

Human Tracking using 3D Surface Colour Distributions

Timothy J. Roberts, Stephen J. McKenna and Ian W. Ricketts

Division of Applied Computing, University of Dundee, DD1 4HN, Scotland
{troberts,stephen,ricketts}@computing.dundee.ac.uk

Abstract

A likelihood formulation for detailed human tracking in real world scenes is presented. In this formulation, the appearance, modelled using feature distributions defined over regions on the surface of an articulated 3D model, is estimated and propagated as part of the state. The benefit of such a formulation over currently used techniques is that it provides a dense, highly discriminatory object-based cue that applies in real world scenes. Multi-dimensional histograms are used to represent the feature distributions and an on-line clustering algorithm, driven by prior knowledge of clothing structure, is derived that enhances appearance estimation and computational efficiency. An investigation of the likelihood model shows its profile to be smooth and broad while region grouping is shown to improve localisation and discrimination. These properties of the likelihood model ease pose estimation by allowing coarse, hierarchical sampling and local optimisation.

Key words:

Human Tracking, Articulated Models, Sequential Estimation, Human Computer Interfaces

1 Introduction

The recent growth of research interest in human body tracking has been innervated by the problem's challenging nature and motivated by the potential solutions' applications in areas such as gestural interfaces, surveillance, monitoring, security, motion capture, games and sport. Tracking people is made particularly difficult by their complex appearance and motion; indeed the two primary problems are how to establish a good, realistic appearance model and how to efficiently estimate the model parameters over time. The focus of this work is to address the former problem, that of modelling human appearance

in real-world scenes. Difficulties in constructing an appearance model for human tracking are that the clothing is not known *a priori*, lies on an irregular 3D surface, is often textured and changing over time and that body parts frequently undergo self-occlusion. Many current state-of-the-art trackers (e.g. [1]) rely upon a simple appearance, such as tight, high contrast, un-textured clothing and/or a simple or known background in order to be successful. Such assumptions are inappropriate for applications such as surveillance and monitoring [2,3]. The difficulty of modelling human appearance is also apparent from work in rendering realistic images of people (see for example [4]).

Human tracking research often uses a generative, high-level model. Firstly, a model is established describing the pose of the human. The aim of the tracking system is then to estimate the (posterior) probability $p(X_t|Y_t)$ of the parameters of this model $X_t = \{x_0, \dots, x_T\}$, based upon the observations $Y_t = \{y_0, \dots, y_T\}$ and a body of prior knowledge $p(X_t)$ (where $x \in \mathfrak{R}^n$, y_t represents an image frame and $t \in [0, T]$). In general, distributions at new times can be found using (2).

$$p(x_t|Y_t) = \int \dots \int p(X_t|Y_t) dx_0 \dots dx_{t-1} \quad (1)$$

$$\propto \int \dots \int p(Y_t|X_t)p(X_t) dx_0 \dots dx_{t-1} \quad (2)$$

The first term on the RHS of (2) represents the likelihood of a path through state space given the image sequence and the second term is the prior over that path. In can be seen that the dimensionality of this integral grows with time as more information is considered and direct evaluation becomes prohibitively expensive. Therefore, it is usual to consider the state evolution to be a Markov process and the distributions can then be found recursively:

$$p(x_t|Y_t) \propto p(y_t|x_t) \int p(x_t|x_{t-1})p(x_{t-1}|Y_{t-1})dx_{t-1} \quad (3)$$

The first term on the RHS of Equation (3) represents a single image likelihood model and the second represents the probability of transitioning from the previous posterior distribution. It is important to realise that the posterior probability distribution is induced by the chosen likelihood model and motion model. Accurate modelling of these terms allows for easier and more accurate estimation. The topic of this paper is the derivation and evaluation of a likelihood model that allows accurate, efficient tracking in real world environments. Since the focus of the paper is on developing and evaluating an efficient, highly discriminatory likelihood we approach the problem as one of recursive maximum likelihood. Prior constraints on pose, such as modelling of human motion, are not discussed in this paper.

