

Towards Multi-camera Human Motion Capture and Low Bit-rate Video

D. Carsky, B. Naidoo, S. McDonald

School of Electrical, Electronic and Computer Engineering,
University of Natal (Durban), South Africa

carskyd@nu.ac.za bnaido@nu.ac.za mcdonalds@nu.ac.za

Abstract

In this paper we describe an approach taken towards applying human motion capture to low bit-rate video. Our objective is to generate parameters defining a person's motion, and with these parameters to create a model of a human performing the same movement. A video sequence of a person walking is obtained from multiple CCD cameras. Key points on the person's body are tracked in 3D without the use of markers or sensors. Spatio-temporal tracking is done by the method of least squares matching. A simple model is fitted to the tracked points in each frame, and the pose is described by the parameters defined in the MPEG-4 standard.

1. Introduction

Human motion capture is a topic that deals with the analysis of images involving humans. The interest in this field has grown significantly in the late nineties, a fact that is evident from the increased number of publications in this field as well as the visible impact the research has had, in particular, within the film industry. From movies such as Terminator 2, Titanic and The Lord of the Rings to name a few, it is evident that human motion capture played a crucial role in creating the special effects that elevated these movies to such a high status. The film industry is but one application area of human motion capture. Applications range from virtual reality and animation, to smart surveillance systems, advanced user interfaces, motion analysis and model-based coding.

The human motion capture domain can be broken down into three sub-categories: face analysis, gesture recognition and body analysis. Face analysis deals with face detection/recognition, face tracking and recognition of facial expressions. The application of face analysis is found in virtual reality and teleconferencing among others. Gesture analysis is in general concerned with the tracking of hands and fingers. A lot of research in this area is directed at developing systems that can understand and generate sign language. Body analysis looks at the large body movements. Some of the applications requiring tracking of the whole body include security systems and counting people. Precise human tracking finds the exact posture of the limbs of a human body, looking at the person as a complex articulated object, rather than a single body. The last area of body analysis deals with recognition of human motion and the analysis. This is particularly relevant in medical and sport applications.

This research focus' on body analysis. Our objective is to extract the pose of the subject in each frame and apply it to low bit-rate video. Instead of sending each frame across a channel, we transmit parameters that fully describe the person's pose. On the receiving end, these parameters are applied to a model that will represent the original pose. The standard MPEG-4 framework and parameters are used to describe the pose. MPEG-4 is object-based and has defined two sets of parameters for the animation of the body.

In this investigation emphasis is placed on the least squares matching (LSM) algorithm to track points in 3D since the tracking of key points is fundamental to the success of the project. This method was introduced by Ackermann [1], and later developed by Gruen [6]. LSM is used to find corresponding points in different views and subsequent frames.

2. System Definition

Several criteria can be used to describe the different approaches to the problem. Most systems are application specific and dependant on assumptions/conditions taken to simplify the problem. In order to establish a system one needs to first determine whether active or passive sensing is to be used. With active sensing, sensors are placed on the subject as well as the surroundings that transmit and receive generated signals respectively. This approach is obtrusive, and in some applications unfeasible. Passive sensing is based on "natural" signal sources, e.g. visible light, and hence no wearable devices are required. Although the amount of processing required increases as the sensor complexity decreases, with today's powerful off the shelf computers, it is possible to implement complex and computationally demanding algorithms.

For our application, we take the route of passive sensing. With passive sensing there are several other criteria that define the system [4]:

- Sensor modality (visible light, infra red, range)
- Sensor mobility (stationery vs moving)
- Sensor placement (centralised vs distributed)
- Sensor multiplicity (monocular vs stereo)
- Dimensionality of tracking (2D vs 3D)
- Model (stick figure, volumetric, statistical)

Three synchronised CCD cameras are used as sensors that are permanently mounted relatively close to each other. As described in [9], one static camera does not allow the recovery

of the 3D structure of the environment. With two cameras, one cannot use the geometric constraint (epipolar constraint) alone, but must introduce other constraints such as continuity, order and unicity to resolve matching ambiguities. The introduction of a third camera reinforces the geometric constraint and greatly simplifies the matching process. Such a stereo system is called trinocular.

