Gait Feature Subset Selection by Mutual Information

Baofeng Guo and Mark S. Nixon, *Associate Member, IEEE*

Abstract—Feature subset selection is an important preprocessing step for pattern recognition, to discard irrelevant and redundant information, as well as to identify the most important attributes. In this paper, we investigate a computationally efficient solution to select the most important features for gait recognition. The specific technique applied is based on mutual information (MI), which evaluates the statistical dependence between two random variables and has an established relation with the Bayes classification error. Extending our earlier research, we show that a sequential selection method based on MI can provide an effective solution for high-dimensional human gait data. To assess the performance of the approach, experiments are carried out based on a 73-dimensional model-based gait features set and on a 64 by 64 pixels model-free gait symmetry map on the Southampton HiD Gait database. The experimental results confirm the effectiveness of the method, removing about 50% of the model-based features and 95% of the symmetry map's pixels without significant loss in recognition capability, which outperforms correlation and analysis-of-variance-based methods.

Index Terms—Biometrics, feature selection, gait recognition, mutual information (MI).

I. INTRODUCTION

HUMAN GAIT recognition is to identify a person through the pattern or style of walking, which becomes quite appealing when it is difficult to get other biometrics information at the specified resolution [1]. The information contained in gait data, no matter the spatial-temporal measurements, such as step length, cycle time, etc., or the kinematic measurements, such as joint rotation, angles of the hip, knee and ankle, etc., can all reflect the human body's unique attributes and differ from each other to some extent. There is also evidence in medicine and psychophysics, showing that human gait can be unique under certain circumstances [1]. Therefore, since the 1990s, gait recognition has received significant attention [2]–[4].

Approaches to extract gait features can be roughly categorized as two classes: model-based and model-free methods. The model-based methods analyze the structures underlying the gait data, and extract measurable parameters, such as the kinematic values. The model-free methods generally consider the motion pattern of the human body holistically, for example, the silhouette of the moving subjects. However,

the extracted features can be high dimensional in both cases. This is partly because of the lack of knowledge about what elements of gait best distinguish between persons, which often leads to a desire to consider as much potentially useful information as possible. For example, the feature vectors in a model-based method [5]–[7] have 73 components, let alone the model-free silhouette-based data [8], [9] that is usually a picture with the size such as 64 by 64 pixels. Research in [10] unwraps the 2-D silhouette picture to a 1-D signal, and then uses the boundary points as the features. This greatly reduced the dimensionality, but the dimensionality is still too high and principal component analysis (PCA) training has been applied subsequently.

Such high-dimensional feature (or data) spaces present several problems to gait recognition. First, it may affect the performance of many classification methods. Although some techniques, such as the k-nn method, may suffer less from this problem, it becomes severe for probability-based algorithms. When the number of training samples is limited, which is quite common for the case of biometric data, parameter estimation, e.g., inverse of covariance matrix, tends to be inaccurate. The mean classification accuracy will at first grow and then decline as the number of gait features increases, i.e., the so-called "curse of dimensionality" in pattern classification [11]. Second, due to the existence of many gait-related factors, e.g., health, age, body size, weight, speed, etc., coupled with limited understanding of the underlying recognition mechanism, many redundant features tend to be included to avoid any loss of useful information. For instance, we may argue that some model-based parameters, such as the hip and knee rotation, are actually correlated with each other to some degree. This is also true in the case of silhouette-based data, where it is known that the majority of pixels in the background area are of little value since they are not in the region of the human torso. Finally, high-dimensional data always imposes requirements for storage space, computational load, and communication bandwidth, which become vital factors in the actual deployment of a real-time system, with realistic cost and time constraints.

It is therefore advantageous to identify the important features for gait recognition, and then to remove those features which convey little or redundant discriminatory information. This task can be implemented by feature selection [12], [13] or other dimensionality reduction techniques. The former technique involves choosing a subset of original features (also called variables or attributes) that will best represent the original data under a certain criterion. In pattern recognition, this criterion might be retaining the discriminatory information as much as possible, and then achieve the best recognition accuracy with the least computing resource. Other dimensionality reduction methods, such as PCA, usually apply a transform to map

The authors are with the School of Electronics and Computer Science, University of Southampton, SO17 1BJ Southampton, U.K. (e-mail: bg@ecs.soton.ac.uk; msn@ecs.soton.ac.uk).
Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

the high-dimensional space into a space spanned by fewer dimensions. Apart from meeting the computational challenges, interpretation of the identified key features can provide valuable clue or insight to understand what is the intrinsic mechanism regarding gait's uniqueness.

Previous researches in gait feature subset selection [2]–[4], [9], [14]–[17] have mainly considered conventional dimensionality reduction or statistical tools, such as PCA and ANalysis Of VAriance (ANOVA). However, there are still some remaining challenges for gait feature selection. PCA is a popular technique in dimensionality reduction, but this technique mainly retains those features that contribute most to variance (energy), and is not optimized for class separability. Therefore, the identified features may not coincide with the discriminatory information that recognition really requires. By incorporating discriminatory analysis, hybrid methods, such as PCA + Linear Discriminant Analysis, can compensate for this problem. However, as we discussed before, dimensionality reduction is only one of the two major objectives in our research, and the second objective, i.e., to identify the key gait features to understand the uniqueness among human gaits, is still difficult to meet. This is because that in the PCA-based methods, the principal components are linear transformations of the original features and the mapped features will lose the original physical meaning. Although there are some strategies [9] proposed to select the original features from the PCA result, the accuracy of selection will be further reduced. The ANOVA technique can help to identify the features based on a null hypothesis test. However, the difference between classes highlighted by ANOVA is carried out by examining whether the population mean vectors are the same and also lacks the explicit or definite relation with the recognition accuracy.

In this paper, we consider feature selection for gait recognition from the point of view of information theory. Like many selection metrics, mutual information (MI) evaluates the statistical dependence between two random variables and so can be used to measure the utility of selected features in recognition [18], [19]. Previous research has applied MI in different areas, such as input selection for neural networks [20]–[22], image registration [23], [24], medical signal selection [25], data visualization [26], etc. [27], [28]. Here, MI is applied to select the model-based and model-free features for gait recognition. In detail, we show that applying the MI to gait feature selection has several important advantages, such as its close relation with the desired classification accuracy (see Section II-C1), a well-founded calculation framework, and efficient implementation strategy (see Section II-C2). Compared to the previous MI-based methods, the method applied in this paper is improved from our earlier research [19], [29], [30] and has derivation and implementation which are distinct to formulations applied in other areas [18], [20]–[22], [26].

II. FINDING THE MOST RELEVANT FEATURES FOR GAIT RECOGNITION

Let $\mathbf{x}' = (X_1', X_2', \ldots, X_P')$ be a P-dimensional gait feature vector, with each component $X_i' \in \mathbb{R}$ representing an observed variable. In gait recognition, this variable could be a parameter in the model-based feature vector or a pixel in the model-free pictures. The objective of feature selection is to find a subset of the aforementioned components, $\mathbf{x} = (X_1, X_2, \ldots, X_Q)$, $Q < P, X_j \in \{X_1', X_2', \ldots, X_P'\}, j = 1, 2, \ldots, Q$, satisfying a certain cost criterion, J, such as

$$J(\mathbf{x}^0) = \min_{\mathbf{x} \in \Xi} J(\mathbf{x}) \quad (1)$$

where Ξ is a set of any Q-dimensional vectors selected from the original P-dimensional vector $\mathbf{x}'$. Here, the number Q can be chosen by meeting some prespecified requirements, such as a preset recognition rate.

