Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2012, Article ID 563864, 35 pages
doi:10.1155/2012/563864

Research Article

Automatic Human Gait Imitation and Recognition in 3D from Monocular Video with an Uncalibrated Camera

Tao Yu[1,2] and Jian-Hua Zou[1,2]

[1] *Systems Engineering Institute, School of Electronic & Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China*
[2] *State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China*

Correspondence should be addressed to Tao Yu, yvt9399@stu.xjtu.edu.cn

Received 26 September 2011; Accepted 16 December 2011

Academic Editor: Yun-Gang Liu

A framework of imitating real human gait in 3D from monocular video of an uncalibrated camera directly and automatically is proposed. It firstly combines polygon-approximation with deformable template-matching, using knowledge of human anatomy to achieve the characteristics including static and dynamic parameters of real human gait. Then, these characteristics are processed in regularization and normalization. Finally, they are imposed on a 3D human motion model with prior constrains and universal gait knowledge to realize imitating human gait. In recognition based on this human gait imitation, firstly, the dimensionality of time-sequences corresponding to motion curves is reduced by NPE. Then, we use the essential features acquired from human gait imitation as input and integrate HCRF with SVM as a whole classifier, realizing identification recognition on human gait. In associated experiment, this imitation framework is robust for the object's clothes and backpacks to a certain extent. It does not need any manual assist and any camera model information. And it is fitting for straight indoors and the viewing angle for target is between 60° and 120°. In recognition testing, this kind of integrated classifier HCRF/SVM has comparatively higher recognition rate than the sole HCRF, SVM and typical baseline method.

1. Introduction

Recently, the investigation on human gait is receiving more and more attention. Especially, in monitoring, human gait is the only one characteristic that can be recognized in long distance without contacting. Hence, the study on it lies at an important status in safeguard area. In medicine, human gait study is valuable in relative diagnosis and modifying bones for patients.

As an important branch of human gait study, human gait imitation does not only help to investigate deeply in real motion—it can acquire the valuable cue of sight characteristic that is hard to get from the raw images directly, making the results more clear. (e.g., some characteristics on human body cannot be detected continually because of the sheltering or by themselves as human is motioning. By the means of imitating human gait, we can adjust the relative motion model at any viewing angle, any time to detect and analyze)—but also owns broad scope of application in virtual reality and 3D TV.

At present, most studies on imitating motion use multicamera to realize reconstruction of motion in 3D [1–5]. These methods share obvious style: solving the problem of sheltering directly in object's motion, and the results are remarkable for various poses of human motion at the expense of complex camera model information and complex computations.

Also, there are many attempts to restore human motion in 3D or to track human motion from monocular video sequence directly. However, most do not only need training to learn and estimate poses but also need manual help originally. For example, [6] needs to initialize the 2D parameter values for the first frame manually by overlaying a model onto the 2D image of the first frame. Reference [7] sets initialization by matching the first frame to six key poses acquired by manual clustering, and the pose having minimal matching error is chosen as the initial pose. Reference [8] involves some prior information which was extracted from some hand-labeled data. In [9], the local model needs information of contour extracted manually from real images.

Some need the information of relative camera, such as [10] which recovers 3D model of humans from just one frame or a monocular video sequence using a simplification of the camera model based on a collinearity condition. In [11], the webcam used needs being calibrated at the phase of preprocessing. Reference [12] needs both to estimate camera information and to locate joints manually in the first frame of video sequences.

Some have other too many demands or constraints. For example, [13] subjects to not only Gaussian prior and Gaussian stabilizers but also the objective time-consuming based on its covariance-scaled sampling. Reference [14] needs to detect pedestrian's motion trajectory and footprints throughout the segmented video sequence by associated clustering technique. [15, 16] need both relative information of edges and textured regions. And [17] needs to preextract state subspace from one sequence of motion capture data for each motion type.

We can see that most motion imitations, whether using multicamera or using single camera, based on numerous continuous characteristics at the cost of expensive accurate instruments or accurate camera model information or manual assist or other extra demands. Are these demands necessary?

For these doubts above, this paper presents a framework. It firstly combines polygon-approximation with deformable template-matching, using knowledge of human anatomy to achieve the characteristics including static and dynamic parameters of real human gait. Then, these characteristics are processed in regularization and normalization. Finally, they are imposed on 3D human model with prior constrains and universal gait knowledge to imitate real human gait, thus realizing the reconstruction of 3D human gait from monocular video directly and automatically. The method is robust for detecting subject's clothes and backpack to a certain extent at 60° ~120° angle of view in stable straight walking in test of CAISIA gait database. Moreover, it does not need any manual assist and any camera model information. In application of this framework, with the dimensionality of time sequences corresponding to curves of human gait imitation reduced by the method of neighborhood preserving embedding (NPE), a kind of classifier which integrates HCRF (hidden conditional random field) with SVM (supported vector machine) using the essential time sequence

acquired from the human gait imitation as input to realize human identification recognition on gait and presents a higher recognition rate than using HCRF or SVM as classifier alone and the typical baseline method in associated experiment as it contains more structural traits of the data to be classified in space and time during dealing with them.

The remainder of this paper is arranged as follows. Section 2 interprets the basic idea and principle of the framework on gait imitation; Section 3 describes the realization of the method in detail; Section 4 gives the application of human gait imitation in identification recognition; Section 5 provides relative experiment and results analysis; Section 6 concludes this paper.

2. Principle

We all know that when people begin to know something, they used to compare the new thing with some general mode. Then, they will carve the mode in detail by the use of characteristics of the real thing, that is, to say, imposing individuation to the mode. Thus, a complete impression on the new thing is formed in their minds, finally.

Here, analogously, during the course of knowing human gait, we have a prior general human motion model firstly. Secondly, we acquire the individual characteristics of some real gait in some way. Then, for the characteristics detected, we update the partial prior values of the relative part in the human model; for the characteristics undetected, we use the prior configuration of the relative part in the human model to compensate in proportion. Thus, integrating the real individual characteristics with prior general information, we form a sufficient and complete expression on the impression of human gait.

This idea of acquiring characteristics heuristically overcomes the phenomenon of sheltering in human motion. It maintains the continuity of motion detection to some extent and has strong adaptation to the situation of some characteristics hard to detect.

The basic principle for the method of gait imitation above can be described in mathematics as both the posterior probability function $P(m, w(m) \mid c)$ in formula (2.1) and entropy function $H(X)$ in formula (2.2). Here, $P(m)$ is the prior knowledge of the human model m ($\in R(m)$) corresponding to some person, $P(w(m) \mid m)$ is the prior knowledge on gait $w(m)$ ($\in R(w)$), likelihood function $P(c \mid w(m), m)$ is the individual characteristic information c ($\in R(c)$) on gait and random variable X whose range is a group of state sequence on gait $\{x_1, x_2, \ldots, x_m\}$ representing gait imitation:

$$P_t(m, w(m) \mid c) = \frac{p(c \mid m, w(m))p(m, w(m))}{p(c)} = \frac{p(c \mid m, w(m))p(w(m) \mid m)p(m)}{\sum_i p\left(c \mid m^i, w(m)^i\right)p\left(w(m)^i \mid m^i\right)p(m^i)}, \tag{2.1}$$

$$H(X) = \sum_{t=1}^{m}\left(P_t \log\left(\frac{1}{P_t}\right)\right) \quad \text{where: } X \in \{x_1, x_2, \ldots, x_m\}. \tag{2.2}$$

In formula (2.1), according to Bayesian rules, at time of t, basing on a certain prior information (note here the prior information is general or universal for most persons) on $P(m)$ and $P(w(m) \mid m)$, the more remarkable the individual characteristic information $P(c \mid w(m), m)$ on gait is, the bigger the value of posterior probability function $P(m, w(m) \mid c)$ will be. Then, at this time, in formula (2.2), according to information theory [18, 19], the value of

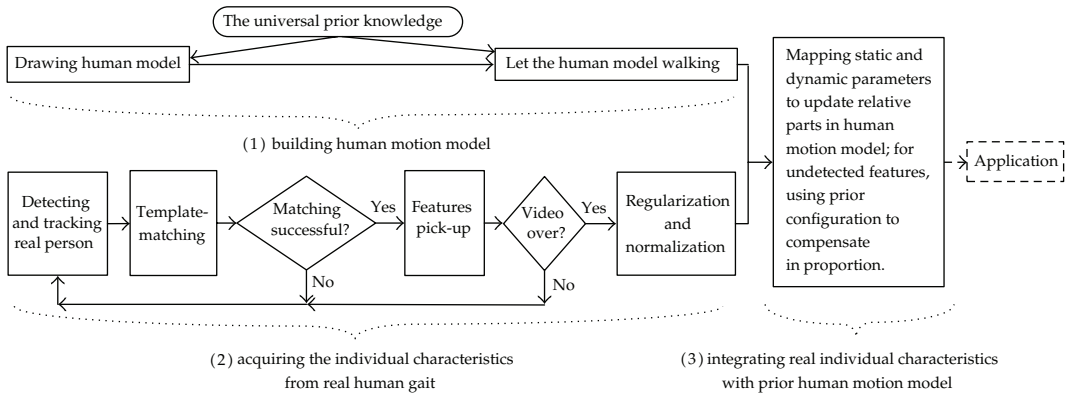


Figure 1: Framework of out approach.

the relative entropy function $H(X)$ will be smaller, showing that the next state closed to the real will be more ensured. At other times, similar analysis is as above. Thus, as a result, the whole imitation will more approximate a real person's gait.

3. Designing

Before designing, we assume that the human gait to be detected is walking in a straight line with a constant speed. According to the idea of Section 2, the designing includes three parts mainly, which is displayed in Figure 1. In the first part—building the human motion model, the universal prior knowledge for drawing human model and letting the human model walking is equivalent to $P(m)$ and $P(w(m) \mid m)$, respectively. The result of second part—acquiring the individual characteristics from real human gait—is equivalent to $P(c \mid w(m), m)$. Both the former two parts based on the knowledge of human anatomy, dynamics and kinematics. Besides, the second part uses mainly a method that combines polygon- approximation and deformable template-matching to track and detect motional object. The third part that integrates real individual characteristics with prior human motion model is equivalent to the principle of formula (2.1) and (2.2), which is realized mainly on updating partial both static and dynamic parameters as well as relative compensations.

3.1. Building Human Model in Prior Knowledge

3.1.1. Drawing for the Whole Human Model

Referring to the standard of H-Anim [20] in VRML (virtual reality modeling language) [21], the human model of this paper is illustrated in Figure 2. Figure 3 gives the relative 3D human model. The general structure of the 3D human model in this paper, with the crotch of the human model in the center of the world coordinates, consists of three parts: trunk, upper limbs, and lower limbs. As can be seen, except for head, hands, and feet, the basic cell is that a sphere plus a cylinder side by just as Figure 4. The cylinder denotes bone, and the sphere denotes joint of bones in the human model. The three parts above are connected with such cells in different direction. In detail, in each joint of the human model, we build a local coordinate to make its z-axis in the same direction of the axis of the next bone. All

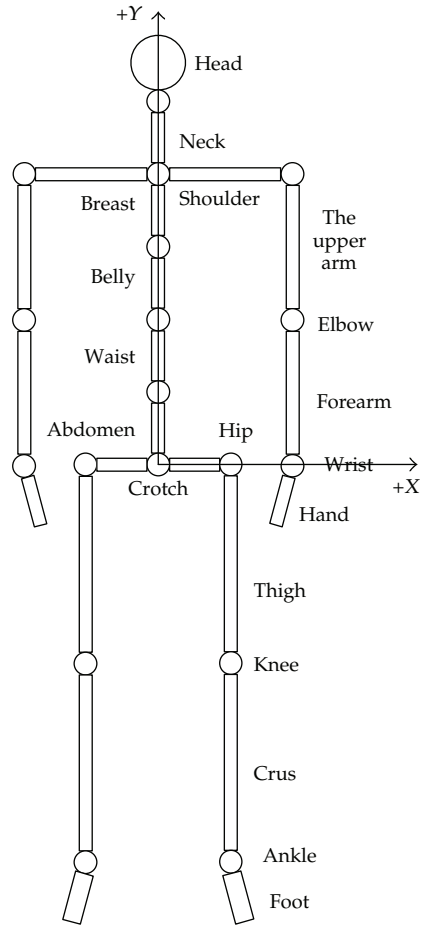


Figure 2: 2D human model.

the parts above begin from the origin of the world coordinates. Firstly, we configure the pose of the local coordinates. Then, we draw a cell in the local coordinates. Next, we move the local coordinates to the end of the cell. Again, we configure the pose of the local coordinates and draw the next cell. Thus, we take that turn repeating until all the parts are drawn over finally. That is shown in Figure 5, where N equals the total number of the cells in each part above. $T_{\text{rotate\_}x}$, $T_{\text{rotate\_}y}$, and $T_{\text{rotate\_}z}$ stand for the rotating matrixes in x, y, and z axes, respectively. $T_{\text{translate\_}x}$, $T_{\text{translate\_}y}$, $T_{\text{translate\_}z}$ stand for the translating matrixes in x, y, and z axes, respectively.