To begin we ask the question: what are the properties of a good likelihood formulation? A good generative model will be able to utilise all the relevant input data, transformed and weighted appropriately and will therefore be able to re-synthesise an appropriate representation of the input data given the solution. A good likelihood model should have characteristics which allow for efficient, accurate estimation, such as a strong, broad response around the solution and large discriminatory power to reduce the effect of false maxima.

A difficulty in the case of visual human tracking, is that, due to the large uncertainty in a foreground and background appearance, single frame likelihood models have poor discrimination. Such models may be constructed by making limiting assumptions but the resulting systems will not allow for accurate estimation in other, more realistic, scenes.

Here we propose to construct a likelihood model by propagating the foreground appearance as part of the state. The state then consists of both a pose component and a texture component, i.e. $x \equiv \{x_{shape}, x_{texture}\}$. Such a *dynamic* likelihood model should have as many of the properties identified above as possible. However, problems arise when trying to estimate a complex, changing appearance given the limited amount of available data and computational resource. Furthermore, model initialisation is essential when using a dynamic likelihood. Existing tracking systems typically require manual initialisation and this approach is similarly adopted in the experiments described here. While this paper investigates the results obtained by applying a dynamic likelihood model, the method presented should ultimately be considered to be part of a larger system that combines static and dynamic likelihood components for automatic (re)initialisation and efficient, iterative tracking.

1.1 Outline

Section 2 discusses previous work on human appearance modelling, dividing the discussion into shape modelling and likelihood modelling. Section 3 describes our method of tracking based upon recursively updating pose *and appearance* parameters. In particular, Section 3.2, the main contribution of the paper, discusses the form of the appearance model, its prior constraints and the foreground appearance grouping procedure. Tracking results are then presented in Section 4 along with an investigation of the form of the likelihood. Finally, Section 5 presents the conclusions of the work and discuss what we consider to be some promising future directions.

2 Background

Gavrila [5] and Aggarwal and Cai [6] provided general reviews of human tracking research up until 1998. Research was classified according to the nature and complexity of the body model and corresponding search space. Moeslund *et al.* [2,7] reviewed the research performed up until 2000 and described research from four system areas: initialisation, tracking, pose estimation and recognition. Appearance modelling can be naturally partitioned into a shape model that describes the pose of the target and a likelihood model that links this model to measurements in the image.

2.1 Shape Models

2D Contour Models Contour body models have been used for tracking humans, e.g. [8]. These systems involve learning the principal shape contour deformations. However, silhouette contour models alone are not always appropriate for the applications under consideration since they do not describe poses such as the hands moving over the body. Recently, models have been proposed that combine the occluding contour with internal pose information [9,10]. Contour models are of particular interest when a model is not known *a priori* or when it is easier to learn the important variations.

2D Articulated Models Since the structure of the body is well studied and relatively simple at a coarse level, it is natural to consider articulated body models. These models capture the kinematic structure using shape primitives connected in a hierarchical fashion. View-based articulated modelling is exemplified by the work of Ju *et al.* [11]. Cham and Rehg [12] described a scaled prismatic model which parameterises foreshortening and avoids certain singularity problems.

3D Articulated Models Modelling the subject in 3D has the advantages of automatically handling self-occlusion and multiple viewpoints. Hierarchical assemblies of 3D components are used as proposed by Marr and Nishihara [13] and adopted in the early work of Hogg [14]. There is a large variation in the complexity of the primitive shape components, but typical choices are cylinders, truncated cones with elliptic cross sections and super-quadrics. For example, Kakadiaris and Metaxas [15] proposed using tapered super-quadrics and fitting these primitives using multiple orthogonal views. In order to reduce the size of the state space, joint angle ranges are often used and some state components, such as the size and shape of the body parts, are considered fixed or well estimated. Spring type joints have been proposed to better model the full range of body motions around the more complex sites such as the shoulder. Bregler and Malik [16] describe the twists and exponential map formulation in which transformations are specified by a rotation around a 3D axis and a translation along that axis. This

formulation linearises the kinematics and removes certain singularity problems. Sminchisescu and Triggs [17] proposed incorporating data to specify a prior on the part sizes and checked pose configurations so as to avoid body part inter-penetration. Karaulova *et al.* [18] used a learnt hierarchical PCA model of the kinematic structure.

