Resolving Ambiguity in Depth Extraction for Motion Capture using Genetic Algorithm

Yin Yee Wai, Chi Kin Chow, Tong Lee
Computer Vision and Image Processing Laboratory
Dept. of Electronic Engineering
The Chinese University of Hong Kong, Shatin, Hong Kong
{yywai, ckchow1, tlee}@ee.cuhk.edu.hk

Abstract - There are considerable interests in motion capture from an image sequence taken from a video camera. However, since the images only consist of 2D information, the distance of an object from the image plane cannot be determined uniquely unless some constraints are imposed. Coplanar motion is a popular constraint adopted for this purpose. If the captured motion is not coplanar, only two possible solutions can be obtained at the best even the actual size of the object is given. Therefore, the motion cannot be captured without resolving the distances of each control point from the image plane. By assuming smooth transition, the problem is formulated as one minimizing the transition between image frames and application of GA to solve this optimization problem is proposed.

I. INTRODUCTION

Motion capture makes animated motion realistic. It is an efficient way to record any human arbitrary motion [2]. Applications include creating 3D humanoid animation, using a pre-generated humanoid 3D model with captured motion, in games, movies, virtual human, simulations and even for human motion analysis. Traditional ways of motion capture include magnetic sensors or markers on real human model to record the motion by multiple cameras from different directions. Herda, Fua, Plankers, Boulic and Thalmann [7] suggested optical motion capture while Moezzi, Tai and Gerard [6] suggested capturing motion by 17 cameras surrounding. However, since the optical and magnetic components are very expensive, such systems are not suitable for low cost applications.

Recently, considerable attention has been given to motion capture from monocular sequence of 2D images. This could largely reduce the production cost. While the 2D coordinates (x and y) are readily available from the 2D images with this approach, a major difficulty is to extract depth information, which is the distance of the object from the image plane. There are some tricks to extract the depths in order to reconstruct the 3D motion. Holt, Netravali, Huang and Qian [3] advocated an approach of decomposition with assumptions of coplanar to determine articulated motion from perspective views. Zhuang, Liu and Pan [5] proposed an approach for video motion capture using feature tracking, such as joint and color model of body part, by Kalmen filter and morph-block similarity algorithm. Taylor [4] presented an algorithm of recovering information about the configuration of an articulated object from point correspondences in a single image. As many as 2048 possible solutions for a model with 11 joints were obtained for human motion capture and the best solution was selected by the user. The unique solution can only be obtained when the coplanar constraints are given. Pan and Ma [2] too adopt coplanar constraint and articulated model assumptions to estimate 3D motion of a human walking. However, if the actual size of an object is known, the number of solutions for a point can be reduced to two. This is illustrated in fig. 1. Given two control points $J_0$ and $J_1$ on an object and the link between them, the depth information can be obtained in a single 2D image from its projection on the image plane because $\Delta x^2 + \Delta y^2 + \Delta z^2 = d^2$ where $\Delta x$ and $\Delta y$ are the projection coordinates of the link while $d$ is the physical distance of the link. However, there are two solutions and if an object has $N$ links, the number of possible solutions is $2^N$.

This paper describes an application of Genetic Algorithm (GA) to determine the motion from a monocular image sequence without coplanar constraint, by resolving the ambiguity in depth extraction for each link. Using a camera with high sampling rate, the captured motion is expected to be smooth and moving in one direction rather than having sudden change of moving direction. Therefore, the absolute values of rotating angles are expected to be small and hence the change of rotating angles from one frame to the next should also be small. Thus we propose to determine the depth information of each control point in each image frame by
finding the best combination that gives a motion with the smoothest transition. Since the depth of each control point at each image frame can have two possible solutions, determining the best combination becomes a combinatorial optimization problem that is most suitably solved using GA.

The remaining of this paper is organized as follows. First, the problem of motion capture is briefly described in the next section. Then the depth extraction problem for motion captured is formulated as an optimization problem and its solution is proposed in section III. Experiments applying the proposed GA for motion captured on two image sequences are reported in section IV. Finally, discussion and conclusion are given in section V and VI respectively.

II. MOTION ANALYSIS

Firstly, we define an articulated humanoid stick model as shown in figure 2. Hence motion capture becomes matching the articulated model with the human posture identified in the image. An assumption is also made that the correspondence between the joints of the model and the feature points in each image. An assumption is also made that the correspondence be measured approximately.

A. Articulated Model

![Articulated Model Diagram]

Fig. 2 The articulated stick model for human motion analysis defined with 13 joints and 5 endpoints.

B. Reference Length

When applying the stick model for motion capturing, the reference length of each link needed to be determined in advance because different persons would have different sizes. They can be measured when all the links are parallel to the image plane. For example, figure 6 shows a stand still human model and the reference lengths are measured in this frame. Generally, it is hard to have the feet absolutely parallel to image plane, so the reference lengths of the feet can only be measured approximately.