Consequently, combining Figure 2 with Figure 3 again, the designing for trunk begins from crotch up, consisting of 6 segments: head, neck, chest, belly, waist, and abdomen. The designing for upper limbs begins from neck down, consisting of 4 segments: shoulders, the upper arms, forearms, and hands. (Here, from crotch to neck, we do not draw any cell, but configure the pose of the local coordinates and move the local coordinates because we have drawn the trunk at first.) The designing for lower limbs begins from crotch down, consisting of 4 segments: hip, thigh, calf, and feet. The last cell of each part above should be replaced with relative part of head, hand, and foot. That is sphere or cuboid. Thus, the human model has 22 segments overall, each of which has 3 degrees of freedom (DOFs). Plus additional

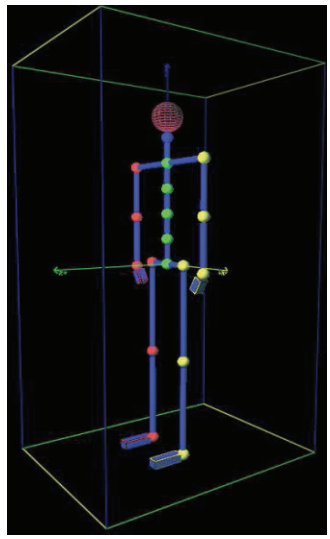


Figure 3: 3D human model.

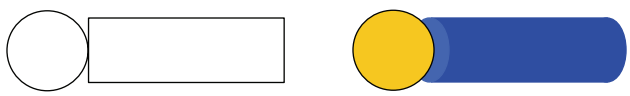


Figure 4: Basic cell in model structure.

3 DOFs for global rotation, so there are $22 \times 3 + 3 = 69$ DOFs for the whole human model. Because hands and feet themselves affect human pose little in motion, by simplification, the overall number of DOFs is $69 - 4 \times 3 = 57$.

When the human model above is in motion, the display of motion for the model is realized on adjusting poses of the cells in the model and updating the display continuously when we draw the cells.

3.1.2. The Universal Prior Knowledge for the Human Motion Model

According to the knowledge of human anatomy [22–24], combining with the H-Anim standard, if we assume the model's height is H, then: head occupies $0.13H$; shoulders' width is $0.26H$; the width of the hip is $0.19H$; when the upper limbs are horizontal, the width between left and right hand equals H; the midpoints of the lower limbs correspond to the knees of the model; the lengths of the upper arm, forearm, and hand are $0.19H$, $0.15H$, and $0.11H$, respectively.

When the model is in walking mode, according to the knowledge of human dynamics and kinematics [25–30], as Figure 6 displaying, a normal walking mode can be described as follows.

The upper limbs swing backwards and forwards on the left and right sides in turn. At one time, so do the lower limbs. The phrases of the motion for the upper limbs and the lower limbs are in contrary. That is, when the upper limbs swing forward on the left side, the lower limbs swing forward on the right side, and when the upper limbs swing forward on the right side, the lower limbs swing forward on the left side, thus to and fro. During walking,

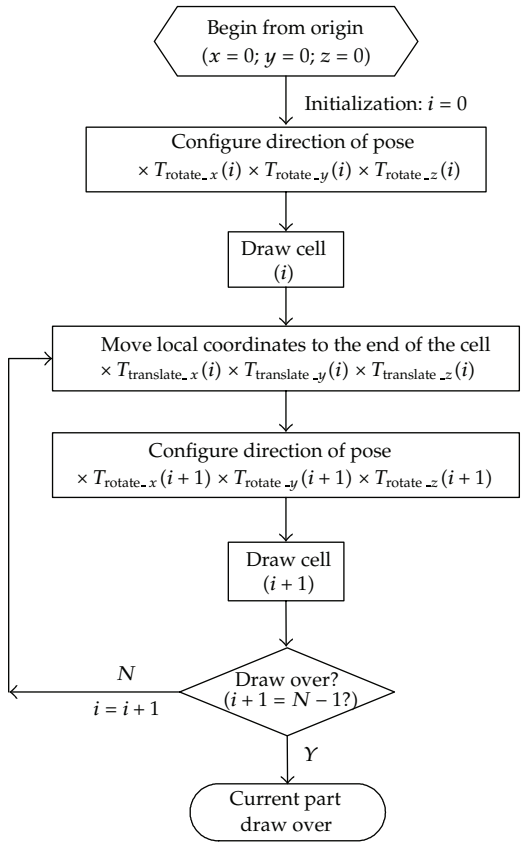


Figure 5: Flow chart for designing every part of the human model.

the neck is slightly bending forward. And especially, the elbows and knees are both in cycles of slight bending and stretching. Simultaneously, the shoulders and hips, corresponding to relative limbs, are also swinging slightly back and forth. With the limbs in motion, the trunk, including head are also swinging wiggly, turning clockwise and anticlockwise, slightly. In addition, there are some constrains: the elbows can bend forward only; the knees can bend backward only, the extent of bending in the elbow increases to maximum when the relative part of upper limbs swings to front end and decreases to minimum when the relative part of upper limbs swings back end; the extent of bending in the knee decreases to minimum when the relative part of lower limbs swings to front end and increases to maximum when the relative part of lower limbs swings back end.

The prior parameters above on structure and motion in model will be updated partially after the actual parameters of some real individual are acquired.

3.2. Acquiring the Individual Characteristics from Motional Object in Reality

3.2.1. Detecting and Tracking Motional Object

This paper processes the video images of real human gait using graying, background modeling and updating, background subtracting, binary conversion, and binary morphological

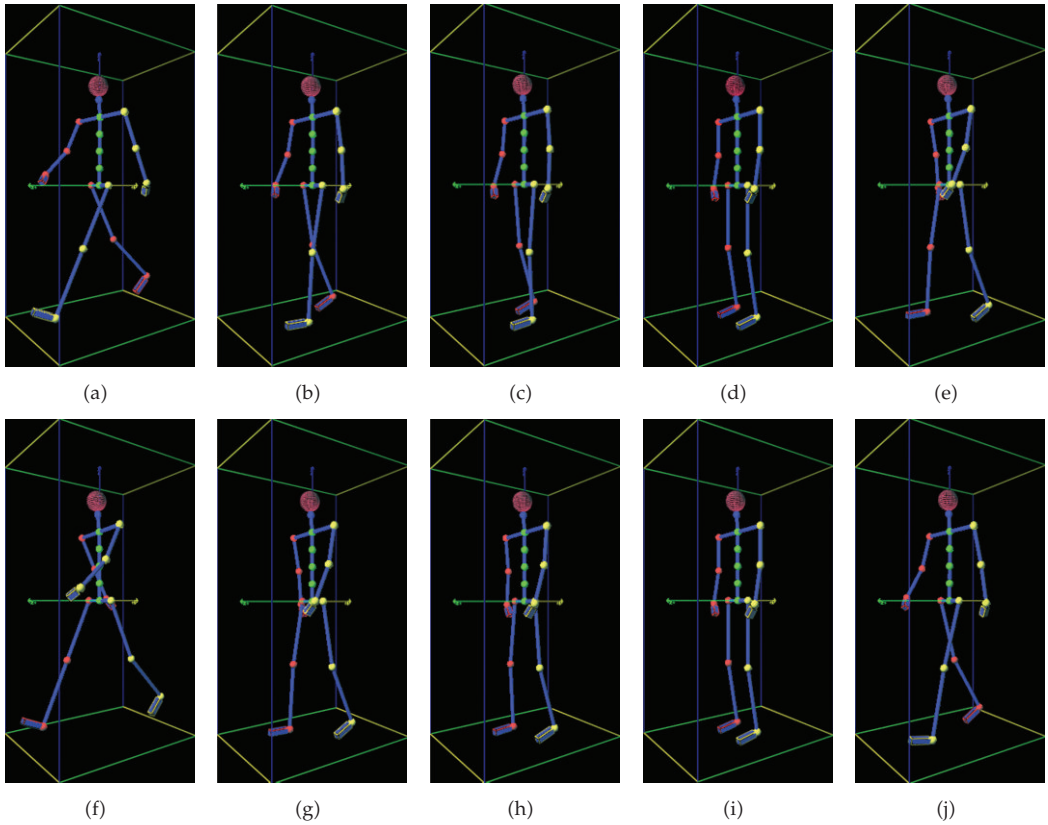


Figure 6: A normal walking mode for the human model.

operation in turn to acquire the actual motional object firstly. Secondly, consulting [31], we use a method of connectivity to retrieve all the contours and reconstruct the full hierarchy of nested contours of motional object. Then, we compress horizontal, vertical, and diagonal segments, leaving only their ending points. Thus, we acquire the external contour of the actual motional object. Finally, referring to [32], the polygon curve is approximated with assigned accuracy. When the overall number of vertexes of the polygon is 4, we begin to carry on a deformable template matching to acquire the individual characteristics of the real human gait. The principle of deformable template-matching is as in Figure 7.

3.2.2. The Principle of Deformable Template-Matching

In Figure 7, the points a, b, c, and d correspond to the vertex of head, crotch, left ankle and right ankle, respectively. Assume that their coordinates are (Xa, Ya), (Xb, Yb), (Xc, Yc), and (Xd, Yd), respectively. So the constraining rules are

$$\begin{gathered} Ya > Yb > Yc, Yd, \\ Xc < Xb < Xd. \end{gathered} \tag{3.1}$$

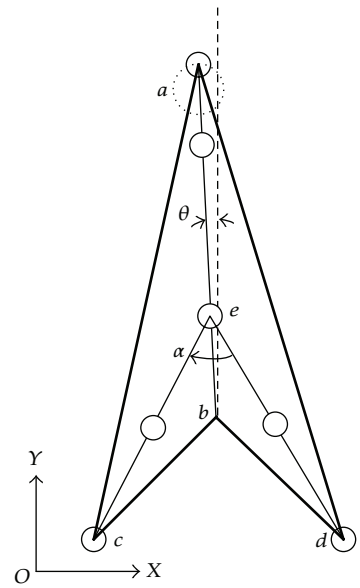


Figure 7: Principle of deformable template-matching.

We can infer that only when the space between lower limbs of the object is the biggest or so, the deformable template-matching can be in effect. In other words, at the moment, the matching is successful probably.

According to the knowledge of human anatomy [22–24], in Figure 7, if we assume the length from the crotch to the top of head in human is h, the neck is about 0.75 h away from the crotch, and the waist is 0.30 h away from the crotch or so. Then, we connect the waist with left ankle and right ankle respectively, so the middle position of the connected parts can be known as knees. Thus, we confirm the motion position of the lower limbs approximately at the moment. Because of the variety of the upper limbs' motion, moreover, some parts of the upper limbs have little effect on gait, the estimation of the structure and motion for the upper limbs is achieved by proportion to the prior human motion model completely. Thus, all the signs above on human object's body are marked automatically, not requiring any manual work, once the deformable template-matching is successful.

Figure 8 shows the actual form of deformable template-matching. Because the prior model sets the crotch and hip in the same horizontal line by simplification and in motion, the model itself has been added with the slight bend of elbows, knees and neck in proportion as universal prior knowledge before imitating real human gait in 3D, consequently, the errors that the measured lengths for limbs are probably shorter than the real lengths because of the bends of real joints can be compensated to a certain extent. This strategy saves processing time effectively.

3.2.3. Regularization and Normalization for Detected Characteristics

Because there is some information on depth-variance for the motion of detected object, to the same part in reality, the lengths detected at the different time, different position differ in values possibly**.** Hence, in order to be uniform approximately, the lengths detected need to

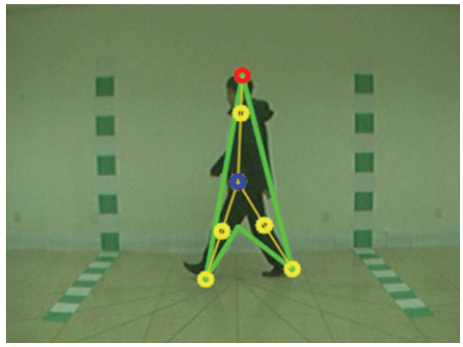


Figure 8: Actual form of deformable template-matching.