In certain applications, such as medical analysis, simple part-based techniques are too crude a model and a more flexible description of the surface of the body is required. Plankers [19] describes a body modelling scheme where rigid body parts are replaced by articulated soft objects. Each of these objects defines a field which in turn specifies the body surface. This model can then be fit to the subject's body using optimisation techniques and used for tracking. Due to the large number of degrees of freedom, these models are less suited to the applications under consideration here.

2.2 Likelihood Models

The range and complexity of clothing and scene conditions complicates the tracking problem. Many trackers rely on simple visual conditions such as tight, textureless clothing and high contrast, uncluttered background to be successful. The following sections present accounts of likelihood models and the trend toward integration of multiple measurements in a robust fashion. An attempt is made to summarise the advantages and disadvantages of different cues.

Edges Matching using edges is popular because the feature is relatively invariant to illumination changes and is fast and easy to extract. The disadvantages are that (i) it has poor discrimination due to textured clothing, background clutter and low contrast part boundaries, and (ii) it only provides a sparse cue. Gavrilu and Davis [20] derived a corrective term from the edge response using a robust Chamfer distance. Wachter and Nagel [21] matched edge filter responses to Gaussian profiles from parts that did not overlap other body parts. Ronfard *et al.* [22] used learnt Gaussian derivative filter responses for body parts and the whole body.

Colour and Texture Colour and texture information, although more susceptible to lighting changes, provide dense cues for tracking. Colour and texture information have been used to model both the foreground and background scene content. Tracking simple targets using colour has been investigated in some detail e.g., [23,24,25]. These techniques however are either not applicable to constrained articulated models (e.g. mean shift) or do not take into account the structure of the target's appearance (clothing), as is the focus here. A special case of foreground modelling is the use of skin colour to find body parts. Wren *et al.* [26] modelled the foreground and background colour statistically and used clustering to find blobs coarsely corresponding to body parts. Forsyth and Fleck [27] used learnt texture distributions to

find naked people in single images. Sidenbladh *et al.* [28] considered learning the principal components of the texture on the surface of a 3D cylindrical body-part model. A disadvantage of this technique is that the appearance is learned off-line.

Optical Flow Cedras and Shah [29] surveyed motion-based recognition systems. Optical flow fields can be used to estimate parametric motion models, providing dense cues that are reasonably insensitive to illumination changes. Several authors have used robust, correlation-based optical flow methods [11,30,16]. However, these systems often assume that only the target is moving. They suffer from accumulation of errors, do not allow single-frame pose estimation and rely upon the brightness constancy assumption. Sidenbladh *et al.* [31] used optical flow in a Bayesian framework with explicit occlusion handling to track a 3D model. Song *et al.* [32] used the position and flow of Lucas-Kanade feature points to detect and label human motion in the presence of simple background movement.

Cue Fusion The aim of cue fusion is to combine multiple complementary sources of information to improve the likelihood model. For example, Wachter and Nagel [21] combined an intensity template found from the previous frame with an edge profile and reasoned that the former stabilised tracking whilst the latter enhanced localisation. Deutscher *et al.* [1] used edge detection and background subtraction from multiple viewpoints to track through various complex motions. Moeslund and Granum [33] used skin colour from a stereo pair to constrain a search for the arm silhouette. Zhao *et al.* [34] integrated line measurements with piecewise surface colour histograms. Sminchisescu and Triggs [17] combined a correlation-based optical flow field with edge features weighted by their proximity to the flow field boundaries. Background modelling to find the silhouette has been combined with edge detection [1,35]. Darrell *et al.* [36] used depth to segment subjects in cluttered environments and colour and texture cues to find identify the head and hands. Okada *et al.* [37] integrated optical flow and depth information to estimate the 3D pose of the subject. Sidenbladh and Black [38] advocated learning filter responses for people in images rather than forming *ad hoc* models. Edge strength, ridge response and optical flow responses were combined to track upper body movement. They concluded that edges and ridges provide sparse information and that colour or texture may provide better results. Park *et al.* [39] used a watershed colour segmentation technique to find regions. Pose estimation was accomplished using region shape and a skin colour classifier.

