A Bag-of-Gait Model for Gait Recognition

Jianzhao Qin, T. Luo, W. Shao, R. H. Y. Chung and K. P. Chow

The Department of Computer Science, The University of Hong Kong, Hong Kong, China

Abstract—In this paper, we propose a novel gait recognition method based on a bag-of-gait model. In the proposed method, the image sequence of a walking person is encoded by a codebook consisting of a list of code words denoting different walking stages; then, this image sequence is represented by a feature vector denoting the existence of the code words, which is further used for classification. Unlike most of previous gait recognition methods, there is no need to estimate gait period in the proposed method. Moreover, the proposed method is capable of recognizing the gait when the observed gait period is incomplete which caused by occlusion or short appearance time. The method is evaluated on a dataset consisting of 151 subjects with four different walking conditions using four-fold cross-validation. The result shows the effectiveness of the proposed method and the state-of-art result is achieved on that dataset.

Keywords: gait recognition; bag-of-gait; biometric; human identification; video surveillance

1. Introduction

In order to enhance the public safety, identifying individuals effectively and reliably in an unobtrusive manner through the video surveillance system has attracted considerable attention in recent years [1]. Unlike the biometric systems based on fingerprint, palm and iris, which require physical contact and subject’s cooperation, the visual information captured by video surveillance system can be utilized without the subject’s cooperation and attention [2]. Gait is one of the most important biometrics. The Psychology experimental results have suggested that humans are capable of identifying individuals solely based on gait [3]. Taking gait as a biometric offers potential for identifying human at low resolution when the subject only consists of few image pixels [4].

The previous proposed methods for gait recognition can be mainly divided into two categories: model-based methods and model-free methods. The basic idea of the model-based methods is that, a structural model of human gait is first generated with parameters describing the position or angle between different human parts; then, given an observed gait video, these parameters are estimated from the video which are further used for classification. In [5], a skeleton model was used to describe the human beings, and four angles from this model were used for recognition. Cunado et.al. [6] modeled the movement of human legs using two pendulums, then the Fourier coefficients were used for classification. In [7], Lee and Grimson modeled the side view of walking people by seven ellipses which represent head/shoulder region, front of torso, back of torso, front thigh, back thigh, front foot and back foot respectively; then, the centroid, aspect ratio of major and minor axis of these ellipses and the orientation of major axis were used to form the feature vector. Wagg and Nixon [8] modeled the human body using two ellipses for the torso and the head, and four line segments for each leg and a rectangle for each foot; then, 45 parameters from the model, e.g. Lower knee width, ankle width and upper knee width, and 18 static parameters, e.g. gait frequency were calculated for classification. Generally speaking, the model based methods are more robust to some variations, such as clothing and carrying condition. But, they are more computational expensive and demands accurate detection of human parts. For the model-free methods, the features are directly extracted from the silhouettes. Kale et.al. [9] proposed a model-free method using the outermost widths of binary silhouettes or the binary silhouettes directly. In their method, the gait cycle of each sequence is estimated firstly; then, for each cycle, the images are divided into several segments; next, the features of the images belonging to different segments and different subjects are clustered to form the status of hidden Markov model; finally, the hidden Markov model is trained for classification. In [10], the similarity of two gaits were calculated by directly comparing two binary silhouettes sequences. Han and Bhanu [11] proposed using the average values of the binary silhouettes (Gait Energy Image (GEI)) from a gait period as the feature to describe human gait property. Tan et.al. [12] proposed to characterize gait signature by the head-torso-thigh part of human silhouettes via thermal infrared imaging. Yu et.al. [13] proposed describing the human gait using the contour of silhouettes. The similarity between the contours is compared using an improved dynamic warping algorithm. In [14], Tan et.al. proposed a pseudoshape model to represent human gait. The proposed pseudoshape is calculated by projecting the silhouette to horizontal direction and vertical direction respectively; then, after normalization, these two projections are concatenated to form the feature vector. Zhang et al. [15] proposed an active energy image (AEI) to represent the gait feature. Unlike the method based on GEI [11], the active energy image is formed by accumulating the difference image between two consecutive silhouette images, which may reduce the variation of the silhouette causing by different clothing and carrying condition. The model-free methods are less computational expensive but are more
sensitive to some variations, such as clothing and carrying condition.

In this paper, we propose a novel gait recognition method based on the bag-of-gait model. In the proposed method, the silhouette sequence of a walking people is encoded by a codebook consisting of a list of code words representing different walking stages and corresponding variations. Then, the silhouette sequence is described by a feature vector which represents the existence of the code words. After that, support vector machine (SVM) classifier is trained to do the classification. In order to handle the situation when the gait cycle is incomplete, we also propose a ‘counting’ classifier, which is more computational efficient and more capable to handle incomplete gait cycle. Unlike most previous gait recognition methods, in the proposed method, there is no need to estimate gait period. Moreover, the proposed method is capable of recognizing complete gait sequence when occlusion is occurred or the observed gait cycle is incomplete.

