人間行動パターン理解のための人体運動追跡

郭　硯　　徐　剛　　辻　三郎

大阪大学　基礎工学部
〒560　大阪府豊中市待兼山町１番１号

あらまし
本論文ではスティックモデル(stick figure model)を用いた人体運動の追跡について新しい手法を提案する。人体の輪郭を元にして作られるポテンシャル場において最小エネルギーを持つスティックモデルを探すことによって、画像にスティックモデルをフィットさせる。このポテンシャル場はモデルと実画像データとの間の橋渡しとなる。探索は画像系列の各画像において、過去の情報を用い、新しい位置を予測しながら、動的計画法によって効率的に行う。

本手法は人間行動パターンの認識のような高次のタスクにとって必要な記述を直接に得ることができ、従来のボトムアップ的な手法と比べて優位性がある。実画像を用いた実験結果は本手法のロバスト性を示す。

TRACKING HUMAN BODY MOTION IN DYNAMIC IMAGE SEQUENCES FOR UNDERSTANDING HUMAN MOTION PATTERNS

Yan GUO　Gang XU　Saburo TSUJI

Department of Control Engineering, Osaka University
1-1, Machikaneyame, Toyonaka, Osaka 560 Japan

Abstract

This paper presents a method for tracking human body motion based on a stick figure model . The fitting of the model to image data is realized by seeking the minimal energy stick figure in a potential field around the contour of the moving person. This potential field bridges the gap between the high level model and low level data. We use a dynamic programming algorithm to search for the optimal stick figure guided by an adaptive predictor which provides a rough estimate for each frame in the image sequence.

This method can provide parameters to meet the needs in high level tasks such as recognition of human motion patterns, thus it has obvious advantages over conventional bottom-up approaches. Experiment results for real world data show the robustness of the method.

1. Introduction

Analysis and understanding of human motion by computer has been a challenging issue for decades ([2] to [12]). Computer-based analysis of human motion has found its applications in sports, training of the disabled, and many other fields. Human motion understanding also has very special significance in the area of motion pattern recognition, and is expected to provide more friendly human-machine interface.

The first step of understanding human motion is to track human body motion in dynamic image sequences. This paper addresses this issue. The very primitive method is to track the human body motion manually, but it is very cumbersome since we usually have to track the motion in a long image sequence. Furthermore, the motion pattern understanding would have less meaning if it was based on manual tracking. A more applicable method is to fix visible marks, such as LEDs, on joint points and tracking these marks by the template matching . This method can be used in prepared experiments, but as for real world images, we cannot readily fix this sort of marks. Thus, it is obvious that an automatic method, with less restriction to subjects and with less interaction with operators, is desired. This kind of attempts can be found in [2], [4], [7], [8], [9], [10].

In general, human motion tracking includes two parts. One is the modeling of human body and the other is the finding of the correspondence between model and real data. Considering the human body as an articulated connection of rigid parts, it gives the idea to model human bodies with stick figures. Many researchers have used this model in their work [5], [8], [11], [12].

In this paper, we present a novel method for tracking human body motion based on a stick figure model. The stick figure model can efficiently represent the human body structure because the human body is a locally deformable object having strong constraints between its parts. It can be regarded as a connection of several rigid parts articulated at joint points. The mathematical description of a stick figure is very compact, requiring only a small number of parameters. The gaps between real data and high level models like the stick figure model are bridged by the concept of the *potential field*. A potential field is generated based on the silhouette of moving figure. After a potential field is built, the correspondence can be established by searching the potential field for the stick figure with the minimal energy. In the search process, we use the dynamic programming algorithm to reduce computation, and apply an adaptive predictor to guide the search in the consecutive image sequence. The experiments for real image sequences show that our method works well. The flowchart of our method is illustrated in Fig. 1.