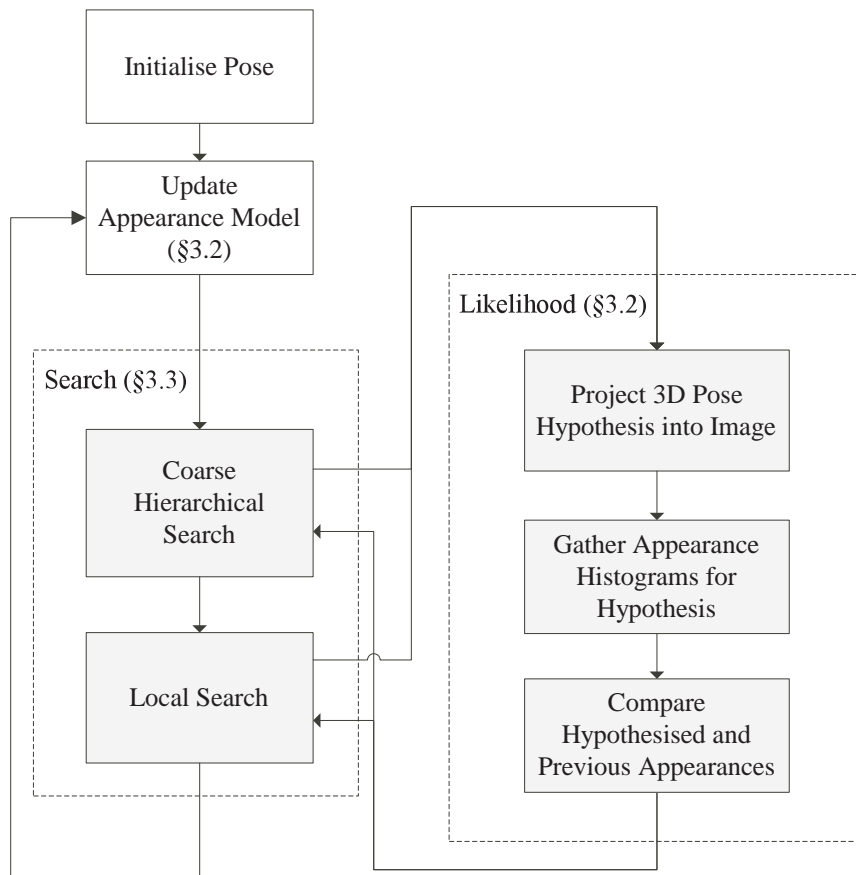


Fig. 1. The flow of control in the tracking system.

3 Method

Many of the systems described above rely upon likelihood models that assume restrictive scene conditions such as tight, high contrast, textureless clothing or a static, simple or known background. The system presented here is less restrictive in that it copes with textured and loose-fitting clothing. This section describes a tracker based upon propagating foreground appearance, adapting the appearance and a simple iterative (maximum likelihood) inference scheme that is adequate for short term tracking (as is often reported in the literature) using this likelihood. The flow of control in the system is illustrated in Figure 1.

3.1 Shape Model

The body is highly deformable and exact modelling of its form is infeasible and unnecessary in this context. Its important properties can be captured using an articulated body model. A 3D articulated shape model is used since it has a low dimensionality, captures the kinematic structure of the body, allows for

