spatio-temporal blocks (\{B_i\}_{i=1}^N). Then, a–bin HOG and b–bin HOF descriptors are computed for all blocks and concatenated into a × N_x × N_y × N_t–element HOG descriptor and b × N_x × N_y × N_t–element HOF descriptor. Finally, the HOG and HOF descriptors are concatenated to generate the point descriptor with (a + b) × N_x × N_y × N_t elements for each STIP.

3.1 A histogram of image gradient (HOG)

HOG [36] is used in this study because local shape information such as gait is often well described by the distribution of local intensity gradients or edge directions. The first step of calculation is to compute a gradient value on each pixel in B_i. To do this, B_i is filtered to obtain x and y derivatives of pixels by using following spatial filter masks. Actually, the more complex spatial filter masks such as 3×3 by using following spatial filter masks. Actually, the

\[ B_i(x, y, t)|_{hog} = \sqrt{I_x^2(x, y, t) + I_y^2(x, y, t)} \]

\[ \angle_{hog} B_i(x, y, t) = \arctan \frac{I_y(x, y, t)}{I_x(x, y, t)} \]

The second step of calculation is to create cell histograms. The a orientation bins (\{O_j\}_{j=1}^a) are used for \([0^\circ, 360^\circ]\) interval. Thus, the a bins are defined as \(O_j = \frac{(j-1)360}{a}, \frac{j360}{a}\). For each pixel’s orientation \(\angle_{hog} B_i(x, y, t)\), the corresponding orientation bin is found and the orientation’s magnitude \(|B_i(x, y, t)|_{hog}\) is voted to this bin, as:

\[ O_j = O_j + |B_i(x, y, t)|_{hog} \]

where

\[ \angle_{hog} B_i(x, y, t) \in O_j : \frac{(j-1)360}{a}, \frac{j360}{a} \]

The last step is a process of histogram normalization. L2-norm normalization is applied on the histogram to generate the a–bin HOG descriptor as:

\[ \frac{\{O_j\}_{j=1}^a}{\sqrt{\sum_{j=1}^a O_j^2}} \]

3.2 A histogram of optical flow (HOF)

In this study, optical flow [38] is applied to extract motion from each pixel B_i(x, y, t) in each spatio-temporal block B_i. It is robust to cluttered backgrounds and variabilities in clothing. Then, b–bin HOF is used to describe B_i.

Assume that B_i(x, y, t) and B_i(x + ∆x, y + ∆y, t + ∆t) correspond to the same body point where ∆x and ∆y are movements in x and y spatial directions respectively along temporal difference ∆t, that is:

\[ B_i(x, y, t) = B_i(x + ∆x, y + ∆y, t + ∆t) \]

Provided ∆x, ∆y, ∆t are not too big, equation (12) can be re-written in a Taylor series expansion as [39]:

\[ B_i(x + ∆x, y + ∆y, t + ∆t) = B_i(x, y, t) + \frac{∂B_i}{∂x} ∆x + \frac{∂B_i}{∂y} ∆y + \frac{∂B_i}{∂t} ∆t + R(\cdot) \]

where \(R(\cdot)\) contains the higher-order terms which are small and can be ignored [38]. Equations (12) and (13) are combined to obtain:

\[ I_x v_x + I_y v_y + I_t = 0 \]

where \(I_x = \frac{∂B_i}{∂x}, I_y = \frac{∂B_i}{∂y}, I_t = \frac{∂B_i}{∂t}, v_x = ∆x = \frac{∂x}{∂t}, v_y = ∆y = \frac{∂y}{∂t}, ∆t = \frac{∂t}{∂t} = 1, v_x \) and \(v_y\) are x and y components of image velocity or optical flow. \(I_x\)
and \(I_o\) can be calculated as in equation (7). Similarly, \(I_t\) can be calculated as:

\[
I_t(x, y, t) = \sum_{j=1}^{1} j \times B_i(x, y, t + j)
\]

Equation (14) can be re-written more compactly as:

\[
\nabla I v^T = -I_t
\]

where \(\nabla I = (I_x, I_y)\) called a spatial intensity gradient, and \(v = (v_x, v_y)\) called an image velocity or optical flow. Then, optical flow \(v\) can be computed by using the 2D Lucas and Kanade method [40] or the 2D Horn and Schunck method [41].

