

Chapter 2

Gait Representations in Video

In this chapter, we first present a spatio-temporal gait representation, called *gait energy image (GEI)*, to characterize human walking properties. Next, a general GEI-based framework is developed to deal with human motion analysis under different situations. The applications of this general framework will be discussed in the next chapter.

2.1 Human Motion Analysis and Representations

Current image-based human recognition methods, such as fingerprints, face or iris biometrics modalities, generally require a cooperative subject, views from certain aspects and physical contact or close proximity. These methods cannot reliably recognize non-cooperating individuals at a distance in the real world under changing environmental conditions. Gait, which concerns recognizing individuals by the way they walk, is a relatively new biometric without these disadvantages. However, gait also has some limitations, it can be affected by clothing, shoes, or environmental context. Moreover, special physical conditions such as injury can also change a person's walking style. The large gait variation of the same person under different conditions (intentionally or unintentionally) reduces the discriminating power of gait as a biometric and it may not be as unique as fingerprint or iris, but the inherent gait characteristic of an individual still makes it irreplaceable and useful in visual surveillance.

In traditional biometric paradigms, individuals of interest are represented by their biometric examples collected in gallery data. In general, the number of examples for each individual is limited to one in gallery data. Then feature vectors are extracted from the gallery examples by various feature extraction algorithms to construct the feature database. In the recognition procedure, the input probe example is processed in the same way to obtain the feature vector that will be compared with those in the database. The recognition decision is made according to the matching scores. This system setup is good for “strong” biometrics such as iris and fingerprint, where the

inherent discriminating features are abundance. Even if the data collection condition changes, there are still enough features to distinguish one individual from others.

This setup may not be appropriate for gait which is supposed to be a “weak” biometric. Human gait properties can be affected by various environmental conditions such as walking surface, carrying objects and environmental temperature, etc. The change of environmental conditions may introduce a large appearance change in the detected human silhouette, which may lead to a failure in recognition. The large gait variation of the same individual under different conditions requires more gallery examples collected from different environmental contexts. However, this requirement is unreal due to the complexity of real-world situations. Due to the difficulty of gait data acquisition, gait gallery examples are generally obtained under similar environmental conditions and the number of examples for each individual is also very limited.

To overcome this problem, we proposed a new gait representation, gait energy image (GEI), for human recognition [59]. Unlike other gait representations [21, 72] which consider gait as a sequence of templates, GEI represents human motion in a single image while preserving temporal information. Considering that human walking is a cyclic motion and the human walking properties are not the same in different cycles, we can obtain a series of GEIs from different cycles of human motion. Moreover, we can generate synthetic GEIs through silhouette distortion analysis. Features are further learned from the GEIs so-obtained for human recognition. In this way, the use of GEI partially overcomes the problem mentioned above and achieves good performance in human recognition by gait.

Human repetitive activity involves a regularly repeating sequence of motion events such as walking, running and jogging. Therefore, GEI can also be used to represent human repetitive activities. In this section, we propose a general GEI-based framework for both human recognition by gait and human activity recognition. Applications of the proposed general framework in different scenarios will be further discussed as case studies in the next chapter.

2.2 Human Activity and Individual Recognition by Gait

In recent years, various approaches have been proposed for human motion understanding. These approaches generally fall under two major categories: model-based approaches and model-free approaches. When people observe human walking patterns, they not only observe the global motion properties, but also interpret the structure of the human body and detect the motion patterns of local body parts. The structure of the human body is generally interpreted based on their prior knowledge. Model-based gait recognition approaches focus on recovering a structural model of human motion, and the gait patterns are then generated from the model parameters for recognition. Model-free approaches make no attempt to recover a structural model of human motion. The features used for gait representation includes: moments of shape, height and stride/width, and other image/shape templates.