Regarding tracking dimensionality, some 2D systems have successfully tracked constrained types of human movement. However, it is unlikely [5] that for more complex and unconstrained human movement, systems using these types of features exclusively would succeed. When dealing with human movement, one will always have to deal with self-occlusion. Self-occlusion together with arbitrary movement makes the 2D tracking problem extremely difficult. As a result, existing systems need to assume *a priori* knowledge of the type of movement and/or the viewpoint under which it is observed.

The key points that are tracked in this research are the locations of the major joints. To these points we will fit a simplified model of the human body, the skeleton model, also known as the stick-man model.

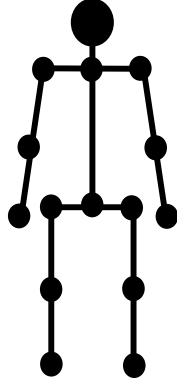


Figure 1 Stick-man model

The human body is a complex articulated object. Should one try to fully describe it mathematically, the result would be very large and difficult to solve system. For this particular reason, simplified models are used. The level to which the model is simplified is subjected to the application. While virtual reality requires detailed models, we are concerned with just the pose, for which the stick-man model is a sufficient representation. Even for detailed animations however, the skeleton is an important part of the model. The multi-layered approach adopted by [3] uses a skeleton, to which ellipsoidal meatballs are attached, simulating muscle and fat tissue. A polygonal surface fits over the meatballs as a representation of skin to produce realistic animations of human bodies. The advantage of the layered method is that once the character is constructed, only the motion of the skeleton needs to be determined for the final animation.

3. Least Squares Matching

The least squares matching technique, as described by [6], is an approach in which the optimum match is defined by a

transformation of one array into another, which minimises the remaining grey value differences.

Although this method is used to match three images, for explanatory purposes we consider only two views, a left and a right view of a stereo pair. Discrete functions are assigned to the two images $f(x,y)$ and $g(x,y)$ representing the left and right view respectively. One view is chosen as the template, say $f(x,y)$, and the other, $g(x,y)$ as a search image. Ideally, when the two image regions are identical, we get

$$f(x,y) = g(x,y) \quad (1)$$

The two images will not be the same due to radiometric (illumination differences) and geometric (different camera orientation) distortions, as well as any noise. Thus a true error vector is added as shown in equation 2.

$$f(x,y) - e(x,y) = g(x,y) \quad (2)$$

A six parameter affine transformation is used to relate the two conjugate patches, allowing for rotation, shifting, scaling and shearing. (a_{11}, b_{11}) are the shifts, (a_{12}, b_{12}) the scales, and (a_{21}, b_{21}) the shears.

$$\begin{aligned} x &= a_{11} + a_{12}x_0 + a_{21}y_0 \\ y &= b_{11} + b_{12}x_0 + b_{21}y_0 \end{aligned} \quad (3)$$

Equation (2) must be linearised to produce the observation equations

$$f(x,y) - e(x,y) = g^0(x,y) + \frac{\partial g^0}{\partial x} \Delta x + \frac{\partial g^0}{\partial y} \Delta y$$

where

$$\begin{aligned} \Delta x &= \frac{\partial x}{\partial p_i} \Delta p_i, \quad \Delta y = \frac{\partial y}{\partial p_i} \Delta p_i \\ i &\in \{da_{11}, da_{12}, da_{21}, db_{11}, db_{12}, db_{21}\} \end{aligned} \quad (4)$$

by defining

$$g_x = \frac{\partial g^0(x,y)}{\partial x}, \quad g_y = \frac{\partial g^0(x,y)}{\partial y} \quad (5)$$

(4) becomes

$$\begin{aligned} f(x,y) - e(x,y) &= g^0(x,y) \\ &+ g_x da_{11} + g_x x_0 da_{12} + g_x y_0 da_{21} \\ &+ g_y da_{11} + g_y x_0 db_{12} + g_y y_0 db_{21} \end{aligned} \quad (6)$$

rearranging and grouping the terms:

$$\begin{aligned}
x^T &= [da_{11}, da_{12}, da_{21}, db_{11}, db_{12}, db_{21}] \\
A_i &= [g_x, g_x x_0, g_x y_0, g_y, g_y x_0, g_y y_0]_i \\
l &= f(x, y) - g^0(x, y)
\end{aligned} \tag{7}$$

we arrive at:

$$-e(x, y) = Ax - l \tag{8}$$

and, assuming

$$E(e) = 0, \quad E(ee^T) = \sigma_0^2 P^{-1} \tag{9}$$

the system can thus be seen to be a Gauss-Markov estimation model.

