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

Antonio Micilotta and Richard Bowden
CVSSP, SEPS, University of Surrey, Guildford, UK
{a.micilotta, r.bowden}@surrey.ac.uk

Abstract

This paper presents a real time approach to locate and track the upper torso of the human body. Our main interest is not in 3D biometric accuracy, but rather a sufficient discriminatory representation for visual interaction. The algorithm employs background suppression and a general approximation to body shape, applied within a particle filter framework, making use of integral images to maintain real-time performance. Furthermore, we present a novel method to disambiguate the hands of the subject and to predict the likely position of elbows. The final system is demonstrated segmenting multiple subjects from a cluttered scene at above real time operation.

1 Introduction

Human-computer interaction (HCI) is a discipline concerned with the design, evaluation and implementation of interactive computing systems for human use With the rapid development of fast, inexpensive PCs, HCI not only has a place in expensive industrial applications, but also in the home where users can control electronic equipment or interact with artificially intelligent constructs. HCI utilising video streams is of great benefit to handicapped or injured people who have limited control of their limbs, as well as enhancing the usability of cumbersome devices.

The overall objective of this work is to create a visual HCI tool using an un-calibrated 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.

Four primary underlying processes have been identified to realise this goal. These processes include background segmentation, human tracking, gesture recognition, and lastly, visual interaction and representation. The focus of this paper however lies within tracking key parts of the human body. For robust HCI, it is essential that the system operate in real-time. This research presents a tracking system running at 48 frames per second on a 2.8 GHz CPU.

2 Related Work

The task of tracking humans has received intense interest within the literature. Such work is too numerous to mention here but the vast majority of techniques fall into 2 areas: 3D

reconstruction of the human or 2D appearance based approaches. Since the seminal work of Hogg [4] articulated models (generally constructed from geometric primitives) have been used in model based approaches to tracking humans. More recent extensions to this approach use particle filtering frameworks, also known as the Condensation Algorithm [6]. Deutscher et al [3] presented the Annealed particle filter with an articulated geometric body model to reconstruct a moving human in 3D. The purpose of the adaptation was to reduce the search time as particle filters tend to be slow in implementation due to the complexity of the approach and dimensionally of the search spaces. The approach also required multiple cameras to overcome ambiguities, resulting in intensive off-line processing and limiting the application to calibrated multi-camera studios. Sidenbladh [9] overcame the constraints of multiple cameras by employing extremely strong motion priors (again in a particle filter framework) to overcome visual ambiguity. Results were presented tracking a walking human in a monocular image sequence. The prior, constructed from motion capture data, makes it difficult to extend to more general applications where the motion of the subject is less well defined.

Particle filters are popular as they provide a probabilistic framework which can support multiple hypotheses, however their susceptibility to converge upon maxima and the large population sizes required for accurate modelling of the posterior mean that typically they are not always suitable and slow in operation. The complexity of the problem also lies in the complexity of the 3D model to be fitted to the image. However, accurate 3D reconstruction is often not necessary and a less accurate 2D model is often sufficient. Ioffe and Forsyth [5] assumed that an image of a human can be decomposed into a set of distinctive segments modelled in 2D. A parallel edge segment detector is used to locate image regions that could possibly be body parts. Once the segments have been detected, they are assembled into groups that could represent a person, using a pre-defined top level classifier.

A comparable example to the overall objective of our work is Pfinder [11], as it makes use of the 4 processes mentioned in Section 1. It uses a blob representation to extract a structurally meaningful segmentation of body parts. Feature vectors are formed at each pixel by adding spatial co-ordinates to the colour components of the image. These vectors are then clustered so that similar properties (location and colour) are combined to form coherent connected regions or 'blobs'. Many approaches make use of colour to identify crucial body parts such as the head and hands. A direct method of doing so is to search for skin tone. This technique is popular, but still proves problematic in cluttered scenes where background objects (such as wooden furniture) is often misclassified as skin. Furthermore, the sole use of simple cues such as colour is insufficient as the head and hands are indistinguishable. Sherrah and Gong [8] take a view-based approach under a Bayesian framework, in which motion, skin colour and intensity based orientation measures are extracted in order to overcome these ambiguities.