In the work of Kale et al. [9], a codebook also is created to represent the hidden states of a hidden Markov model. However, before the hidden states creation, the gait cycle of the image sequence has to be estimated; and the images in each cycle have to be manually aligned to keep the images have the same states order, which is a very time-consuming procedure. Moreover, since the gait is modeled by a hidden Markov model in [9] which represents the transitions between all the hidden states, their method cannot handle the situation when the unknown gait sequence is incomplete. Roy et.al. [16] proposed a gait recognition method using pose kinematics and pose energy image. In their method, key poses are extracted using the K-means algorithm, which is similar to our codebook creation procedure. However, the gait image sequences are represented by the frequencies of the key poses during a gait cycle, which makes the method invalid or vulnerable to the condition when the walking cycle is incomplete.

The proposed method is evaluated on a dataset consisting of 151 subjects with four different walking conditions, i.e. normal walking, walking at low speed, walking at high speed and walking with bag [14] . The result shows the effectiveness of the proposed method and the state-of-art performance is achieved on that dataset. We also simulate the incomplete gait cycle situation by randomly select a number of silhouette from a complete gait cycle from the dataset. By using only 17% silhouettes (4 from 25) of a complete walking period, the proposed method can reach 90% accuracy rate, which illustrates the effectiveness of the proposed method in handling the situation when the observed gait period is incomplete.

The rest of this paper is organized as follows. In Section 2, we introduce the proposed method for gait recognition including silhouette feature extraction, gait codebook generation, the bag-of-gait model forming procedure and classifier training. Section 3 represents the experimental results. This paper is concluded in Section 4.

2. Proposed method

2.1 Framework

Figure 1 illustrates the framework of the proposed method. In the training stage, the silhouettes are firstly extracted from the training sequences and normalized to a predefined size; secondly, features which are used to describe the characteristic of the silhouette are extracted from the normalized silhouettes; thirdly, category-specific clustering is performed to create a gait codebook which consists of a list of code words describing different walking stages; fourthly, based on the codebook, the training sequences are represented by a bag-of-gait model; finally, a classifier is trained for classifying unknown sequence. In the test stage, given an unknown sequence, following the similar procedure, this sequence is represented by the bag-of-gait model, which is further input to the classifier for giving the final classification result.

2.2 Silhouette feature extraction

We directly use the silhouette image to form the feature vector. For a given image, its foreground image can be obtained using foreground extraction algorithms, such as Gaussian Mixture Modeling [17]. Figure 2 depicts a sample of foreground image. Then, a minimum rectangle region which contains all the foreground pixels is extracted from the foreground image. After that, this rectangle region is normalized to form the silhouette image. Figure 3 shows the samples of silhouette images. Taking the white pixel as 1 and black pixel as 0, the feature vector is formed by concatenating the silhouette image in a row-wise manner, i.e., \( f = [r_1; r_2; \ldots; r_{nrow}] \) where \( r_i \) is the vector representing the values of the \( i \)th row of the silhouette image.

2.3 Gait codebook generation

We employ K-means algorithm to generate the codebook of gait from the silhouette images extracted from the training set. Denote the code words in the codebook as \( c_1, c_2, c_3, \ldots, c_k \) where \( k \) is the number of code words, and denote the features extracted from silhouette as \( f_1, f_2, f_3, \ldots, f_n \) where \( n \) is the total number of features. The code words are set by solving the following optimization problem: \( \min_{c_i} \sum_{f_j \in c_i} ||f_j - c_i|| \). The codebook generation is performed in a category-specific manner, i.e. the clustering is performed on the silhouette features extracted from the silhouettes belonging to a single subject; the created code words for different subjects are concatenated to form the final codebook. This category-specific codebook generation is capable of enhancing the discriminative ability of the generated code words.
2.4 Gait codebook generation

Given a silhouette sequence $S$ consisting of a set of silhouettes $s_1, s_2, s_3, \ldots, s_m$. The bag-of-gait representation $x$ is formed by the following steps:

Step 1: Extract the features $f_1, f_2, f_3, \ldots, f_m$ from the silhouettes.

Step 2: Fill all the elements of $x$ with 0.

Step 3: For the $i$th feature $f_i$, calculate its Euclidean distance with the code words in the codebook by Eq.1:

$$d_j = ||f_i - c_j||, j = 1, \ldots, k$$

Step 4: Repeat step 3 until all the features $f_i, i = 1, 2, \ldots, m$ are handled.