In pattern recognition, a natural selection metric will be the classification accuracy or inversely the Bayes classification error. However, direct minimization of the Bayes error is difficult to perform analytically, and a wide range of alternative statistics that are easier to evaluate have been considered. Two typical measures used in gait feature selection are introduced as follows:

A. Feature Selection by One-Way ANOVA

In statistics, ANOVA is a general method for studying sampled-data relationships. The observed variance is partitioned into components due to different explanatory variables, such as between-class and within-class variance. "One-way ANOVA" is the simplest case, and its purpose is to test for significant differences between class means.

Let x_j^i be a sample of the random variable representing a gait feature $X \in \{X_1, X_2, \ldots, X_N\}$, where superscript $i = 1, 2, \ldots, M$ stands for the subject class, and $j = 1, 2, \ldots, N_i$ denotes the sample number of the subject i. ANOVA can be used to investigate whether the population means are the same (i.e., the null hypothesis of equality of means $H_0 : \mu_1 = \mu_2 = \cdots = \mu_M$, $\mu_i = (1/N_i) \sum_{j=1}^{N_i} x_j^i$), and if not, to what extent the means differ. The result of this hypothesis can be quantitatively given by the following F-test with rejection of H_0 at level of α:

$$F = \frac{\text{SS}_\text{b}/(M-1)}{\text{SS}_\text{w}/\left(\sum_{i=1}^{M} N_i - M\right)} > F_{M-1, \sum_{i=1}^{M} N_i - M}(\alpha) \quad (2)$$

where $F_{M-1, \sum_{i=1}^{M} N_i - M}(\alpha)$ is the upper (100α)th percentile of the F-distribution with $M-1$ and $\sum_{i=1}^{M} N_i - M$ degrees of freedom. The between class variance $\text{SS}_\text{b} = \sum_{i=1}^{M} N_i(\bar{x}^i - \bar{x})^2$ is the sum of square variance among classes (or groups), where $\bar{x}^i$ and $\bar{x}$ are the group and the overall sample means, respectively. The within-class variance $\text{SS}_\text{w} = \sum_{i=1}^{M} \sum_{j=1}^{N_i} (x_j^i - \bar{x}^i)^2$ is the total within-class sum of square variance.

The F-test in (2) gives a statistic which indicates to what extent the samples of variable X are drawn from a distribution of the same means, i.e., no differences between means of different classes. It can be shown that this number, linked with the between- and within-class variance (akin to the Fisher's linear discriminant), determines whether the groups are actually

different by the measured characteristic. This can provide useful information for feature selection. However, the ANOVA-based feature selection methods are based on the assumptions of normality of the distributions and homogeneity of variances. It can be argued that such assumptions cannot be always guaranteed in biometric applications, and therefore, the ANOVA statistics may not be sufficient to evaluate the relationship between the features, particularly in the case where the aforementioned assumptions do not hold.

B. Feature Selection by Correlation Metric

Correlation indicates the strength and direction of a linear relationship between two random variables [31]. In detail, correlation refers to the departure of two variables from linear independence. This metric is useful to investigate how far a feature departs from the class label, and in a broad sense was used for feature selection. The conventional Pearson correlation coefficient is defined by

$$p_{X,Y} = \frac{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y} \qquad (3)$$

where x_i and y_i are samples of two random variables X and Y; $\bar{x}$ and $\bar{y}$ are their sample means; and σ_x and σ_y their standard deviations.

According to the feature selection framework discussed in (1), the selection metric should ideally quantify the causality between the class label and features, to effectively reflect the prediction capability for recognition. Although it is popular in many applications and easily calculated, correlation actually does not imply causation and cannot be validly used in the above feature selection frame. In detail, a strong correlation is necessary but not sufficient to establish a causal relationship. Therefore, using correlation for feature selection cannot be totally and logically justified.

From the aforementioned discussions, we contend that the existing techniques lack explicit and definite relation with the classification accuracy or Bayes classification error, and using them for gait feature selection may deviate from the ideal solution. Hence, it is preferable to consider other metrics to select features for gait recognition.

C. Feature Selection Based on Mutual Information Analysis

In information theory, MI measures the statistical dependence between two random variables and so can be used to evaluate the relative utility of each feature to classification [18]–[22], [29], [30]. Given the relationship between MI and classification error, the feature selected by MI analysis is related to a criterion of optimizing the classification error (see Section II-C1 for the bound relation and its proof). Moreover, the implementation of MI imposes relatively lower computational requirement. For example, the calculations of the matrix inverse and determinant are not needed, which makes it a convenient alternative to approaches based on other separability metrics.
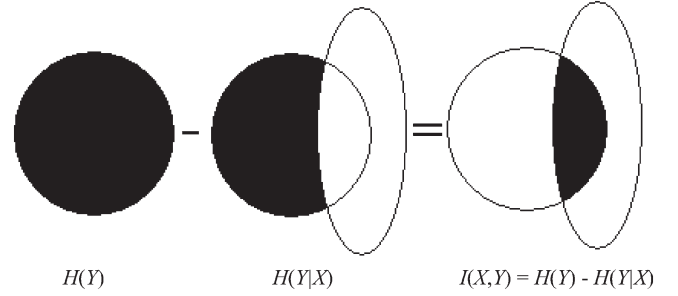

Fig. 1.   Illustration of MI.

1) Mutual Information and Entropy: The MI is a quantity that measures the mutual dependence of the two variables, and is defined as

$$I(X,Y) = \int_Y \int_X p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dx\, dy \qquad (4)$$

where $p(x,y)$ is the joint probability density function of continual random variables X and Y, and $p(x)$ and $p(y)$ are the marginal probability density functions, respectively. Given the Shannon entropy (discrete) defined as

$$H(X) = -\sum_X p(x) \log p(x). \qquad (5)$$

MI is related to entropy as

$$
\begin{aligned}
I(X,Y) &= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X) \\
&= H(X) + H(Y) - H(X,Y) \qquad (6)
\end{aligned}
$$

where $H(X)$ and $H(Y)$ are the entropy of X and Y, $H(X,Y)$ their joint entropy, and $H(X|Y)$ and $H(Y|X)$ the conditional entropies of X given Y and of Y given X, respectively. For discrete variables, the joint and conditional entropies can be written as

$$H(X,Y) = -\sum_{x \in X}\sum_{y \in Y} p(x,y) \log p(x,y) \qquad (7)$$

$$H(X|Y) = -\sum_{x \in X}\sum_{y \in Y} p(x|y) \log p(x|y) \qquad (8)$$

where $p(x|y)$ is the conditional probability.

