Multiple Cues used in Model-Based Human Motion Capture

Thomas B. Moeslund and Erik Granum
Laboratory of Computer Vision and Media Technology
Institute of Electronic Systems, Aalborg University
Niels Jernes Vej 14, DK-9220 Aalborg East, Denmark
{tbm,eg}@vision.auc.dk

Abstract

Human motion capture has lately been the object of much attention due to commercial interests. A ”touch free” computer vision solution to the problem is desirable to avoid the intrusiveness of standard capture devices. The object to be monitored is known a priori which suggest to include a human model in the capture process. In this paper we use a model-based approach known as the analysis-by-synthesis approach. This approach is powerful but has a problem with its potential huge search space. Using multiple cues we reduce the search space by introducing constraints through the 3D locations of salient points and a silhouette of the subject. Both data types are relatively easy to derive and only require limited computational effort so the approach remains suitable for real-time applications. The approach is tested on 3D movements of a human arm and the results show that we successfully can estimate the pose of the arm using the reduced search space.

1. Introduction

Human motion capture has lately been the object of much attention due to commercial interests, especially from the entertainment industry. Different devices are developed to capture human motion [10] but because of their intrusiveness great efforts have been invested into finding a ”touch free” computer vision solution to the problem.

Human motion capture can be seen as a special case of motion analysis where the object to be monitored is known a priori. There are various ways of exploiting a priori knowledge. One is to structure measured data into an efficient model description according to the expected object [4][7][9][13][16]. Another is to start with a model described as detailed as required and let the measured data be used only to adjust those parameters which may change during observation, e.g. position and pose [3][5][6][8][15].

A more comprehensive discussion of different ways of applying a priori knowledge and examples of systems using them may be found in [10] and [11].

2. The approach

We propose to use the latter approach and let the current state of the model represent the output of the system. The model is updated from each image by investigating which synthesised model configuration provides the best match to the input data. This is known as the ”analysis-by-synthesis” (AbS) approach. The configuration or pose of the human model is described by the current values of the variables representing the different degrees of freedom in the model. Hence the configuration is represented in a coordinate system spanned by the different degrees of freedom. This is known as the configuration space or search space.

The AbS approach has some advantages: 1) It is easy to incorporate a priori knowledge about the human, 2) A plausible output is guaranteed since it comes from the configuration space rather than from the noisy image data, 3) A body pose can be represented as one point in the configuration space yielding a very efficient representation, 4) The format of the output is fixed and identical to the main goal of most motion capture systems, i.e. to estimate the location and orientation of the joints. The result can directly be used in a classification scheme or in an interface application, e.g. controlling an avatar, and 5) The configuration space is scalable in the sense that dimensions can be added or deleted to obtain just the required accuracy and detail.

The major problem of using the AbS approach is that the search space for the right configuration to synthesise may grow beyond what is feasible to handle in real-time systems.

2.1. Reducing the search space

Luckily the AbS approach contains straight forward ways to reduce the search space: 1) Its dimensionality is set according to the current application. If a system only is interested in capturing the head pose of a subject then the space can be reduced to only the degrees of freedom associated with the head, 2) Kinematic constraints of the human motor system are also used to constrain the search space, e.g. the leg cannot bend forwards at the knee. This can also be seen in a temporal context where the movement of different body parts have velocity and acceleration constraints, and 3) Collision constraint. Two body parts can not occupy the same space at the same time.

Even after using the above constraints the search space may still be rather large and therefore we introduce additional constraints to reduce the search space further. Our approach is to focus on simple cues which will require only limited computational effort so the approach remains suitable for real-time applications:

- Real-time measurements of the 3D position of salient points of the human body allow us to exploit the structural constraints of the model for the current configuration. We are initially looking for the head and hands, and suggest that even rather coarse 3D positions provide useful information.

- Spatial image features may help to reduce the search space further. We are initially using the silhouette of the subject which is fairly simple to derive.

2.2. Scenario and model complexity

We address the application area of man-machine-interactions. This allows an assumption of the subject to be monitored where head and torso are frontoparallel with respect to the camera, i.e. the position of the head in 3D will allow estimation of the 3D position of other salient points such as the shoulders. To focus our attention on the approach we chose to work with *motion capture of the left arm* with the hand being a ”stiff” extension of the lower arm along its axis. This is discussed further in section 6.