2.2.1 Human Recognition by Gait

2.2.1.1 Model-Based Approaches

Niyogi and Adelson [126] make an initial attempt for gait-based recognition in a spatio-temporal (XYT) volume. They first find the bounding contours of the walker, and then fit a simplified stick model on them. A characteristic gait pattern in XYT is generated from the model parameters for recognition. Yoo et al. [193] estimate hip and knee angles from the body contour by linear regression analysis. Then trigonometric-polynomial interpolant functions are fitted to the angle sequences, and the parameters so-obtained are used for recognition. In Lee and Grimson's work [96], human silhouette is divided into local regions corresponding to different human body parts, and ellipses are fitted to each region to represent the human structure. Spatial and spectral features are extracted from these local regions for recognition and classification.

In these model-based approaches, the accuracy of human model reconstruction strongly depends on the quality of the extracted human silhouette. In the presence of noise, the estimated parameters may not be reliable. To obtain more reliable estimates, Tanawongsuwan and Bobick [160] reconstruct the human structure by tracking 3D sensors attached on fixed joint positions. However, their approach needs lots of human interaction. Wang et al. [181] build a 2D human cone model, track the walker under the Condensation framework, and extract dynamic features from different body part for gait recognition. Zhang et al. [202] use a simplified five-link biped locomotion human model for gait recognition. Gait features are first extracted from image sequences, and are then used to train hidden Markov models for recognition.

2.2.1.2 Model-Free Approaches

Moments of shape is one of the most commonly used gait features. Little and Boyd [107] describe the shape of human motion with a set of features derived from moments of a dense flow distribution. Shutler et al. [156] include velocity into the traditional moments to obtain the so-called velocity moments (VMs). A human motion image sequence can be represented as a single VM value with respect to a specific moment order instead of a sequence of traditional moment values for each frame. He and Debrunner's [65] approach detects a sequence of feature vectors based on Hu's moments of each motion segmented frame, and the individual is recognized from the feature vector sequence using hidden Markov models (HMMs) [152]. Sundarsen et al. [159] also proposed HMMs for individual.

BenAbdelkader et al. [11] use height, stride and cadence as features for human identification. Kale et al. [82, 85] choose the width vector from the extracted silhouette as the representation of gait. Continuous HMMs are trained for each person and then used for gait recognition. In their later work [83], different gait features

are further derived from the width vector and recognition is performed by a direct matching algorithm.

To avoid the feature extraction process which may reduce the reliability, Murase and Sakai [118] propose a template matching method to calculate the spatio-temporal correlation in a parametric eigenspace representation for gait recognition. Huang et al. [72] extend this approach by combining transformation based on canonical analysis, with eigenspace transformation for feature selection. BenAbdelkader et al. [10] compute the self-similarity plot by correlating each pair of aligned and scaled human silhouette in an image sequence. Normalized features are then generated from the similarity plots and used for gait recognition via eigenspace transformation. Wang et al. [180] generate the boundary distance vector from the original human silhouette contour as the template, which is used for gait recognition via eigenspace transformation. Similarly, Boulgouris et al. [24] extract gait signatures through angular analysis, and identity recognition and verification are based on the matching of time normalized walking cycles.

Han and Bhanu [59] proposed a new gait representation, called gait energy image (GEI), which represents human motion in a single image while preserving temporal information. Gait signatures are computed from original and derived gait sequences through principal component and discriminant analysis. Xu et al. [184] compute gait signatures from GEI through a matrix based analysis, named coupled subspace analysis and discriminant analysis with tensor representation, instead of conventional principal component and discriminant analysis based on vector representation. Alternative gait feature extraction methods include Radon transform and linear discriminant analysis [23], marginal Fisher analysis [185], discriminant locally linear embedding [104], general tensor discriminant analysis and Gabor features [161, 162], multilinear principal component analysis [111, 112], factorial HMM and parallel HMM [28], etc. Lam et al. [94] propose the motion silhouette contour templates (MSCTs) and static silhouette templates (SSTs), which capture the motion and static characteristic of gait, and fuse them for human recognition.