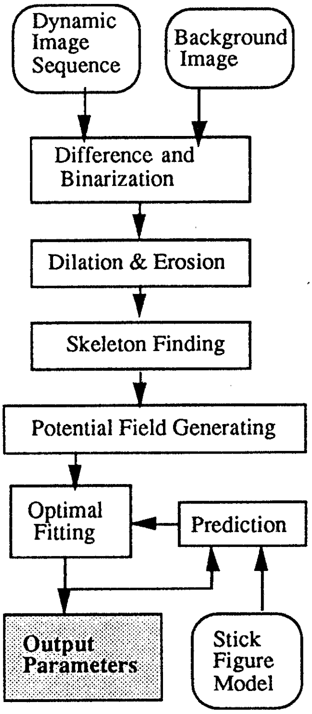


Fig. 1 Flowchart of Our Approach

2. Stick Figure Model

In our work, the scene of a person walking or running on a treadmill placed on the horizontal plane is videotaped with the camera axis parallel to the horizontal plane and vertical to the person's moving direction. By using a treadmill, we can get long image sequences without moving our camera. Under this condition, we can model the human body structure by a 2D stick figure model shown in Fig. 2.

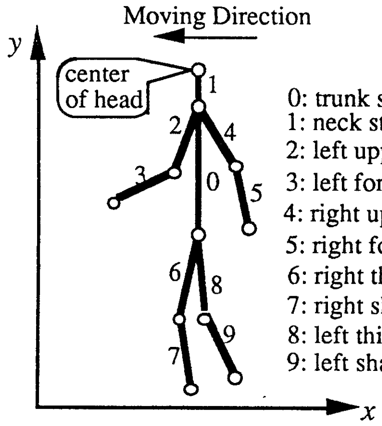


Fig. 2 2D Stick Figure Model

The x axis of the coordinate system is on the horizontal plane and parallel to the person's moving. The y axis is vertical to the horizontal plane. The stick figure model can be described by a vector:

$$S = [x, y, d, h_t, h_a, h_l, h_n, \alpha_0, \alpha_1, \ldots, \alpha_9]$$

where its components are:

(x, y): position of the center of head

d: equal to 1 for forward moving
equal to - 1 for backward moving

h_t: length of the trunk stick

h_a: length of the arm sticks

Here we assume that the forearms and the upper arms have the same length.

h_l: length of the leg sticks

Also we assume that the thighs and the shanks have the same length.

h_n: length of the neck stick

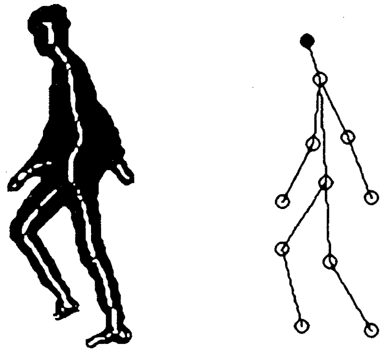
α_0: angle between stick 0 and y axis

$\alpha_1, \ldots, \alpha_9$: angles between adjacent sticks

Our goal now can be expressed as finding the optimal figure vector for each image frame in the image sequence. We leave the description to the following sections.

3. Potential Field

It is not difficult to extract the silhouette and skeleton of a human figure from the real scene image, as shown in Fig. 3 (a). In our experiment, we assume that the scene contains only a single person and the static background image is memorized in advance. Subtracting the background image from each frame, and binarizing it with an appropriate threshold, we can get the binary silhouette of the person in motion. The skeleton of the silhouette is obtained by a classical method given in [1]. The skeleton is represented by the white lines within the silhouette in Fig. 3 (a).



(a) (b)

Fig.3 (a) Silhouette and skeleton of the figure
(b) Stick figure fitting to the skeleton

The correspondence between the stick figure model and the silhouette can be established by fitting the stick figure, such as that shown in Fig. 3(b), to the skeleton. To formulate the fitting, we define notations first.

K = {k }, k: a point on the skeleton

T = { t }, t: a point on the stick model

where T can be computed from stick figure model vector S.

d(x,y): Euclidean distance between two points in the image plane.

we now define the energy function to evaluate the fitting of T to K in the sense of the least mean square as:

$$E(K,T) = \sum_{t \in T} \min_{k \in K} d^2(k,t) \qquad (3.1)$$

To minimize the total energy, we have to make use of some kind of search algorithm. For every search, we have to compute the energy defined in (3.1). The computation is immense because the search space for finding the optimal solution is very large. To reduce this computation, we rewrite (3.1) to:

$$E(K,T) = \sum_{t \in T} P(t) \qquad (3.2)$$

where

$$P(x) = \min_{k \in K} d^2(k,x) \qquad (3.3)$$

and x is an arbitrary point in the image. This P(x) can be generated before the search procedure and can be used in every search for the same frame. In other words, we only need to compute P(x) once for each frame and the computation can be implemented in parallel. We call P(x) defined in (3.3) the potential field. In terms of the potential field, the original correspondence problem can be redescribed as finding a stick figure which has minimal energy in the potential field.