Our approach brings together various elements of this work to track humans in cluttered scenes. Background suppression forms a crude segmentation of foreground from background, then a particle filter is used to locate and track the human. We demonstrate how an integral image can be used to provide the real-time operation of a particle system. From this initial estimate of gross body location, a person specific colour model is learned to model the head and hands and tracked with 3 further dedicated particle filters. A probabilistic model of body configurations is then used in a novel manner to both disambiguate the hands and provide an estimate for the elbow positions.

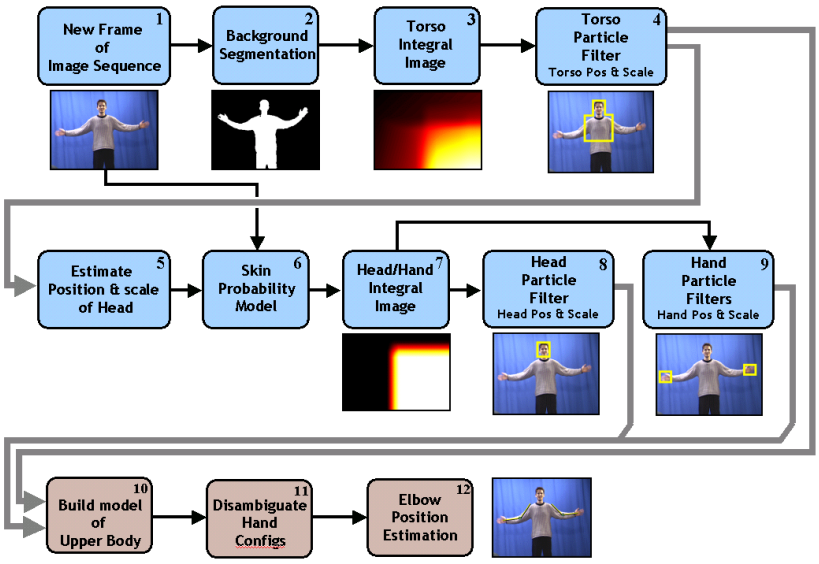


Figure 1: System Overview

3 System Overview

Our approach to the tracking of body parts includes various steps which are inter-dependent. With reference to Figure 1, a brief overview of the steps is listed below:

1. Acquire a colour image of the user using a single un-calibrated camera.
2. Segment the subject from the background.
3. Create an integral image (see section 5) from the binary image to improve system performance in terms of robust tracking and speed.
4. Initialise a particle filter system to track the torso, thereby providing the scale and position information of the user. These outputs form the prior data that is fed back into step 1 to continue tracking of the torso.
5. Estimate the scale and position of the user's head and hands based on the torso data of step 4.
6. Build a skin probability model from colour information of the located head region.
7. Detect regions of skin tone using the on-line skin model. An integral image of the detected regions is created as in Step 3.

8, 9. Initialise 3 particle filter systems for the head and hands.

10. From the outputs of the particle filters, build a model of the upper body.
11. Using the prior model of body configurations, disambiguate the hands from each other.
12. From the prior, infer the elbow positions of the disambiguated model.

4 Tracking of Body Parts

Figure 2 (a) illustrates the Renaissance subdivision of the human figure into eight lengths, each of which is equal to the length of the head measured from the chin to the top of the skull. For the purpose of this paper, these parts will be referred to as *skeletal units*.

We have chosen to use four particle filter systems, each dedicated to tracking a specific part (*torso*, *head*, *left hand* and *right hand filters*), as this reduces the dimensionality of the search space, and therefore the number of particles required to estimate the posterior. These are applied in a hierarchical manner such that the gross location of the torso influences the predicted priors of the latter filters. In addition to position, the size of the head and hand masks is also dependent on that of the torso mask, as once the torso model has been fitted, an estimate of the skeletal unit is obtained.