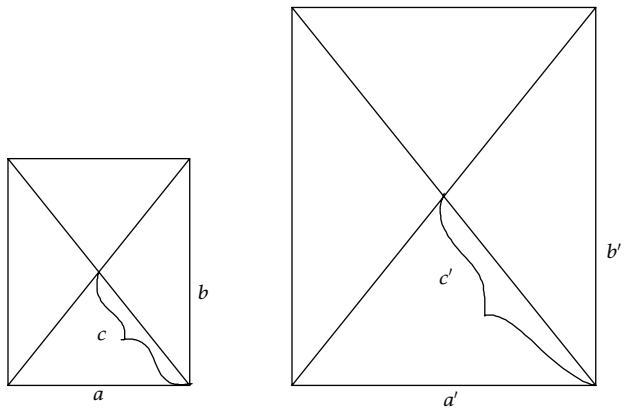


Figure 9: Principle of regularization.

be regularized and normalized. The principle of regularization in this paper is illustrated in Figure 9:

$$\frac{c}{\sqrt{ab}} = \frac{\sqrt{cc}}{\sqrt{ab}} = \sqrt{\frac{c}{a}} \times \sqrt{\frac{c}{b}} = \sqrt{\frac{c'}{a'}} \times \sqrt{\frac{c'}{b'}} = \frac{\sqrt{c'c'}}{\sqrt{a'b'}} = \frac{c'}{\sqrt{a'b'}}. \tag{3.2}$$

In Figure 9, the two rectangles are alike in shape and different in size. We assume that this phenomenon results from the same object being detected in different distances or depths from view. The line segments c and c' differ in direct measurement, but according to the analysis of formula (3.2), as long as we divide them by the square root of area of the relative smallest circumscribable rectangle (or other circumscribable shapes), the two different measurements in directivity will be transformed into equivalent forms, that is, the changed forms can act as the expressions of real size. Figure 10 shows the form of actual deformable template-matching with circumscribable rectangle for real motional object.

This paper regularizes the lengths of every part of detected object in the light of the principle above. Then, we divide the accumulated regularized lengths of each part in the whole sample process by respective sample count to achieve the mean value of each part. Next, we divide the mean values by the height of the object to acquire the proportions

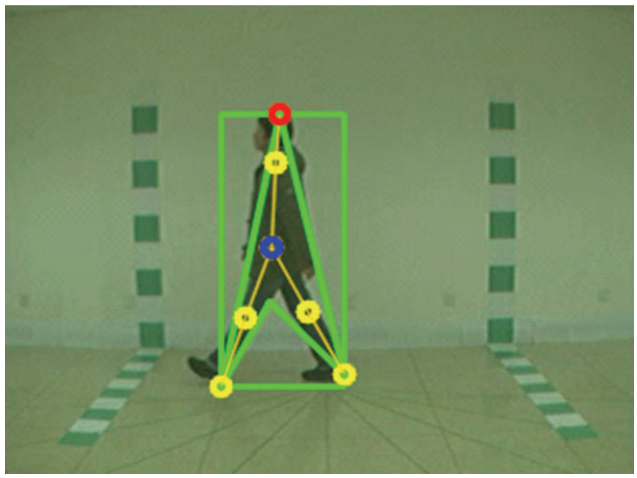


Figure 10: Form of the actual deformable template-matching with rectangle inside for real motional object.

of all the parts, realizing the normalization. In addition, in Figure 7, the step is measured based on the maximum angle α between two thighs; so for this variable, we do not need any regularization and normalization, but use law of cosines and the relative inverse trigonometric function to achieve its value when the space between two thighs is maximum. As for the obliquity θ of trunk, it is measured with inverse tangent function computed by the differences of x-coordinates and y-coordinates each between the hip and the top of head, furthermore, acquiring its mean. The actual obliquity θ of trunk may lean forward or backward, which is decided by the sign of result. The corresponding formula is in (3.3), where l equals length of relative part; X and Y stand for x-coordinate and y-coordinates of relative position; N_{sample} is sample amount; $A_{\text{rectangle}}$ is the area of the smallest circumscribable rectangle for object. Finally, all the parameters above are regarded as the final results of real detected individual characteristics:

$$
\begin{aligned}
l_{\text{mean}} &= \frac{\sum_i^{N_{\text{sample}}} \left(l_{\text{temp}}/\sqrt{A_{\text{rectangle}}}\right)_i}{N_{\text{sample}}}, \\
l_{\text{proportion}} &= \frac{l_{\text{mean}}}{l_{\text{ae\_mean}} + [(l_{\text{ec\_mean}} + l_{\text{ed\_mean}})/2]}, \\
\alpha &= \max\left(\arccos\left(\frac{l_{\text{ec}}^2 + l_{\text{ed}}^2 - l_{\text{cd}}^2}{2 \times l_{\text{ec}} \times l_{\text{ed}}}\right)\right), \\
\theta &= \frac{\sum_i^{N_{\text{sample}}} \left[\text{arctg}((X_b - X_a)/(Y_a - Y_b))\right]_i}{N_{\text{sample}}}.
\end{aligned}
\tag{3.3}
$$

3.2.4. Transforming from General Viewing Angle to Silhouette

In theory, any angle of view can be decomposed into two parts in horizontal and vertical directions, that is, any viewing angle can be regarded as the synthesizing from the two kinds

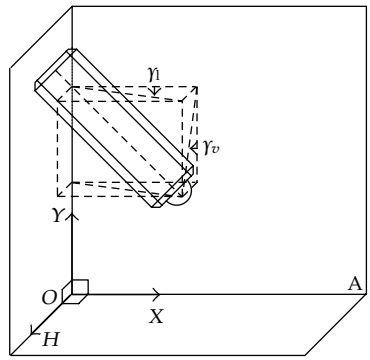


Figure 11: The decomposing principle of a general viewing angel for a common camera.

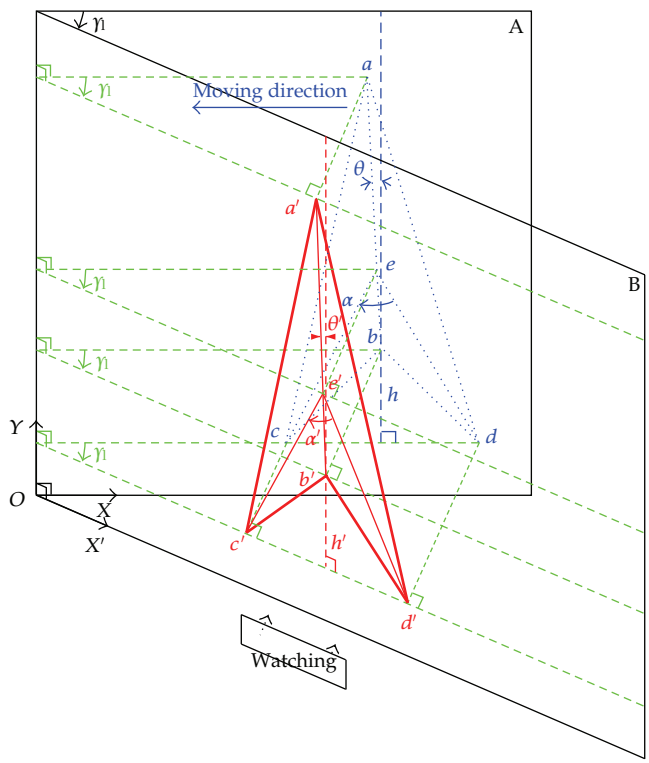


Figure 12: The projection of the deformable template in horizontal obliquity direction.

of viewing angles. Figure 11 gives the decomposing principle of a general viewing angel for a common camera, where γ_l and γ_v are the obliquities of the camera against the flat A in horizontal and vertical directions. Basing on these parameters, Figures 12 and 13 display the projections of the deformable template mentioned in horizontal and vertical directions, where the blue figure is the real gait detecting in flat A, the red figure is the watching result in flat B, and the green lines are relative assistant lines. In Figures 12 and 13, according to the relationships of the coordinates between the projection in flat A and the original graph in flat

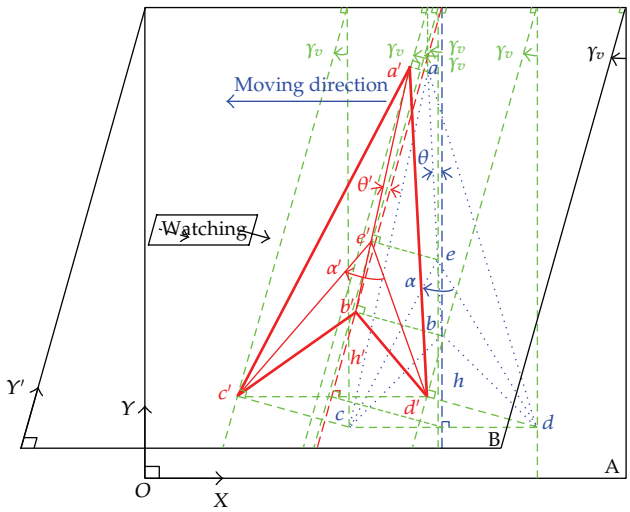


Figure 13: The projection of the deformable template in vertical obliquity direction.

B, the transforming computation and distance computation of coordinates to a general angle of view can be inferred as formula (3.4):

$$x_{\text{in flat } A} = \frac{x_{\text{at viewing angle } r}}{\cos(r_l)}, \qquad y_{\text{in flat } A} = \frac{y_{\text{at viewing angle } r}}{\cos(r_v)},$$
$$l_{MN \text{ in flat } A} = \sqrt{\left(x_{M \text{ in flat A}} - x_{N \text{ in flat A}}\right)^2 + \left(y_{M \text{ in flat A}} - y_{N \text{ in flat A}}\right)^2}. \tag{3.4}$$

In this paper, the angle of view is defined as that the facing direction of detected object circles anticlockwise from walking direction to watching screen, forming the angle. Namely,

$$r_l = 90 - r_{\text{viewing angle}},$$
$$r_v = 0. \tag{3.5}$$

From formula (3.5) above, it can be found that when the viewing angle $r_{\text{viewing angle}}$ is about 90 degree or so, the associated cosine function approximates 1 in formula (3.4). So, at this time, in formula (3.4), the viewing angle can be ignored for the relative transforming computation at a certain extent.

And in reality, strictly speaking, the viewing angle for a real camera is changing because of the limited range of eyesight by the camera itself even if the camera fixes position, just as Figure 14.

In Figure 14, at the fixed leaning angle r for a real camera, when the motion object (the red rectangle) is moving from position A to B and C, we can see that the relative viewing angle is changing and the relative transforming computation should base on the associated viewing angle at special some position according to formula (3.4). Especially, when the object is moving to the position B, the relative viewing angle is 90 degree. At this moment, according

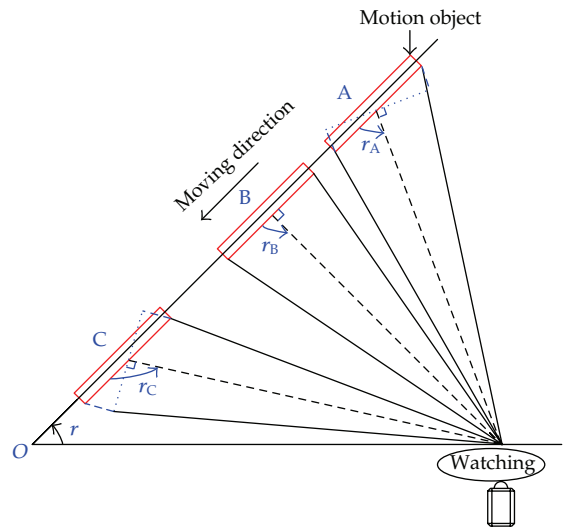


Figure 14: The real viewing angle changes with the motion object at fixed camera.

to the analysis above, the angle can be ignored during the transforming computation. This paper, in subsequent experiment, we will find that the detecting deformable template-matching is often successful at this angle or so because the triangle template can be in effect only when the both shoulders of subject, nearly in overlapping or in approximate verticality over viewing. And considering the normalization and average in formula (3.3), thus, the viewing angle can be omitted during transforming computation to a great extent.

3.3. Integrating Real Individual Characteristics of Detected Motional Object with Prior Human Motion Model

The process that this paper maps for both the parameters of structural proportion and motion for the real motional object above onto the 3D human model is as follows.