3. The implemented system

The approach and constraints outlined above are implemented in the system architecture shown in figure 1.

First the hand and head are found in two stereo images and combined into two 3D points. The possible constraints are applied to reduce the search space before the data is synthesised. The silhouette is found and compared to each synthesised pose and the best match is used to update the model.

Below follows a description of each of the functional blocks of the figure. Results of the operational system are described through typical examples and perspectives of the approach are discussed and an occlusion given.

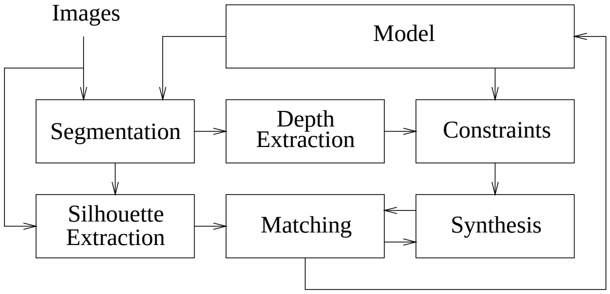

Figure 1. The functional block diagram of the proposed system.

3.1. Segmenting the hand and the head

The hand and head can be segmented using colour information. Original RGB-colours are sensitive to the intensity of the lighting. Therefore we use chromatic colours which are normalised according to the intensity. The chromatic colours are defined as $r = R/c$, $g = G/c$, and $b = B/c$, where $c = R + G + B$.

Since the three chromatic colours always sum to 1, two components are sufficient to represent a colour image. Empirically an upper and lower threshold for skin colour of the two chromatic colours have been found. Applying these thresholds to the image in figure 2.A results in the image in figure 2.B, where skin-pixels are white and non-skin-pixels black. To increase efficiency a 2nd order predictor is used to limit the areas wherein the hand and head are searched [1].

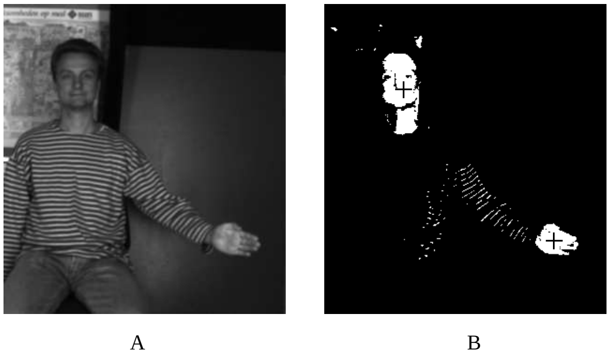

Figure 2. A: An input image. B: A segmented image. The ’+’s indicate the centre of mass for the head and hand, respectively.

The skin-pixels within each search area are labelled into coherent blobs. To speed up the processing the image is down sampled with a factor of eight. The blobs with the correct size and shape are identified as the hand and the head, and their centres of mass are found, see figure 2.B.

3.2. Depth extraction

Two segmentation processes are run in parallel on two stereo images. The output is two sets of 2D points which are used to calculate the 3D position of the hand and head. To do so, the acquisition processes must be synchronised and the transformation between the two cameras (images planes) needs to be known, i.e. camera calibration.

The calibration is carried out using Tasi's calibration technique. The synchronisation is obtained using a master-slave structure which negotiate time stamp difference when the system is started up, and a hardware synchronisation signal (between the cameras) [1].

Triangulation of the centres of mass for the hand and head in the two images is used to obtain the 3D positions of the hand and head.

3.3. Configuration space and constraints

A human arm and two local coordinate systems are shown in figure 3. The elbow-coordinate system is oriented according to the upper arm. It is defined as a translated version of the shoulder coordinate system when the upper and lower arm are pointing upwards. One of the degrees of freedom in the shoulder[1] has been moved to the elbow for convenience. This results in two degrees of freedom in the shoulder and two in the elbow, i.e. the configuration space is four-dimensional. The θ-angles are defined as the angles between the Z-axis and upper and lower arm, respectively.

The ϕ-angles are the angles between the X-axis and the projection of the lower and upper arm onto their respective XY-planes.