The bag-of-gait representation represents the existences of the code words in current sequence. It can be considered as a generalization of the GEI feature to different walking stages. Comparing with other gait features, the advantages of the bag-of-gait representation are that (1) unlike the extraction of many other gait features, there is no need to estimate the period of gait in the proposed method; and (2) The proposed method can handle the situation when the observed gait cycle is incomplete (e.g., the subject only appears for a very short time or the subject is occluded frequently during walking. For the GEI and AEI features, the features are calculated by averaging the silhouette images over a complete walking cycle. If the walking cycle is incomplete, it will distort the extracted features, which results in degraded performance.

2.5 Classifier

In this sub-section, we introduce two classifiers, the SVM classifier and a ‘counting’ classifier which makes the classification based on the number of existence code words in the proposed method.

2.5.1 SVM classifier

In the training process, images in the training set can be represented as a set of $k$ dimensional features, $\{x_1, x_2, x_3, \ldots, x_L\}$, where $L$ denotes the number of training images. To formulate it into a SVM classifier, given a training set of labeled data:

$$\{(x_i, y_i) | (x_i, y_i) \in \mathbb{R}^n \times \{\pm1\}, i = 1, \cdots, L\}$$

where $x_1, x_2, \cdots, x_L$ are the dimensional features that have been labeled as $y_1, y_2, \ldots, y_L$, the training of SVM classifier
with linear kernel can be formulated as the following optimization problem for a 2-class classification problem:

$$\min_{w, \eta} P \sum_{i=1}^{L} \eta_i + \frac{1}{2} \|w\|_2^2$$

s.t.  

$$y_i w^T [x_i^T \ 1]^T + \eta_i \geq 1, \quad \eta_i \geq 0, i = 1, \ldots, L$$  \hspace{1cm} (3)

where $P > 0$ is a penalty parameter ($P = 1$ in this paper), and $\eta_i$ is the slack variable that represents the classification error of $x_i$. In the classification stage, given a feature vector $x_i$, the sign of $w^T [x_i^T \ 1]^T$ determines the class of this vector. The 2-class SVM classifier can be extended to a multi-class scenario using the one-against-all method, that is, we take the labels of samples from one class as $1$ and the labels of samples from other classes as $-1$, then solve the above optimization problem for $C$ (the number of classes) times. From the formulation of the SVM classifier, we can see that the absolute value of the elements of $w$ determines the importance of the corresponding code word for classification.

### 2.5.2 ‘Counting’ classifier

For the ‘counting’ classifier, the class of the unknown gait sequence is set as the class which has the maximum number of code words which has been selected to represent the input images. The predicted class $y_{pred}$ can be selected by the following equation:

$$y_{pred} = \max_y \sum_{\text{label}(x_i) \in y} x_i$$  \hspace{1cm} (4)

Since only addition operation is needed for the ‘counting’ classifier, it can be efficiently calculated by a digital computer. Furthermore, the classification of the ‘counting’ classifier does not depend on the training data. Therefore, training procedure is not needed. This property of the classifier brings us another advantage: it is more robust to the situation when the given walking period is incomplete. Because, if the given walking period is incomplete, the similarity between the feature vector of test sample and the feature vectors of training samples, which have the same category, may decrease. It will deteriorate the performance of the classifiers (e.g., distance classifier, SVM classifier, neural network) training from training samples. The classification procedure of the ‘counting’ classifier, however, does not depend on the similarity measurement between the test sample and training samples.

### 3. Experimental result

The performance of the proposed method is evaluated on the CASIA Infrared Night Gait Dataset [14]. The foreground images of the dataset can be directly downloaded from "http://www.cbsr.ia.ac.cn/english/Gait%20Databases.asp". This dataset contains 153 subjects (Only 151 subjects’ silhouettes are usable after unzipping the downloaded file.). Four walking conditions, i.e. normal walking, slow walking, fast walking, and normal walking with a bag, are included in this dataset in order to test the robustness of the gait recognition algorithm under different walking speed and carrying conditions. Each subject contains four "normal walking" sequences, two "slow walking" sequences, two "fast walking" sequences and two "walking with a bag" sequences. Figure 4 depicts several samples of foreground images from the CASIA Infrared Night Gait Dataset. We follow the same experiment setting as in [14]. In experiment setting A, the training set consists of three "normal walking" sequences for each subject and the test set consists of one remaining "normal walking" sequence for each subject. Four-fold cross-validation is performed, i.e., the experiment is repeated four times with different combination of training and test sequences. The average accuracy rate from the four-fold cross-validation is reported. In experiment setting B, the training set consists of three "normal walking" sequences for each subject; and the test set consists of two "fast walking" sequences for each subject. The experiment is also repeated four times for different combination of training sequences. In experiment setting C and D, the training set setting is the same as in experiment setting A and B. The test set consists of two "slow walking" sequences and two "normal walking" sequences for experiment setting C and D respectively. We train 8 code words for each subject in the dataset. That is, the final gait codebook consists of 1208 code words in total. Table 1 summaries the experiment settings. Table 2 compares the proposed method with recent gait recognition methods which report results on this dataset including the pseudoshape based method [14], active energy image based method [15] and template matching based method [12]. This results show the superiority of the proposed method.