We omit the radiometric shift parameter r_s in equation (6) to avoid radiometric correction at this time, as it would interfere with original radiometric calibration of the system [7]. The function $g(x, y)$ contains stochastic quantities, a fact which is ignored in order to allow the use of the Gauss-Markov model. This should not have significant effects on the results [6].

Minimising the sum of the square of the differences between the grey levels in the template and search patches, the unknown parameters in x are determined in equation 10.

$$\hat{x} = (A^T P A)^{-1} A^T P l \tag{10}$$

with the variance factor and residual vector given by:

$$\begin{aligned}
\sigma_0^2 &= \frac{1}{r} v^T P v \\
v &= A \hat{x} - l
\end{aligned} \tag{11}$$

The final solution has to be attained iteratively due to the nonlinearity in (2). After each step, parameters are updated,

$$a_1 = a_0 + da \tag{12}$$

the search image is re-sampled and the matrix A and vector l are re-evaluated. It is sufficient to resample the image using bilinear interpolation from the surrounding 4 pixel values [1]. Convergence occurs when $\Delta \rightarrow 0$.

4. Input Data

The video sequence is acquired with three synchronised CCD cameras, resulting in a triplet of grey-level images per frame. Multiple views allow us to infer 3D information of the scene. Increasing the number views increases the system accuracy and robustness.

In order to be able to attain accurate 3D information from the three views, we need to calibrate the system, so that we can obtain the relationship between image coordinates and world

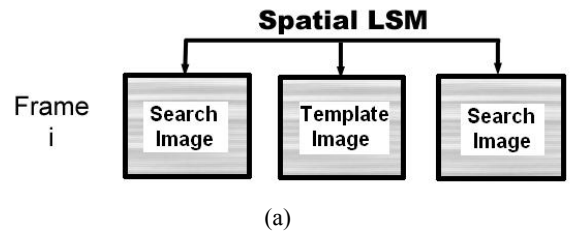
coordinates. The calibration process finds the necessary intrinsic and extrinsic parameters of each camera (intrinsic parameters – focal length, effective pixel size, image centre coordinates and radial distortion coefficient; extrinsic parameters – rotation matrix and translation vector). The method we have adopted is that of [8]. The method requires the cameras to observe a planar pattern from at least two different orientations. The procedure consists of a closed-form solution, followed by a nonlinear refinement based on maximum likelihood criterion.

Finally, the pixel coordinates of an image point are linked to the world coordinates of the corresponding 3D point by the perspective (pinhole) model. The world coordinate system is the 3D position relative to a predetermined reference point.

5. Tracking

The least squares algorithm is used to refine an initial estimate of correspondences to sub-pixel accuracy. In the first frame, key points are selected manually. The initial estimate of correspondences is achieved by performing correlation matching along epipolar lines to provide an approximation within a few pixels. This initial approximation is important, because both the matching reliability and convergence of least squares algorithm depend on it. With three views, the centre image is chosen as the template, and the left and right as search images.

The tracking algorithm is similar to that of [2]. Considering a single point somewhere in the template image, we define a small patch (15x15 pixels) around it. Using spatial LSM we find the corresponding points in the left and right search patch (Figure 2a). Looking at each view separately, we predict the positions in the subsequent frame. With temporal LSM we find the exact position of the point in each of the three views in the next frame (Figure 2b). Finally, working with the point in frame $i+1$ of the template image, spatial LSM is performed to find corresponding points in the search images again (Figure 2c). If this result is acceptably close to the result of the temporal LSM of the left and right view, that point is considered to be exactly tracked. However, if the difference is too large, the matching process repeats with modified parameters, e.g. larger search area.