The magnitude \(|B_i(x, y, t)|_{hof}\) and the orientation \(\angle_{hof}B_i(x, y, t)\) of the optical flow are computed as:

\[
|B_i(x, y, t)|_{hof} = \sqrt{v_x^2 + v_y^2}
\]

\[
\angle_{hof}B_i(x, y, t) = \arctan\left(\frac{v_y}{v_x}\right)
\]

The next step of calculation is to create cell histograms. The \(b\) orientation bins \(\{(F_j)_{j=1}^b\}\) are used for \([0^\circ, 360^\circ]\) interval. Thus, the \(b\) bins are defined as \(F_j : \frac{(j-1)360}{b}, \frac{j360}{b}\). For each pixel’s orientation \(\angle_{hof}B_i(x, y, t)\), the corresponding orientation bin is found and the orientation’s magnitude \(|B_i(x, y, t)|_{hof}\) is voted to this bin, as:

\[
F_j = F_j + |B_i(x, y, t)|_{hof}
\]

where

\[
\angle_{hof}B_i(x, y, t) \in F_j : \left[\frac{(j-1)360}{b}, \frac{j360}{b}\right)
\]

The last step is a process of histogram normalization. L2-norm normalization is applied on the histogram to generate the \(b\)-bin HOF descriptor as:

\[
\frac{\sum_{j=1}^{b} F_j}{\sum_{j=1}^{b} F_j^2}
\]

\section{Gait feature extraction}

Given a probe gait \(G_0\) and a gallery gait dataset \(\{G_{i}\}_{i=1}^{N}\), from sections 2 and 3, a set of STIP descriptors \(\{D_{j,n}\}_{n=1}^{N}\) is extracted from each gait video sequence \(G_{j}\), \(0 \leq j \leq N\), where \(D_{j,n}\) is the descriptor of the \(n^{th}\) STIP detected from \(G_j\) and \(N_j\) is the total number of STIPs detected from \(G_j\).

In this paper, the bag-of-words (BoW) model [42] is applied on each set of STIP descriptors to compute a gait feature. This process contains two main steps including: 1) histogram quantization and 2) histogram computation. In the histogram quantization step, the \(k\)-means clustering [43] is applied to all STIP descriptors \(D_{j,n}, 1 \leq j \leq N, 1 \leq n \leq N_j\), from the gallery dataset. It clusters these interest points into \(k\) groups, in which the center \(v_m\) of each group is used as the visual word to define each bin \(H_m\) of the histogram. To capture the behavior of quantization, the function class \(C\) is designed as follows.

\[
C = \{f(D_{j,n})|f(D_{j,n}) = \arg\min_{1 \leq m \leq k} d(D_{j,n}, v_m)\}
\]

where \(d\) is a chosen similarity measurement function e.g. Euclidean distance [44].

In the histogram computation step, for each probe gait \(G_0 : \{D_{0,n}\}_{n=1}^{N_0}\), its gait feature \(\{H_m\}_{m=1}^{k}\) is extracted by assigning each \(D_{0,n}\) to the corresponding bin and accumulating this bin’s value by one. Let \(H_m\) denotes the normalized number of STIPs in the gait video that are mapped to visual word \(v_m\).

\[
H_m = \frac{1}{N_0} \sum_{n=0}^{N_0} \delta(f(D_{0,n}), m)
\]

where \(\delta(a, b) = 1\) if \(a = b\), otherwise \(\delta(a, b) = 0\).
TABLE 2
Comparisons with the baseline method [24] on the Experiment Set B.