The use of MI for feature selection can be intuitively justified by the following arguments: Let Y be a variable representing the class label (e.g., the subject's identity), and X a variable denoting a gait feature. The entropy $H(Y)$ is known to be a measure of the amount of uncertainty about Y (i.e., the objective of recognition), while $H(Y|X)$ is the amount of uncertainty left in Y when knowing an observation X. From (6), it is seen that $I(X,Y)$ is the reduction in the uncertainty of class label Y by the knowledge or measurement obtained at gait feature X. Hence, MI can be interpreted as the amount of information that the measurement at gait feature X contains about the class label Y (see Venn diagram in Fig. 1). In other words, MI can indicate the amount of information that a gait

feature X contains about the class label Y. Since the variable defined by class label Y is the required classification result, the MI measures the capability (in the sense of information theory) of using this gait feature to predict the class label or the subject's identity, i.e., the objective of gait recognition.

The important theoretical evidence that can support using MI for feature selection is given by various bounds between MI and the classification error. Although several papers cited some examples of these bounds [26], its proof is either absent or based on some special cases, such as the Renyi entropy [32] or the two-class problem [33]. For the convenience of understanding, a full proof from a general classification perspective is presented as follows:

First, the conditional entropy $H(Y|X)$ is related to the Bayes classification error by the following inequality (i.e., the Fano's inequality):

$$H(Y|X) \leq H(P_e) + P_e \log(l-1) \tag{9}$$

where the error probability $P_e = P\{f(X) \neq Y\}$; f is a classification function; and l is the number of classes. Its proof is as follows:

Let the class label $Y \in \{c_1, c_2, \ldots, c_l\}$, the feature vector $X \in \mathbb{R}^n$, and the classification result $f(X) \in \{c_1, c_2, \ldots, c_l\}$. Z is the Bayes cost, i.e.,

$$Z = \begin{cases} 1 & \text{if } f(X) \neq Y \\ 0 & \text{if } f(X) = Y. \end{cases} \tag{10}$$

According to (8), we have

$$H(Y|X) = \sum_x p(x) H(Y|X = x)$$
$$= \sum_x p(x) \sum_y p(y|x) \log \frac{1}{p(y|x)}$$
$$= \sum_{x,y} p(x,y) \log \frac{1}{p(y|x)}$$
$$= \sum_{x,y,z} p(x,y,z) \log \frac{1}{p(y|x)}$$
$$= \sum_z p(z) \sum_{x,y} \frac{p(x,y,z)}{p(z)} \log \frac{1}{p(y|x)}. \tag{11}$$

Applying Jensen's inequality (i.e., $\mathbb{E}\{\varphi(Y)\} \geq \varphi(\mathbb{E}\{Y\})$) [34] with $\varphi(\cdot) = -\log(\cdot)$) to (11), we get

$$H(Y|X) \leq \sum_z p(z) \log \left[\frac{1}{p(z)} \sum_{x,y} \frac{p(x,y,z)}{p(y|x)} \right]$$
$$= \sum_z p(z) \log \frac{1}{p(z)} + \sum_z p(z) \log \sum_{x,y} \frac{p(x,y,z)}{p(y|x)}$$
$$= \sum_z p(z) \log \frac{1}{p(z)} + \sum_z p(z) \log \sum_{x,y} p(x) p(z|x,y)$$
$$= H(P_e) + \sum_z p(z) \log \sum_{x,y} p(y) p(z|x,y). \tag{12}$$

Now, considering the definition of Bayes cost in (10), the last term of the right side of (12), denoted as L, becomes

$$L = \sum_z p(z) \log \sum_{x,y} p(x) p(z|x,y)$$
$$= \sum_z p(z) \log \left[\sum_x \sum_{\{y|y=f(x)\}} p(x) p(z|x,y) \right.$$
$$\left. + \sum_x \sum_{\{y|y\neq f(x)\}} p(x) p(z|x,y) \right]$$
$$= P(Z=0) \log \left[\sum_x \sum_{\{y|y=f(x)\}} p(x) \right]$$
$$+ P(Z=1) \log \left[\sum_x \sum_{\{y|y\neq f(x)\}} p(x) \right]$$
$$= P(Z=0) \log \left(\sum_x p(x) \right)$$
$$+ P(Z=1) \log \left[(l-1) \sum_x p(x) \right]$$
$$= P_e \log(l-1). \tag{13}$$

Combining the results of (12) and (13), we get (9), i.e., Fano's inequality.

Given $C_1 = H(P_e)$ and $C_2 = \log(l-1)$, and using the second row of (6), the inequality of (14) is derived as follows:

$$\frac{H(Y) - I(X,Y) - C_1}{C_2} \leq P_e. \tag{14}$$

From (14), it is seen that the classification error P_e has a lower bound tied by a term decided by $I(X,Y)$ and $H(Y)$. Thus, given a fixed $H(Y)$ (e.g., the same prior distribution), maximization of MI can optimize the lower bound of classification error, and using MI for feature selection is actually indirectly guided by recognition accuracy. Comparing with the statistical tools introduced in Section II, we can see that the MI has more clear and definite relation with the recognition objective. From the point of view of statistics and information theory, the MI measures more the general dependencies in the data, and therefore may lead to better capability in feature selection.

2) Calculation of Mutual Information for Feature Selection: Applying the MI to the feature selection framework in (1), we can carry out a new gait feature subset selection based on information theory. To implement this method, there are two obstacles to overcome:

1) how to evaluate multidimensional MI;
2) how to search for the maximum.

Aiming to solve these problems, we apply a gradient ascent optimization strategy improved from our earlier research [19], [29], [30], described as follows. First, we show that a multidimensional MI can be decomposed into a series of 1-D MIs:

Let $\mathbf{x} = (X_1, X_2, \ldots, X_M)$ be a random vector representing the selected features X_i, $i = 1, 2, \ldots, M$, and Y the random variable corresponding to the class label. The MI between them can be written as

$$I(\mathbf{x}, Y) = I\left((X_1, X_2, \ldots, X_M), Y\right). \tag{15}$$

If $\mathbf{x}$ only has two components, i.e., $\mathbf{x} = (X_1, X_2)$, (15) becomes

$$
\begin{aligned}
I(\mathbf{x}, Y) &= I\left((X_1, X_2), Y\right) \\
&= H(X_1, X_2) - H(X_1, X_2|Y).
\end{aligned} \tag{16}
$$

From (6), we can derive the following two equations:

$$H(X_1, X_2) = H(X_1) + H(X_2) - I(X_1, X_2) \tag{17}$$

$$H(X_1, X_2|Y) = H(X_1|Y) + H(X_2|Y) - I(X_1, X_2|Y). \tag{18}$$

Substituting $H(X_1, X_2)$ and $H(X_1, X_2|Y)$ of (17) and (18) into (16), we get

$$I(\mathbf{x}, Y) = \sum_{i=1,2} I(X_i, Y) - I(X_1, X_2) + I(X_1, X_2|Y). \tag{19}$$