However, only the skeleton information is used if we compute the potential field using (3.3). There would be no problem if the skeleton points were really on the axes of the parts of the human body. Unfortunately, the skeleton points obtained from the silhouette are not always on the axes because the human body in motion occludes itself very frequently. When the self-occlusion occurs, the skeleton obtained from the silhouette deviates from the body axes, and two or more axes may merge to one skeleton line. Moreover, the skeleton may not be connected or unneeded skeleton points may exist due to the image noise or shadows despite we take several measures to obtain more reliable skeletons.

To overcome the problems mentioned above, we improve the definition of potential field (3.3) to the form of (3.4):

$$P(x) = \min_{k \in K} \{1-\exp(-d^2(k,x)/2\sigma_k^2)\} \quad (3.4)$$

where

$\sigma_k = \{\text{nearest distance from k to contour}\}/3$

and x is an arbitrary point in image plane. Fig.4 shows the comparison of the square type function in (3.3) and the Gaussian type function in (3.4).

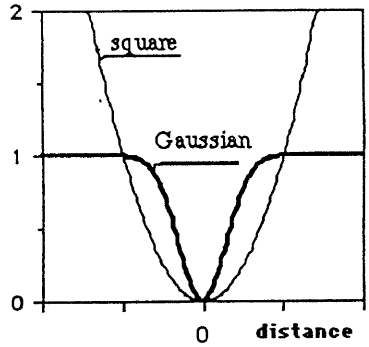


Fig.4. Comparison of the square type function and the Gaussian type function

We see from Fig. 4 that the potential value of each skeleton point is normalized to a range of [0,1) in the Gaussian type function, and the contribution of each skeleton point to the whole potential field is limited to a local scope that can be controlled by σ_k. Here we let σ_k be equal to the third of the minimal distance, obtained in the stage of computing skeletons, from skeleton point k to the contour of the silhouette. The valleys of the potential field defined by (3.4) are very similar in shape with the original silhouette. In definition (3.4), not only skeleton information but also the information about the silhouette width is contained. This gives us the following merits:

(1) The search becomes insensitive to the deviation of the skeleton from the axes of body parts because the potential field there usually appears like a wide valley. The deviation can be remedied by searching global minimal energy for the whole stick figure.

(2) Continuous potential valley may be obtained even in the place where the skeleton is not connected.

(3) The influence of noise skeleton points are limited to small local scopes because these points often have very small σ_k values.

We now consider the problem of finding the optimal stick figure which minimizes the total energy in the potential field. We find that the dynamic programming algorithm fits our goal. Details are given in the next section.

4. Dynamic Programming Algorithm

The dynamic programming algorithm proposed to solve the minimal path problem has been found to be a powerful tool to the problems in the field of pattern recognition and computer vision. We can change our problem to a minimal path problem if we regard the joint points and terminal points in our stick figure model as the nodes, and let the energy of each stick in the potential field be the arc length. Here we have one start node (head node) and four end nodes (two wrist nodes, two ankle nodes). We arrange our path structure as that shown in Fig. 5.

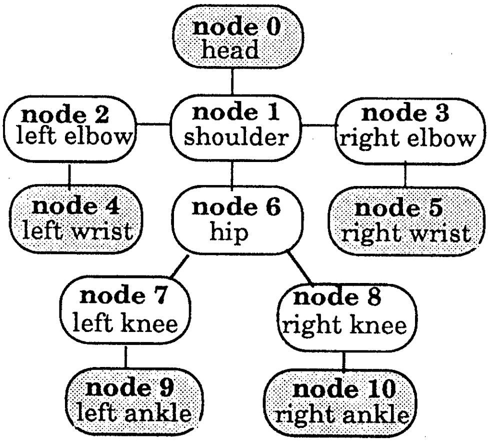


Fig. 5. Path Structure for Dynamic Programming

The five terminal nodes are node 0, node 4, node 5, node 9 and node 10, as shown in Fig. 5. They are nodes that has only one arc connected to the other nodes. We select the head node, that is, node 0 as start node, and the other terminal nodes as end nodes just for our convenience. The other choices will give the same solution.