As a direct template matching approach, Phillips et al. [137, 150] measure the similarity between the gallery sequence and the probe sequence by computing the correlation of corresponding time-normalized frame pairs. Similarly, Collins et al. [32] extract key frames and the similarity between two sequences is computed from normalized correlation. Tolliver and Collins [166] cluster human silhouettes/poses of each training sequence into k shapes. In the recognition procedure, the silhouettes in a testing sequence are also classified into k prototypical shapes which are compared to prototypical shapes of each training sequence for similarity measurement. Liu and Sarkar [108] use a generic human walking model, derived from a population, to generate a dynamics normalized gait sequence. The dissimilarity between gait sequences are measured by the summation of the distance of corresponding stances in the feature space, which is transformed from the silhouette space by principal component analysis and linear discriminant analysis.

2.2.2 Human Activity Recognition

2.2.2.1 Model-Based Approaches

Guo et al. [56] represent the human body structure in the silhouette by a stick figure model. The human motion characterized by a sequence of the stick figure parameters is used as the input to a neural network for classification. Fujiyoshi and Lipton [46] analyze the human motion by producing a star skeleton, determined by extreme point estimation, obtained from the extracted silhouette boundaries. These cues are used to recognize human activities such as walking or running. Sappa et al. [149] develop a technique for human motion recognition based on the study of feature points' trajectories. Peaks and valleys of points' trajectories are first detected to classify human activity using prior knowledge of human body kinematics structure together with the corresponding motion model. In model-based approaches, the accuracy of human model reconstruction strongly depends on the quality of the extracted human silhouette. In the presence of noise, the estimated parameters may not be reliable.

2.2.2.2 Model-Free Approaches

Polana and Nelson [139] analyze human repetitive motion activity based on bottom up processing, which does not require the prior identification of specific parts. Motion activity is recognized by matching against a spatio-temporal template of motion features. Rajagopalan and Chellappa [141] develop a higher-order spectral analysis-based approach for detecting people by recognizing repetitive motion activity. The stride length is determined in every frame, and the bispectrum which is the Fourier transform of the triple correlation is used for recognition. Bobick and Davis [21] propose motion-energy image (MEI) and motion-history image (MHI) for human movement type representation and recognition. Both MEI and MHI are vector-images where the vector value at each pixel is a function of the motion properties at this location in an image sequence. Vega and Sarkar [172] discriminate between motion types based on the change in the relational statistics among the detected image features. They use the distribution of the statistics of the relations among the features for recognition. Davis [39] proposes a probabilistic framework to address the issue of rapid-and-reliable detection of human activities using posterior class ratios to verify the saliency of an input before committing to any activity classification.

2.3 Gait Energy Image (GEI) Representation

In this section, we only consider individual recognition by activity-specific human motion, i.e., regular human walking, which is used in most current approaches of individual recognition by gait.

2.3.1 Motivation

Regular human walking can be considered as cyclic motion where human motion repeats at a stable frequency. While some gait recognition approaches [72] extract features from the correlation of all the frames in a walking sequence without considering their order, other approaches extract features from each frame and compose a feature sequence for the human walking sequence [32, 107, 150]. During the recognition procedure, these approaches either match the statistics collected from the feature sequence, or match the features between the corresponding pairs of frames in two sequences that are time-normalized with respect to their cycle lengths. The fundamental assumptions made here are: (a) the order of poses in human walking cycles is the same, i.e., limbs move forward and backward in a similar way among normal people; (b) differences exist in the phase of poses in a walking cycle, the extent of limbs, and the shape of the torso, etc. Under these assumptions, it is possible to represent the spatio-temporal information in a single 2D gait template instead of an ordered image sequence.