4.1 Background Suppression

Like Pfinder [11] where a single Gaussian was used to model the variance of background pixels, we assist tracking via a static camera assumption and background suppression. Although this is not the novelty of the paper it is state of the art and as such a brief description is included for completeness. Our background removal algorithm was originally developed for exterior visual surveillance and relies upon modelling the colour distribution with a Gaussian mixture model on a per pixel basis. This is learnt in an online fashion using an iterative approximation to expectation maximisation. The results of which are a binary mask where black denotes background and white the foreground object of interest. For further details the interested reader is directed to [7]. This binary mask is then converted to an integral image and passed to the torso particle filter for tracking.

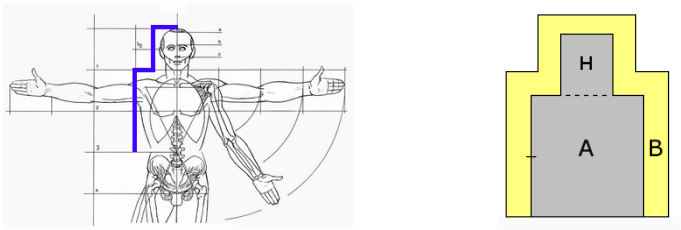
4.2 Tracking the Torso

Referring to Region A of Figure 2(b), the torso mask is constructed to accommodate the head and shoulders, and extends down to the waist, where major spinal rotation occurs. A secondary outer mask, Region B, is created such that Area(A) = Area(B).

The motivation behind using such a coarse body-shaped mask lies in the application of integral images (see section 5). The use of integral images allows for large particle population sizes for robust tracking, while maintaining real-time performance.

This mask is applied to the segmented binary image of the user: the sum total of all non-zero valued pixels in each region is computed, and with a slight abuse of notation, is denoted $\sum A$ and $\sum B$. The *net* result is $\sum A - \sum B$ (with negative values set to 0), which is then normalised by Area(A) to give a fitness score $S = \frac{1}{Area(A)} |\sum A - \sum B|$, in the range [0,1]. This score is computationally inexpensive to calculate, contributing to the real-time aspect of the particle filter tracking system. Note that the bottom section of both the inner and outer windows coincide, preventing the lower torso from influencing S.

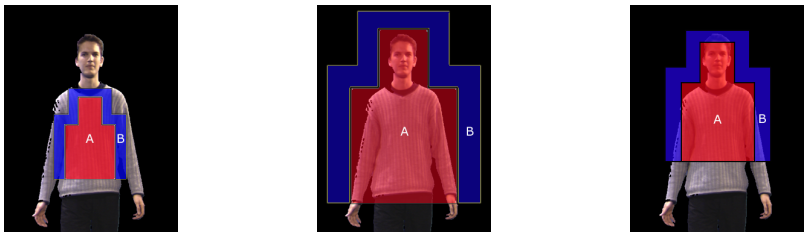
Figure 3 illustrates situations that yield extremes of S. In Figure 3(a), the entire mask lies within the subject. Since Area(A) = Area(B), $\sum A \simeq \sum B$, resulting in $S \simeq 0$. The mask in Figure 3(b) is very large, providing a large value for $\sum A$. However, S will be low due to the normalisation by Area(A). The highest value for S arises in a situation demonstrated by Figure 3(c). The subject occupies the majority of Region A, with little overlap into Region B. Here $\sum A \simeq Area(A)$ and $\sum B \simeq 0$, resulting in a high score for S.



(a) (b)

Figure 2: (a) Virtuvian Man (b) Torso mask

The polygon is fitted to the image using a similarity transform to minimise S, allowing the scale, position and rotation of the torso to be extracted. However, the use of the integral image precludes rotation from the estimate. This is acceptable as we assume the subject remains upright. (See Section 8 for further discussion). The scale plays a particularly useful role in the reconstruction of the user as the *skeletal unit* can now be determined. A combination of this information leads to a reliable estimation of the position and size of the head - it is assumed that the head exists in region H of Figure 2 (b). Arm length and hand size can also be determined, all of which is fed to latter filters.