To extend (19) to more than two components, we can replace X_1 of (19) with X_1' and redefine $X_1' = (X_1, X_3)$. Then, by using the above decomposition, we have

$$
\begin{aligned}
&I(\mathbf{x}, Y) \\
&= I\left(X_1', Y\right) + I(X_2, Y) - I\left(X_1', X_2\right) + I\left(X_1', X_2|Y\right) \\
&= \left[\sum_{i=1,3} I(X_i, Y) - I(X_1, X_3) + I(X_1, X_3|Y)\right] + I(X_2, Y) \\
&\quad - \left[\sum_{i=1,3} I(X_i, X_2) - I(X_1, X_3) + I(X_1, X_3|X_2)\right] \\
&\quad + \left[\sum_{i=1,3} I(X_i, X_2|Y) - I(X_1, X_3|Y) + I(X_1, X_3|X_2, Y)\right] \\
&= \sum_{i=1}^{3} I(X_i, Y) - \sum_{i=1,3} I(X_i, X_2) - I(X_1, X_3|X_2) \\
&\quad + \sum_{i=1,3} I(X_i, X_2|Y) + I(X_1, X_3|X_2, Y).
\end{aligned} \tag{20}
$$

It is known that given variables A, B, and C, $I(A, B|C) \leq I(A, B)$ and $I(A, B) = I(B, A)$. Therefore, (21) becomes

$$
\begin{aligned}
I(\mathbf{x}, Y) &= \sum_{i=1}^{3} I(X_i, Y) - \sum_{i=1,3} I(X_i, X_2) - I(X_1, X_3|X_2) \\
&\quad + \sum_{i=1,3} I(X_i, X_2|Y) + I(X_1, X_3|X_2, Y)
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{3} I(X_i, Y) - \sum_{i=1,3} I(X_i, X_2) - I(X_1, X_3) \\
&\quad + \sum_{i=1,3} I(X_i, X_2|Y) + I(X_1, X_3|Y) \\
&= \sum_{i=1}^{3} I(X_i, Y) - \sum_{i=1}^{3}\sum_{j>i} I(X_i, X_j) \\
&\quad + \sum_{i=1}^{3}\sum_{j>i} I(X_i, X_2|Y).
\end{aligned} \tag{21}
$$

After repeating this decomposition, we can get the following approximation:

$$
\begin{aligned}
I(\mathbf{x}, Y) &\approx \sum_{i} I(X_i, Y) - \sum_{i}\sum_{j>i} I(X_i, X_j) \\
&\quad + \sum_{i}\sum_{j>i} I(X_i, X_j|Y).
\end{aligned} \tag{22}
$$

Based on (22), a fast approach to maximize $I(\mathbf{x}, Y)$ can be implemented as follows: the first variable is chosen as:

$$X_1^0 = \max_i I(X_i, Y) \tag{23}$$

where X_1^0 represents the result of maximization at step 1).

Then, the second variable is chosen as

$$X_2^0 = \max_{X_i \neq X_1^0} \left[I(X_i, Y) - I\left(X_i, X_1^0\right) + I\left(X_i, X_1^0|Y\right)\right]. \tag{24}$$

The remaining variables are chosen in the same way

$$X_n^0 = \max_{X_i \neq X_j^0} \left[I(X_i, Y) - \sum_j I\left(X_i, X_j^0\right) + \sum_j I\left(X_i, X_j^0|Y\right)\right] \tag{25}$$

where X_j^0, $j = 1, 2, \ldots, n-1$ are the variables already selected. This selection is repeated until the prespecified number, N, of variables is reached. For a general classification task, the number N can be decided by the tolerance of degradation for detection accuracy or false alarm probability. Moreover, other application-related requirements, such as computing capability, storage space, communication bandwidth, etc., can also be used to decide this number.

To calculate 1-D MI, we can treat each gait feature component as a random variable X with continuous measured values, and its class category as Y with discrete identity labels $\omega_1, \omega_2, \ldots, \omega_n$. Thus, MI between X and Y can be evaluated as follows (with a similar formula for $I(X_i, X_j)$):

$$
\begin{aligned}
I(X, Y) &= -\int_X p(x) \log p(x)\, dx - \sum_y P(y) \log P(y) \\
&\quad + \sum_y \int_X p(x, y) \log p(x, y)\, dx.
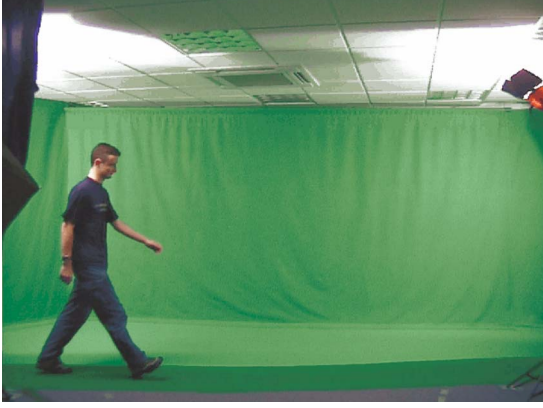\end{aligned} \tag{26}
$$

Fig. 2. Illustration of one frame of a sequence of Southampton HiD Gait database.

In the estimation of the MI from the data, the probability densities can be estimated based on histograms, and a flexible binning scheme is applied in this paper. In detail, to represent the class label's histogram, we usually know the number of classes (i.e., the number of subjects in a gait gallery). Therefore, the number of bins for the variable Y can be chosen by this *a priori* knowledge. For the feature variable X, an adaptive binning is adopted, where the number of bins n_b is decided by the number of data samples n_s, i.e., $n_b = \sqrt{n_s}$. Other binning methods that can optimize the probability density estimation may also be applied, but this discussion is not in the scope of this paper.

III. EXPERIMENTS

To assess the MI-based gait feature selection, an indoor Southampton HiD Gait database [35] (http://www.gait.ecs. soton.ac.uk/database) is used. This database consists of 2163 sequences from 116 subjects walking both to the left and to the right. Each subject was filmed from a fronto-parallel viewpoint at a resolution of 720×576 pixels, in controlled laboratory conditions. The sequence is recorded at a rate of 25 frames per second with approximately 90 frames per gait sequence. An example of one frame of such a sequence is shown in Fig. 2.

A. Gait Feature Sets

Based on this data set, two approaches to obtaining a raw set of features are used. The first approach is a model-based method [5], which extracts a 73-dimensional measurement vector for each gait sequence, based on techniques such as adaptive model and deformable contours. The head and torso are modeled using ellipses; each leg is modeled by two tapered pairs of lines, and the foot by a rectangle. The extracted feature components include 18 model parameters based on joint rotation models for the hip, knee and ankle, and 28 parameters describing the subjects' speed, gait frequency, body proportions, etc. Fig. 3(a) shows a list of human body model parameters and shows how they are measured [36].

The second approach is a model-free method based on analyzing the symmetry of human motion [8]. This method is supported by the psychologists' view that human gait is

a symmetrical pattern of motion, and therefore suggesting symmetry is suitable for gait recognition. The generalized symmetry operator was applied to locate features according to their symmetrical properties without relying on the borders of a shape or on general appearance. For each sequence, the obtained gait signature is a 64 by 64 pixel symmetry map, such as Fig. 3(b).

B. Experimental Results

After gait features have been extracted, tests of recognition accuracy were carried out to assess the performance of the feature selection methods. Three methods of feature selection are used, the one proposed here (i.e., MI-based) and two comparison methods introduced in Sections II-A and B. To calculate the selection metrics, i.e., the correlation coefficients $P_{X,Y}$, one-way ANOVA's f-statistics, and the values of $I(X_i, Y)$ and $I(X_i, X_j)$, half of the sequences from each subject were randomly chosen as the training set. This training set is assumed as the known data samples, and then can be used for multiple purposes, such as to learn the selection metrics and validate classifier's parameters. The remaining 50% sequences are assumed as the unseen data, and form the testing set on which the recognition accuracy was assessed. To avoid bias, random sampling was used to generate the testing set and training set, and repeated ten times to allow an estimate of the error inherent in this sampling process.