C. Perspective Scale

Although the reference lengths of each link can be obtained using a reference frame, the links would appear to have different lengths at different frames because the object may move away or closer to the camera. However, the lengths of each link should only differ by a scale factor. This scale factor can be found with the aid of the triangular-linked joints J0-J1-J2. As the reference lengths of the 3 links are known, the scale factor, s, in a particular frame can be calculated by trigonometry. The steps are shown as follow:

(a) Measure the reference lengths of links J0-J1 (R01), J1-J2 (R12), and J0-J2 (R02). See figure 3 (a).

(b) Similarly, from one particular frame, measure the projected lengths of the 3 links as r01, r12 and r02 in the 2D image. See figure 3 (b).

(c) r is in fact the projection of scaled R. See figure 3 (c).

(a) and (b) show two possible depths of J1, one is behind and one is in front of J0.

Let

\[ J_0 = (x_0, y_0, z_0), \]

\[ J_1 = (x_1, y_1, z_1), \]

\[ J_2 = (x_2, y_2, z_2). \]

In (a) and (b)

\[ z_1 = z_0 \pm \sqrt{R_{01}^2 - \left(\frac{r_{01}}{s}\right)^2} \] \hfill (1)

where s is the scale due to perspective projection.

z2 is obtained in the similar way and to be

\[ z_2 = z_0 \pm \sqrt{R_{12}^2 - \left(\frac{r_{12}}{s}\right)^2} \] \hfill (2)

In order to eliminate z0, R12 is considered. Similar to (a) and (b), but with J0 replaced by J2, the following relation can be obtained.
Then, $s$ can be derived by solving equations (1) to (3) since $R$ are the reference lengths and $r$ can be obtained from each frame.

$$R_{12} = (z_1 - z_2)^2 + \left(\frac{r_{12}}{s}\right)^2$$  \hspace{1cm} (3)

### D. Depth Extraction

When a 3D object is projected onto a 2D image, as shown in fig. 4, the displacement of the object on the image plane can be measured from the 2D coordinates ($x$, $y$) on the image plane but the depth cannot be determined uniquely. There are two possibilities corresponding to swing forward position $(x_{\text{front}}, y_{\text{front}})$ or backward position $(x_{\text{back}}, y_{\text{back}})$. Figure 4 shows possible depths of joint 10 where the reference length of link $J_8$-$J_{10}$ is known. The possible depths can be found by trigonometry,

$$z = \sqrt{(sR)^2 - r^2}$$ \hspace{1cm} (4)

$$\theta = \cos^{-1} \frac{r}{sR}$$ \hspace{1cm} (5)

where $r$ represents the projected length of body part (measured from the 2D image) and $sR$ represents the scaled reference length. “$z$” is the depth relative to the head and “$\pm \theta$” is the relative rotating angle.

Therefore the depth of link $i$ can be defined by $\theta_i$ which could take either positive or negative values. So for $N$ links in the model, there will be $N \theta_i$ and each could take either positive or negative values. Therefore, given each value of $\theta_i$ for each link in each frame, the objective of our work is to determine the most suitable set of signs of $\theta_i$ describing the motion to be captured. We propose to solve this problem by a GA in the following section.

### III. OPTIMIZATION BY GENETIC ALGORITHM

We intend to derive the required solution by finding the set of $\theta$ for each link in each frame such that a smooth motion can be obtained. Therefore, the change of angles should not be too much along the image sequence. Following is an example illustrating the assumption. Considering a sample human motion shown in figure 5. Assume the model standing with his hands pointing down parallel to the image plane in the first frame. In the second frame, possibly, he may swing his hands upward in front of or behind the body, and the same for the third frame. However, the motion is supposed to be smooth, such that, the hands should swing either in front of or behind the body all the way. It is not expected to have the hands to swing in front of the body now but behind after a few mini-seconds then back to front immediately. If so, the resulting absolute value of change of rotating angle will be very large and in different directions compared to the adjacent frames.

Therefore, we have the following total change of rotating angle criterion:

$$F = \sum_{j=1}^{M-1} \sum_{i=1}^{N-1} |\alpha^j_{i+1} \theta^j_{i+1} - \alpha^j_i \theta^j_i|$$ \hspace{1cm} (6)
M and N are the total numbers of joints and frames. \( \theta_j^i \) stands for the relative rotating angle of joint \( i \) in frame \( j \). \( \alpha_j^i \) is the scale factor of the rotating direction which is either 1 for \( +ve \) rotation and -1 for \( -ve \) rotation. So we seek the values of \( \alpha_j^i \) to minimize Eq. (6). With Eq. (6) as the fitness function, the directions of rotation (\( \alpha_j^i \)) can be derived using GA such that \( F \) is minimized. Therefore, after all \( \theta_j^i \) have been obtained from the image sequence, we define the chromosome with \( \alpha_j^i \) for all \( i \) and \( j \). Table I shows an example of the chromosome with “1” or “-1” representing one gene.

