Kinematic-based Human Motion Analysis in Infrared Sequences

Bir Bhanu and Ju Han
Center for Research in Intelligent Systems
University of California, Riverside, California 92521, USA
{bhanu,jhan}@cris.ucr.edu

Abstract

In an infrared (IR) image sequence of human walking, the human silhouette can be reliably extracted from the background regardless of lighting conditions and colors of the human surfaces and backgrounds in most cases. Moreover, some important regions containing skin, such as face and hands, can be accurately detected in IR image sequences. In this paper, we propose a kinematic-based approach for automatic human motion analysis from IR image sequences. The proposed approach estimates 3D human walking parameters by performing a modified least squares fit of the 3D kinematic model to the 2D silhouette extracted from a monocular IR image sequence, where continuity and symmetry of human walking and detected hand regions are also considered in the optimization function. Experimental results show that the proposed approach achieves good performance in gait analysis with different view angles with respect to the walking direction, and is promising for further gait recognition.

1. Introduction

Current human recognition methods, such as fingerprints, face or iris biometrics, generally require a cooperative subject, views from certain aspects and physical contact or close proximity. These methods can not reliably recognize non-cooperating individuals at a distance in real-world situations with changing environmental conditions. Moreover, in various applications of personnel identification, many established biometrics can be obscured. Gait, which concerns recognizing individuals by the way they walk, has been used as a important biometric without the above-mentioned disadvantages.

In recent years, some approaches have already been employed for automatic gait recognition from regular cameras. Niyogi and Adelson [9] make an initial attempt in a spatiotemporal (XYT) volume. They first find the bounding contours of the walker, and then fit a stick model on them. A characteristic gait pattern in XYT is generated from the model parameters for recognition. Little and Boyd [7] propose a model-free approach making no attempt to recover a structural model of human motion. Instead they describe the shape of the motion with a set of features derived from moments of a dense flow distribution. Similarly, He and Debrunner's [2] approach detects a sequence of feature vectors based on Hu's moments of motion segmentation in each frame, and the individual is recognized from the feature vector sequence using hidden Markov models. To avoid the feature extraction process which may reduce reliability, Murase and Sakai [8] propose a template matching method to calculate the spatio-temporal correlation in a parametric eigenspace representation for gait recognition. Huang et al. [4, 3] extend this approach by combining canonical space transformation (CST) based on canonical analysis, with eigenspace transformation (EST) for feature selection.

Unlike regular cameras which detect visible light, a long wave infrared (IR) camera records electromagnetic radiation emitted by objects in a scene as thermal images whose pixel values represent temperature. Therefore, in an IR image that consists of human in a scene, a simple thresholding can extract the human silhouette from the background regardless of lighting conditions and colors of the human surfaces and backgrounds, because the temperatures of the human body and background are different in most cases. In cool weather or without direct sunshine, a person will normally be warmer than the background due to body's internal heat. In most other cases, such as with the sun shining on a person's back, the ground may be warmer than the person. Moreover, by setting the temperature range to approximate human body temperature, image regions corresponding to human skin can be easily and reliably detected in IR images [1, 5, 10].

In this paper, we propose a kinematic-based approach to analyze gait from IR image sequences. The proposed approach automatically estimates 3D human model parameters by performing a least squares fit of the 3D kinematic model to the 2D silhouette extracted from an IR image sequence. Moreover, continuity and symmetry of human walking and detected hand regions are considered in the optimization function. Our approach also eliminates the assumption of human walking frontoparallel to the image plane in most existing gait recognition approaches. Thus, in our approach, people can walk at different angles with respect to the image plane.

2. Image Analysis

2.1. Silhouette Extraction

Assuming that people are the only moving objects in the scene, they can be extracted by a simple background subtraction method. Given an IR image sequence containing moving people and the corresponding background IR image, for each frame I_i in the sequence, the difference $\Delta p_i(x, y) = ||p_i(x, y) - p_b(x, y)||$

Proceedings of the Sixth IEEE Workshop on Applications of Computer Vision (WACV'02)
0-7695-1858-3/02 $17.00 © 2002 IEEE

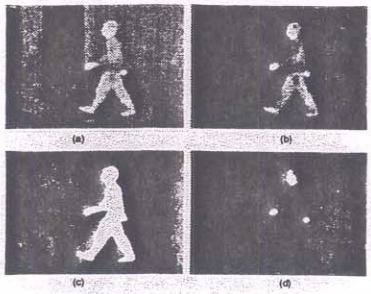

Figure 1. An example of silhouette extraction results: (a) original IR image; (b) background-subtracted image; (c) segmented human silhouette; (d) segmented face and two hands.