3.3.1. Mapping Static Parameters

This paper configures the height of human model used for gait imitation in some constant prior value. Then, integrating with the proportions of all the parts on trunk and lower limbs in Section 3.2, we figure out the relative lengths. For lengths of all the parts of the upper limbs, we default the normal prior values of the human model.

3.3.2. Mapping Dynamic Parameters

This paper substitutes the maximum angle α between two thighs and the obliquity θ of trunk detected in reality both for relative prior values. When the limbs of the human model in walk are swinging backward and forward, the condition of switching between left side and right side is based on step, so, the span of the upper limbs swinging to and fro is in proportion to the step of the lower limbs.

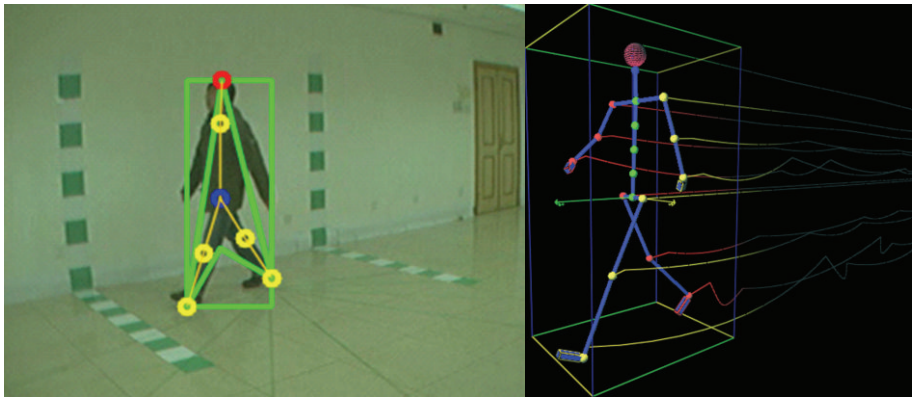


Figure 15: Final form of motion imitation.

3.4. Interpretation and Relative Analysis for the Final Motional Form of the Mapped Human Model

Figure 15 shows the final form of motion for the mapped human model and draws the curves of motion for the model's main joints including ankles, knees, hips, wrists, elbows, shoulders and the top of head in real time. According to analysis, the asymmetry exists in the curves because there are cycles of slight bow in elbows and knees when the limbs swing backward and forward on both sides. Besides, the curves of the upper limbs in reality need the curves in the lower limbs to synthesize. Note that in Figure 15, the size of video image for detecting motional object is 320 × 240, and the size of image for imitating motion is 362 × 522 initially. Here, because of the limitation in page size, we decrease sizes of the images by unchanging proportion between width and height of image. And we set angles of view for the human motion model and the real motional object to be same for the convenience of observation in contrast (similar intention for the subsequent same kinds of images). In fact, we can transform the viewing angle of the human motion model freely to observe after it achieves the actual individual characteristics on motion and updates by itself.

In order to study deeply, we draw the relative curves accurately in MATLAB tool by using the data corresponding to the real-time curves acquired from the development environment of VC6.0 in Figure 15. This is shown in Figure 16. Next, we combine Figure 15 with Figure 16 to understand the characteristics of the curves from important parts in the motional model.

For ankles, when the corresponding lower limb rises from the last position of body and the knee bends backward, the highest position of the ankle's curve is reached. Then, this lower limb begins to swing forward, and the curve begins to decline. After the limb's motion passes the trunk, the curve begins to rise. When the limb swings to front end, because the knee can only bend backward and can not bend forward furthermore, at one time, the extent of the bending is the smallest compared with the whole motion process, the curve rises to the higher position. Then, the limb begins to fall to the ground. Because the horizontal displacement from front end to the ground is quite short, the curve begins to decline with higher slope than before. As the landing of the limb is finished, the curve declines to the lowest position. Next, it is turn of the part in opposite side of body to motion in the same style, thus, time after time. In respect of wrists, similarly, the differences from ankles' motion exist in that the elbows can only bend forward and cannot bend backward. So, when part of

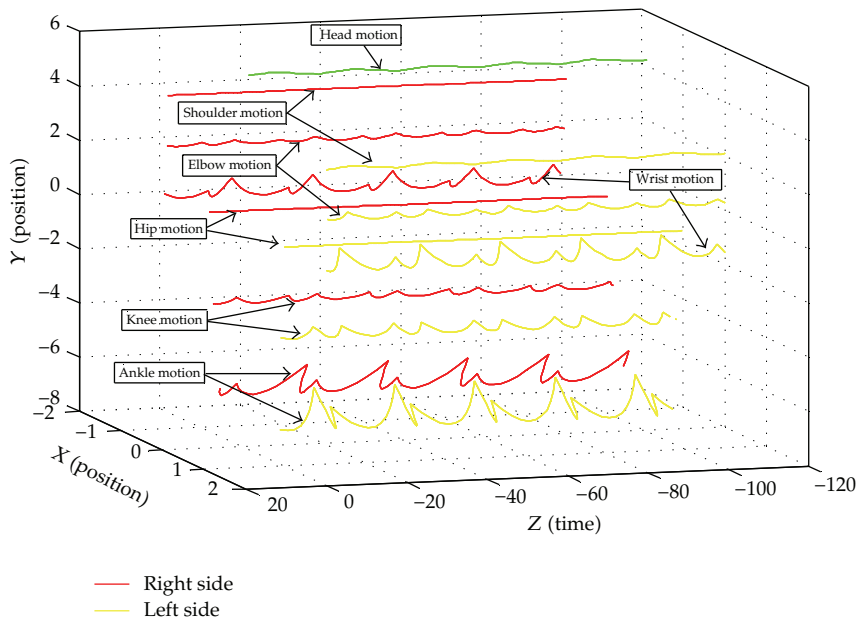


Figure 16: Curves of human gait imitation.

the upper limbs swings to front end, the relative wrist' curve arrives at the highest position, and when the upper limb swings to back end, the curve arrives at the higher position. While the upper limb's motion passes by the trunk, the curve arrives at the lowest position. Notice that in reality, although the curves of the upper limbs need the curves in the lower limbs to synthesize, both curves change with the same trend generally in horizontal and vertical directions. Therefore, they affect each other with the same tendency. Specifically, when the upper limbs swing to front end and back end, corresponding to the highest position, and the higher position, the lower limbs swing also to front end and back end. Moreover, at one time, for the lower limbs, the part of front end does not bend and the part of back end does not rise yet, as a result, the body's barycenter tends to rise. While the upper limbs swing parallel with the trunk, the lower limbs swing parallel with the trunk too and slightly bend in order to exchange the phase of the swinging. So, the body's barycenter tends to decline.

For knees, during the swinging of lower limbs, in fact, the motional displacement from last position to front end for one lower limb is approximately two times that from front end to last position for another lower limb compared with ground; therefore, in motion, the length of curve of swinging forward is bigger than that of swinging backward for knees, which conforms to reality basically. As for elbows, similarly, furthermore, the tendencies of changes are the same in both curves of upper limbs and lower limbs' relative parts from description above, so the length of elbow's curve of swinging forward is also bigger than that of swinging backward in motion, which consists with reality too.

For shoulders, when relative part of the upper limbs swings forward, with trunk's twist, the corresponding shoulder produces a short offset forward and upward and when the relative part swings backward, with trunk's reverse twist, the shoulder produces a short offset backward and downward. To hip, also, its motion style is just as shoulders except that the phases are opposite with shoulders on the same side and the extents of motion are smaller than shoulders.

For head, during the process of swinging in both sides of the upper limbs, because of both the inertia of arms dragging the trunk and twisting motion existed in trunk itself, head swings slightly leftward and rightward, always towards the part of upper limbs swinging backward.

The above analysis is mainly based on the characteristics of human anatomy. According to the analysis, the curves on imitating motion are basically consistent with the actual characteristics of human gait. In terms of the results of subsequent experiment, different individuals differ mainly in the values of parameters of these curves, but general shapes of the curves do not change essentially.

4. Application to Identification Recognition on Gait

As an application, we will utilize the features of curves acquired from the framework of human gait imitation proposed in this paper, combining with associated classifier, to realize human identification recognition. Figure 17 displays the principle of this gait recognition. Here, firstly, because of the periodicity for human gait, only one gait cycle is used to study by letting the human model walk for two steps. Then, all the curves are arrayed in sequences respectively. Thus, each curve of human gait imitation can be regarded as a time sequence. Next, the problem of recognition is transformed into the problem of dealing with all the data of time sequences.

4.1. NPE for Reducing the Dimensionality of Time-Sequences

In reality, the lengths of the time-sequences are very long and perceptually meaningful structure of the sequences is of much lower dimensionality, so dimensionality reduction is needed. Considering that all the time-sequences of the curves are correlated with each other for the same person, and these correlations are important individual characteristics, so, these structural correlations should be preserved as much as possible while carrying on reducing dimensionality.

Referring to [33], the method of neighborhood preserving embedding (NPE) aims at preserving the local neighborhood structure on the data manifold and is less sensitive to outlier than principal component analysis (PCA). Comparing to the recently proposed manifold learning algorithms such as Isomap and locally linear embedding, NPE is defined everywhere, rather than only on the training data points. So, we adopt this method to reduce dimensionality:

$$\min \sum_i \left\| x_i - \sum_j W_{ij} x_j \right\|^2 \quad \text{with constraints: } \sum_j W_{ij} = 1,\ j = 1, 2, \ldots, k, \tag{4.1}$$

$$X(I - W)^T (I - W) X^T a = \lambda X X^T a \quad \text{where } X = (x_1, \ldots, x_k),\ I = \operatorname{diag}(1, \ldots, 1), \tag{4.2}$$

$$y_i = A^T x_i = (a_0, a_1, \ldots, a_{m-1})^T x_i \quad \text{where } y_i \text{ is a } m\text{-dimensional vector.} \tag{4.3}$$

The method in detail is just as [33], whose main principle is displayed in formulas (4.1)–(4.3). Before applying [33], note, here, each time-sequence i ($i = 1, \ldots, 13 \times 3 = 39$) produced by associated curve is regarded as data point x_i. To preserve the property of

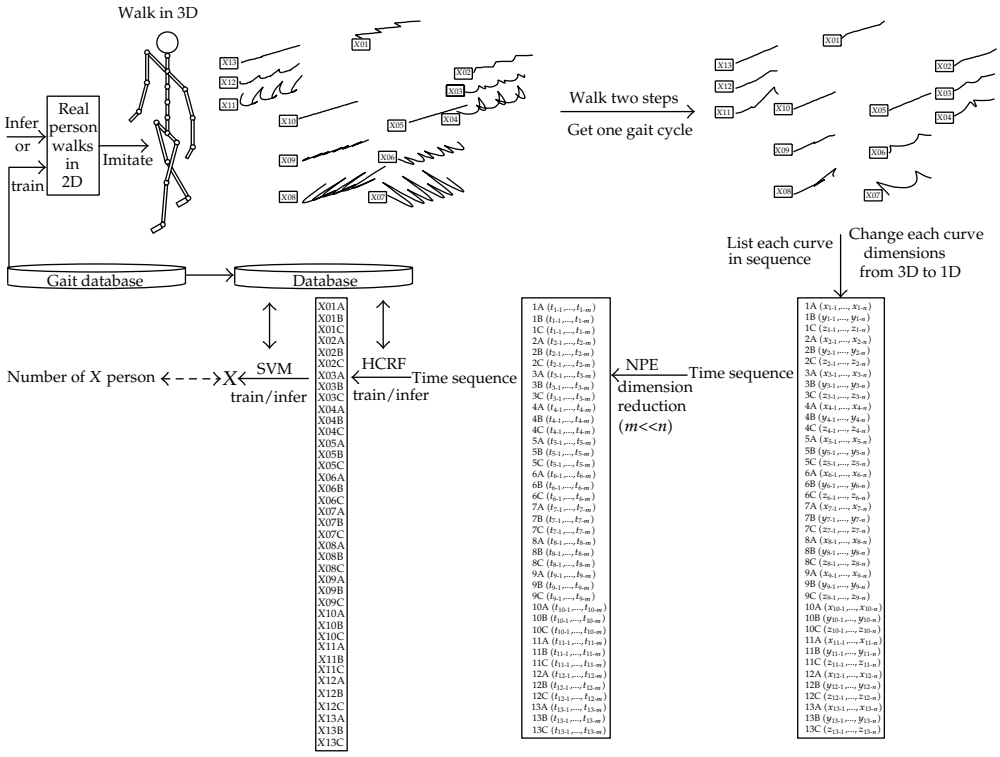


Figure 17: The principle of integrated classifier for gait recognition based on gait imitation.

correlation among curves, the way to construct adjacency graph is KNN, and K is set 39. That is to say, each time-sequence is reconstructed by adjacent other 38 time-sequences in motion model. And it is reasonable to assume that each local neighborhood is linear although these data points might reside on a nonlinear submanifold. Then, we use formula (4.1) to compute the weight matrix W of the structural relation existed in the data points, use formula (4.2) to compute the projections on reducing dimensionality and use formula (4.3) to realize the final transformation of reducing dimensionality for each data point x_i, in turn.