### TABLE I

<table>
<thead>
<tr>
<th>Joint 1</th>
<th>Joint 2</th>
<th>⋅</th>
<th>Joint 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame 1</td>
<td>1</td>
<td>⋅</td>
<td>-1</td>
</tr>
<tr>
<td>Frame 2</td>
<td>⋅</td>
<td>1</td>
<td>⋅</td>
</tr>
<tr>
<td>Frame 3</td>
<td>-1</td>
<td>⋅</td>
<td>-1</td>
</tr>
</tbody>
</table>

A. Reproduction

With the chromosomes defined above, new offspring is generated using the following cross-over and mutation operations.

1) Cross-Over

A multi-point cross-over operator is adopted here. Given two chromosomes, the number of genes go through the cross-over operation is randomly generated. The location of the cross-over point is also randomly generated such that all genes have the same probability of going through cross-over. With the human model we defined in figure 2, which has 18 joints, the expected number of genes that will go through the cross-over operation is 9.5.

2) Mutation

After cross-over, the chromosome will go through the mutation process. Again, the number of genes that will go through the mutation is randomly generated with equal probability. Therefore the expected number of genes to be mutated is also 9.5. Then the location of genes that will go through mutation is randomly generated next. As shown in Table I, each gene takes on the value of either “1” or “-1”. The gene to be mutated will have its value change from “-1” to “1” or vice versa.

B. Selection

After reproduction, the new population is doubled (from \( n \) to \( 2n \)) in size including the original population. A new generation is formed by selecting the best \( n \) chromosomes from the new population. To do the selection, we examine the fitness of each chromosome defined in Eq. (6). Since our objective is to minimize the change of angles, chromosome with smaller fitness is the preferred solution. Therefore, chromosomes with large \( F \) are omitted so that the population is reduced to the original \( n \).

C. Termination

The process (reproduction → fitness function → selection → reproduction → ……) will continue until the solution converges. This is done by checking the resulting \( F \) in Eq. (6); if it reaches stable state, it converges.

### IV. EXPERIMENTAL RESULTS

The algorithm has been applied to two real image sequences. One is about a human pretending to pick up something on the ground and another one is about a singer walking freely on the stage. Both of them were recorded by a single monocular camera while the first one was taken from the front and the second from the side. The images were captured with 0.3 seconds difference in-between every two frames. 6 frames of the image sequences are shown in figures 8 and 10 while the estimated motions in figures 9 and 11 respectively. Since we assumed that the joint locations are available, they are marked manually.
Comparing figures 8 and 9, the estimated poses were very much like the real image sequence.

B. Singer Walking Freely On Stage Captured At Side View

For this sequence, no stand-still reference frame is available. The reference lengths were estimated from the last frame because the upper part of the human model was approximately parallel to the image plane, just similar to the stand still pose. The results were shown in figure 7. The frame also showed the side view of the feet, from which, the reference length of the feet could be estimated. This approximation step may introduce error to the joint location and degrade the result but good results were also obtained as depicted in fig. 11. Comparing figures 10 and 11, the estimated poses were also very much similar to the real image sequence.

V. DISCUSSION

The current implementation assumes the correspondence between the joints of the real human body and the feature points on the 2D image are provided. In practice, this is not applicable and automatic feature point mapping should be applied. As the performance of the algorithm highly depends on how accurate the feature points are being marked in the images, an accurate automatic feature extraction system should be developed to mark the joints. Alternatively, colored markers or Morph-block algorithm in [5] may be applied to track the joints. More constraints may be considered to obtain more accurate solution, like the kinetic constraints applied in [8].

VI. CONCLUSION

This paper describes an application of GA for motion capture. Since the chromosome consists of binary genes, from the Schema Theorem [9], such problems can be solved using GA. Our implementation showed that the proposed procedure was effective. When implemented on a 450MHz Pentium III PC, the average operation time for the two motions is 2 minutes. However, each image sequence only consists of 6 frames. In practice, an image sequence would consist of many frames. Subsequently, the number of parameters to be tuned by the GA is huge. Extension of the proposed approach for handling long image sequence is under consideration.

ACKNOWLEDGEMENT

The project is partially supported by the CUHK Direct Grant 01/02.

REFERENCE


![Fig. 8 The first real image sequence.](image_url)
Fig. 9  Estimated motion of the first image sequence

Fig. 10  Image sequence of singer walking onto stage.

Fig. 11  Estimated motion of the second image sequence