<table>
<thead>
<tr>
<th>(Gallery, Probe)</th>
<th>[24] (%)</th>
<th>The proposed method (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0°, 180°)</td>
<td>8.1</td>
<td>2.4</td>
</tr>
<tr>
<td>(0°, 0°)</td>
<td>24.6</td>
<td>26.2</td>
</tr>
<tr>
<td>(0°, 18°)</td>
<td>6.9</td>
<td>5.6</td>
</tr>
<tr>
<td>(18°, 0°)</td>
<td>4.4</td>
<td>6.5</td>
</tr>
<tr>
<td>(18°, 18°)</td>
<td>27.0</td>
<td>29.0</td>
</tr>
<tr>
<td>(18°, 36°)</td>
<td>18.5</td>
<td>15.7</td>
</tr>
<tr>
<td>(36°, 18°)</td>
<td>8.5</td>
<td>8.1</td>
</tr>
<tr>
<td>(36°, 36°)</td>
<td>30.2</td>
<td>34.3</td>
</tr>
<tr>
<td>(36°, 54°)</td>
<td>16.5</td>
<td>16.9</td>
</tr>
<tr>
<td>(54°, 36°)</td>
<td>10.1</td>
<td>13.3</td>
</tr>
<tr>
<td>(54°, 54°)</td>
<td>30.6</td>
<td>36.7</td>
</tr>
<tr>
<td>(54°, 72°)</td>
<td>5.6</td>
<td>13.3</td>
</tr>
<tr>
<td>(72°, 54°)</td>
<td>7.7</td>
<td>14.9</td>
</tr>
<tr>
<td>(72°, 72°)</td>
<td>31.0</td>
<td>42.7</td>
</tr>
<tr>
<td>(72°, 90°)</td>
<td>21.8</td>
<td>26.2</td>
</tr>
<tr>
<td>(90°, 72°)</td>
<td>20.6</td>
<td>33.1</td>
</tr>
<tr>
<td>(90°, 90°)</td>
<td>32.7</td>
<td>52.0</td>
</tr>
<tr>
<td>(90°, 108°)</td>
<td>16.5</td>
<td>21.4</td>
</tr>
<tr>
<td>(108°, 90°)</td>
<td>27.8</td>
<td>32.7</td>
</tr>
<tr>
<td>(108°, 108°)</td>
<td>30.2</td>
<td>39.1</td>
</tr>
<tr>
<td>(108°, 126°)</td>
<td>9.3</td>
<td>18.5</td>
</tr>
<tr>
<td>(126°, 108°)</td>
<td>18.5</td>
<td>21.8</td>
</tr>
<tr>
<td>(126°, 126°)</td>
<td>26.2</td>
<td>39.9</td>
</tr>
<tr>
<td>(126°, 144°)</td>
<td>8.9</td>
<td>15.3</td>
</tr>
<tr>
<td>(144°, 126°)</td>
<td>18.1</td>
<td>19.8</td>
</tr>
<tr>
<td>(144°, 144°)</td>
<td>30.6</td>
<td>33.5</td>
</tr>
<tr>
<td>(144°, 162°)</td>
<td>1.2</td>
<td>2.4</td>
</tr>
<tr>
<td>(162°, 144°)</td>
<td>4.0</td>
<td>5.0</td>
</tr>
<tr>
<td>(162°, 162°)</td>
<td>27.0</td>
<td>31.5</td>
</tr>
<tr>
<td>(162°, 180°)</td>
<td>6.5</td>
<td>7.5</td>
</tr>
<tr>
<td>(180°, 162°)</td>
<td>12.5</td>
<td>11.2</td>
</tr>
<tr>
<td>(180°, 180°)</td>
<td>27.4</td>
<td>23.0</td>
</tr>
<tr>
<td>(180°, 0°)</td>
<td>9.3</td>
<td>8.1</td>
</tr>
<tr>
<td>Average</td>
<td>17.5</td>
<td>21.4</td>
</tr>
</tbody>
</table>

Then, the simple but widely adopted Euclidean distance \(d\) is used to measure the gait dissimilarity \(d\) as follows:

\[
d(H^1, H^2) = \sqrt{\sum_{m=0}^{k} (H^1_m - H^2_m)^2}
\]

where \(H^1\) and \(H^2\) are gait features of \(G_1\) and \(G_2\) respectively, and the smaller value of \(d\), the more possibility that the two gaits belong to the same subject.

5 Experiments

The CASIA gait database B [24] is used in our experiments. 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).

![Sample gait images from the CASIA gait database B.](image)

Fig. 3. Sample gait images from the CASIA gait database B. The first row shows gait under various views (0°, 36°, 108°). The second row shows gait of normal walking (nm), walking when carrying a bag (bg), and walking when wearing a coat (cl), under 90°.