Originally, each of the motional curves in the human model for one gait cycle has 90 3D space samples. By using NPE, the corresponding time-sequences' dimensionality reduces from 90 dimension of one gait cycle to 39 dimension and the local manifold structure is preserved in low-dimensional space with an optimal embedding.

4.2. Classifier Integrates SVM with HCRF for Classifying All the Time-Sequences

During the key phase of recognition, we integrate the hidden conditional random field (HCRF) with supported vector machine (SVM) to construct classifier. This kind of classifier owns both merits of HCRF and SVM. On one hand, for each time-sequence, a set of latent variables conditioned on local features can be learned, while the observations need not be independent and may overlap in space and time [34]. On the other hand, the separating margins of final decision boundaries on classification are maximized in the high-dimensional

space called feature space [35]. That is, it can resolve the problem of classification for the whole multisequence existed in the same course of time.

4.2.1. HCRF for Marking All the Time-Sequences

In Figure 17, number of X ($X = 1, 2, \ldots$) person, according to that different joints in human motion model corresponds to different motion curves, hence, all the curves on his motion model, further, all the corresponding time-sequences are marked with X01A, X01B, X01C, X02A, X02B, X02C, ..., X13A, X13B, X13C differently, respectively.

Since all the time-sequences are correlated with each other, these observations are not conditional independence of course. Referring to [34, 36], hidden conditional random field (HCRF) which uses intermediate hidden variables to model latent structure of input domain and defines a joint distribution over class label and hidden state labels conditioned on the observations, with dependencies between the hidden variables expressed by an undirected graph, does not need observations to be independent and may overlap in space and time. And it can also model sequences where the underlying graphical model captures temporal dependencies across frames and incorporate long range dependencies. So, using HCRF to mark these sequences is a reasonable mode for describing them. And the mapping between the sequences and the corresponding labels is conducted by HCRF's training or inferring. Here, the HCRF method which is just as [37] in principle is similar with [36]. The associated formulas are as (4.4) and (4.5):

$$\begin{aligned}
\arg\max_{y\in Y}\left(P\left(y \mid x, \theta^*, w\right)\right) &= \arg\max_{y\in Y}\left(\sum_h P\left(y, h \mid x, \theta^*, w\right) a\right), \\
&= \arg\max_{y\in Y}\left(\frac{\sum_h e^{\Psi(y,h,x;\theta^*,w)}}{\sum_{y'\in Y, h\in H^m} e^{\Psi(y',h,x;\theta^*,w)}}\right) \\
&\text{where}: \ \Psi\left(y, h, x; \theta^*, w\right) = \sum_{j=1}^{n} \varphi(x, j, w) \cdot \theta_h^*\left[h_j\right] \\
&\qquad + \sum_{j=1}^{n} \theta_y^*\left[y, h_j\right] + \sum_{(j,k)\in E} \theta_e^*\left[y, h_j, h_k\right],
\end{aligned} \tag{4.4}$$

$$\theta^* = \arg\max_{\theta}\ (L(\theta)) = \arg\max_{\theta}\left(\sum_{i=1}^{n} \log P\left(y_i \mid x_i, \theta, w\right) - \frac{1}{2\sigma^2}\|\theta\|^2\right). \tag{4.5}$$

Formula (4.4) describes the principle of inferring label y from HCRF model given the observation x, the HCRF model's parameters θ^* and the window parameter w which is used to incorporate long-range dependencies. And $h = \{h_1, h_2, \ldots, h_m\}$ is a vector of latent variables, which are not observed on training examples and where each h_j is a member of a finite set H of possible hidden states in the HCRF model. Intuitively, each h_j corresponds to a hidden state of x_j with some member of H, which may correspond to "component" structure in an observation. $\Psi(y, h, x, \theta^*, w)$ is a potential function parameterized by θ^* and w. The graph E is a chain where each node corresponds to a hidden state variable at time t; $\boldsymbol{\phi}(\mathbf{x}, j, w)$ is a vector that can include any feature of the observation x for a specific

window size w. The inner product $\boldsymbol{\phi}(\mathbf{x}, j, w) \cdot \theta_h{}^*[h_j]$ measures the compatibility between the observation x and hidden state h_j at window size w. Each parameter $\theta_y^*\ [y, h_j]$ measures the compatibility between hidden state h_j and a label y. Each parameter $\theta_e^*\ [y, h_j, h_k]$ measures the compatibility between an edge with states h_j and h_k and the label y.

Formula (4.5) describes the estimation for HCRF model's parameters θ^*. The first term is the logarithmic likelihood of the trained data. The second term is the log of a Gaussian prior with variance σ^2, that is, $P(\theta) \sim \exp(\|\theta\|^2/(2\sigma^2))$. Combing with [37], the typical method of conjugate gradient in [38] is used to estimate the parameters θ^*.

In this paper, the number of hidden states is set 10 and the window size w is set at 0,1,2 in turn to test; there are too many time-sequences in kinds and numbers to fit for the one-versus-all HCRF model or the muti-class HCRF model directly, so the compromise between one-versus-all HCRF model and muti-class HCRF model is adopted. In training, each type of articular time-sequence for all the persons and its corresponding labels are learned with a separate HCRF model. In inferring, the tested sequence is run with the HCRF model producing the same type of articular time-sequence. The class label with the biggest probability corresponds to the label of the test sequence.

4.2.2. SVM for Classifying All the Signs of Multisequence

In Figure 17, since from HCRF's training, a group of labels and their corresponding motion curves, namely, corresponding time-sequences of a specific human model are learned and these labels' definitions or values differ from person to person, thus, these labels can be regarded as a group of features for a specific person. However, as there are some similarities existed in most people's walking styles, it is possible that from HCRF's inferring, sometimes, a few of these features among some persons are identical although HCRF can overcome the overlapping of space and time to a certain extent. At this time, part of these features is overlapping. Referring to [35], according to SVM's property that the decision boundaries are determined directly by the training data, as a result, the separating margins of decision boundaries are maximized in high-dimensional space called feature space. Thus, most nonseparable data in low-dimensional space becomes separable possibly in high-dimensional space by mapping. So, here, SVM is used as final classifying means to recognize person by using the marked features associated with his motion imitation as input.

Here, it is the problem of multiclass classification. On multiclass SVM, there are many methods at present. According to relative comparison [39], the "one-against-one" approach [40, 41] is suitable for practical use. So, we use this method, which is just as [42]. According to this method, if the amount of training persons in database is K, $K(K-1)/2$ binary SVM classifiers are needed. Combing with [42], each classifier adopts C-support vector classification (CSVC) model with RBF kernel, in which two parameters are considered: C and γ. The two parameters are selected by using typical cross validation via parallel grid search, and all the $K(K-1)/2$ decision functions share the same (C, γ) finally.

For training data from the ith and the jth classes, formula (4.6) displays the binary classification problem to be solved. In (4.6), the training data x_i is mapped into a higher space by the function Φ and C is penalty parameter, ζ is relaxation parameter. Minimizing $((w^{ij})^T w^{ij})/2$ means to maximize $2/\|w^{ij}\|$, the margin between the ith and the jth classes of data. When data are not linear separable, the penalty term $C(\sum_t (\zeta^{ij})_t)$ manages to balance between the regularization term $((w^{ij})^T w^{ij})/2$ and reducing the number of training errors. In addition, during final classifying, voting strategy suggested in [40], in which if

$\operatorname{sign}(((w^{ij})^T\Phi(x_t)) + b^{ij})$ infers x is in the ith class, then the vote for the ith class is added by one, otherwise, the jth class is increased by one, is used to predict x is in the class with the largest vote. For the case that two classes have identical vote, the one with smaller index is selected:

$$\begin{aligned}
&\min_{w^{ij},b^{ij},\zeta^{ij}} \quad \left(\frac{1}{2}\left(w^{ij}\right)^T w^{ij} + C\left(\sum_t \left(\zeta^{ij}\right)_t\right)\right) \quad \text{where } \zeta_t^{ij} \geq 0, \\
&\text{subject to} \quad \left(\left(w^{ij}\right)^T \phi(x_t)\right) + b^{ij} \geq 1 - \zeta_t^{ij}, \quad \text{if } x_t \text{ in the } i\text{th class}, \\
&\qquad\qquad \left(\left(w^{ij}\right)^T \phi(x_t)\right) + b^{ij} \leq -1 + \zeta_t^{ij}, \quad \text{if } x_t \text{ in the } j\text{th class}.
\end{aligned} \tag{4.6}$$

5. Performance Evaluation

5.1. Evaluation Setup and Dataset

In order to prove the ability of the application in general environment for the method proposed in this paper, all the experiments are conducted in the environment of Microsoft visual c++6.0 at the platform of Pentium 1.73 GHz personal computer.

We test the framework proposed by this paper with the videos of CASIA gait database [43]. There are 3 subsets in this database: dataset A, dataset B, and dataset C. Dataset A (viz.: NLPR) consists of 20 subjects. Each subject has 12 walking sequences, which include 3 walking directions (that make an angle of 0°, 45°, 90°, resp., with image plate) and in each direction, there are 4 walking sequences. The length of each image sequence varies from person to person in speed.The sum of frames in each sequence is between 37 and 127. Dataset B is a large scale of database with multiangle of view. The subset includes 124 subjects, each of whom has 11 angles of view (covering: 0°, 18°, 36°, 54°, 72°, 90°, 108°, 126°, 144°, 162°, and 180°) and walks in 3 conditions (involving: thin coat, thick coat, and backpack). Dataset C is a large scale of database screened with infrared photography in night, in which there are 153 subjects walking in 4 conditions (involving: common walk, quick walk, slow walk, and backpack walk).

5.2. Experimental Procedures

5.2.1. For Gait Imitation

This paper uses the original videos in CASIA gait database (viz. Dataset B) to test the proposed framework on gait imitation. We experiment with all the 124 subjects who walk in thin coat, thick coat, and backpack, respectively, under different angles of view in the dataset.

$$DD_{I,J}^{\text{Euclidean}} = \operatorname{sqrt}\left(\sum_{n=1}^{6}(i_n - j_n)^2\right), \tag{5.1}$$

$$DD_{I,J}^{\text{Mahalanobis}} = \operatorname{sqrt}\left[(I-J)^T * \operatorname{cov}^{-1}(I,J) * (I-J)\right]. \tag{5.2}$$

In experiment, the whole different extent of detected parameters among the different working conditions for the same person at the same angle of view is measured from two kinds of distance measure: the Euclidean distance in formula (5.1) and the Mahalanobis distance in formula (5.2), where I, J are any two groups of different detected data above and i, j are elements in the I, J, respectively. Finally, all the results of the testing subjects are drawn in 3D space with MATLAB R2009b tool to analyze.

5.2.2. For Gait Recognition Based on Motion Imitation

In the experiment of the application on gait imitation, we use the original videos of dataset B in CASIA gait database to test the recognition framework which is just as Figure 17 based on the integrated HCRF/SVM classifier at different window size w.

We adopt the leave-one-out cross validation to train/infer gait identification. In detail, for each person, we take 5 out 6 videos with thin coat, take 1 out 2 videos with thick coat, and take 1 in 2 videos with backpack for training. And we take the remainder one video in thin coat, thick coat and backpack for recognizing or inferring. Next, we change the order and repeat the train/infer experiment above until all the videos have chances of inferring. At last, we compute the average value of these recognition rates and regard it as the final result. The associated computation is as formula (5.3), where function inference can be regarded as the whole function corresponding to the relative recognizing system of HCRF/SVM when the ith subject is testing object x_i ($x_i > 0$). And only when the function's result equals input x_i, the result is correct at this time:

$$R_{\text{recognition rate}} = \frac{1}{N}\sum_{i=1}^{N} \& \left(x_i - \text{function\_infer}_i^{\text{hcrf/svm}}(x_i)\right),$$

$$\text{where } \& \text{ is the unit pulse function and } \& (m) = 1 \text{ if } m = 0, \text{ zero otherwise.} \tag{5.3}$$

Combing with Section 4, in training for HCRF, the method of conjugate gradient as [38] is used to estimate associated parameters and in training for SVM, the method of cross-validation via parallel grid search as [42] is used to estimate associated parameters. According to Section 4, since there are 124 subjects and each subject corresponds to 13 3D time-sequences, there are $13 \times 3 = 39$ HCRF models and $124 \times (124 - 1)/2 = 7626$ SVM classifiers altogether. When the training for all the subjects in the dataset B of CASIA gait database is finished, all the trained parameters for associated HCRF models and SVM classifiers are saved as another database.