2.3.2 Representation Construction

We assume that silhouettes have been extracted from original human walking sequences. The silhouette extraction and preprocessing procedure will be introduced in Sect. 2.4.1 in detail. Given the normalized and aligned binary gait silhouette images $B_t(x, y)$ at time t in a sequence, the grey-level gait energy image (GEI) is defined as follows

$$G(x, y) = \frac{1}{N} \sum_{t=1}^N B_t(x, y), \quad (2.1)$$

where N is the number of frames in the complete cycle(s) of a silhouette sequence, t is the frame number in the sequence (moment of time), x and y are 2D image coordinates. Figure 2.1 shows the sample silhouette images in a gait cycle from two people, and the right most image is the corresponding GEI. As expected, GEI reflects major shapes of silhouettes and their changes over the gait cycle. We refer to it as gait energy image because: (a) each silhouette image is the space-normalized energy image of human walking at this moment; (b) GEI is the time-normalized accumulative energy image of human walking in the complete cycle(s); (c) a pixel with higher intensity value in GEI means that human walking occurs more frequently at this position (i.e., with higher energy).

2.3.3 Relationship with MEI and MHI

Bobick and Davis [21] propose motion-energy image (MEI) and motion-history image (MHI) for human movement recognition. Both MEI and MHI are vector-images

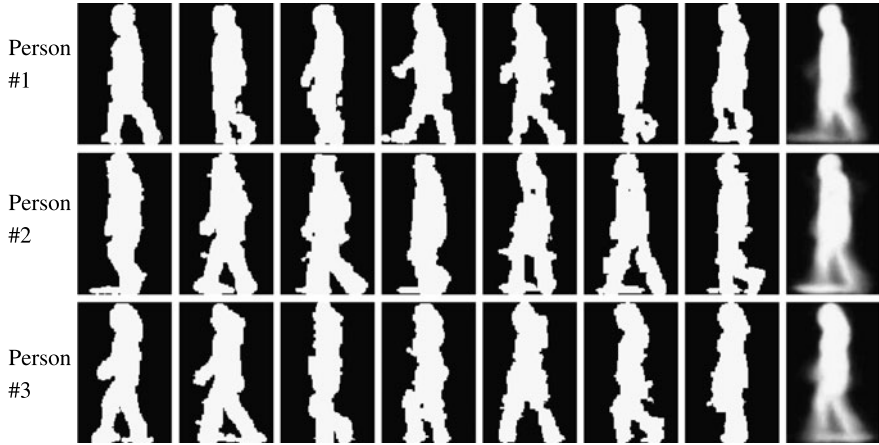


Fig. 2.1 Examples of normalized and aligned silhouette frames in different human walking sequences. The rightmost image in each row is the corresponding gait energy image (GEI)

where the vector value at each pixel is a function of the motion properties at this location in an image sequence. As compared to MEI and MHI, GEI targets specific normal human walking representation and we use GEI as the gait template for individual recognition.

MEI is a binary image which represents where motion has occurred in an image sequence:

$$E_{\tau}(x, y, t) = \bigcup_{i=0}^{\tau-1} D(x, y, t - i), \quad (2.2)$$

where $D(x, y, t)$ is a binary sequence indicating regions of motion, τ is the duration of time, t is the moment of time, x and y are values of 2D image coordinate. To represent a regular human walking sequence, if $D(x, y, t)$ is normalized and aligned as $B(x, y, t)$ in (2.1), MEI $E_N(x, y, N)$ is the binary version of GEI $G(x, y)$.

MHI is a grey-level image which represents how motion in the image is taking place:

$$H_{\tau}(x, y, t) = \begin{cases} \tau, & \text{if } D(x, y, t) = 1; \\ \max\{0, H_{\tau}(x, y, t - 1) - 1\}, & \text{otherwise.} \end{cases} \quad (2.3)$$

2.3.4 Representation Justification

In comparison with the gait representation by binary silhouette sequence, GEI representation saves both storage space and computation time for recognition and is less sensitive to silhouette noise in individual frames. Consider a noisy silhouette image $B_t(x, y)$ that is formed by the addition of noise $\eta_t(x, y)$ to an original silhouette image $f_t(x, y)$, that is, $B_t(x, y) = f_t(x, y) + \eta_t(x, y)$, where we assume that

at every pair of coordinates (x, y) the noise at different moments t is uncorrelated and identically distributed. Under these constraints, we further assume that $\eta_t(x, y)$ satisfies the following distribution:

$$\eta_t(x, y) = \begin{cases} \eta_{1t}(x, y): P\{\eta_t(x, y) = -1\} = p, P\{\eta_t(x, y) = 0\} = 1 - p, \\ \quad \text{if } f_t(x, y) = 1, \\ \eta_{2t}(x, y): P\{\eta_t(x, y) = 1\} = p, P\{\eta_t(x, y) = 0\} = 1 - p, \\ \quad \text{if } f_t(x, y) = 0. \end{cases} \quad (2.4)$$