Elbow candidates

Given the assumption that the subject's head and torso are frontoparallel with respect to the camera, the position of the shoulder can be found based on the position of the head and the length between the neck and the shoulder. Using the shoulder, the hand, and the length of the upper and lower arms, reduces the possible elbow positions to a circle (in 3D) perpendicular to the line spanned by the shoulder and hand. This circle is illustrated in figure 4.A. Given one point on the circle and rotating this point around the shoulder-hand-line all points on the circle are specified.

We chose the point with the highest possible z-value which lie in the plane containing the shoulder and hand points, and the Z-axis. This point is shown in figure 4 as E and can be found in the following way:

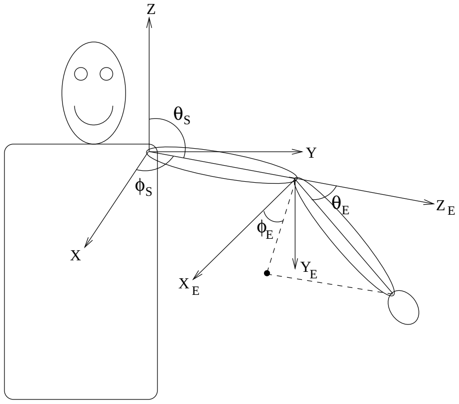

Figure 3. The coordinate systems for the shoulder and the arm.

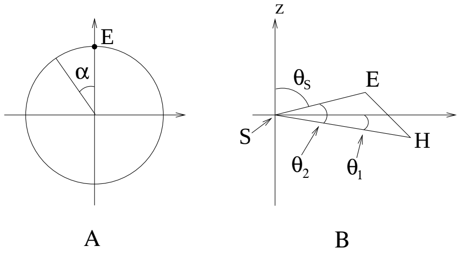

Figure 4. A: The possible elbow positions. B: A triangle with vertices in the shoulder S, elbow E, and hand H, when $\alpha = 0$.

$$\begin{aligned}
\theta_s &= 90^\circ - (\theta_2 - \theta_1) && (1)\\
\theta_1 &= arcsin\left(\frac{|H_z|}{|SH|}\right) && (2)\\
\theta_2 &= arcsin\left(\frac{2\omega}{|SE||SH|}\right) && (3)\\
\omega &= \sqrt{\tau(\tau - |SE|)(\tau - |SH|)(\tau - |EH|)} && (4)\\
\tau &= \frac{1}{2}(|SE| + |SH| + |EH|) && (5)
\end{aligned}$$

Since the shoulder and hand are in the same plane which includes the Z-axis, their ϕ-values in the shoulder coordinate system will be equal:

$$\phi_s = \phi_{hand} = arctan\left(\frac{H_y}{H_x}\right) \qquad (6)$$

The 3D coordinates of the elbow can now be calculated:

[1] The shoulder is modelled as a socket joint with three degrees of freedom but in real life it (known as the shoulder complex [2]) is much more complicated. However, for the accuracy needed in this work it is sufficient to use a socket joint.

$$E_x = |SE| sin(\theta_s) cos(\phi_s) \quad (7)$$
$$E_y = |SE| sin(\theta_s) sin(\phi_s) \quad (8)$$
$$E_z = |SE| cos(\theta_s) \quad (9)$$

This point, E, is located on the circle. All other points on the circle are found by rotating E α degrees around the line $|SH|$, using the rotation matrix R:

$$R = \begin{bmatrix} x^2 + (1-x^2)cos(\alpha) & xy(1-cos(\alpha)) + zsin(\alpha) & zx(1-cos(\alpha)) - ysin(\alpha) \\ xy(1-cos(\alpha)) - zsin(\alpha) & y^2 + (1-y^2)cos(\alpha) & yz(1-cos(\alpha)) + xsin(\alpha) \\ zx(1-cos(\alpha)) + ysin(\alpha) & yz(1-cos(\alpha)) - xsin(\alpha) & z^2 + (1-z^2)cos(\alpha) \end{bmatrix}$$

where (x, y, z) is the direction unit vector of $|SH|$.