In the area of gait recognition, baseline algorithm in [44] is a kind of typical method which estimates silhouettes by background subtraction and performs recognition by temporal correlation of silhouettes. So, we will compare this method with the framework presented in this paper at recognition property. Here, during realizing [44], the silhouettes including the actual motional object are acquired from the CASIA gait dataset B videos by the detecting phase mentioned in Section 3. Because of the various viewing angles in the gait database, the associated gait period is detected by computing the ratio of the number of pixels in the silhouettes to the relative smallest circumscribable rectangle. And the leave-one-out cross validation is also adopted with the average recognition rates computed as the final identification rates.

In addition, at the phases of just after acquiring the normalized parameters of real detected object and just after reducing the dimensionality of associated time-sequences, we use these temporal associated data as input with HCRF, SVM used as classifier solely to recognize human gait. At this situation, one HCRF model corresponds to one subject, namely, there are 124 HCRF models, and the amount of SVM classifiers is as before. Of course, another database on associated trained parameters is produced from dataset B of CAISIA gait database after training. At last, we conduct the comparisons of the recognition rates with the results of recognition framework proposed by this paper at same window parameter w.

5.3. Experimental Results and Associated Analysis

5.3.1. For Gait Imitation

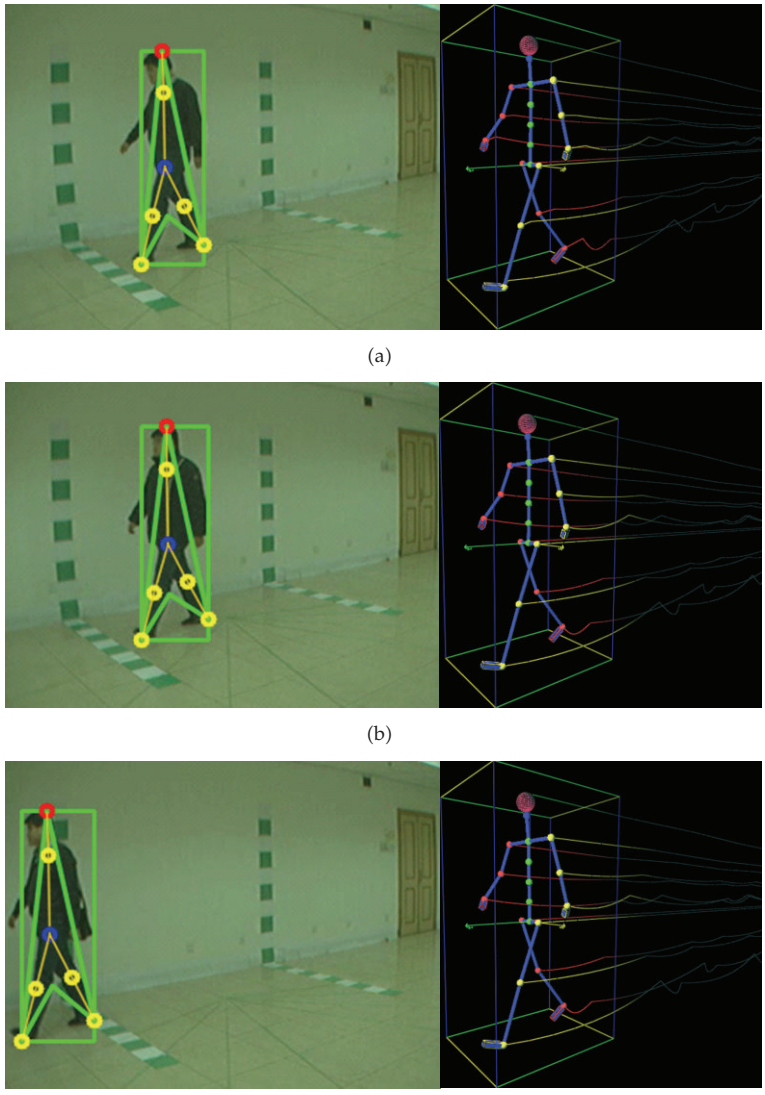
Figure 18 displays motion imitation for a same subject who wears thin coat, thick coat and backpack, respectively, at angle of view of 54°. Observing from relative emulational model and motion curves, the three impressions are quite similar with each other. The parameters on real motion characteristics of the three types of walking in Figure 18 list in Table 1. The relative measuring results of the two kinds of distances: Euclidean distance and Mahalanobis distance are shown in Table 2. By comparing and analyzing, all the values in Table 2 are very small universally. Namely, the whole different extent between the congener values in Table 1 is very small universally. Thus, we can infer that the detected parameters' values listed in Table 1 are quite close to each other.

That is to say, for the same person at the three different walking conditions, the detected proportions, key poses of limbs, trunk in human body are almost unchanged, which is the essential reason why the three types of walking motion imitation are alike.

Figure 19 shows the experiment of 3D gait imitation for some subjects at other angles of view with three types of walking. By observation, comparison and relative measurement as above, the effect of these motion imitations each quite resembles the form of relative real individual objects in motion.

If we take columns of Table 2 as points in 3D space of Euclidean distance and the Mahalanobis distance, respectively, we can draw a point in each of the two spaces, which corresponds to a subject's measurement for different extent in different walking conditions. Similarly, Figure 20 displays the testing results of 124 subjects' measurement on different extent of different walking conditions at viewing angle of 54°, 72°, 90°, 108°, and 126°, respectively, in the CAISIA database. From Figure 20, we can see that most points in the two 3D spaces near origin. Whether in Euclidean or Mahalanobis measurement, the distances between thin coat and backpack for same subject are universally slightly bigger comparing with the other distances and because the Mahalanobis measurement includes extracovariance matrix comparing with Euclidean measurement, the former is slightly bigger than the latter, but they all are in acceptable ranges from the whole result. So, according to inferring as before, it means that for each of the 124 subjects, the whole different extent of walking at different walking conditions is comparatively small universally.

Thus, we find that the method proposed by this paper is robust in clothes and backpack for the motional persons to a certain extent. Notice, here, the robust means that the forms of gait imitation for the same object walking in different conditions are consistent with each other to a great extent. Thereby, some latent constant essential characteristics for gait are shown to some extent.



(a)

(b)

(c)

Figure 18: Motion imitation at three types of walking in 54° angle of view for the same subject (a) in thin coat, (b) in thick coat, and (c) in backpack.

With regard to the experiment at other angles of view, because it is very hard for the shape of contour of detected object to be uniform with the deformable template proposed by this paper, or in other words, the errors are too big, the effect of relative gait imitation is failed in those situations.

The final quality of the results for the framework of human motion imitation proposed in this paper based mainly on analysis for the relative characteristics of human anatomy and curves of gait, besides, the observation, comparison and associated measurement in the experiment of CASIA gait database. It can be seen that the extent of similarity between motion imitation and real motional object is rather large. As the motion imitation of this paper uses

Table 1: Parameters on motion imitation of three types of walking for same object.

Measurement		Object Three types of walking In thin coat	In thick coat	In backpack
Proportion parameters to whole stature	Calvaria to neck	1.814569×10^{-1}	1.820641×10^{-1}	1.832075×10^{-1}
	Neck to hip	3.224494×10^{-1}	3.229364×10^{-1}	3.206368×10^{-1}
	Hip to knee	2.491963×10^{-1}	2.488364×10^{-1}	2.495140×10^{-1}
	Knee to ankle	2.468975×10^{-1}	2.461631×10^{-1}	2.466417×10^{-1}
Angle (radian)	Step crossing angel	4.278397×10^{-1}	4.233917×10^{-1}	4.247243×10^{-1}
	Trunk bend angle	3.107436×10^{-2}	2.483856×10^{-2}	2.570835×10^{-2}

Table 2: The whole different extent of parameters on motion imitation at three types of walking for same object.

Comparison in different types of walking	Whole different extent in measurement Euclidean distance	Mahalanobis distance
Between in thin coat and thick coat	7.742374×10^{-3}	4.588733×10^{-3}
Between in thick coat and backpack	3.133021×10^{-3}	1.392382×10^{-1}
Between in thin coat and backpack	6.709422×10^{-3}	2.663143×10^{-1}

the detected data directly come from preceding detected real object and does not execute any prediction on poses of motion, the motion imitation is not carried out simultaneously. Here, synchronization is not our main object. In this paper, after collecting enough information of real person's gait, we attempt to reconstruct human gait in 3D for other studies.

Up to now, it can be seen that the method of the deformable template-matching in this paper not only can apply in many angles of view and is robust in clothes, backpack for the motional persons to a certain extent, but also not need any manual work and any model information. And it does not need fitting the motion model in each video frame unless the outer template-matching at the key states is successful in some frames and it does not need considering bending at elbows, knees, and neck during fitting, but compensates in proportion as universal prior knowledge before imitating real gait in 3D finally, which improves detecting efficiency greatly. In comparison, [7, 8] study motion imitation in 2D and not only need manual assistant originally, but also mainly aim at silhouettes with 90° angle of view and need fitting relative motion model in each frame, which is a very limited range of angle and objective time-consuming.

5.3.2. For Gait Recognition Based on Motion Imitation

Table 3 gives the results of associated comparisons of the integrated HCRF/SVM classifiers based on recognition application of gait imitation at different window size w and different viewing angle with typical baseline method. Figure 21 gives the bar graph associated with data in Table 3.

From Figure 21, combing with Table 3, we can see that the recognition rate differs from different window size w and when w equals 1, the recognition rates are universally higher than at other window sizes. At every tested w, when the viewing angle near 90°, including 108°, the recognition rates are universally higher than at other viewing angles

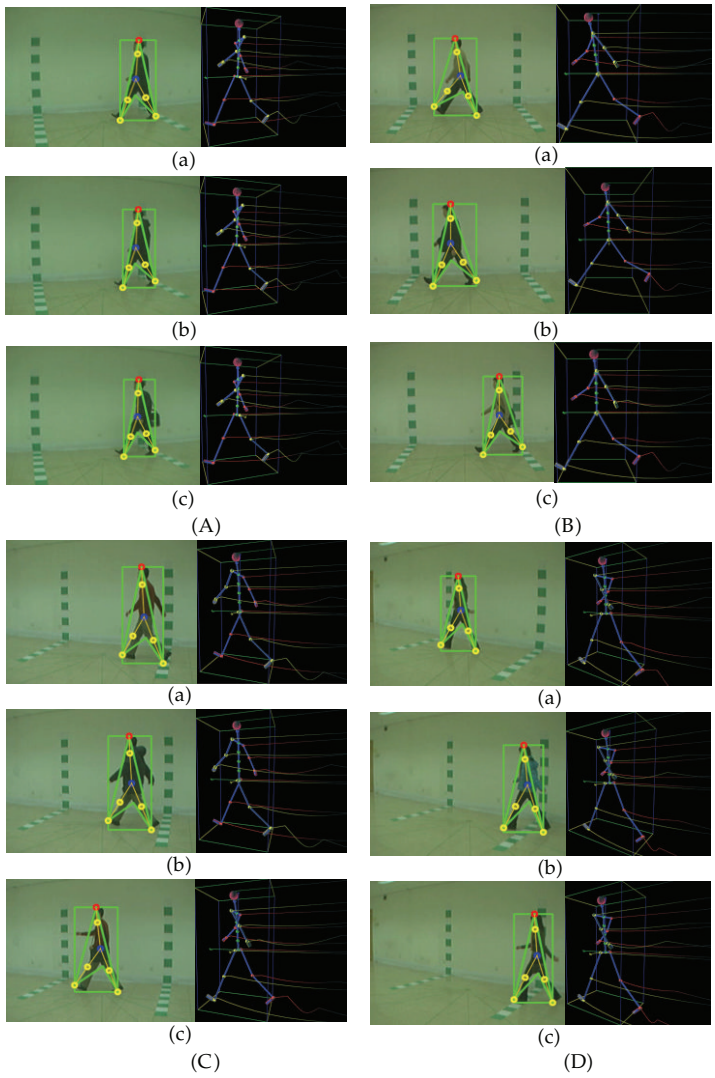


Figure 19: Form of motion imitation in other angles of view. (A) 72° angle of view; (B) 90° angle of view, (C) 108° angle of view; (D) 126° angle of view (a) in thin coat, (b) in thick coat, and (c) in backpack.