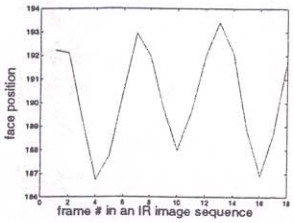

Figure 2. An example of face position (row position from the bottom of the image) variation in an IR image sequence.

is computed for each pixel, where $p_i(x,y)$ and $p_b(x,y)$ are the intensity (temperature) values of the pixel at (x,y) in the ith frame and background image, respectively. Given a threshold t, if $\Delta p_i(x,y) > t$ the pixel at (x,y) is determined to be part of the moving objects; otherwise, it is part of the background. In addition, by setting the temperature range to approximate human body temperature, image regions corresponding to human skin, such as face and hands, can be easily and reliably detected in IR images. An example of silhouette extraction results is shown in Figure 1.

2.2. Pre-processing

In a human walking sequence, the hand on the camera side is always visible and the other hand is occluded by the body in some frames. According to the walking direction, hand regions in each frame, and the continuity of the hand position in adjacent frames, we can determine which hand is the left hand in each frame. This is very important in our 3D model matching. Given a 2D human silhouette, there would be many additional corresponding solutions of 3D model parameters if the order of left and right body parts were unknown.

Once the order of left and right body parts in each frame is known, a sequence can be divided into motion cycles. The motion cycle can be detected by exploiting the face position (row position from the bottom of the image) variation in a sequence as shown in Figure 2. In each walking cycle, the silhouette with highest local face position means that person stands straight in that frame; the silhouette with lowest local face position means the person is in mid-stride. Our approach will be carried out during these walking cycles.

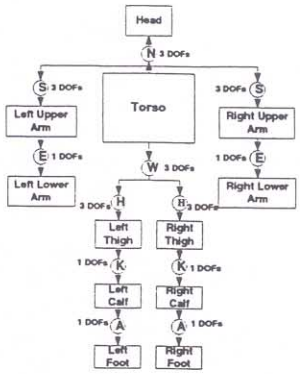

Figure 3. 3D Human Kinematic Model (the hand is viewed as a part of the lower arm).

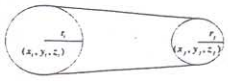

Figure 4. Body part geometric representation.

After the silhouette has been cleaned by a filtering procedure, its height, width and centroid can be easily extracted for subsequent motion analysis. In addition, the moving direction of the walking person is determined as follows

$$\theta = \begin{cases} tan^{-1}\frac{f(h_1-h_N)}{h_1 y_N - h_N y_1}, & \text{if } y_1 > y_N; \\ tan^{-1}\frac{f(h_1-h_N)}{h_1 y_N - h_N y_1} + \pi, & \text{otherwise.} \end{cases} \quad (1)$$

where f is the camera focal length, y_1 and y_N are the horizontal centroid of the silhouette in the first and last frames of the walking cycle, and h_1 and h_N are the height of the silhouette in the first and last frames of the walking cycle.

3. Technical Approach

3.1. 3-D Human Kinematic Model

A human body is considered as an articulated object, consisting of a number of body parts. The body model adopted here is shown in Figure 3, where a circle represents a joint and a rectangle represent a body part (N: neck, S: shoulder, E: elbow, W: waist, H: hip, K: knee, and A: ankle).

Most joints and body part ends can be represented as spheres, and most body parts can be represented as cones. The whole human kinematic model is represented as a set of cones connected by spheres [6]. Figure 4 shows that most of the body parts can be approximated well in this manner. However, the head is approximated only crudely by a sphere and the torso is approximated by a cylinder with two spheroid ends.