(a) (b) (c)

Figure 3: (a) Low Score (b) Low Score (c) High Score

4.3 Tracking the Head and Hands

With a reasonable estimate of head size and position, a skin model specific to the user's face is built on-line. Skin tone of the hands is assumed to be the same as that of the face, and hence the same skin model is used for hand detection.

A prior bivariate Gaussian (Hue and Saturation) was constructed from skin samples taken from a selection of natural images. This produces a model far too general for use in robust segmentation, but does however allow us to discard pixels that are outliers. From these inliers in region H, a new bivariate is learned, creating a refined skin model.

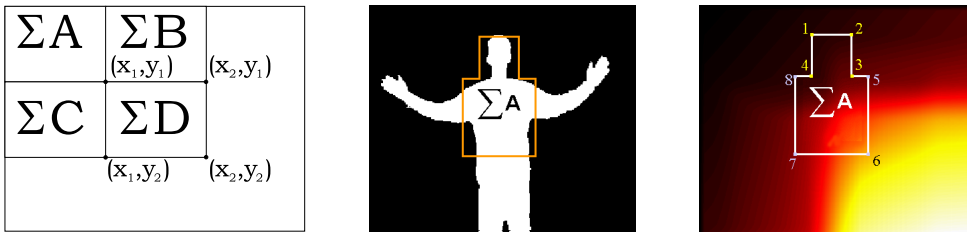
The head and hand filters function in a similar manner to the torso filter, except in this case skin-tone valued pixels contribute to $\sum A$ and $\sum B$.

5 Integral Images for Real-Time Performance

Rectangular features can be computed quickly using an intermediate representation called an *integral image* [10], also known as a *summed area table* in texture mapping [2]. The value of the integral image at point (x,y) is the sum of all the pixels above and to the left of (x,y), inclusive:

$$II(x,y) = \int_{i=0}^{x} \int_{j=0}^{y} I(i,j)\mathrm{d}j\ \mathrm{d}i \tag{1}$$

where II is the integral image, and I is the original image. In the case of a discrete image, this is approximated by $\sum\sum I(i,j)$.



(a) (b) (c)

Figure 4: Integral Image Computation

The entire II can be computed in this manner for all (x,y), however it is more efficient to determine the integral image incrementally, rather than conducting a double summation of the previous pixels:

$$\forall x,y\ II(x,y) = I(x,y) + II(x,y-1) + II(x-1,y) - II(x-1,y-1) \tag{2}$$

Using the integral image, any rectangular sum can be computed in three mathematical operations. In Figure 4(a), The value of the II at point (x_1,y_1) is the sum of the pixels in Region A ($\sum A$)of the original image, $II(x_2,y_1) = \sum A + \sum B$, etc. $\sum D$ can therefore be computed as

$$\sum D = \int_{i=x_1}^{x_2} \int_{j=y_1}^{y_2} I(i,j)\mathrm{d}j\ \mathrm{d}i \equiv II(x_2,y_2) - II(x_2,y_1) - II(x_1,y_2) + II(x_1,y_1) \tag{3}$$

The body part masks of Section 4 have been constructed from rectangular shapes such that their sums can be computed in the same manner mentioned above. Figure 4(c) visually represents the integral image of Figure 4(b). Using the integral image, $\sum A$ can be computed as $(3-2-4+1)+(6-5-7+8)$.

Integral images applied within a particle filter framework allow for real-time performance as the incremental summation of non-zero pixels to determine $\sum A$ would prove to be too computationally expensive. For illustrative purposes, let the best fit torso mask

of the user occupy 12 500 (100x100 + 50x50) pixels. In order to determine $\sum A$, 25 000 operations(1 query and 1 addition) would have to be conducted. A typical particle filter may contain $>$ 500 masks, resulting in a total of 25 000 x 500 = 12.5 million operations. In addition, the greater the image resolution, and the closer the user is to the camera, the greater the size of the mask, making the computation of $\sum A$ even more costly.