because the relative detected parameters including proportions of trunk and lower limbs are more accurate than at other viewing angles. Although any viewing angle can be mapped into 90 degree viewing angle according to the transform computation mentioned, the little error still cannot be escaped. We can also see that, at same window size w and same viewing angle, usually the recognition rate in thin coat is slightly higher than in thick coat and the recognition rate in thick coat is slightly higher than in backpack. Here, after all, whether in thick coat or in backpack, more or less, the sheltering affects the detecting accuracy to a certain extent. When the walker is in backpack, not only sheltering but also the disturbance of additions on body affects the accuracy. So in this situation, the recognition rates are the smallest comparing with other walking conditions. Of course, from the whole effect, the results of the recognition framework proposed by this paper are satisfied, and this method overcomes the limitations

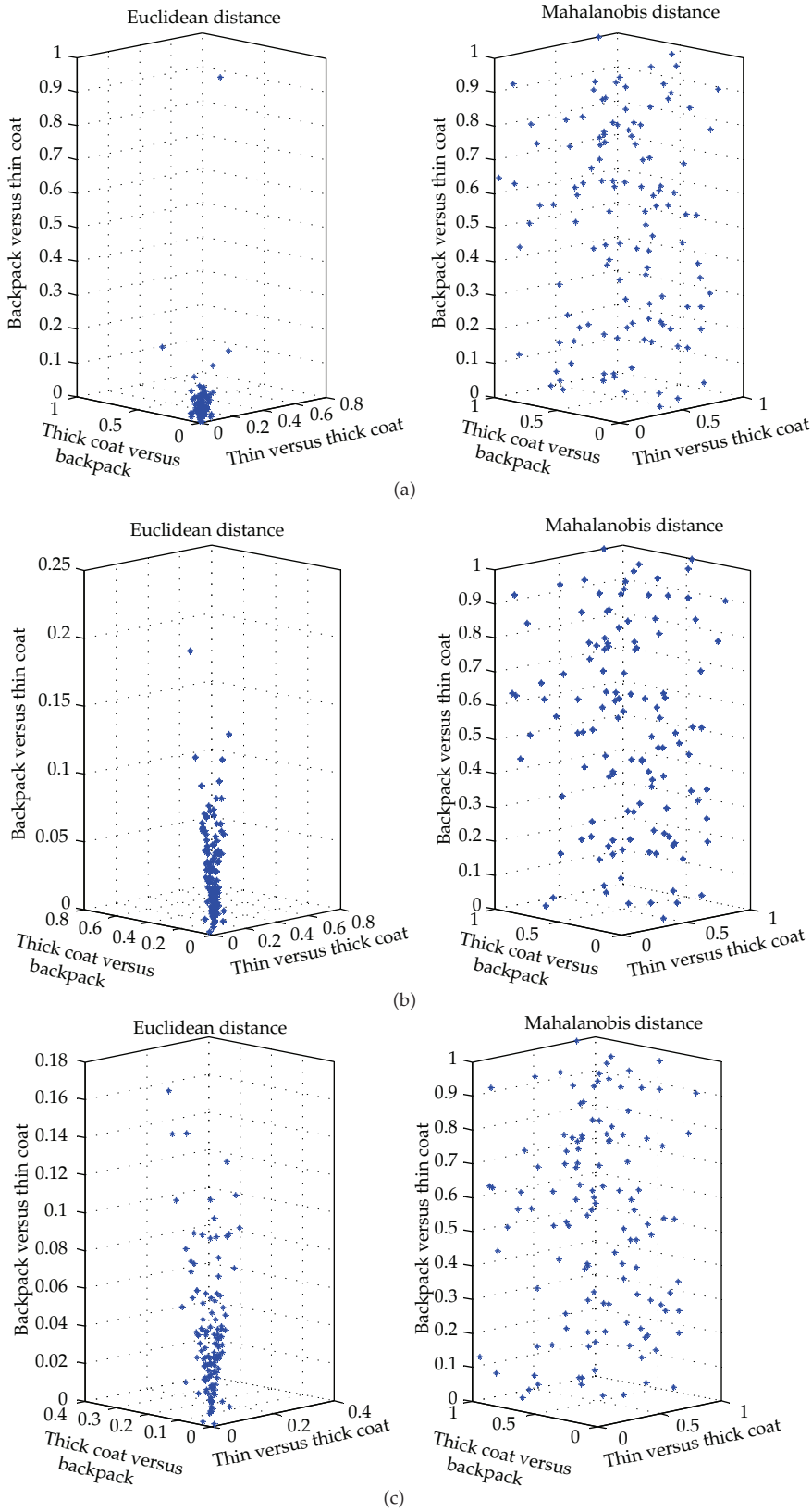
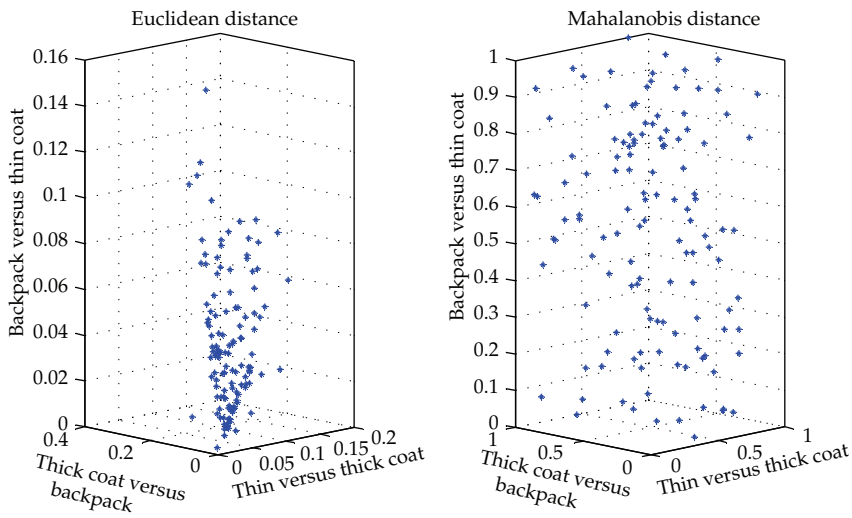
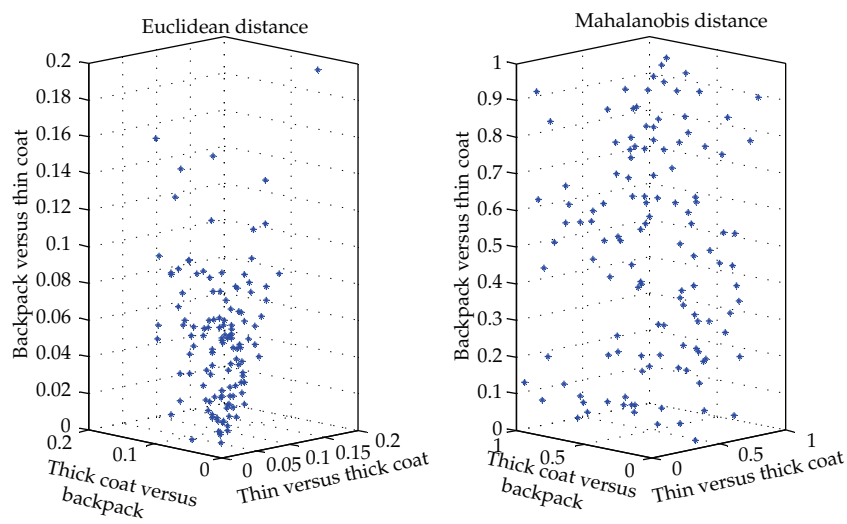


Figure 20: Continued.



(d)



(e)

Figure 20: Different extent of different walking conditions at some viewing angles for each of 124 subjects. (a) 54° viewing angle; (b) 72° viewing angle; (c) 90° viewing angle; (d) 108° viewing angle. (e) 126° viewing angle.

to some extent. That is, the recognition framework, as the gait imitation framework above, is robust to subject's coat or backpack to a certain extent.

Not depicting gait information in 3D, baseline method mainly studies silhouettes gait sequences in 2D, which makes this kind of gait information's volume and accuracy more limited than the framework of this paper. So, its identification rates are lower than this paper's framework universally. In addition, at same walking conditions, its identification rate at about 90 degree viewing angle is slightly lower than at other angles because the gait information on frontal in silhouettes at this angle is less than at other angles, which improves the possibility of identical detecting results among the testing samples. At the same viewing

Table 3: Associated comparisons of the integrated HCRF/SVM classifiers based on recognition application of gait imitation at different window size w and different viewing angle with baseline method.

Viewing angle	Types											
	HCRF (w = 0) + SVM			HCRF (w = 1) + SVM			HCRF (w = 2) + SVM			Baseline method		
	Recognized rate (%)											
	Thin coat	Thick coat	Backpack	Thin coat	Thick coat	Backpack	Thin coat	Thick coat	Backpack	Thin coat	Thick coat	Backpack
54°	89.2	86.1	82.3	91.4	87.3	84.8	87.0	83.4	80.9	81.6	75.1	68.5
72°	91.3	85.5	80.5	92.4	88.2	83.6	88.1	85.3	79.0	80.2	74.4	67.7
90°	92.1	88.2	82.4	94.5	92.0	85.2	90.1	86.6	81.2	78.5	72.9	65.4
108°	92.8	89.2	82.7	95.8	93.1	88.3	89.3	85.2	81.5	79.3	73.9	68.1
126°	87.7	85.4	81.8	89.7	86.4	84.5	85.2	82.8	80.7	82.6	76.5	69.2

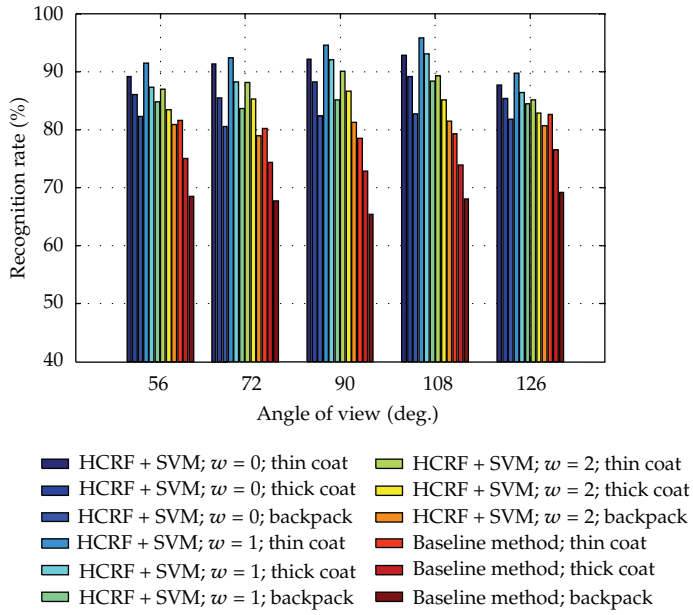


Figure 21: Associated comparisons of the integrated HCRF/SVM classifiers based on recognition application of gait imitation at different viewing angle and different window parameter w with baseline method.

angles, the identification rate in thin coat is higher than in thick coat, and the rate in thick coat is higher than in backpack, whose associated reasons are also the disturbances of relative sheltering and additions on body.

Table 4 presents associated comparisons of the sole SVM or HCRF classifier on gait recognition at same window size w and different viewing angles. Figure 22 gives the bar graph associated with data in Table 4 and data of the integrated HCRF/SVM classifier at same window size w in Table 3.

In Figure 22, at each tested visual angel, the recognition rate of HCRF + SVM is comparatively higher than HCRF or SVM solely. According to the analysis in Section 4, when the method of NPE reduces the dimension of the time-sequences, the local neighborhood structure on the data manifold is preserved; when the HCRF trains or infers the signs of the time-sequences, the sequences where the underlying graphical model captures temporal dependencies across frames is modeled and incorporates long range dependencies and when the SVM trains or infers the final identification of the relative signs, the separating margins of decision boundaries on classification are maximized as the data is mapped into high-dimensional space. Thus, comparing with the HCRF or SVM solely, the HCRF + SVM contains more structural traits of the data to be classified during dealing with the data, which make the recognition more sufficiently. At the phase of just after acquiring normalized parameters of real detected object, the recognition rate of SVM is a little higher than HCRF because these detected parameters at the same time instant are not correlated sequence in time or space, so merits of HCRF could not be presented fully. Namely, taking the independent parameters as correlated sequences to do with is unreasonable to some extent. And SVM is more fitting for classifying multi-dimensional data at the same time instant than HCRF. Contrarily, at the phase of just after reducing dimensionality of associated time-sequences, using similar

Table 4: Associated comparisons of the sole SVM or HCRF classifier on gait recognition at same window size w and different viewing angles.