We have

$$E\{\eta_t(x, y)\} = \begin{cases} -p, & \text{if } f_t(x, y) = 1, \\ p, & \text{if } f_t(x, y) = 0 \end{cases} \quad (2.5)$$

and

$$\sigma_{\eta_t(x, y)}^2 = \sigma_{\eta_{1t}(x, y)}^2 = \sigma_{\eta_{2t}(x, y)}^2 = p(1 - p). \quad (2.6)$$

Given a walking cycle with N frames where $f_t(x, y) = 1$ at a pixel (x, y) only in M frames, we have

$$\begin{aligned} G(x, y) &= \frac{1}{N} \sum_{t=1}^N B_t(x, y) = \frac{1}{N} \sum_{t=1}^N f_t(x, y) + \frac{1}{N} \sum_{t=1}^N \eta_t(x, y) \\ &= \frac{M}{N} + \bar{\eta}(x, y). \end{aligned} \quad (2.7)$$

Therefore, the noise in GEI is

$$\bar{\eta}(x, y) = \frac{1}{N} \sum_{t=1}^N \eta_t(x, y) = \frac{1}{N} \left[\sum_{t=1}^M \eta_{1t}(x, y) + \sum_{t=M+1}^N \eta_{2t}(x, y) \right]. \quad (2.8)$$

We have

$$\begin{aligned} E\{\bar{\eta}(x, y)\} &= \frac{1}{N} \left[\sum_{t=1}^M E\{\eta_{1t}(x, y)\} + \sum_{t=M+1}^N E\{\eta_{2t}(x, y)\} \right] \\ &= \frac{1}{N} [M(-p) + (N - M)p] = \frac{N - 2M}{N} p \end{aligned}$$

and

$$\begin{aligned} \sigma_{\bar{\eta}(x, y)}^2 &= E\{[\bar{\eta}(x, y) - E\{\bar{\eta}(x, y)\}]^2\} \\ &= \frac{1}{N^2} E\left\{ \left[\sum_{t=1}^M [\eta_{1t}(x, y) - E\{\eta_{1t}(x, y)\}] \right. \right. \\ &\quad \left. \left. + \sum_{t=M+1}^N [\eta_{2t}(x, y) - E\{\eta_{2t}(x, y)\}] \right]^2 \right\} \\ &= \frac{1}{N^2} [M\sigma_{\eta_{1t}(x, y)}^2 + (N - M)\sigma_{\eta_{2t}(x, y)}^2] = \frac{1}{N} \sigma_{\eta_t(x, y)}^2. \end{aligned}$$

Therefore, the mean of the noise in GEI varies between $-p$ and p depending on M while its variability ($\sigma_{\bar{\eta}(x,y)}^2$) decreases. If $M = N$ at (x, y) (all $f_t(x, y) = 1$), $E\{\bar{\eta}(x, y)\}$ becomes $-p$; if $M = 0$ at (x, y) (all $f_t(x, y) = 0$), $E\{\bar{\eta}(x, y)\}$ becomes p . At the location (x, y) , the mean of the noise in GEI is the same as that in the individual silhouette image, but the noise variance reduces so that the probability of outliers is reduced. If M varies between 0 and N at (x, y) , $E\{\bar{\eta}(x, y)\}$ also varies between p and $-p$. Therefore, both the mean and the variance of the noise in GEI are reduced compared to the individual silhouette image at these locations. At the extreme, the noise in GEI has zero mean and reduced variance where $M = N/2$. As a result, GEI is less sensitive to silhouette noise in individual frames.