3.2. Matching 3D Model with 2D Silhouette

The matching procedure determines a parameter vector $\mathbf{x}$ so that the proposed 3D model fits the given 2D silhouette as well as possible. For that purpose,

Proceedings of the Sixth IEEE Workshop on Applications of Computer Vision (WACV'02)
0-7695-1858-3/02 $17.00 © 2002 IEEE

two transformations transform human body local coordinates (x, y, z) into image coordinates (x', y')[11]: the first transformation transforms local coordinates into camera coordinates; while the second transformation projects camera coordinates into image coordinates.

In the human body local coordinates (x, y, z), the x axis is the human vertical axis, the y axis is along the walking direction, the z axis is perpendicular to the (x, y) plane, and the origin is the human centroid. In the image coordinates (x', y'), the x' and y' axes are the vertical and horizontal axes, respectively. In the camera coordinates (x'', y'', z''), assuming that the camera axis is parallel to the ground, the z'' axis is the camera axis, and the x'' and y'' axes are along the x' and y' axes in the image coordinate, respectively. Assuming the origin of the human local coordinate at a specific moment in the camera coordinate (x'', y'', z'') is (x_0'', y_0'', z_0'') and the walking angle with respect to the image plane of the camera is θ given by Equation (1), we have $x'' = x_0'' + x$, $y'' = y_0'' + y\cos\theta - z\sin\theta$, $z'' = z_0'' + y\sin\theta + z\cos\theta$, and $x' = fx''/z''$, $y' = fy''/z''$, where f is the camera focal length.

Each 3D human body part is modeled by a cone with two spheres $\mathbf{s}_i$ and $\mathbf{s}_j$ at its ends, as shown in Figure 4 [6]. Each sphere $\mathbf{s}_i$ is fully defined by 4 scalar values, (x_i, y_i, z_i, r_i), which define its location and size. Given these values for two spheroid ends (x_i, y_i, z_i, r_i) and (x_j, y_j, z_j, r_j) of a 3D human body part model, its projection $P_{(ij)}$ onto the image plane is the convex hull of the two circles defined by (x_i', y_i', r_i') and (x_j', y_j', r_j').

Therefore, given a set of 3D sphere parameters $\mathbf{x} = \{\mathbf{x}_i : (x_i, y_i, z_i, r_i)\}$ and camera parameters, its projection $P_{\mathbf{x}}(x', y')$ is uniquely determined, where x', y' are 2D image plane coordinates.

3.3. Model Parameter Selection

Human motion is very complex due to so many degrees of freedom (DOFs). To simplify the matching procedure, we use the following reasonable assumptions: (1) the camera is stationary; (2) people are walking before the camera at a distance; (3) people are moving in a constant direction; (4) the swing direction of arms and legs parallels to the moving direction. According to these assumptions, we do not need to consider the waist joint, and only need to consider one DOF for each other joint. Therefore, the elements of the parameter vector of the 3D human kinematic model are defined as: (a) Radius $r_i(11)$: torso(3), shoulder, elbow, hand, hip, knee, ankle, toe, and head; Length $l_i(9)$: torso, inter-shoulder, inter-hip, upper arm, forearm, thigh, calf, foot, and neck; (b) Location $(x, y)(2)$; Angle $\theta_i(11)$: neck, left upper arm, left forearm, right upper arm, right forearm, left thigh, left calf, left foot, right thigh, right calf, and right foot. With 33 stationary and kinematic parameters, the projection of the human model can be completely determined.

3.4. Model Parameter Estimation

If the 2D human silhouette is known, we may find the related 3D body parts locations and orientations with the knowledge of camera parameters. A simple method is to perform a least squares fit of the 3D human model to the 2D human silhouette. That is, to estimate the set of sphere parameters $\mathbf{x} = \{\mathbf{x}_i : (x_i, y_i, z_i, r_i)\}$ by choosing $\mathbf{x}$ to minimize

$$J = error(\mathbf{x}; I) = \sum_{x', y' \in I} (P_{\mathbf{x}}(x', y') - I(x', y'))^2, \quad (2)$$

where I is the silhouette binary image, $P_{\mathbf{x}}$ is the binary projection of the 3D human model to image plane, and x', y' are image plane coordinates.