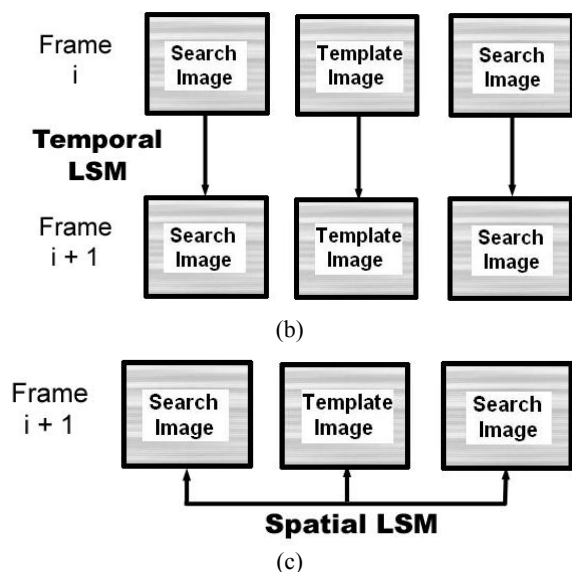


Figure 2 a) Step one, b) step two, c) step three of the tracking process

Once the point is tracked, its trajectory, velocity and acceleration can be obtained. Despite good matching results, false trajectories may occur. To filter the false trajectories out, the velocity and acceleration information is used. If, instead of a single point we track whole surfaces for the key parts of the human body, we get clusters of trajectories. As we consider the human body as an articulated moving object, the resulting vector field of trajectories must be locally uniform, i.e. the velocity vector must be nearly constant in sufficiently small regions at a particular time. Thus single trajectories are compared to local mean values of the velocity vector. Should the difference be too large, the trajectory is disregarded.

6. MPEG-4 for Human Body Animation

The MPEG-4 standard has defined two sets of parameters for the animation of the human body. These are the Body Definition Parameter set (BDP), and the Body Animation Parameter set (BAP). The BDP set defines the set of intrinsic properties of the person. It transforms the default body to a customised body with its body surface, body dimensions and if required the texture. The BAP set defines the extrinsic properties that determine the pose of the body. As the bitstream of BAP's is received, animation of the body is produced.

We intend to use the BAP set which contains the parameters that define the skeleton. Although the MPEG-4 defined skeleton is of much higher complexity than our stick-man model, it is our intention to work within the MPEG-4 standard when investigating the applicability of our system to low bit-rate video.

7. Conclusion

In this paper we have presented the approach that was taken towards tracking key points in 3D to capture human movement,

and applying it to low bit-rate video. A key element in our research is the method of least squares image matching. The input to the system consists of a video sequence with three views per frame, and the desired output is the model generated from derived pose parameters. Future work may look at model based tracking, where the model aids the tracking process. This helps the system deal with occlusion. With this approach the model would have to be more realistic, i.e. volumetric, to estimate the person's dimensions. These parameters are defined in the BDP set. In order to improve the automation of the system, an initialisation stage needs to be added to make rough initial estimates that will then be refined by the LSM method.

8. Acknowledgements

This research is sponsored by Thales, Armscor and the Department of Labour (South Africa).

References

- [1] Ackerman F., High Precision Digital Image Correlation. In Proc. 39th Photogrammetric Week, Institut fur Photogrammetrie, Universitat Stuttgart, Stuttgart, Germany, 1983.
- [2] D'Apuzzo N. and Plankers R., Human Body Modelling from Video Sequences. *International Archives of Photogrammetry and Remote Sensing*, Vol. 32, Part 5-3W12, pp 133-140, Onuma, Japan, 1999.
- [3] Fua P., Plankers R. and Thalmann D., From Synthesis to Analysis: Fitting Human Animation Models to Image Data. *In Proc. Computer Graphics International*, Anmore, Alberta, Canada, 1999.
- [4] Gavrilu D.M., The Visual Analysis of Human Movement: A Survey. *Computer Vision and Image Understanding*, 73(1), January 1999.
- [5] Gavrilu D.M. and Davis L.S., 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] Gruen A., Adaptive Least Squares Correlation: A Powerful Image Matching Technique. *South African Journal of Photogrammetry, Remote Sensing and Cartography*, Vol. 14, Part 3, June 1985.
- [7] Maas H.G., Stefanidis A. and Gruen A., From Pixels to Voxels: Tracking Volume Elements in Sequences of 3-D Digital Images. *International Archives of Photogrammetry and Remote Sensing*, Vol. 30, Part 3/2, 1994
- [8] Zhang Z., A Flexible New Technique for Camera Calibration. Technical Report MSR-TR-98-71, Microsoft Corporation, 2002.

- [9] Zhang Z. and Faugeras O., 3D Dynamic Scene Analysis: A Stereo Based Approach. Springer-Verlag, 1992.