Taking the integral image into account, $\sum A$ is computed by $(3-2-4+1)+(6-5-7+8)$, a mere 7 mathematical operations. The initial overhead of the integral image calculation is dependent on image resolution, with a total of $\approx$330 000 operations when working with $\frac{1}{2}$PAL images (384x288). However, once the integral image has been determined, the score calculation for each mask is independent of image resolution and subject size. For a population of 500 masks, the total number of operations = 333 000 + (500 x 7) = 336 500, 2.7% of that of the incremental summation method. The larger the population size, the greater the benefit of the integral image.

6 Prior Data for Pose Estimation

The converged particle filter systems provides the position of 8 points on the body, namely the hips, shoulders, hands and head (chin and hairline). Two prior models (ϕ and ψ) of body configurations have been built from 4647 representative examples from image sequences of subjects performing deaf sign language. Coloured gloves and elbow bands were used for ease of acquisition of groundtruth data for prior model construction. ϕ contains concatenated coordinates for the 8 points mentioned above, while ψ contains 10 points as the elbow positions are also included.

With these prior data sets, the left and right hands can be disambiguated, and the undetermined elbow positions of the tracked subject can be inferred.

6.1 Disambiguating the Hands

A 16x16 covariance matrix Cov_ϕ, is formed from data set ϕ, where $Cov_\phi = \frac{1}{N-1}(\phi_i - \mu_\phi)(\phi_i - \mu_\phi^T)$. The 8 x,y coordinates of the tracked user are concatenated to form two feature vectors $\mathbf{Y}'_1, \mathbf{Y}'_2 \in \Re^{16}$, where the coordinates of the hands are swapped in each vector. To determine which vector fits the prior model best, the Mahalanobis distance d is determined. $\mathbf{Y}'_j$ is chosen according to the smaller of the two distances:

$$j = argmin_{i=1}^{2} \left(\sqrt{(\mathbf{Y}'_i - \mu_\phi) Cov_\phi^{-1} (\mathbf{Y}'_i - \mu_\phi)^T} \right) \tag{4}$$

Awkward poses, for example arms crossed over each other, yield a large d, thereby indicating that the pose is unnatural. In such a situation, the hand filters are either switched or re-initialised according to the size of d.

6.2 Estimation of Elbow Positions

Image cues for the detection of elbow positions 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 prove to be cumbersome in 2D applications, and also offer multiple solutions as the arm length 'changes' due to perspective. An approach that makes use of the prior model offers a relatively accurate starting point for each elbow.

Eigen decomposition of $Cov_\psi(\Re^{20})$ yields an eigenvector Matrix $\mathbf{P}$, and a corresponding set of eigenvalues $\mathbf{b}$. In matrix form, any body configuration $\mathbf{X}$ can be reconstructed as the mean plus the weighted sum of the eigenvectors:

$$\mathbf{X} = \bar{\mathbf{X}}_\psi + \mathbf{P}\mathbf{b}^{\mathbf{new}} \quad (\mathbf{X}, \mathbf{b}^{(\mathbf{new})} \in \Re^{20}) \quad where \quad -3\sqrt{b_i} < b_i^{new} < 3\sqrt{b_i} \tag{5}$$

$\bar{\mathbf{X}}_\psi$ is the mean vector of data set ψ. $\mathbf{Y}'_j$ represents the disambiguated 16D model of the tracked user, and we attempt to construct the vector $\mathbf{Y} \in \Re^{20}$, the original model including estimates for the elbow positions. To achieve this, the elbow data is stripped from $\mathbf{P}$, giving a 16x20 matrix $\mathbf{P}'$. Similarly, the mean elbow positions are stripped from $\bar{\mathbf{X}}_\psi$, giving $\bar{\mathbf{X}}'_\psi \in \Re^{16}$. A new set of projection weights $\mathbf{b}_{\mathbf{new}}$ can be calculated by rearranging Equation 5:

$$\mathbf{b}^{\mathbf{new}} = \mathbf{P}'^{-1}(\mathbf{Y}' - \bar{\mathbf{X}}'_\psi) \tag{6}$$

With this new set of weights, the estimated model $\mathbf{Y}$ can now be determined:

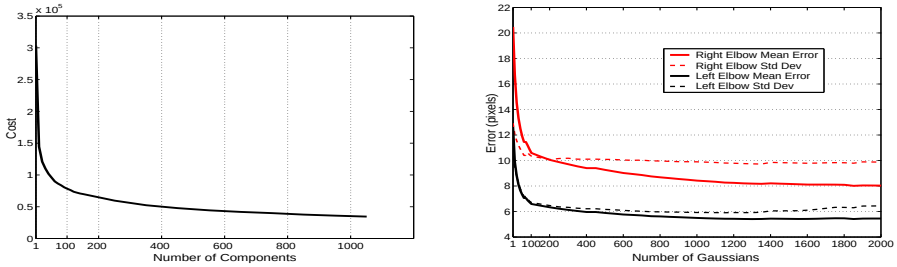
$$\mathbf{Y} = \bar{\mathbf{X}}_\psi + \mathbf{P}\mathbf{b}^{\mathbf{new}} \quad (\mathbf{Y} \in \Re^{20}) \tag{7}$$

7 Results

Following background removal and the formation of an integral image, a k-means clustering of foreground pixels is performed to extract an initial estimate and size for the position of each user. k is selected by hand to be the number of subjects to track, alternatively a connected component analysis of the binary image can be used to provide an estimate. Over estimation is of minor concern, as incorrectly initialised particle systems will diminish quickly. We choose to concurrently run k distinct particle filters, initialised with Gaussian distributions from the k component estimations. In terms of modelling the posterior, this makes little difference but it provides a convenient partitioning of the population samples for monitoring and re-initialisation. In order to cope with user occlusions, the means of each particle system are compared to each other to detect when more than one particle filter has converged upon a single solution – at this point it is reinitialised. In the examples, each torso filter consists of 500 particles with a Gaussian white noise drift term. The mean of the top 10% of particles from each system is selected as the likely estimate of position (shown in figures 7 and 6). From this mean hypothesis the head region is estimated and is used to build the HS bivariate colour model to represent skin.

A data set of 7600 representative body configurations was extracted from a video sequence consisting of a user wearing coloured gloves and elbow bands for ease of groundtruthing. This data set was then separated into 4647 training examples and 2953 unseen test examples. PCA was performed upon the training set and the test data was used to estimate the elbow positions as detailed in section 6. Figure 5(b) shows the mean and std dev error for the estimates. It is obvious that the manifold on which the data set lies is unlikely to be linear and is therefore not well approximated by a single Gaussian. We therefore use a Gaussian mixture model to represent the training set. However, the optimum number of components must be selected. Following [1] we use the cost function from k-means to estimate the number of components. Figure 5(a) shows how the cost

decreases as the number of components increases. The natural number of clusters is said to be the number for which further increases does not produce significant gain in overall cost. From the figure we would estimate this to be in the region of 100 components. This is confirmed by Figure 5(b) showing the reconstruction error for mixture models with an increasing number of Gaussians. Again the highest benefit in terms of error minimisation is in the region of 100 Gaussians. Further increases reduce the generalisation of the model as it will eventually regress to a nearest neighbour approach [1].



(a) (b)

Figure 5: (a) Cost (b) Error

Figure 6 shows a subject performing deaf sign language, with the torso, head and hands tracked. Using a Gaussian mixture model of 100 components, the estimated elbow positions are indicated by P_R and P_L. The body configurations were also used to test the validity of the hand disambiguation method (Section 6.1), and the hands were correctly disambiguated in 98% on the test examples.