However, model fitting in a single frame is unreliable, especially when the segmentation result is not good. A better method is to consider all the frames in a walking cycle as a whole. In this way, we can consider not only the continuity and symmetry of model parameters in a walking cycle but also the extracted hand regions as important references. We accordingly modify Equation (2) as follows

$$J' = E + \alpha C + \beta S + \gamma H. \quad (3)$$

- $E = \sum_{i=1}^{N} error(\mathbf{x}_i; I)$ is the matching error over the complete walking cycle. $error(\mathbf{x}_i; I)$ is the matching error between the ith projected model image and corresponding silhouette image which is defined in Equation (2).
- C is the discontinuity error of the kinematic parameters on the complete walking cycle. Discontinuity means that the variation of the motion parameters is not smooth during the complete walking cycle.
- S is the non-symmetry error of the kinematic parameters on the complete walking cycle. Non-symmetry means that the motion parameters of the left body parts in the first half cycle do not coincide with the motion parameters of the right body parts in the second half cycle.
- H is the lower arm matching error over the complete walking cycle. In our approach, it is the standard deviation of the average pixel value of the lower arm (including hand) on the camera side. This makes use of the detected hand regions in the IR images.

In Equation (3), α, β and γ are corresponding weights which will be explained in next section.

To estimate the model parameters over the whole walking cycle, we first initialize these parameters according to average human statistical information. Then, the set of parameters is estimated from these initial parameters by choosing a parameter vector $\mathbf{x}$ to minimize the optimization function J' in equation (3) with respect to some human kinematic constraints.

A Genetic Algorithm (GA) is appropriate to solve this optimization problem. GA provides a learning method motivated by an analogy to biological evolution. Rather than search from general-to-specific hypotheses, or from simple-to-complex, GA generates successor hypotheses by repeatedly mutating and recombining parts of the best currently known hypotheses. At each step, a collection of hypotheses called

Proceedings of the Sixth IEEE Workshop on Applications of Computer Vision (WACV'02)
0-7695-1858-3/02 $17.00 © 2002 IEEE

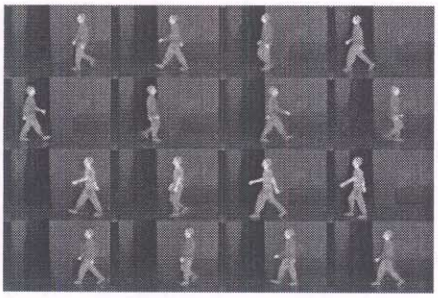

Figure 5. Four IR sequences used in our experiments with walking angles of -7, 182, 3, and 9 degree of arc, respectively. Each row represents one sequence with four example frames shown.

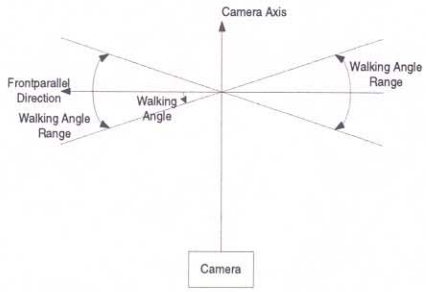

Figure 6. Illustration of the walking direction angle from top view.

the current population is updated by replacing some fraction of the population by offspring of the most fit current hypotheses. After a large number of steps, the hypotheses having the best fitness are considered as solutions. In our approach, we use 7 bits to represent each body part length and 5 bits to represent each angle in the given range (totally $7*20+5*13*10=790$ bits for a 10-frame cycle); fitness function is $(1-(J'-J'_{min})/(J'_{max}-J'_{min}))$; population size is 100; crossover rate is 0.5; crossover method is uniform crossover; mutation rate is 0.2; the GA will terminate if the fitness values have not changed for 15 successive steps.

4. Experimental Results

The IR data used in our experiments are real human walking data from the same person recorded in an indoor environment. In the four testing sequences, one person walks in different directions as shown in Figure 5. In each sequence, the walking direction θ remains the same, ranging from $-10°$ to $10°$ along the image plane of the camera as illustrated in Figure 6. The size of IR image is 240×320.