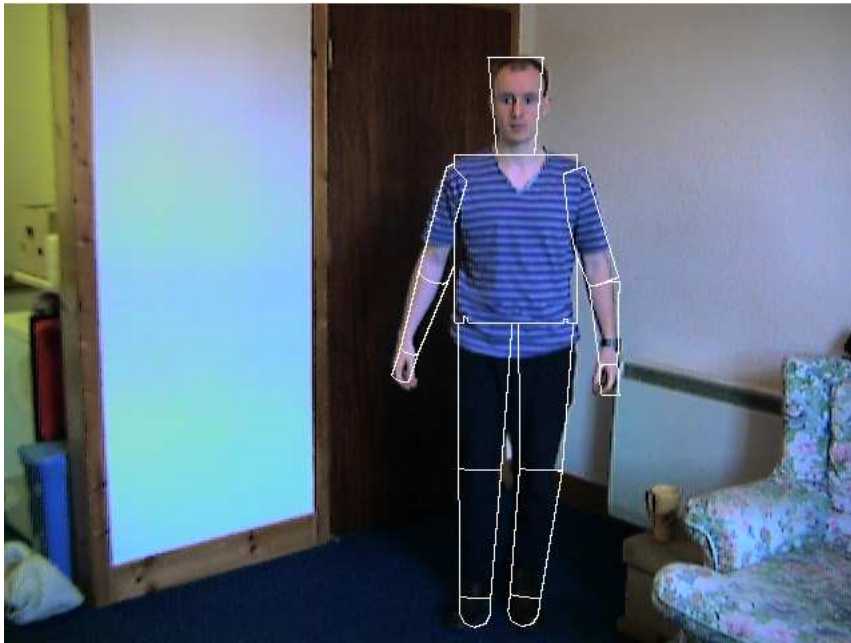


Fig. 2. The model overlaid on a frame from a waving gesture sequence used throughout the paper to illustrate ideas. Notice the approximate alignment of the edges.



Fig. 3. Frames 0, 10 and 26 from a waving sequence used throughout to illustrate ideas.

easy encoding of prior knowledge such as joint limits, automatically handles self-occlusion and enables changes in body part appearance due to rotation in depth to be handled explicitly.

The shape component of the state space, x_{shape} , then in general becomes the relative position and orientation of the primitives, their shapes and their sizes. Each of the shape primitives, indexed by $b \in \{1 \dots B\}$, has a surface that is naturally described using some co-ordinate system, a point in which is denoted by ω_b . For example, the surface of a fixed size cylinder is conveniently described by a length and an angle, i.e. $\omega_b = (l, \theta)$. A point on the subject is then specified by the pair (b, ω_b) . In order to project a surface point onto the image plane, the co-ordinates are first converted to Cartesian form. Homogeneous, relative transformations are then chained together to transform up the kinematic tree into world co-ordinates and finally, using a camera model, to project onto the image plane.

In the particular implementation described here, the body was represented using elliptic cross-section cylinders with constant, manually initialised sizes and shapes. The camera was modelled using an orthographic projection since the sequences under consideration did not contain strong perspective effects and the likelihood model was relatively insensitive to small changes in shape. However, the extension to perspective projection is straightforward. Independent movement of the head, hands and feet was not modelled, leaving a total of 22 degrees of freedom, encoded as 3 root translations, T_{Torso} , and 19 rotational degrees of freedom, $\{R_b\}$, four for each limb and three for the torso. Prior knowledge on joint angles was encoded using a quadratic ramp function at boundaries. The shape component, which is illustrated projected onto an example image in Fig. 2, can thus be written as:

$$x_{shape} = \{T_{Torso}, \{R_b\}\} \quad (4)$$

3.2 Likelihood Model

Likelihood evaluations are performed by projecting the model onto the image and comparing observed image features with expected values. Features are denoted here by \vec{q} . The expected features at a point, such as pixel colour or local filter responses, will have a (multi-modal) probability distribution rather than a single value due to body and clothing model inaccuracies, discretisation and noise. In general, these feature distributions are not known *a priori*. Here we consider a *dynamic* likelihood model where the features for regions on the surface of the 3D articulated model are initialised in the first frame and then propagated as part of the state:

$$x_{texture} = \{p_{(b,\omega_b)}(\vec{q}), \omega_b \in \Omega_b\}. \quad (5)$$

Using a dynamic likelihood model can greatly improve discrimination. However, it is well known that errors in appearance estimates can cause a tracker to diverge. In order to reduce such errors the foreground appearance is only updated using points that are very different from the foreground and by grouping points based upon prior expectations of clothing structure. Nevertheless, the reliability of long term tracking will be improved by incorporating static likelihood models. Since the focus of this paper is the dynamic likelihood this is left to future work.