Copyright (c) 2014 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.
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 the experiments, as mentioned, STIPs are detected at multiple scales based on a regular sampling of the scale parameters [33][34]. The standard parameters are set as $h = 0.0005$, $a^2 = \{4, 8\}$, $r^2 = \{2, 4\}$. 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.

The experiments are divided into two main scenarios. The first scenario is the gait recognition with no walking variations. That is, probe and gallery gaits 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 gaits 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.

The proposed method is first compared with the baseline method [24] which uses the most well known gait feature i.e. gait energy image (GEI) [13], under three variations including 1) viewing angle changes; 2) clothing changes; and 3) carrying condition changes. Then, the proposed method is compared with other existing methods for six cases of (gallery, probe): ($nm, nm$), ($bg, bg$), ($cl, cl$), ($nm, bg$), ($nm, cl$), and ($cl, bg$), under side walk ($90^\circ$).

### 5.1 Comparisons with the baseline method under three variations

As set up in [24], 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, normal walking ($nm$) under one view are used as probe gaits and normal walking ($nm$) 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, walking when wearing a coat ($cl$) under one view are used as probe gaits and normal walking ($nm$) under another view are used as gallery gaits. Experiment Set $C$ is for investigating the effects of carrying condition on the gait recognition performance and the algorithms' robustness to carrying condition change. Thus, walking when carrying a bag ($bg$) under one view are used as probe gaits and normal walking ($nm$) under another view are used as gallery gaits.

The gallery sets of Experiment Set $B$ and $C$ are the same with $A$, but the probe sets are different. All the sequences of walking with a coat are put into the probe set of Experiment Set $B$, and all those of walking with a bag are put into the probe set of Experiment Set $C$. These three experiment sets are used to evaluate the algorithms' robustness to view changes, clothing changes, and carrying condition changes.

From Table 1, 2, and 3, many key aspects have been investigated. 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 [24] for most views but slightly

<table>
<thead>
<tr>
<th>Probe-Gallery</th>
<th>$nm-nm$</th>
<th>$bg-bg$</th>
<th>$cl-cl$</th>
<th>$bg-nm$</th>
<th>$cl-nm$</th>
<th>$bg-cl$</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>GEI+2DLPP [18]</td>
<td>95.5</td>
<td>-</td>
<td>-</td>
<td>55.7</td>
<td>44.4</td>
<td>-</td>
<td>n/a</td>
</tr>
<tr>
<td>EGEI+2DLPP [18]</td>
<td>95.7</td>
<td>-</td>
<td>-</td>
<td>48.8</td>
<td>48.4</td>
<td>-</td>
<td>n/a</td>
</tr>
<tr>
<td>The baseline method [24]</td>
<td>97.6</td>
<td>-</td>
<td>-</td>
<td>52.0</td>
<td>32.7</td>
<td>-</td>
<td>n/a</td>
</tr>
<tr>
<td>GEI+CDA [45]</td>
<td>99.4</td>
<td>-</td>
<td>-</td>
<td>60.2</td>
<td>22.0</td>
<td>-</td>
<td>n/a</td>
</tr>
<tr>
<td>RF+FSS [46]</td>
<td>91.9</td>
<td>-</td>
<td>-</td>
<td>37.9</td>
<td>42.3</td>
<td>-</td>
<td>n/a</td>
</tr>
<tr>
<td>RF+FSS+CDA [46]</td>
<td>100.0</td>
<td>-</td>
<td>-</td>
<td>50.0</td>
<td>33.1</td>
<td>-</td>
<td>n/a</td>
</tr>
<tr>
<td>RF+FSS+MDA [46]</td>
<td>99.6</td>
<td>-</td>
<td>-</td>
<td>46.0</td>
<td>33.9</td>
<td>-</td>
<td>n/a</td>
</tr>
<tr>
<td>$M_{j}+ACDA$ [47]</td>
<td>99.6</td>
<td>-</td>
<td>-</td>
<td>54.9</td>
<td>33.0</td>
<td>-</td>
<td>n/a</td>
</tr>
<tr>
<td>LF+AVG [25]</td>
<td>71.4</td>
<td>63.1</td>
<td>60.7</td>
<td>13.1</td>
<td>20.2</td>
<td>11.9</td>
<td>40.1</td>
</tr>
<tr>
<td>LF+DTW [25]</td>
<td>61.9</td>
<td>17.9</td>
<td>0.0</td>
<td>21.4</td>
<td>25.0</td>
<td>22.6</td>
<td>24.8</td>
</tr>
<tr>
<td>LF+oHMM [25]</td>
<td>63.8</td>
<td>31.8</td>
<td>21.4</td>
<td>19.7</td>
<td>22.6</td>
<td>9.1</td>
<td>28.1</td>
</tr>
<tr>
<td>LF+iHMM [25]</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>GEI+PCA+LDA [48]</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>GPPE [49]</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>GEnI [50]</td>
<td>92.3</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>95.4</td>
<td>73.0</td>
<td>70.6</td>
<td>60.9</td>
<td>52.0</td>
<td>29.8</td>
<td>63.6</td>
</tr>
</tbody>
</table>
worse for frontal views. This is because the method in [24] uses global shape information which is visible more clearly than 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 [24] for most views but comparable for frontal views. This is because view changes significantly alter available visual features for both global shape information used in [24] and local motion information used in the proposed method. For the case of clothing variations (see Table 2), the proposed method significantly outperforms the GEI-based method [24] because GEI relies heavily on shape information which is significantly altered by cloth type changes. For the case of carrying variations (see Table 3), the proposed method performs slightly better than the GEI-based method [24]. This is because, in the CASIA gait database B, a bag in the carrying condition is only a small bag which does not affect a visual walking pattern significantly.