Currently-popular support vector machines (SVMs) [37]–[39] were chosen as the classifiers in these experiments since they are less sensitive to the data's dimensionality. Although SVMs are used here, the proposed method is not limited to this particular classifier and other classification algorithms are also applicable. The relevant SVM formulas are listed as follows:

Let $\mathbf{x}_i = (X_1, X_2, \ldots, X_M)$ be an M-dimensional gait feature (data) vector, $y_i \in (+1, -1)$ be the class label (i.e., the subject identity), $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_N)$, be the Lagrange multipliers, N the number of examples and b a threshold. The SVM classifier can be represented as

$$f(\mathbf{x}) = \text{sgn}\left(\sum_{i=1}^{M} y_i \alpha_i K(\mathbf{x_i}, \mathbf{x}) + \mathbf{b}\right)$$

where $K(\mathbf{x}, \mathbf{x}') = \Phi(\mathbf{x})^T \Phi(\mathbf{x}')$ is an appropriate kernel function which has a corresponding inner product expansion Φ. The commonly used functions are polynomials and Gaussian radial basis functions

$$K(\mathbf{x}, \mathbf{x}') = \left(\mathbf{x}^T \mathbf{x}' + 1\right)^d \tag{27}$$

$$K(\mathbf{x}, \mathbf{x}') = \exp\left\{-\gamma \|\mathbf{x} - \mathbf{x}'\|^2\right\}. \tag{28}$$

Since SVMs are inherently binary (two-class) classifiers, a "one-against-one" scheme was used with subsequent majority voting to give a multiclass result. The kernel function used is an inhomogeneous polynomial [i.e., (27)], and the SVMs' parameters, i.e., polynomial order d and the penalty parameter C are chosen by a validation procedure based on the training data (i.e., to further split the training set for cross-validation).

Static Parameters		Static Parameters	
HW	Head width	TL*	Thigh length
HH	Head height	SL*	Shin length
HDX	Head x offset	PW*	Pelvis width
HDY	Head y offset	**Dynamic Parameters**	
TW	Torso width	$X(t)$	x centre of torso
TH	Torso height	$Y(t)$	y centre of torso
LWH	Leg width at hip	$\Theta_{px}(n)$	x pelvis rotation
LWKU	Leg width at knee (upper)	$\Theta_{py}(n)$	y pelvis rotation
LWKL	Leg width at knee (lower)	$\Theta_h(n)$	Hip rotation
LWA	Leg width at ankle	$\Theta_k(n)$	Knee rotation
HPDY*	Hip y offset	$\Theta_a(n)$	Ankle rotation
FW*	Foot width	ω	Gait frequency
FH*	Foot height	ϕ	Gait phase

(a)   (b)
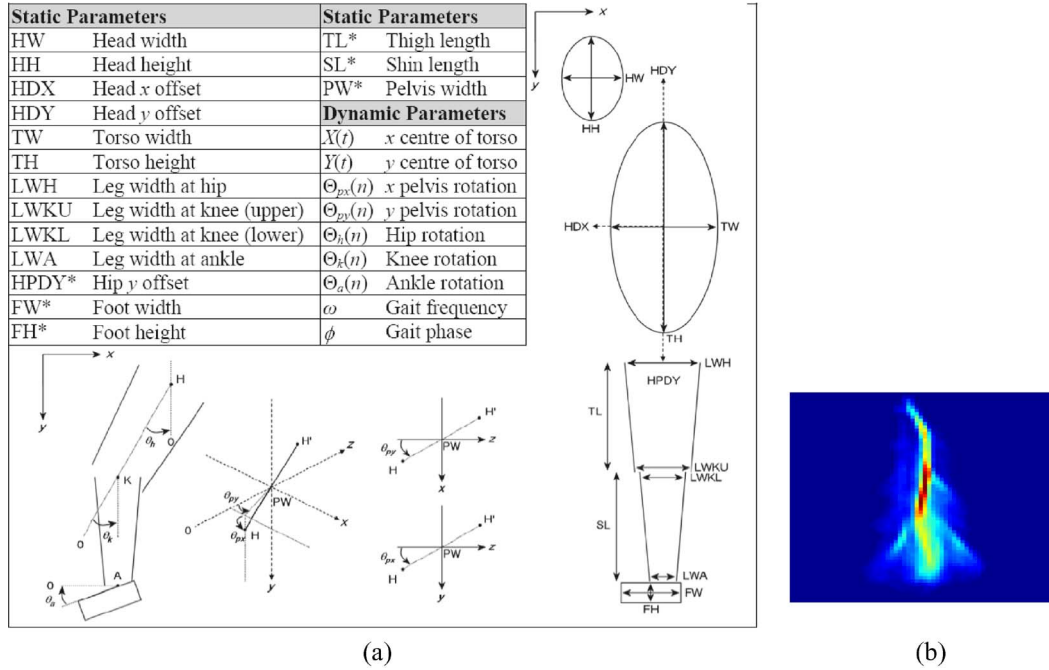
Fig. 3.   (a) List of some geometric human body model parameters and their illustration [36]. (b) Example of a gait symmetry map.
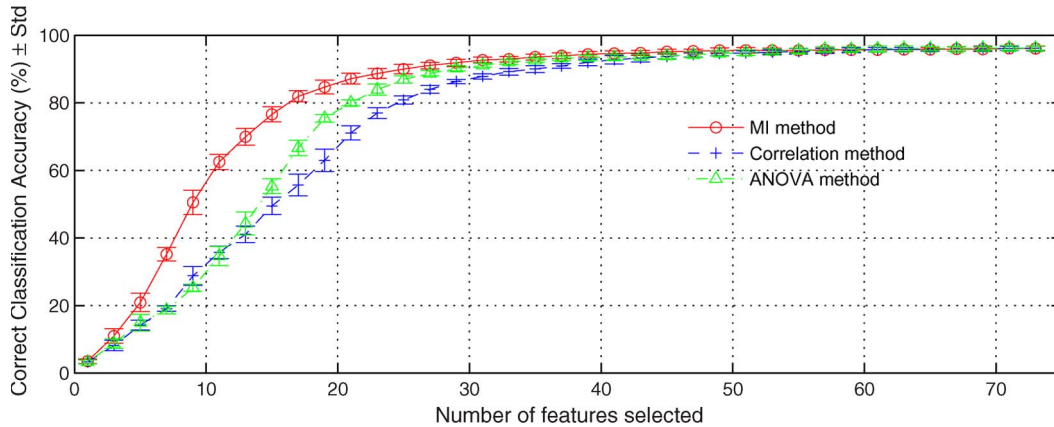

Fig. 4.   Cumulative recognition rates as the model-based features are progressively selected, accuracy (%) ± standard deviation; two features increment at each step.

Then, experiments are carried out to test the aforementioned three feature selection methods on the test set to obtain the corresponding recognition rate [i.e., Correct Classification Recognition (CCR)] on the reduced feature subsets.