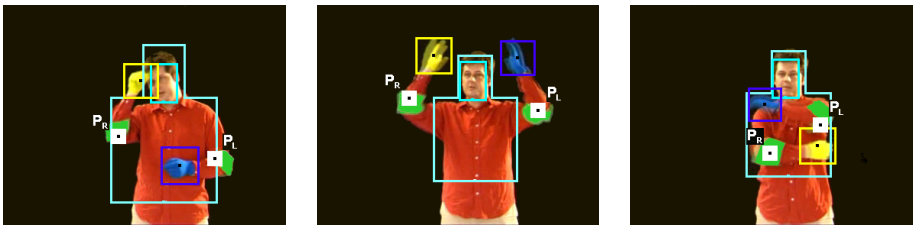


Figure 6: Estimation of elbow positions

Figure 7 shows snapshots of an image sequence of two users moving in a cluttered scene. It is apparent how the torso filter of each system contends with the two differently sized users, and their forward movement with respect to the camera. The hand filters are colour coded indicating the distinction between left and right hands after disambiguation.

8 Conclusions and Future Work

Knowledge of Da Vinci's human figure schematic has allowed for the use of coarse body shape estimates. These estimates, combined with the invaluable speed benefits of integral

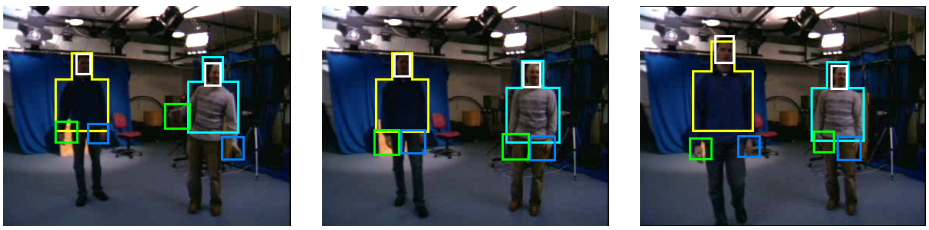


Figure 7: Tracking multiple people in a complex, cluttered scene

images, have made a real-time particle filter system achievable. Tracking has proven to be even more robust as the particle population sizes can be increased significantly with minimal influence on system performance. A small limitation of the integral image is that it cannot be used in the detection of rotated body parts, however future work aims to conduct integral images on rotated versions of the image sequence frames in order to overcome this constraint.

Relatively accurate elbow positions are estimated based on the current prior model on body configurations. A prior built from more extensive data should prove to be able to cope with a wider variety of users, and offer even more accurate positional estimates.

References

[1] R. Bowden. *Learning Non-Linear Models of Shape and Motion*. PhD thesis, Brunel University, Systems Engineering, October 1999.

[2] F. Crow. Summed-area tables for texture mapping. In *Proc of the 11th annual conf. on Computer graphics and interactive techniques*, pages 207 – 212, 1984.

[3] J. Deutscher, A. Blake, and I.Reid. Articulated body motion capture by annealed particle filtering. In *In Proc CVPR*, volume 2, pages 2126–2133, Columbia, USA, 2000.

[4] D.C. Hogg. Model-based vision: A program to see a walking person. In *Image and Vision Computing*, volume 1, pages 5–20, 1983.

[5] S. Ioffe and D. Forsyth. Probabilistic methods for finding people. *International Journal of Computer Vision*, 43(1):45–68, 2001.

[6] M. Isard and A. Blake. Visual tracking by stochastic propagation of conditional density. In *Proc. Conf of Autmatic Face and Gesture Recognition*, volume 1, pages 38–44, 1996.

[7] P. KaewTraKulPong and R. Bowden. A real-time adaptive visual survelance system. *Journal of Image and Vision Computing*, 21(10):913–929, September 2003.

[8] J. Sherrah and S. Gong. Resolving visual uncertainty and occlusion through probabilistic reasoning. In *Proc. of the BMVC*, volume 1, pages 252–261, Bristol, UK, 2000.

[9] H. Sidenbladh. *Probabilistic Tracking and Reconstruction of 3D Human Motion in Monocular Video Sequences*. PhD thesis, Royal Institute of Technology, CVAPL, November 2001.

[10] P. Viola and M. Jones. Robust real-time object detection. *ICCV Workshop on Statistical and Computation Theories of Vision*, 2001.

[11] C. Wren, A. Azarbayejani, T. Darrel, and A. Pentland. Pfinder: Real-time tracking of the human body. In *IEEE PAMI*, volume 19, pages 780–785, 1999.