Now we formulize the dynamic programming algorithm applied to our problem. Let A(n) be the search area for node n, and $e(P_i,P_j)$ be the energy of a stick from point P_i to point P_j, where P_i is an arbitrary point in A(i) and P_j is an arbitrary point in A(j). We denote the minimal energy at a point P_i of A(i) to end nodes node t_1, ..., node t_N as $E(P_i|t_1,...,t_N)$, thus the minimal total energy of the optimal stick figure can be expressed as $E(P_0|4,5,9,10)$, which can be determined by the following steps:

Step 0:

$E(P_i|i) = 0$, for each P_i in A(i); i is 4,5,9,10.

Step 1:

Step 1.1:

$$E(P_8|10) = \min_{P_{10}} \{e(P_8,P_{10}) + E(P_{10}|10)\}$$

$$E(P_6|10) = \min_{P_8} \{e(P_6,P_8) + E(P_8|10)\}$$

$$E(P_7|9) = \min_{P_9} \{e(P_7,P_9) + E(P_9|9)\}$$

$$E(P_6|9) = \min_{P_7} \{e(P_6,P_7) + E(P_7|9)\}$$

$$E(P_6|9,10) = E(P_6|9) + E(P_6|10)$$

$$E(P_1|9,10) = \min_{P_6} \{e(P_1,P_6) + E(P_6|9,10)\}$$

Step 1.2:

$$E(P_3|5) = \min_{P_5} \{e(P_3,P_5) + E(P_5|5)\}$$

$$E(P_1|5) = \min_{P_3} \{e(P_1,P_3) + E(P_3|5)\}$$

Step 1.3:

$$E(P_2|4) = \min_{P_4} \{e(P_2,P_4) + E(P_4|4)\}$$

$$E(P_1|4) = \min_{P_2} \{e(P_1,P_2) + E(P_2|4)\}$$

Step 2:

$$E(P_1|4,5,9,10) = E(P_1|9,10)+E(P_1|4)+E(P_1|5)$$

$$E(P_0|4,5,9,10) = \min_{P_1} \{e(P_0,P_1)+E(P_1|4,5,9,10)\}$$

Above procedure yields the minimal energy for an arbitrary point P_0 in the start node area. Backtracking the point which gives minimal energy in every step, we can obtain the optimal path that minimizes the total energy. It is very simple to transform this node expression to the vector expression S we defined in Section 2. This is the optimal stick figure model vector we wish to find.

To add the constraints embedded in the stick figure model to the search procedure, we check if there is any conflictions with the model for each stick between two points P_i and P_j when computing $e(P_i,P_j)$. The check is to make sure if both the stick length and the stick angle with respect to its adjacent stick are within the plausible ranges. If the stick passes the check, $e(P_i,P_j)$ equals to the sum of the potential values of every point on the stick; otherwise, we assign it to a very large value. This protects us from choosing unreasonable stick positions in the programming procedure.

The start node (the head node) position P_0 is determined by the template matching technique. Because the head is seldom occluded by the other parts, the template matching can give a reliable start node position.

We need an initial search area A(i) for each node i to implement the algorithm. These areas (including sizes and positions) are given by an interactive manner for the first frame in the image sequence. However, from the second frame, we can use the stick figure vectors readily obtained to predict the stick figure vector for the next frame. Changing the estimated vector to the node positions, we can seek the optimal stick figure for the next frame from these predicted positions using the same algorithm. In the following Section, we discuss the prediction method.

5. Adaptive Predictor Algorithm

There are many methods, for instance, Kalman filter method, to predict future values using the present values and the past values of a stochastic process. Most of these methods, the second-order moments of all processes are required. However, we have no knowledge of prior statics of our processes, and we do not like to be involved with the complex estimation problem. Thus, we only choose a very simple adaptive ARMA filter as our predictor.

In the stick figure vector S, only stick angle parameters are predicted, because the stick length parameters change very less and the head point is tracked by the template matching. Further, we assume these angles are independent of each other. Let $\alpha[n]$ be an arbitrary angle component in time sequence S[n], the problem under consideration is phrased as to predict $\alpha[n+1]$ in terms of the sum

$$\hat{\alpha}[n+1]=\sum_{k=0}^{N-1} a_k[n]\alpha[n-k] \qquad (5.1)$$

where coefficients $a_k[n]$ are time-dependent and they are so chosen as to reduce somehow the instantaneous error