5.2 Comparisons with other existing methods under changes of clothing and carrying condition

The proposed method is further compared with other existing methods in the literature, under changes of clothing and carrying condition. Table 4 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 \((bg-bg, cl-cl)\), the proposed method significantly performs better than all other existing methods in the literature. Also, for the case of large variations \((bg-nm, cl-nm, bg-cl)\), the proposed method can achieve the best performance.

6 Conclusion

This paper has proposed a new method for gait recognition. It constructs a new gait feature directly from a raw video without a pre-processing of foreground-background segmentation. The proposed gait feature is extracted in the spatio-temporal domain. The Space-Time Interest Points (STIPs) are detected from a raw gait video sequence. They represent significant movements of human body along both spatial and temporal directions. Then, HOG and HOF are used to describe each detected STIP.

Finally, a gait feature is constructed by applying BoW on a set of HOG/HOF-based STIP descriptors from each gait sequence. It can be seen that the proposed gait feature relies on local motion information which is more robust to walking variations than global shape information used in most of existing methods for recognizing gaits. In this paper, when compared to the other methods in the literature, the proposed method has been reported to achieve the promising performance for the case of no variation and to achieve the significantly better performance for the case of large variations caused by clothing and carrying condition changes.

References


Worapan Kusakunniran (M’10) received the B.Eng. degree in computer engineering from the University of New South Wales (UNSW), Sydney, Australia, in 2008, and the Ph.D. degree in computer science and engineering from UNSW, in cooperation with the Neville Roach Laboratory, National ICT Australia, Kensington, Australia, in 2013.

He is currently a Lecturer with the Faculty of Information and Communication Technology, Mahidol University, Nakhon Pathom, Thailand. He is the author of several papers in top international conferences and journals. His current research interests include biometrics, pattern recognition, image processing, computer vision, multimedia, and machine learning.

Dr. Kusakunniran served as a Program Committee Member for the ACCV2014 Workshop on Human Gait and Action Analysis in the Wild: Challenges and Applications, and the IEEE Workshop on the Applications of Computer Vision (WACV) 2013. He has also served as a Reviewer for several international conferences and journals, such as the International Conference on Pattern Recognition (ICPR), the IEEE International Conference on Image Processing (ICIP), the IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS), the Pattern Recognition (PR), the IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics (TSMCB), the IEEE Transactions on Image Processing (TIP), the IEEE Transactions on Information Forensics and Security (TIFS), and the IEEE Signal Processing Letters (SPL). He was a recipient of ICPR Best Biometric Student Paper Award in 2010.