Kinematic constraints

A human has certain limitations to his/her movements, e.g. the elbow can only bend so much and move so fast. These constraints are known as kinematic constraints. For each value of α (at a finite resolution) the corresponding 3D position is calculated, converted into the four angles in figure 3, and tested against the kinematic constraints. The two shoulder angles can be found as:

$$\theta_s = arccos\left(\frac{E_z}{|SE|}\right) \quad (10)$$
$$\phi_s = arctan\left(\frac{E_y}{E_x}\right) \quad (11)$$

The elbow angles are found in the same way, *after* the hand coordinates have been transformed to the elbow coordinates system. The transformation is carried out by first translating the hand coordinates and then carefully rotating them around two of the axes depending on which quadrant the elbow is located in.

For an α-value to be accepted it's shoulder and elbow angles have to be within the ranges shown in table 1. The ranges vary between people so the data of one of the authors have been used for table 1.

Table 1. The ranges of the different angle parameters for the shoulder and elbow.

	θ_s	ϕ_s	θ_E	ϕ_E
Pos. min.	0	-45	0	-90
Pos. max.	180	135	145	45

Besides angle values also the angle velocity and acceleration yield constraints. Their ranges depend on the activity carried out. We have considered the maximum velocity of "normal" movement to be $400°/s$ and assumed that a subject can accelerate the upper and lower arm to their maximum velocity within one tenth of a second, i.e. the maximum acceleration $= 4000°/s^2$.

Collision constraint

The collision constraint states that two body parts can not occupy the same space at the same time. To test this a human model is needed. Preferably a deformable model which in appearance is very similar to the actual subject. But for simplicity we use simple stiff segments to model the human. The torso and head are both modelled using elliptic cylinders. The arm is modelled by two lines representing the upper and lower arm. The reason for not using volumetric shapes to model the arms is first of all to make the algorithm tractable to real time implementation, but also to account for the lack of flexibility in the head and torso models. Using lines for the arms and comparing them with a stiff torso/head model, yields approximately the same result as comparing a flexible volumetric arm with a deformable torso/head model.

Each line is represented in parametric form:

$$\mathbf{P} = \mathbf{P}_0 + t \cdot \Delta \;\Rightarrow\; \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix} + t \cdot \begin{bmatrix} l \\ m \\ n \end{bmatrix} \quad (12)$$

where $\mathbf{P}$ is a point on the line, $\mathbf{P}_0$ is the starting point (shoulder for upper arm and elbow for lower arm) and Δ is the slope of the line. For the upper arm $0 \geq t \geq |SE|$ and for the lower arm $0 \geq t \geq |EH|$.

For each point on the two lines (at a given resolution) it is calculated whether the point is within the torso or head, i.e. a collision. If so this α-value can be eliminated from the solution space (circle).

3.4. Extracting the silhouette

Before the system is started an image of the background without a subject is stored. During processing each image is subtracted from the background image and converted, through a threshold, into a binary image of the silhouette and background. Due to, especially, the textured clothes worn by different subjects the silhouette images will contain "noise" which effects the matching process. Therefore the silhouette image is filtered using the morphologic closing-operator.

The part of the image from where the silhouette should be extracted can be reduced using the position of the head and hand, and the assumption of the head and torso being frontoparallel with respect to the camera. In figure 6.A the region of interest of the extracted and filtered silhouette of the arm in image 2.A is shown.

4. Synthesis and matching

The circle in figure 4.A has through the constraints been reduced to an arc with α-values producing legal poses of the arm. Each of the legal poses needs to be synthesised and later compared with the image silhouette. To synthesis a pose means to project the modelled human arm into one of the stereo images. This is done via the transformation matrices found during calibration.

After the modelled arm has been projected into the image it is also represented as a silhouette for comparison with the image silhouette. This matching process is carried out through an AND operation and the result is a similarity measure. Formally we define it as:

$$S(\alpha) \quad = \quad \frac{1}{L} \sum_{t \in \text{arm}} \sum_{w=0}^{\text{Width}} Sil(Syn(\mathbf{P}(\alpha, t, w))) \quad (13)$$

where α represent the pose, $L = \text{Width}(|SE| + |EH|)$, $Sil()$ is the value (zero or one) of the extracted silhouette at the position given by the synthesised point, $Syn()$, which originates from the 3D arm model for a given pose α at position t given by equation 12. The width of the arm is given by *Width* (in pixels) and the closer this value is to the subject's actual arm width the more precise the model is. On the other hand the larger the width the more processing is required. To investigate which width to use we calculate the similarity measure for different widths for a large number of images. Generally the similarity measure, as a function of α, for one pose is a bell-shaped curve. Depending on the constraints for a given pose the outer areas of the curve might be eliminated. In figure 5 the similarity measures for three different widths are shown for the none-constrained α-values calculated based on the image in figure 2.A.