$$\varepsilon[n+1] = \alpha[n+1] - \hat{\alpha}[n+1] \qquad (5.2)$$

To determine $a_k[n]$, we evaluate $a_k[n]$ recursively according to the Widrow algorithm [13]

$$a_k[n] = a_k[n-1] + \mu\bar{\varepsilon}[n]\bar{\alpha}[n-k] \qquad (5.3)$$

In the above, μ is a positive constant and $\bar{\varepsilon}[n]$ and $\bar{\alpha}[n]$ are quantized versions of $\varepsilon[n]$ and $\alpha[n]$ respectively. The *adaptation constant* μ is dictated by two conflicting requirements: If μ is small, the adaptation is slow and might not follow rapid changes; if μ is large, then the error might increase.

In our experiments, we use the present value and only one past value to predict the next value, that is, N in (5.1) equals to 2. The initial values of $a_k[n]$ are set to

$a_0[0] = 2$ and $a_1[0] = -1$

The prediction have two main merits. The first is that the computation is decreased because we can search less areas around the predicated positions. The second, and more important, is it permits us to overcome the self-occlusion problem. During the period when the two legs overlap each other, it is difficult to recognize the left leg from the right one without the history information. The predicted positions follow the changes based on the history information. A real example is given in the next section.

6. Experiments and Results

In our experiments, the image sequences are recorded at video speed, that is, 30 frames / sec. We preserve only the odd line data to remove the motion blur due to interlace scanning. The image size is 256*240.

Fig.6 to Fig.8 show an example. Fig.6 is the background image, The first column of Fig.7 are images of a walking man, arbitrarily selected from a 30 frame long image sequence. The second column of Fig.7 are the silhouette images obtained by subtracting the background. To remove the isolated points outside the silhouette and fill the holes within it, we erode the images once, then dilate them twice and then erode them one more time. The third column of Fig.7 are improved silhouette images. The skeletons are calculated by Montanari algorithm and non-significant skeleton points are suppressed by assigning the threshold K to a value about 0.5. Obtained skeletons are shown in the fourth column of Fig.7. Their potential fields calculated by (3.4) are shown in the fifth column of Fig.7.

The first image of Fig. 8 shows the stick figure fitted manually to the potential field, the others are tracked automatically. It can be seen that our method tracks the two legs successfully even during the period when two legs overlap each other. However, the tracking of arms is omitted at present because they are occluded by the truck silhouette most of the time.

7. Conclusions and discussions

In this paper, we presented a new method for tracking human motion based on a stick figure model. The key idea we proposed is the using of a potential field to bridge over the gap between the model and real data. Our final objective is to recognize the human motion patterns. Experimental results have shown that the method can work stably under different conditions, and is able to provide us with model parameters to achieve the final goal. Our future work will focus on the human motion recognition problems. Meanwhile, we will continue to improve the present method, in such areas as the rough estimation of the stick figure position for the first frame and the extension of the method to 3D models.

References

[1] U.Montanari, "A Method for Obtaining Skeletons Using a Quasi-Euclidean Distance," Journal of the Association for Computing Machinery, Vol.15, No.4, pp. 600-624, October 1968.

[2] Joseph O'Rourke and Norman I. Badler,"Model-Based Image Analysis of Human Motion Using Constraint Propagation," IEEE Trans. on PAMI, Vol.2, No.6, pp.522-536, November 1980.

[3] Richard F.Rashid, "Towards a System for Interpretation of Moving Light Displays," IEEE Trans. on PAMI, Vol.2, No.6, pp.574-916, November 1980.

[4] Toshifumi Tsukiyama and Yoshiaki Shirai, "Detection of the Movements of Persons From A Sparse Sequence of TV Images," Pattern Recognition Vol.18, Nos. 3/4, pp.207-213, 1985.

[5] Hsi-Jian Lee and Zen Chen, "Determination of 3D Human Body Postures from a Single View," Computer Vision, Graphics, and Image Processing, Vol.30, pp.148-168, 1985.

[6] S.Tsuji, A. Morizono and S. Kuroda, "Understanding a Simple Cartoon Film By a Computer Vision System," Proc. 5th Int. Joint Conf. Artificial Intell., pp.609-610, 1977.

[7] S.Tsuji, M.Osada and M.Yachida, "Tracking and Segmentation of Moving Objects in Dynamic Line Images," IEEE Trans. on PAMI, Vol.2, No.6, pp.516-522, Nov.1980.