The main objective of feature selection is to choose the most "informative" or "relevant" features and achieve the possible highest recognition accuracy on the reduced feature set. The experiments were thus designed to assess the change of recognition accuracy as features are progressively selected (i.e., the cumulative recognition rate). The feature selection results for the model-based feature set are shown in Fig. 4, where the points marked with "O" represents the results of the MI-based method, benchmarked by that of the correlation-based method (the points marked with "+") and the ANOVA-based method (the points marked with "Δ"). Data points are at two feature increments at each step, and the corresponding recognition rates

with their error bars (i.e., CCR ± stand deviation) are also shown in Fig. 4.

From Fig. 4, it is shown that the feature subset selected by MI achieves a better recognition rate with fewer model parameters than for the correlation and ANOVA based methods. This is particularly significant when up to 35 features are selected. For example, the MI-based method achieves 90% correct recognition rate with 25 features, compared with 29 features for the ANOVA-based method and 37 features for the correlation-based method. For the MI-based method, when more than 37 features (accounting for about 50% of all 73 features) are selected, the recognition rate then improves only slightly. This indicates that for this particular model-based feature set, about 50% of parameters are almost redundant and can be removed without significant loss of recognition accuracy. When more and more features are selected, the differences between the

TABLE I
TOP 15 SELECTED MODEL-BASED PARAMETERS BY USING THE THREE METHODS

| Rank	MI method	Correlation method	ANOVA method
1	Ankle Width LWA	Leg Width 08	Leg Width 08
2	Leg Width 08	Leg Width 09	Leg Width 09
3	Hip Rotation Amplitude	Knee Width LWKU	Leg Width 04
4	Leg Width 01	Ankle Width LWA	Leg Width 05
5	Head dx HDX	Leg Width 07	Leg Width 07
6	Y Motion Amplitude Ay	Head Height HH	Leg Width 06
7	Torso Width TW	Knee Rotation angle(12)	Ankle Width LWA
8	Knee Width LWKU	Knee Rotation Mean	Knee Width LWKL
9	Knee Rotation angle(02)	Head dx HDX	Leg Width 03
10	Gait Frequency omega	Knee Width LWKL	Leg Width 01
11	Knee Rotation angle(11)	Knee Rotation angle(04)	Leg Width 00
12	Ankle Rotation angle(07)	Ankle Rotation angle(12)	Leg Width 02
13	Hip Rotation Mean	Knee Rotation angle(03)	Gait Frequency omega
14	Knee Rotation angle(13)	Leg Width 06	Y Motion Amplitude Ay
15	Leg Width 03	Ankle Rotation angle(06)	Head dx HDX
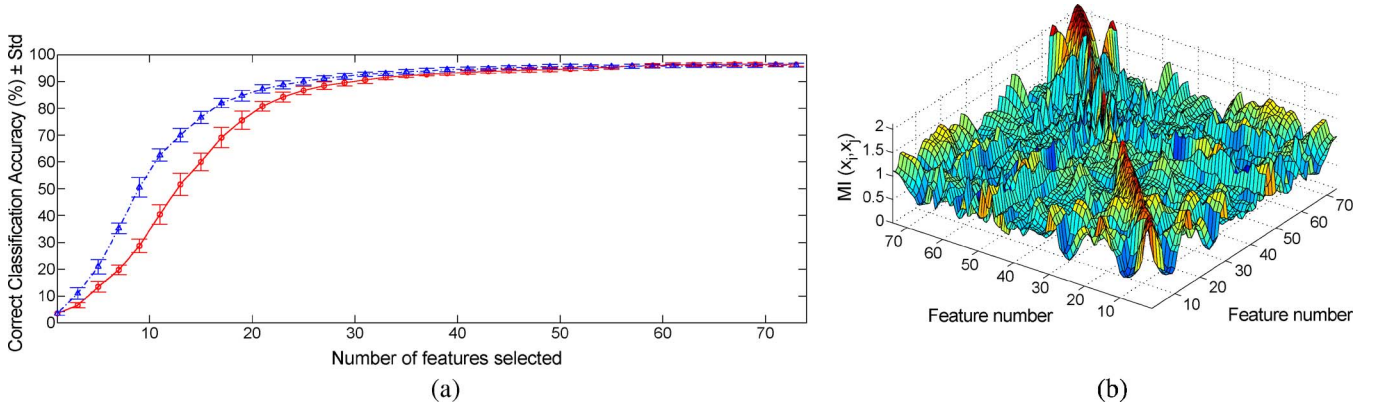
Fig. 5. (a) Comparison of cumulative recognition rates for two MI maximization schemes: "O" stands for the results of maximizing $\sum_i I(X_i, Y)$ and "Δ" for maximizing $I(\mathbf{x}, Y)$ of (22). (b) Three-dimensional illustration of the MI matrix, i.e., $I(X_i, X_j)$, which is used for maximizing $I(\mathbf{x}, Y)$ of (22).

three methods tend to be small. This is because at this stage, most of the essential features have already been included. The results in Fig. 4 also suggest that the ANOVA-based method is generally better than the correlation-based method.

Table I further lists the names of the 15 top model-based features selected by the MI, the correlation and the ANOVA-based methods, respectively. It is intuitively suggested that the features selected by the MI seem to contain the wider range of gait information and more dynamics than other two methods. For example, the results show that the MI-based method includes information from more body parts, such as ankle, leg, hip, head, knee, as well as their dynamic characters, such as rotation angle, motion amplitude, etc. The other two methods concentrate more on the major body parts, such as leg, knee, and its static information, such as width, etc. Given the complexity of human anatomy, the selection using MI is quite appealing for recognition purpose.

To justify the MI maximization method presented in Section II-C2, two different optimization schemes are also compared. The first one is based on the features' independent MI ranking and is to maximize the term $\sum_i I(X_i, Y)$ of (22). This scheme only needs to calculate the MI between each feature and the class label, which is faster and can be implemented easily. The second scheme considers the combinatorial MI of a group features $\mathbf{x} = (X_1, X_2, \ldots)$, i.e., maximization of

$I(\mathbf{x}, Y)$. It needs to calculate the MI between each pair of features $I(X_i, X_j)$ as well, and search for the maximum of the combinatorial term. This requires extra computational load but gives a more accurate result. Fig. 5(a) shows the cumulative recognition rates for these two schemes.

It is shown from Fig. 5(a) that the second scheme (the points marked with "Δ") is better than the first one (the points marked with "O"). This result indicates that the feature-independent assumption, on which the first scheme is based, may not hold in this application. Fig. 5(b) further shows the MI between each pair of features. It is shown that except the most significant values appear in the leading diagonal, other considerably higher values can also be found in other areas [as in the many peaks in Fig. 5(b)]. This observation reinforced the aforementioned conclusion regarding the independence assumption.