What can be seen in figure 5 is, not surprisingly, that the shape of the curves becomes sharper as the width increases. More important it is to realize that the actual location of the ”peak” approximately stays the same. This means that the width of the arm is not critical when finding the pose. It is sufficient to use the simplest arm model where *Width*=1, i.e. w and the second sum in equation 13 can be removed.

During operation we find the ”peak” using a starting point (the middle of the remaining α-values) and the slope of the curve.

5. Results

In figures 6.B and C the result of two images from a sequence processed by the system is shown. The images have been cropped form the original images to reduce their size. The result of the algorithm is superimposed onto the images and it can be seen that the correct position of the elbow and thereby the skeleton have been found. Note that even though there is a significant occlusion between the hand and torso the system have found the correct pose.

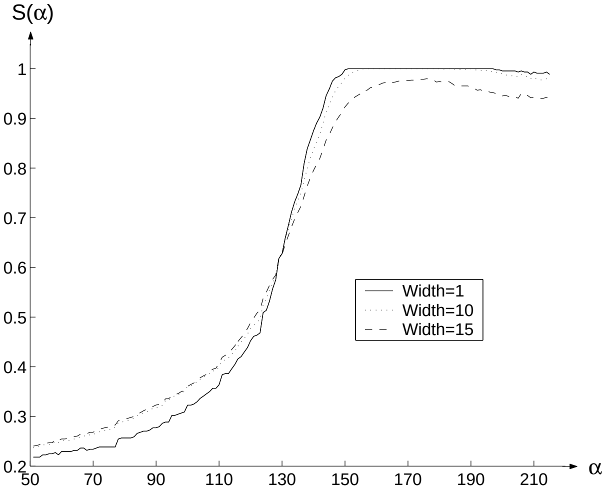

Figure 5. The similarity measure for three different widths of the arm. The curves are shown only for the α-values not eliminated by the constraints.

When calculating the 3D location of the hand and head (shoulder) an amount of uncertainty have been observed. This transforms the α-circle into a torus where each point inside the torus is an elbow candidate. An investigation into this showed that not much can be gained using this, more correct, pose candidate selection. Also, the uncertainty due to the variation of the clothes is larger than the extra precision gained by using a torus instead of a circle. Therefore it was decided to use the less complex model of the elbow candidates.

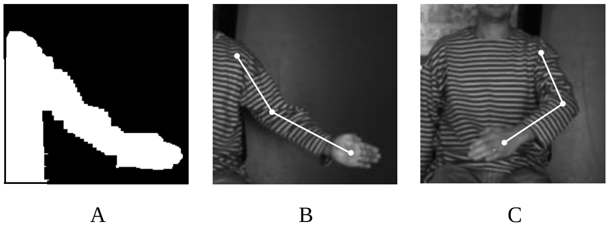

Figure 6. A: The extracted and filtered silhouette of the arm. B and C: Cropped input images with the estimated arm pose superimposed.

6. Discussion

We use the assumption that the subject is facing the camera with his/her head fixed with respect to the shoulder. This might be a reasonable assumption in many man-machine-interfaces, but it is not a general valid assumption. To expand our system to handle this we need to increase the dimensionality of the configuration space by adding new degrees of freedom for the head and shoulder. Also new constraints should be incorporated to reduce the search space. This would of course make the problem more complex but our approach can still be used. The same can be said about the assumption of the hand being part of the arm.

The clothes of the subject obviously affects the estimated pose. To actually model this effect will require heavy computational methods and knowledge about the clothes worn by the subject. It is clear that the tighter a subject's clothes are the better the pose will be estimated. However, even with rather ”loose” clothes, as in figure 2, a reasonable result can normally be obtained, and that is without the normal constraints imposed on a subject's clothes (tight-fitting [17], special coloured [5], special textured [14], or markers [12]).