[8] K.Akita, "Image Sequence Analysis of Real Word Human Motion," Pattern Recognition, Vol.17, No.1, pp.73-83, 1984.

[9] Maylor K.Leung and Yee-Hong Yang, "Human Body Motion Segmentation in a Complex Scene," Pattern Recognition, Vol.20, No.1, pp.55-64, 1987.

[10] Maylor K.Leung and Yee-Hong Yang, "A Region Based Approach for Human Body Motion Analysis," Pattern Recognition, Vol.20, No.3, pp.321-339, 1987.

[11] Masanobu Yamamoto and Kazutada Koshikawa, "Human Motion Analysis Based on A Robot Arm Model," Proceedings of CVPR'91, pp. 664-665, June 3-6, 1991.

[18] N.Sasaki and I.Namikawa, "Measuring and Display System for a Marathon Program by Realtime Image Processing," IEICE Trans. Vol. E 74, No.10, October 1991.

[13] Athanasios Papoulis, "Probability, Random Variables, and Stochastic Processes," Second Edition, MacGraw-Hill.

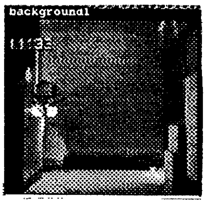


Fig. 6. Background image

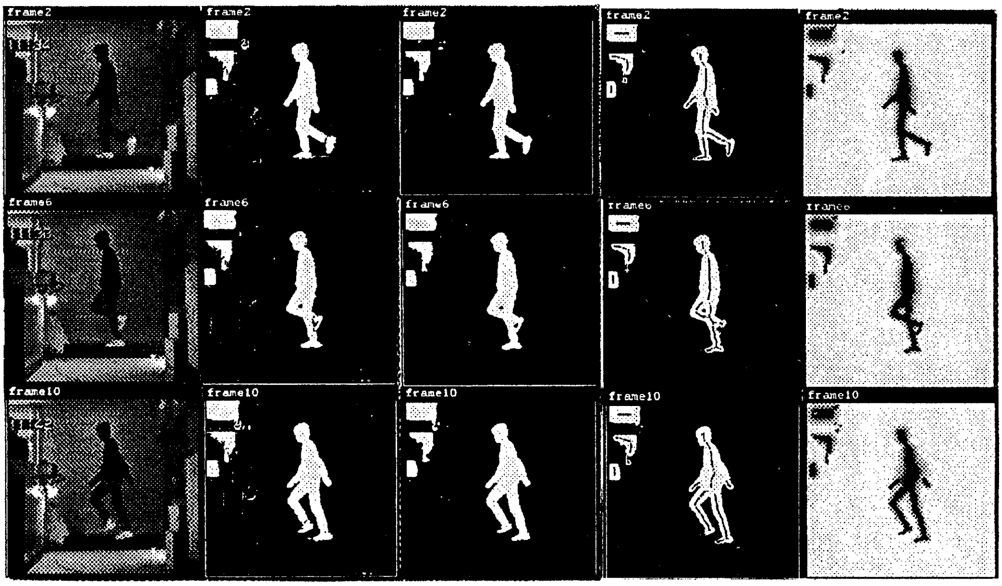


Fig. 7. Original images and intermediate results (see text)

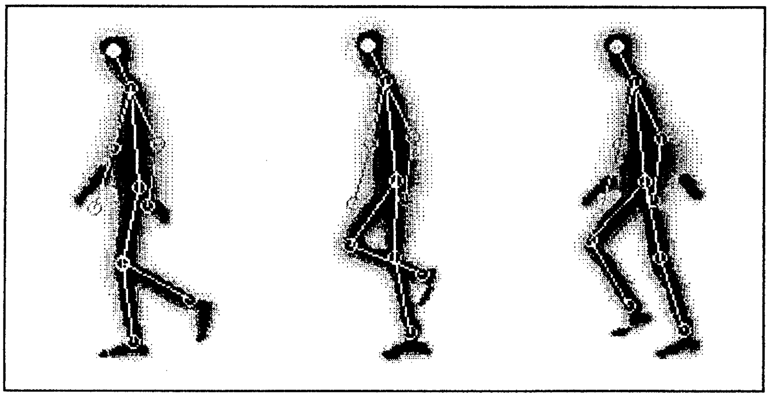


Fig. 8. Stick figures fitted to the potential fields