2.4 Framework for GEI-Based Recognition

In this section, we describe a GEI-based general framework for the purpose of recognition. First, human silhouettes are extracted and processed to detect the base frequency and phase for each frame. GEIs are then constructed from difference human motion cycles according to different recognition purpose. Next, features are learned from training GEI templates by statistical analysis. The recognition is performed by matching features between probe and gallery templates. The system diagram is shown in Fig. 2.2.

2.4.1 Silhouette Extraction and Processing

The raw silhouettes are extracted by a simple background subtraction method. This method is relatively easy to implement and requires less computational time. Once a Gaussian model of the background is built using sufficiently large number of frames, given a foreground frame, for each pixel, we can estimate whether it belongs to the background or foreground by observing the pixel value compared with the mean and standard deviation values at this pixel. The raw silhouettes so-obtained are further processed by size normalization (proportionally resizing each silhouette image so that all silhouettes have the same height) and horizontal alignment (centering the upper half silhouette part with respect to its horizontal centroid) [150]. In a silhouette sequence so-obtained, the time series signal of lower half silhouette part

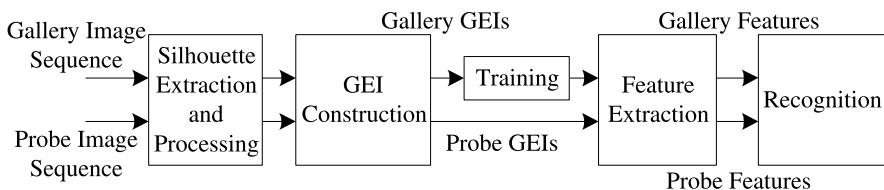
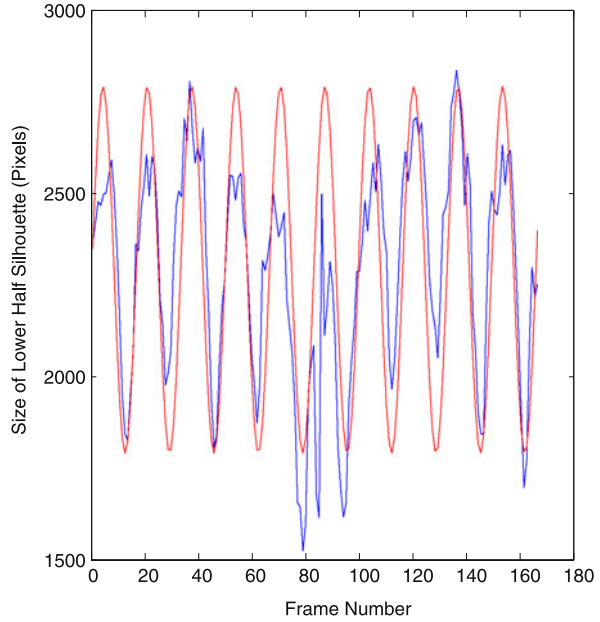


Fig. 2.2 Diagram of the proposed GEI-based framework for the purpose of recognition

Fig. 2.3 Frequency and phase estimation of human walking



size from each frame indicates the motion frequency and phase information. The obtained time series signal consists of few cycles and lots of noise, which lead to sidelobe effect in the Fourier spectrum. To avoid this problem, we estimate the motion frequency and phase by maximum entropy spectrum estimation [107] from the obtained time series signal as shown in Fig. 2.3.

2.4.2 Feature Extraction

Once we obtain a series of training GEI templates for each person (class), the problem of their excessive dimensionality occurs. There are two classical linear approaches for finding transformations for dimensionality reduction—principal component analysis (PCA) and multiple discriminant analysis (MDA) that have been effectively used in face recognition [9]. PCA seeks a projection that best represents the data in the least-squares sense, while MDA seeks a projection that best separates the data in the least-squares sense. Huang et al. [72] combine PCA and MDA to achieve the best data representation and the best class separability simultaneously. In our approach, the learning procedure follows this combination approach.