### Table 1: Experiment settings

<table>
<thead>
<tr>
<th>Experiments</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gallery (training)</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
</tr>
<tr>
<td>Probe (test)</td>
<td>Normal</td>
<td>Fast</td>
<td>Slow</td>
<td>Bag</td>
</tr>
<tr>
<td>Nr Gallery</td>
<td>453</td>
<td>453</td>
<td>453</td>
<td>453</td>
</tr>
<tr>
<td>Nr Prob</td>
<td>151</td>
<td>302</td>
<td>302</td>
<td>302</td>
</tr>
</tbody>
</table>
Table 2 also shows that the proposed ‘counting’ classifier has comparable performance with more complex SVM classifier when the sequence contains complete walking period.

The proposed method is also evaluated under the situation that the walking cycle is incomplete. We simulate this incomplete situation by randomly selecting a few silhouette images from the given test image sequence from “normal walking”. The number of the selected silhouette image is smaller than the number of images of a complete walking cycle/period. We take the first $n$th images from each sequence of the dataset for classification using the ‘counting’ classifier. Figure 5 shows the number of images versus the accuracy rates (The complete walking period is around 25 in the dataset). From this figure, we can see that, for the ‘counting’ classifier, when one image is used for classification, 74.01% accuracy rate is achieved; when four images are used for classification, 90.57% accuracy rate can be achieved; when night images are used for classification, we can achieve 97.19% accuracy rate; after 13 images are used, the accuracy rate approaches 100%. This result suggests that, even the observed walking sequence is incomplete, the proposed method is still able to achieve certain accuracy rate, which enables the proposed method can be used for the condition when occlusion is occurred or the subject only appeared for a short time. Figure 5 also shows that the ‘counting’ classifier can give much better performance than the SVM classifier when the number of given images is small. For four given images, the SVM classifier only can get a 77.49% accuracy rate while the ‘counting’ classifier can reach 90.57%. Because, the classification result of SVM classifier is based on the kernel measurement between the test feature vector and support vectors. The ‘counting’ classifier, however, does not depends on training data. Therefore, it is more robust to the incomplete situation.

4. Conclusions

In this paper, we have proposed a bag-of-gait model based method for gait recognition. In the proposed method, a codebook is firstly created by performing clustering on the features extracted from the silhouette images; then, based on this codebook, a given sequence of gait silhouette images is represented by a feature vector which represents the existence of different code words. After that, the feature vector is used for further classification. The major advantages of the proposed methods are (1) there is no need to estimate the period of gait and (2) it can handle the situation when the observed gait cycle is incomplete. The proposed method was evaluated on a dataset consisting of 151 subjects under different walking conditions using four-fold cross-validation. The experimental results show that the proposed method can achieve the state-of-the-art performance on a dataset consisting of 151 subjects using four-fold cross-validation. Moreover, the method can reach 90% accuracy rate using only 20% number of images of a whole walking period, which indicates that it is capable of handling the gait recognition problem when the given gait sequence is incomplete.

5. Acknowledgment

This research was sponsored by the Innovation and Technology Commission of the Government of the Hong Kong Special Administrative Region, under the Grant ITS/082/11.

References

### Table 2: COMPARISON WITH OTHER METHODS ON CASIA DATASET C

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>99.84%</td>
<td>99.32%</td>
<td>98.4%</td>
<td>88.89%</td>
<td>94%</td>
</tr>
<tr>
<td>B</td>
<td>94.21%</td>
<td>95.78%</td>
<td>93.7%</td>
<td>90.20%</td>
<td>88%</td>
</tr>
<tr>
<td>C</td>
<td>91.23%</td>
<td>90.48%</td>
<td>91.3%</td>
<td>89.22%</td>
<td>85%</td>
</tr>
<tr>
<td>D</td>
<td>89.82%</td>
<td>88.41%</td>
<td>24.7%</td>
<td>79.74%</td>
<td>51%</td>
</tr>
</tbody>
</table>