Viewing angle	Types											
	Just after acquiring normalized parameters of real detected object						Just after reducing dimensionality of associated time sequences					
	HCRF1 (w = 1)			SVM1			HCRF2 (w = 1)			SVM2		
	Recognized rate (%)											
	Thin coat	Thick coat	Backpack	Thin coat	thick coat	Backpack	Thin coat	Thick coat	Backpack	Thin coat	Thick coat	Backpack
54°	63.0	58.2	55.7	69.3	64.5	60.9	73.2	70.3	61.2	69.3	67.1	55.2
72°	64.2	60.5	54.8	72.1	65.5	55.4	75.4	70.5	62.4	72.2	65.3	58.1
90°	66.1	61.3	56.7	75.3	67.8	58.5	78.3	72.5	65.5	75.1	66.2	58.6
108°	73.4	67.5	55.5	74.9	69.0	58.9	80.8	78.1	66.8	78.0	66.3	59.9
126°	65.5	61.6	57.2	68.4	64.2	56.5	74.4	71.9	64.3	70.2	63.0	56.4

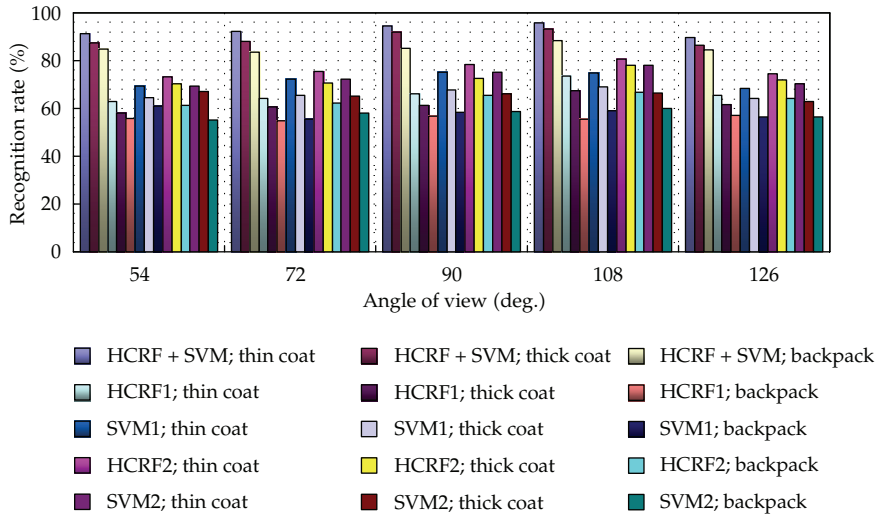


Figure 22: Associated comparisons of different classifiers in recognition application based on gait imitation at same window size w and different viewing angles.

analysis above, the recognition rate of HCRF is a little higher than SVM of course. In addition, for whether the sole HCRF or the sole SVM, the trends of recognition rates with viewing angle changing is consistent with the integrated HCRF/SVM classifier above.

6. Conclusion

This paper mainly proposes a framework of human gait imitation, analogous with human cognitive process, which integrates individual gait characteristics in reality with general prior knowledge of human motion to realize the reconstruction of human gait in 3D from monocular video of an uncalibrated camera directly and automatically. According to the results of the experiment with the CASIA gait database and relative measurement, analysis, the method in this paper is reasonable and robust to object's clothes and backpack to a certain extent.

In the application of this framework, firstly, all the imitated motion curves are transformed into time-sequences with limited lengths by the means of extracting one gait cycle, arraying in lines, reducing dimensionality with the method of NPE, in turn. Then, a kind of classifier which integrates HCRF with SVM is used to classify the multisequences by marking time-sequences with relative labels and classifying the labels in turn, realizing identification recognition on human gait. At associated experiment, this kind of integrated classification displays better properties than using HCRF or SVM solely and the typical baselined method because it contains more structural traits of the data to be classified in space and time during dealing with the data.

We do not think that the adopted integrated classification in this paper is the only one and the best one suited to this human gait imitation proposed in this paper. It is just one way and maybe there are other good classifying methods fitting for recognizing the gait imitation, too. We do not also think that the adopted integrated classification can only deal with this kind of dataset in this paper. Maybe it is also suitable for other kinds of datasets. Of course, that needs new testing in other situations.

In the next work, we will investigate the imitated walking curves deeply in other different speeds or different moods for persons and search the relative constant properties among these curves for identification recognition. In fact, the recognition process of this paper is two-stage of classifying process. This produces the possibility that the error aroused from former classifier affects the last classifier, forming accumulated error. And the training time and testing time is about 2-3 hours and 30 seconds or so, respectively. So, we will study the two kinds of classifier more deeply in theory and manage to search only one stage of classifying process which equals the two stages of classifying process above in theory, thus uniting the two classifiers into one classifier is essentially to improve the accuracy and speed of recognition.

Acknowledgments

The authors wish to thank to the National Laboratory of Pattern Recognition, Institute of Automation of Chinese Academy of Sciences for supplying investigators who study gait with free-CAISIA gait database. This brings us much convenience in the relative experiment of this paper.

References

[1] D. M. Gavrila and L. S. Davis, "3-D model-based tracking of humans in action: a multi-view approach," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '96)*, pp. 73–80, June 1996.

[2] C. Bregler and J. Malik, "Tracking people with twists and exponential maps," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 8–15, Santa Barbara, CA , USA, June 1998.

[3] N. D'Apuzzo, R. Plänkers, P. Fua, A. Gruen, and D. Thalmann, "Modeling human bodies from video sequences," in *Proceedings of the SPIE's Electronic Imaging*, vol. 3641, pp. 36–47, San Jose, Calif, USA, January 1999.

[4] J. Deutscher, A. Blake, and I. Reid, "Articulated body motion capture by annealed particle filtering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 126–133, June 2000.

[5] J. Darby, B. Li, and N. Costen, "Tracking human pose with multiple activity models," *Pattern Recognition*, vol. 43, no. 9, pp. 3042–3058, 2010.

[6] N. R. Howe, M. E. Leventon, and W. T. Freeman, "Bayesian reconstruction of 3d human motion from single-camera video," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 820–826, 1999.

[7] H. Ning, T. Tan, L. Wang, and W. Hu, "Kinematics-based tracking of human walking in monocular video sequences," *Image and Vision Computing*, vol. 22, no. 5, pp. 429–441, 2004.

[8] Z. Ziheng, A. Prügel-Bennett, and R. I. Damper, "A bayesian framework for extracting human gait using strong prior knowledge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1738–1752, 2006.

[9] W. Wang, X. Deng, X. Qiu, S. Xia, and Z. Wang, "Learning local models for 2D human motion tracking," in *Proceedings of the IEEE International Conference on Image Processing (ICIP '09)*, pp. 2589–2592, November 2009.

[10] F. Remondino and A. Roditakis, "Human figure reconstruction and modeling from single image or monocular video sequence," in *Proceedings of the Fourth International Conference on 3-D Digital Imaging and Modeling (3DIM '03)*, 2003.

[11] U. Gaur, A. Jain, and S. Goel, "Towards real-time monocular video-based avatar animation," *Lecture Notes in Computer Science*, vol. 5359, no. 2, pp. 949–958, 2008.

[12] B. Zou, S. Chen, C. Shi, and U. M. Providence, "Automatic reconstruction of 3D human motion pose from uncalibrated monocular video sequences based on markerless human motion tracking," *Pattern Recognition*, vol. 42, no. 7, pp. 1559–1571, 2009.

[13] C. Sminchisescu and B. Triggs, "Covariance scaled sampling for monocular 3D body tracking," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. I447–I454, December 2001.

[14] A. D. Sappa, N. Aifanti, S. Malassiotis, and M. G. Strintzis, "3D gait estimation from monoscopic video," in *Proceedings of the International Conference on Image Processing (ICIP '04)*, pp. 1963–1966, October 2004.

[15] R. D. Green and L. Guan, "Quantifying and recognizing human movement patterns from monocular video images—part I: a new framework for modeling human motion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 2, pp. 179–190, 2004.

[16] R. D. Green and L. Guan, "Quantifying and recognizing human movement patterns from monocular video images—part II applications to biometrics," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 2, pp. 191–198, 2004.

[17] X. Zhao and Y. Liu, "Generative tracking of 3D human motion by hierarchical annealed genetic algorithm," *Pattern Recognition*, vol. 41, no. 8, pp. 2470–2483, 2008.

[18] R. M. Gray, *Entropy and Information Theory*, Springer, New York, NY, USA, 1990.

[19] F. Escolano, P. Suau, and B. Bonev, *Information Theory in Computer Vision and Pattern Recognition*, 2009.

[20] H-ANIM WEB SITE, http://www.h-anim.org/.

[21] VRML WEB SITE, http://www.vrmlsite.com/.

[22] D. A. Winter, *Biomechanics and Motor Control of Human Movement*, Wiley-Interscience, 2nd edition, 1990.

[23] J. H. Yoo, M. S. Nixon, and C. J. Harris, "Extracting human gait signatures by body segment properties," in *Proceedings of the 5th IEEE Southwest Symposium on Image Analysis and Interpretation*, 2002.

[24] W. T. Dempster and G. R. L. Gaughran, "Properties of body segments based on size and weight," *American Journal of Anatomy*, vol. 120, no. 1, pp. 33–54, 1967.

[25] R. Boulic, N. M. Thalmann, and D. Thalmann, "A global human walking model with real-time kinematic personification," *The Visual Computer*, vol. 6, no. 6, pp. 344–358, 1990.

[26] M. P. Murray, "Gait as a total pattern of movement," *American Journal of Physical Medicine*, vol. 46, no. 1, pp. 290–333, 1967.

[27] M. Murray, A. B. Drought, and R. C. Kory, "Walking patterns of normal men," *The Journal of Bone and Joint Surgery*, vol. 46, pp. 335–360, 1964.

[28] M. P. Murray, R. C. Kory, B. H. Clarkson, and S. B. Sepic, "Comparison of free and fast speed walking patterns of normal men," *American Journal of Physical Medicine*, vol. 45, no. 1, pp. 8–23, 1966.

[29] P. A. Hageman and D. J. Blanke, "Comparison of gait of young women and elderly women," *Physical Therapy*, vol. 66, no. 9, pp. 1382–1387, 1986.

[30] K. M. Ostrosky, J. M. VanSwearingen, R. G. Burdett, Z. Gee, and M. Eastlack, "A comparison of gait characteristics in young and old subjects," *Physical Therapy*, vol. 74, no. 7, pp. 637–646, 1994.

[31] G. Bradski and A. Kaehler, *Learning OpenCV: Computer Vision with the OpenCV library*, O'Reilly Media, Inc., 2008.

[32] D. H. Douglas and T. K. Peucker, "Algorithms for the reduction of the number of points required to represent a digitized line or its caricature," *Cartographica*, vol. 10, no. 2, pp. 112–122, 1973.

[33] X. He, D. Cai, S. Yan, and H. J. Zhang, "Neighborhood preserving embedding," in *Proceedingsof the 10th IEEE International Conference on Computer Vision (ICCV '05)*, pp. 1208–1213, October 2005.

[34] A. Quattoni, S. Wang, L. P. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1848–1853, 2007.

[35] S. Abe, *Support Vector Machines for Pattern Classification*, Springer, London, UK, 2nd edition, 2010.

[36] S. B. Wang, A. Quattoni, L. P. Morency, D. Demirdjian, and T. Darrell, "Hidden conditional random fields for gesture recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, pp. 1521–1527, June 2006.

[37] L. P. Morency, A. Quattoni, C. M. Christoudias, and S. Wang, *Hidden-state Conditional Random Field Library*. User Guide, 2008.

[38] W. W. Hager and H. Zhang, "A new conjugate gradient method with guaranteed descent and an efficient line search," *SIAM Journal on Optimization*, vol. 16, no. 1, pp. 170–192, 2005.

[39] C. W. Hsu and C. J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[40] J. Friedman, "Another approach to polychotomous classification," Tech. Rep., Department of Statistics, Stanford University, 1996.

[41] O. Chapelle, P. Haffner, and V. N. Vapnik, "Support vector machines for histogram-based image classification," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1055–1064, 1999.
[42] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," *Science*, vol. 2, no. 3, pp. 1–39, 2001.
[43] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR '06)*, pp. 441–444, August 2006.
[44] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 162–177, 2005.