

# View-based Location and Tracking of Body Parts for Visual Interaction

Antonio Micilotta

Richard Bowden

Centre for Vision, Speech and Signal Processing, University of Surrey

## Abstract

*The purpose of this research is to provide a coarse estimate of body pose. Our main interest is not in 3D biometric accuracy, but rather a sufficient discriminatory representation for visual interaction. The algorithm employs a general approximation to body shape, applied within a condensation [1] framework, while making use of an integral image to maintain near real-time performance.*

*Furthermore, we seek to locate the head and hands in a hierarchical manner. Hand position is disambiguated using a prior on body configurations. This prior is further used to estimate the location of other key body components using statistical methods. We demonstrate the system tracking within a complex, cluttered environment.*

## 1. Introduction

The final objective of this work is to create a visual human-computer interaction tool using an uncalibrated monocular camera system in a cluttered environment. Current progress locates and tracks key body parts of the upper torso in a hierarchical manner. Thereafter, pose and movement of these components will be used for gesture recognition purposes for visual interaction.

## 2. Methodology

Following initial background segmentation, a torso shaped mask is applied to the image using a particle filter [2]. Knowledge of the position and scale of the torso is used to locate the position of the head, after which a statistical skin model can be determined on the fly. With the assumption that skin tone of the face and hands is similar, the same skin model is used to locate the hands. Using a pre-learned prior on likely body configurations, the left and right hand are disambiguated. As elbow positions are not easily distinguishable from other body parts, this prior is also used to predict likely positions of the elbows. These estimates are then used to aid the location of the elbows in the video sequence.

### 2.1. Background Segmentation

Chroma-keying offers a simple solution to background segmentation in uncluttered, uniform backgrounds. Adaptive background segmentation algorithms prove to be essential in cluttered scenes where multiple individuals are to be isolated.

### 2.2. Location of Torso

With the user segmented from the background, detection and tracking of body parts can commence. A mask in the shape of a human torso (see Figure 1) has been designed based on the ratio of Da Vinci's Vitruvian Man. Under a CONDENSATION framework, the best fit mask offers key body measurements such as height and arm length that play a crucial role in the tracking of the head and hands.

Owing to the complete removal of the background, an integral image can also be used to maintain near real-time performance. Processing speed is greatly improved because the number of queries to be processed per frame is reduced by a factor of approximately 2500. A trade-off for this enhancement is that rotation of the torso cannot be determined.

### 2.3. Tracking of Head and Hands

The information obtained in isolating the upper torso is then used to locate the position of the head and hands in a hierarchical manner. A skin model is built from the facial region using robust statistics, and is then used for further location and tracking of both the face and hands. Rotation of the head and hands is also taken into account as it may play a useful role in the development of gestures for human computer interaction.

### 2.4. Disambiguating hands

A prior of body configurations has been formed from a training set of actors with hand-labelled body part positions. Once tracking of the hands commences, this prior can be used to disambiguate the left hand from the right in the event that the hands cross over each other. The Mahalanobis metric is used to compare the body configurations of the incoming and prior data. Awkward poses, whereby arms are crossed over, yield a large Mahalanobis distance, thereby indicating that the pose is unnatural.

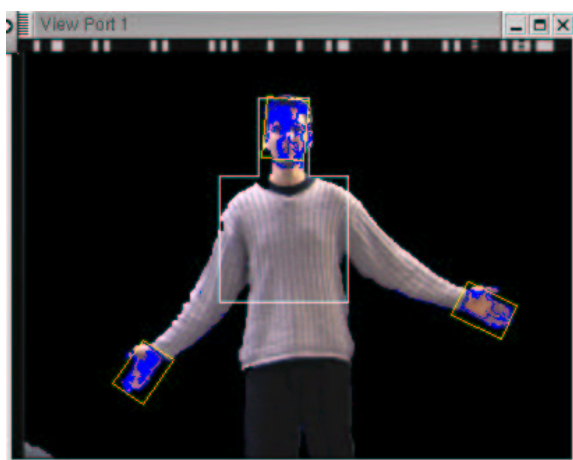


Figure 1: Tracking of Key Body Parts



Figure 2: Predicting Elbow Positions

### 2.5. Estimation of Elbow Positions

The inclusion of the elbows in a hierarchical manner may help to disambiguate the hands in situations where they occlude each other. Image cues for the detection of elbows are not apparent, and predictive methods need to be employed in order to offer a starting point with which to search the image space. Inverse kinematics proves cumbersome in a 2D application and multiple solutions are offered as the arm length 'changes' due to perspective. An exemplar approach that makes use of the prior, as mentioned in Section 2.4, proves to offer a relatively accurate starting point for each elbow. The hypotheses will however be limited to poses contained within the training set, and thus a full repertoire of training gestures would prove too costly for a real-time application. Statistical methods have therefore been investigated, whereby a statistical representation of the entire dataset can be pre-computed offline. With the aid of Principal Component Analysis, unknown elbow positions are inferred from the partial dataset obtained from the video data. The black markers of Figure 2 illustrate the locations of body parts that are tracked in the video sequence. The white markers represent the predicted positions of the elbows, as obtained via statistical analysis. Although not entirely accurate, these estimates do reduce the search space greatly.

## 3. References

[1] A. Blake and M. Isard. *Active Contours*. Springer Verlag, 1998.

[2] J. Deutscher, A. Blake and I.Reid. Articulated Body Motion Capture by Annealed Particle Filtering. *In Proceedings of Computer Vision and Pattern Recognition*, volume 2, pages 2126-2133, Columbia, USA , 2000.