In our experiments, we first set the range of body parts length and angle variation from human statistical data. With regard to the weights in Equation (3), they cannot be large because the matching error should be the dominant factor in J'. Also, they should be chosen in such a way that all the three additional factors have a similar contribution to J'. In our experiments, we choose $\alpha = 1/6$, $\beta = 1$ and $\gamma = 6$ for all

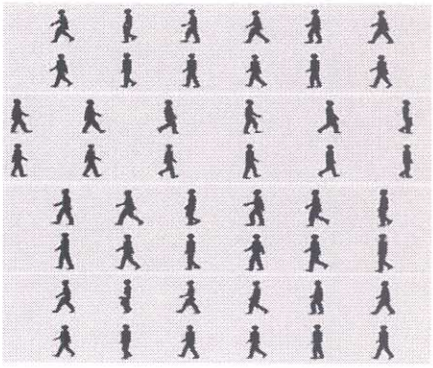

Figure 7. Matching results for the four sequences. Each block represents one sequence with six example frames shown. In each block, the first row consists of the segmented silhouette and the second row show the projected model images.

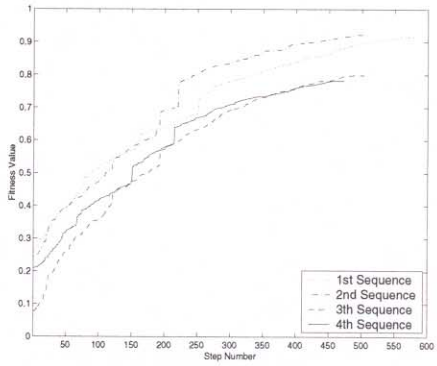

Figure 8. Fitness value variation in a typical run of matching four sequences.

the four sequences according to the ratio of the average values of these factors. Figure 7 shows typical matching results for some frames in the four image sequences. Our approach achieves good matching performance between the 2D silhouette and the projected 3D model, although they may not seem *perfect* in some frames on close inspection due to the non-zero matching errors. The fitness value variation during the matching procedure for the four sequences is shown in Figure 8. It demonstrates that our GA-based matching algorithm converges after $500 \sim 600$ steps for the complete cycle. The size of the search space is approximately $2^{790} \approx 10^{237}$, where 790 is the number of bits in the encoded string of a 10-frame cycle. The efficiency of GA as a tool for global optimization is clear by examining the ratio of the search space examined divided by the size of the search space. Note that the size of the search space is greatly reduced by the high degree dependency of the parameters in adjacent frames.

The estimated angle variation curves from the four sequences are shown in Figure 9. Figure 9 is obtained by normalizing each walking cycle to 20 frames so that walking at different speeds does not affect the motion

Proceedings of the Sixth IEEE Workshop on Applications of Computer Vision (WACV'02)
0-7695-1858-3/02 $17.00 © 2002 IEEE

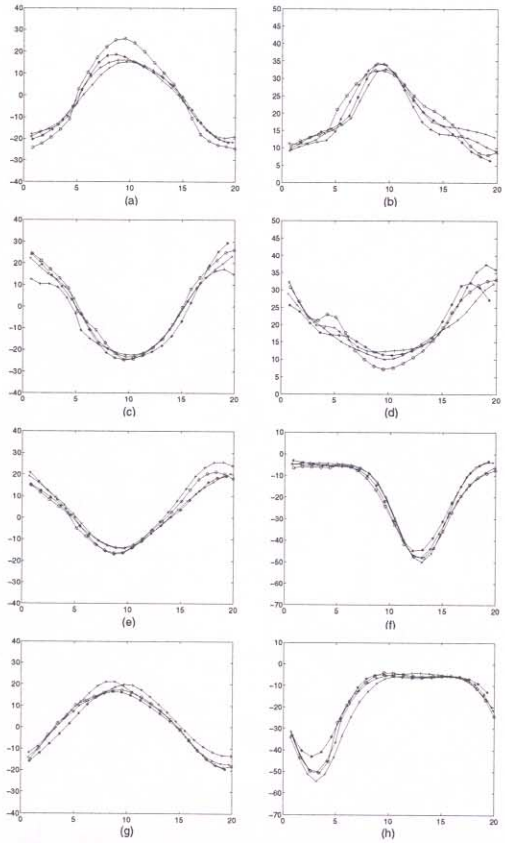

Figure 9. Estimated angle variation curves from the 4 normalized cycles from different sequences for (a) left upper arm, (b) left forearm, (c) right upper arm, (d) right forearm, (e) left thigh, (f) left calf, right thigh, (g) right calf, and (h) right foot, respectively. The horizontal axis is the normalized frame number, and the vertical axis is the angle.

curve. Table 1 shows the maximum, minimum and average values of the standard deviation over the complete normalized sequence. Each value of the standard deviation are computed from the four angle values at each frame. The experimental results show that the curves coincide well even when walking in different directions. This means that our approach can compute very similar human motion patterns from different sequences of the same person.