The feature selection results for the model-free gait symmetry pictures are shown in Fig. 6. The points marked with "O" are the results of the MI-based method, comparing with the correlation-based method (the points marked with "+") and the ANOVA-based method (the points marked with "Δ"). It is seen that for the case of the MI-based method, the recognition rate changes little when more than 100 important symmetry pixels are selected, which compares to 150 important pixels selected by the ANOVA-based method. Moreover, the correlation-based method cannot reach the 95% accuracy rate that is achieved by
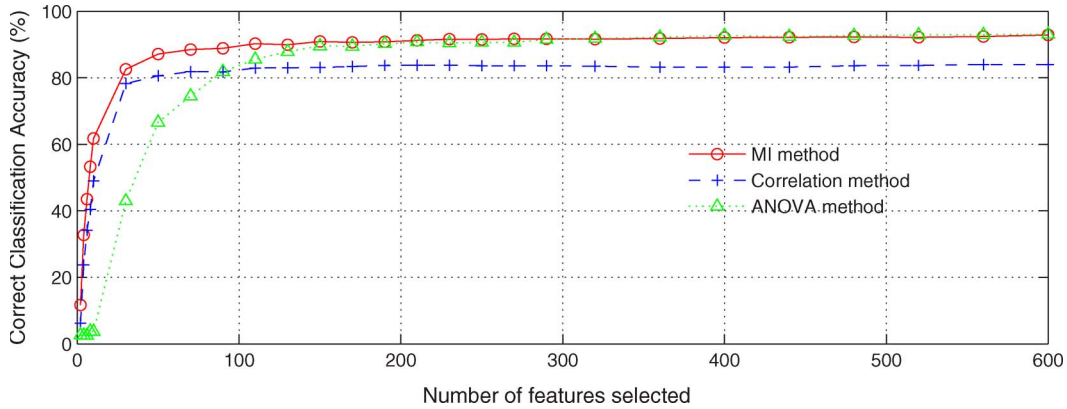
Fig. 6.   Cumulative recognition rates as the model-free symmetry pixels are progressively selected.

both the MI and the ANOVA-based methods, even after using about 200 pixels in the symmetry maps (these account for about 5% in all 64×64 pixels).

The cumulative recognition results on the model-free gait data shown in Fig. 6 are consistent with our previous findings in the model-based feature set shown in Fig. 4, with the best recognition rate achieved by the MI-based method, followed by ANOVA and correlation-based methods. These results confirm the effectiveness of using MI as the feature selection metric with its suitability to approximate the classification accuracy, which is in contrast to the correlation coefficient that is based on a linear rather than the causal relationship, and the ANOVA, which is built on the assumptions of normality distribution and homogenized variances. These empirical findings are based on the specific data set, and many application-related factors should be taken into account to extend them, such as the variability of data set, noise on the symmetry maps, etc.

The above experiments used half of the sequences from each subject as the training set, which covered the general variability within data (e.g., how far the sequences were apart for each subject). In biometric applications, it is also important to assess the performance with respect to temporal variability, such as on the data taken on the order of weeks or months. To address this scenario, further experiments are carried out by separating data into three sets: a training set accounting for 50% of the subjects to find the important features, a validation set accounting for 25% of the subjects to validate the classifier's parameters, and a test set accounting for the remaining 25% of the subjects to assess classification accuracy. In this case, since the training and the test data set are drawn from different subjects, they are person-disjoint. Meanwhile, this is a suitable configuration to increase the considerable amount of time variability based on the current data set. The classification results on selecting several typical numbers of features are listed in Table II for the model-based feature set and in Table III for the model-free symmetry feature set, respectively. The results shown in Tables II and III reconfirmed our previous finding, i.e., better performance has been achieved by the MI-based method. Compared to the earlier results shown in Figs. 4 and 6, it can be found that the numbers in Tables II and III present a better situation. This is because the new experiments used 50% and 25% of the subjects for the training and validation,

TABLE II
CLASSIFICATION ACCURACY (%) BY USING THE THREE METHODS;
BASED ON THE MODEL-BASED PARAMETERS SEPARATED BY SUBJECTS

Number of features	5	15	25	35	45	55
MI-method	86.17	96.95	98.29	98.73	98.61	98.66
ANOVA-method	74.06	94.64	97.23	98.13	98.28	98.45
Correlation-method	64.06	90.13	95.08	96.38	97.47	98.13

respectively, but the former experiments used half of the sequences from each subject for these purposes. Therefore, more training data has been used in the new experiments, showing improved accuracies. The different configurations are mainly used to check different variability. This may produce slightly different results, but will not affect the overall conclusions on performance comparison.

In this section, results are presented for each of the two types of raw feature sets (i.e., the model-based and model-free gait features), comparing the performance of the MI-based method to that of the reference correlation and ANOVA-based methods. The MI-based method is found better in the range of a lower number of features selected. With enough features, all three methods reach the same performance level for the model-based feature set and MI and ANOVA-based methods reach the same performance for the model-free feature set.

IV. CONCLUSIONS

We have presented a feature selection method for gait recognition based on MI analysis. This feature selection technique has a definite relation with the classification accuracy, and can effectively identify the most pertinent features to improve system's efficiency. To assess the proposed method, experiments were carried out on a model-based gait feature set and a model-free symmetry gait data. The results showed that the MI-based method has good application capability in gait feature selection, with fewer features that can obtain good recognition accuracy. It also outperformed the correlation-based and the ANOVA-based methods. Further research is ongoing to explore the selected results to understand the underlying distinctiveness among human gaits.

Currently, the algorithms were tested on the indoor Southampton HiD Gait database, which was filmed in

TABLE III
CLASSIFICATION ACCURACY (%) BY USING THE THREE METHODS; BASED ON THE MODEL-FREE SYMMETRY FEATURES SEPARATED BY SUBJECTS

| Number of features	5	10	20	30	40	50	60	110	190
MI-method	54.78	80.34	90.22	92.14	93.67	94.11	94.45	95.66	95.34
ANOVA-method	49.94	64.02	74.70	79.11	81.91	83.80	85.78	91.50	94.08
Correlation-method	39.69	64.63	79.83	84.88	87.04	88.23	89.37	91.88	93.89

controlled laboratory conditions. To obtain more general results for different surface types, times, clothing, etc., future research will be carried out to extend the algorithm on gait data sets with other variability, such as the HumanID gait data set [4].