Given n d -dimensional training GEI templates $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, PCA minimizes the function

$$J_{d'} = \sum_{k=1}^n \left\| \left(\mathbf{m} + \sum_{i=1}^{d'} a_{ki} \mathbf{e}_i \right) - \mathbf{x}_k \right\|^2, \quad (2.9)$$

where $d' < d$, $\mathbf{m} = \frac{1}{n} \sum_{k=1}^n \mathbf{x}_k$, and $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{d'}\}$ is a set of unit vectors. $J_{d'}$ is minimized when $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{d'}$ are the d' eigenvectors of the scatter matrix S having the largest eigenvalues, where

$$S = \sum_{k=1}^n (\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^T. \quad (2.10)$$

The d' -dimensional feature vector \mathbf{y}_k is obtained from the \mathbf{x}_k as follows

$$\mathbf{y}_k = M_{\text{pca}} \mathbf{x}_k = [a_1, \dots, a_{d'}]^T = [\mathbf{e}_1, \dots, \mathbf{e}_{d'}]^T \mathbf{x}_k, \quad k = 1, \dots, n. \quad (2.11)$$

Suppose that the n d' -dimensional principal component vectors $\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n\}$ belong to c classes. MDA seeks a transformation matrix W that maximizes the ratio of the between-class scatter matrix S_B to the within-class scatter matrix S_W :

$$J(W) = \frac{|\tilde{S}_B|}{|\tilde{S}_W|} = \frac{|W^T S_B W|}{|W^T S_W W|}. \quad (2.12)$$

The within-class scatter S_B is defined as

$$S_W = \sum_{i=1}^c S_i, \quad (2.13)$$

where

$$S_i = \sum_{\mathbf{y} \in \mathcal{D}_i} (\mathbf{y} - \mathbf{m}_i)(\mathbf{y} - \mathbf{m}_i)^T \quad (2.14)$$

and

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{y} \in \mathcal{D}_i} \mathbf{y}, \quad (2.15)$$

where \mathcal{D}_i is the training template set that belongs to the i th class and n_i is the number of templates in \mathcal{D}_i . The within-class scatter S_B is defined as

$$S_B = \sum_{i=1}^c n_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T, \quad (2.16)$$

where

$$\mathbf{m} = \frac{1}{n} \sum_{\mathbf{y} \in \mathcal{D}} \mathbf{y}, \quad (2.17)$$

and \mathcal{D} is the whole training template set. $J(W)$ is maximized when the columns of W are the generalized eigenvectors that correspond to the largest eigenvalues in

$$S_B \mathbf{w}_i = \lambda_i S_W \mathbf{w}_i. \quad (2.18)$$

There are no more than $c - 1$ nonzero eigenvalues, and the corresponding eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_{c-1}$ form the transformation matrix. The $(c - 1)$ -dimensional feature vector \mathbf{z}_k is obtained from the d' -dimensional principal component vector \mathbf{y}_k :

$$\mathbf{z}_k = M_{\text{mda}} \mathbf{y}_k = [\mathbf{v}_1, \dots, \mathbf{v}_{c-1}]^T \mathbf{y}_k, \quad k = 1, \dots, n. \quad (2.19)$$

For each training gait template, its gait feature vector is obtained as follows

$$z_k = M_{\text{mda}} M_{\text{pca}} x_k = T x_k, \quad k = 1, \dots, n. \quad (2.20)$$

The obtained feature vectors represent the n templates for individual recognition.

2.5 Summary

In this chapter, we presented a new spatio-temporal gait representation, called the gait energy image (GEI). Unlike other gait representations which consider gait as a sequence of templates (poses), GEI represents human motion sequence in a single image while preserving temporal information. Moreover, we described a general GEI-based framework for human motion analysis. Applications of the proposed general framework in different motion scenarios will be further discussed, as case studies, in detail in the next chapter.



<http://www.springer.com/978-0-85729-123-3>

Human Recognition at a Distance in Video

Bhanu, B.; Han, J.

2010, XXV, 253 p., Hardcover

ISBN: 978-0-85729-123-3