7. Conclusion

We have in this paper showed how the search space in the analysis by synthesis approach can be reduced beyond the usually methods. This is carried out using multiple cues: 3D points and silhouette data. Both data types are fairly simple but we have showed that they together reduce the search space and make it possible to estimate the 3D pose of a human arm. This result is achieved without the assumptions used in similar systems: special coloured/textured/tight-fitting clothes, and markers.

In future work we plan to extend our approach to the entire human body. Clearly this requires a larger configuration space and a number of new constraints. We want to investigate other ”simple” cues and how to obtain and use depth data for the entire human. Possibilities are to locate the 3D position of the feet and other salient points, e.g crutch and arm pits.

Acknowledgement

We would like to thank the authors of [1] who contributed to the segmentation and depth calculation, and the Danish National Research Councils, who through the project: ”The Staging of Virtual Inhabited 3D Spaces” founded this work.

References

[1] B. Andersen, T. Dahl, M. Iversen, M. Pedersen, and T. Søndergaard. Human Motion Capture. Technical report, Laboratory of Image Analysis, Aalborg University, Denmark, January 1999.

[2] N. Badler, C. Phillips, and B. Webber. *Simulating Humans - Computer Graphics Animation and Control*. Oxford University Press, 1993.

[3] C. Bregler and J. Malik. Tracking People with Twists and Exponential Maps. In *International Conference on Computer Vision and Pattern Recognition*, 1998.

[4] H. Fujiyoshi and A. Lipton. Real-Time Human Motion Analysis by Image Skeletonization. In *Workshop on Applications of Computer Vision*, 1998.

[5] D. Gavrila and L. Davis. 3-D Model-Based Tracking of Humans in Action: A Multi-View Approach. In *Conference on Computer Vision and Pattern Recognition*, San Francisco, USA, 1996.

[6] L. Goncalves, E. Bernardo, E. Ursella, and P. Perona. Monocular Tracking of the Human Arm in 3D. In *International Conference on Computer Vision*, Cambridge, Massachusetts, 1995.

[7] I. Haritaoglu, D. Harwood, and L. Davis. W^4: Who? When? Where? What? - A Real Time System for Detecting and Tracking People. In *International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, 1998.

[8] D. Hogg. Model-Based Vision: A Program to See a Walking Person. *Image and Vision Computing*, 1(1), February 1983.

[9] S. Iwasawa, K. Ebihara, J. Ohya, and S. Morishima. Real-Time Estimation of Human Body Posture from Monocular Thermal Images. In *Conference on Computer Vision and Pattern Recognition*, 1997.

[10] T. Moeslund. Computer Vision-Based Human Motion Capture - A Survey. Technical report, Laboratory of Image Analysis, Aalborg University, Denmark, 1999.

[11] T. Moeslund. Summaries of 107 Computer Vision-Based Human Motion Capture Papers. Technical report, Laboratory of Image Analysis, Aalborg University, Denmark, 1999.

[12] O. Munkelt, C. Ridder, D. Hansel, and W. Hafner. A Model Driven 3D Image Interpretation System Applied to Person Detection in Video Images. In *International Conference on Pattern Recognition*, 1998.

[13] S. Niyogi and E. Adelson. Analyzing and Recognizing Walking Figures in XYT. In *Conference on Computer Vision and Pattern Recognition*, 1994.

[14] R. Plänkers, P. Fua, and N. D'Apuzzo. Automated Body Modeling from Video Sequences. In *International Workshop on Modeling People at ICCV'99*, Corfu, Greece, 1999.

[15] K. Rohr. *Human Movement Analysis Based on Explicit Motion Models*, chapter 8, pages 171–198. Kluwer Academic Publishers, Dordrecht Boston, 1997.

[16] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-Time Tracking of the Human Body. *Transactions on Pattern Analysis and Machine Intelligence*, 19(7), July 1997.

[17] C. Yaniz, J. Rocha, and F. Perales. 3D Region Graph for Reconstruction of Human Motion. In *Workshop on Perception of Human Motion at ECCV*, 1998.