REFERENCES

[1] A. Kale, A. Sundaresan, A. N. Rajagopalan, N. P. Cuntoor, A. K. Roy-Chowdhury, V. Kruger, and R. Chellappa, "Identification of humans using gait," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1163–1173, Sep. 2004.
[2] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer, "The humanID gait challenge problem: Data sets, performance, and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 162–177, Feb. 2005.
[3] M. S. Nixon and J. N. Carter, "Automatic recognition by gait," *Proc. IEEE*, vol. 94, no. 11, pp. 2013–2024, Nov. 2006.
[4] Z. Liu and S. Sarkar, "Improved gait recognition by gait dynamics normalization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 6, pp. 863–876, Jun. 2006.
[5] D. K. Wagg and M. S. Nixon, "Automated markerless extraction of walking people using deformable contour models," *Comput. Animation Virtual Worlds*, vol. 15, no. 3/4, pp. 399–406, Jul. 2004.
[6] C. Yam, M. Nixon, and J. Carter, "Automated person recognition by walking and running via model-based approaches," *Pattern Recognit.*, vol. 37, no. 5, pp. 1057–1072, May 2004.
[7] A. I. Bazin and M. S. Nixon, "Probabilistic combination of static and dynamic gait features for verification," in *Proc. Biometric Technol. Human Identification II, SPIE Defense Security Symp.*, 2005, vol. 5779, pp. 23–30.
[8] J. Hayfron-Acquah, M. S. Nixon, and J. N. Carter, "Automatic gait recognition by symmetry analysis," *Pattern Recognit. Lett.*, vol. 24, no. 13, pp. 2175–2183, Sep. 2003.
[9] G. Veres, L. Gordon, J. N. Carter, and M. Nixon, "What image information is important in silhouette-based gait recognition?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2004, vol. 2, pp. 776–782.
[10] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette analysis-based gait recognition for human identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1505–1518, Dec. 2003.
[11] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 1, pp. 55–63, Jan. 1968.
[12] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 2, pp. 153–158, Feb. 1997.
[13] A. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artif. Intell.*, vol. 97, no. 1/2, pp. 245–271, Dec. 1997.
[14] B. Bhanu and J. Han, "Human recognition on combining kinematic and stationary features," in *Proc. 4th Int. Conf. AVBPA*, 2003, pp. 600–608.
[15] A. Veeraraghavan, R. Chellappa, and A. Roy Chowdhury, "Role of shape and kinematics in human movement analysis," in *Proc. IEEE Conf. CVPR*, 2004, vol. 1, pp. 730–737.
[16] Z. Liu, L. Malave, A. Osuntogun, P. Sudhakar, and S. Sarkar, "Toward understanding the limits of gait recognition," *Proc. SPIE*, vol. 5404, pp. 195–205, 2004.

[17] M. S. Nixon, T. N. Tan, and R. Chellappa, *Human Identification Based on Gait*. New York: Springer-Verlag, 2005, ch. 5, pp. 45–104.
[18] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
[19] B. Guo, S. R. Gunn, R. I. Damper, and J. D. B. Nelson, "Band selection for hyperspectral image classification using mutual information," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 4, pp. 522–526, Oct. 2006.
[20] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Netw.*, vol. 5, no. 4, pp. 537–550, Jul. 1994.
[21] N. Kwak and C.-H. Choi, "Input feature selection by mutual information based on Parzen window," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1667–1671, Dec. 2002.
[22] N. Kwak and C.-H. Choi, "Input feature selection for classification problems," *IEEE Trans. Neural Netw.*, vol. 13, no. 1, pp. 143–159, Jan. 2002.
[23] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Trans. Med. Imag.*, vol. 16, no. 2, pp. 187–198, Apr. 1997.
[24] J. Pluim, J. Maintz, and M. Viergever, "Mutual-information-based registration of medical images: A survey," *IEEE Trans. Med. Imag.*, vol. 22, no. 8, pp. 986–1004, Aug. 2003.
[25] M. Deriche and A. Al-Ani, "A new algorithm for EEG feature selection using mutual information," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, vol. 2, pp. 1057–1060.
[26] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," *J. Mach. Learn. Res.*, vol. 3, no. 3, pp. 1415–1438, 2003.
[27] S. Razniewski and M. Strzelecki, "Evaluation of texture features based on mutual information," in *Proc. 4th Int. Symp. Image Signal Process. Anal.*, 2005, pp. 233–238.
[28] X. Zhou, J. Zhu, K. Liu, B. Sabatini, and S. Wong, "Mutual information-based feature selection in studying perturbation of dendritic structure caused by TSC2 inactivation," *Neuroinformatics*, vol. 4, no. 1, pp. 81–94, Mar. 2006.
[29] B. Guo and M. S. Nixon, "Gait feature subset selection by mutual information," in *Proc. 1st IEEE Int. Conf. Biometrics: Theory, Appl., Syst.*, 2007, pp. 1–6.
[30] B. Guo, R. I. Damper, S. R. Gunn, and J. Nelson, "A fast separability-based feature selection method for high-dimensional remotely-sensed image classification," *Pattern Recognit.*, vol. 41, no. 5, pp. 1670–1679, 2008.
[31] H. Abdi, *Encyclopedia of Measurement and Statistics*, N. J. Salkind, Ed. Thousand Oaks, CA: Sage, 2007, ch. Coefficients of correlation, alienation and determination.
[32] D. Erdogmus and J. Principe, "Lower and upper bounds for misclassification probability based on Renyi's information," *J. VLSI Signal Process. Syst.*, vol. 37, no. 2/3, pp. 305–317, 2004.
[33] M. E. Hellman and J. Raviv, "Probability of error, equivocation, and the Chernoff bound," *IEEE Trans. Inf. Theory*, vol. IT-16, no. 4, pp. 368–372, Jul. 1970.
[34] E. Weisstein, "Jensen's inequality," *MathWorld—A Wolfram Web Resource*. [Online]. Available: http://mathworld.wolfram.com/JensensInequality.html
[35] J. Shutler, M. Grant, M. S. Nixon, and J. N. Carter, "On a large sequence-based human gait database," in *Proc. 4th Int. Conf. Recent Advances Soft Comput.*, 2002, pp. 66–72.
[36] D. Wagg, "Local and global models for articulated motion analysis," Ph.D. dissertation, School Electron. Comput. Sci., Univ. Southampton, Southampton, U.K., 2006.
[37] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th Annu. Workshop Comput. Learn. Theory*, 1992, pp. 144–152.
[38] C. Cortes and V. N. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.
[39] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discovery*, vol. 2, no. 2, pp. 121–167, Jun. 1998.
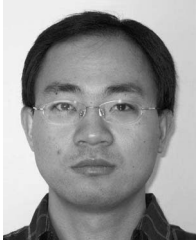
Baofeng Guo received the B.Eng. degree in electronic engineering and M.Eng. degree in signal processing from Xidian University, Xi'an, China, in 1995 and 1998, respectively, and the Ph.D. degree in signal processing from the Chinese Academy of Sciences, Beijing, China, in 2001, respectively.

From 2002 to 2004, he was a Research Assistant with the Department of Computer Science, University of Bristol, Bristol, U.K. Since 2004, he has been Research Fellow with the School of Electronics and Computer Science, University of Southampton, Southampton, U.K. His current research interests include pattern recognition, image processing, and machine learning.

Mark S. Nixon (M'04–A'05) is currently a Professor in computer vision with the University of Southampton, Southampton, U.K. His team develops new techniques for static and moving shape extraction which have found application in automatic face and automatic gait recognition and in medical image analysis. His team were early workers in face recognition, later came to pioneer gait recognition and more recently joined the pioneers of ear biometrics. Among research contracts, he was Principal Investigator with John Carter on the Defense Advanced Research Projects Agency supported project Automatic Gait Recognition for Human ID at a Distance. He chaired British Machine Vision Conference 98 and, with Josef Kittler, chaired the Audio Visual Biometric Person Authentication 2003. He was Publications Chair for the International Conference on Pattern Recognition 2004 and the IEEE 7th International Conference on Face and Gesture Recognition 2006. His vision book, cowritten with Alberto Aguado, *Feature Extraction and Image Processing* was published in 2002 by Butterworth, and with Tieniu Tan and Rama Chellappa, his new book *Human ID based on Gait*, which is part of the new Springer Series on Biometrics, was published in 2005. His research interests are in image processing and computer vision.