5. Conclusions

In this paper, we proposed an approach to estimate 3D human motion from a monocular IR image sequence for automatic gait recognition. The proposed approach performs a least squares fit of a complex 3D human model to the 2D human silhouette. Continuity and symmetry of human walking and detected hand regions are also considered in the optimization function. Experimental results show that the proposed approach achieves good performance in gait analysis with different view angles with respect to the walking direction.

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
max.	4.1	3.3	5.4	3.2	4.7	5.6	3.3	6.6
min.	0.7	0.6	0.5	0.5	0.2	0.4	0.6	0.2
ave.	2.4	1.6	2.5	1.8	1.9	1.7	1.9	2.3

Table 1. The maximum, minimum and average values (degree of arc) of the standard deviation over the complete normalized sequence.

In the future with large database, we plan to evaluate the parameter estimation in the presence of intra-personal variation. The angle differences among different persons might be more than 10° of arc and the average intra-personal variation value is about 2° (see Table 1). Thus, we expect to extract promising gait features from these curves. In addition, we plan to deal with complex situations like the 2D silhouette is not properly extracted (due to partial occlusions or the person wearing a thick coat), the direction of motion of the person changes, or the hands are not accurately identified.

Acknowledgment

This work was supported in part by grants F49620-97-1-0184, F49620-02-1-0315 and DAAD19-01-0357; the contents and information do not necessarily reflect the position or policy of U.S. Government.

References

[1] H. Arlowe. Thermal detection contrast of human targets. in *Proc. IEEE International Carnahan Conference on Security Technology*, pages 27–33, 1992.

[2] Q. He and C. Debrunner. Individual recognition from periodic activity using hidden markov models. in *Proc. IEEE Workshop on Human Motion*, pages 47–52, 2000.

[3] P. Huang. Automatic gait recognition via statistical approaches for extended template features. *IEEE Trans. SMC, Part B*, 31(5):818–824, 2001.

[4] P. Huang, C. Harris, and M. Nixon. Recognizing humans by gait via parameteric canonical space. *Artificial Intelligence in Engineering*, 13:359–366, 1999.

[5] S. Iwasawa, K. Ebihara, J. Ohya, and S. Morishima. Real-time estimation of human body posture from monocular thermal images. in *Proc. IEEE Conference on CVPR*, pages 15–20, 1997.

[6] M. Lin. Tracking articulated objects in real-time range image sequences. in *Proc. ICCV*, pages 648–653, 1999.

[7] J. Little and J. Boyd. Recognizing people by their gait: the shape of motion. *Videre: Journal of Computer Vision Research*, 1(2):1–32, 1998.

[8] H. Murase and R. Sakai. Moving object recognition in eigenspace representation: gait analysis and lip reading. *Pattern Recognition Letters*, 17(2):155–62, 1996.

[9] S. Niyogi and E. Adelson. Analyzing and recognizing walking figures in xyt. in *Proc. IEEE Conference on CVPR*, pages 469–474, 1994.

[10] Y. Sato, Y. Kobayashi, and H. Koike. Fast tracking of hands and fingertips in infrared images for augmented desk interface. in *Proc. IEEE Conf. on Automatic Face and Gesture Recognition*, pages 462–467, 2000.

[11] S. Wachter and H.-H. Nagel. Tracking of persons in monocular image sequences. in *Proc. IEEE Workshop on Nonrigid and Articulated Motion*, pages 2–9, 1997.

Proceedings of the Sixth IEEE Workshop on Applications of Computer Vision (WACV'02)
0-7695-1858-3/02 $17.00 © 2002 IEEE