In this paper we consider only colour distributions, the primary reasons being their quasi-invariance to viewpoint and ease of implementation. Since clothing is often textured these distributions can be multi-modal. Therefore, (non-robust) template tracking is inappropriate. Matching using distributions provides greater discrimination than matching individual measurements. We pro-

pose using normalised multi-dimensional histograms to represent the feature distributions and denote them by $H_{(b,\omega_b)}$. Other possible distribution statistics include, for example, cumulative histograms and moments.

In order to find the likelihood of a hypothesised pose, rays are cast into the scene at each pixel to determine the point of intersection, if any, with the shape model. Hypothesized histograms, $H'_{(b,\omega_b)}$, are collected for each region on the articulated model from the current image. These are compared to the propagated appearance using a distribution similarity measure, S . The likelihood is then defined, in Equation (6), as the sum of similarities weighted by the visibility of the region in the image, where V denotes the set of pixels corresponding to the body.

$$p_{dynamic}(y_t|x_t) \propto \frac{\sum_{\{V\}} S(H'_{(b,\omega)}, H_{(b,\omega)})}{|V|} \quad (6)$$

3.2.1 Clustering using split and merge

Estimating the feature distributions at points on the body is difficult given the limited image data available. Furthermore, the distributions are varying over time due to illumination changes and clothing movement. However, we observe that many of the points on the surface of the body belong to the same piece of clothing and will therefore often have similar distributions. A clustering routine can group points on the 3D model to improve estimation. We use a computationally efficient, iterative grouping scheme. Consider estimating a ‘grouped’ histogram, \tilde{H} , using all the ‘raw’ foreground histograms, $\{H\}$ (by marginalising over parts and positions on the parts and only considering pairwise terms):

$$\tilde{H}_{(b,\omega_b)}(q|\{H\}) = \sum_{b'} \sum_{\omega'_b} H_{(b',\omega'_b)}(q) p(b', \omega'_b | H_{(b',\omega'_b)}, H_{(b,\omega_b)}, b, \omega_b) \quad (7)$$

We model the second term on the RHS in a Bayesian fashion using a likelihood given by a similarity measure, S , on the ‘raw’ histogram bins and a prior determined from knowledge of clothing structure:

$$p(b', \omega'_b | H_{(b',\omega'_b)}, H_{(b,\omega_b)}, b, \omega_b) \propto S(H_{(b',\omega'_b)}, H_{(b,\omega_b)}) P(b', \omega'_b | b, \omega_b) \quad (8)$$

Incorporating prior knowledge on clothing structure in this way should make the tracker less prone to drift due to errors in foreground appearance estimation. Direct use of the sum in (7) is computationally expensive since it involves summing over all points on the body for each unestimated histogram bin. It can be seen that large contributions to the sum must be similar to the histogram in question and therefore similar to each other. Therefore, the sum is

| b | ω_b | b' | ω'_b | $p(b', \omega'_b b, \omega_b)$ |
|-----------|-------------|-----------|----------------------------|----------------------------------|
| Upper Arm | l, θ | Upper Arm | $l, \theta + \delta\theta$ | 0.9 |
| Head | l, θ | Hand | - | 0.7 |
| Upper Arm | l, θ | Upper Leg | - | 0.3 |

Table 1
Example histogram merge priors

reasonably well approximated by the average bin value taken from the group of similar regions.

To perform region merging, a threshold K is introduced. It controls the level of detail represented by the system and encodes the model order. When K is large, the system behaves like a template tracker by preserving individual regions. When K is small, the system behaves like a blob tracker, ultimately representing the person using a single distribution. For a particular sequence, with a particular image resolution, target size and level of noise, there will be an optimal choice of threshold for tracking that balances appearance estimation with excessive loss of local structure. The merging decision criterion then becomes:

$$S(H_{(b', \omega'_b)}, H_{(b, \omega_b)}) > \frac{K}{p(b', \omega'_b | b, \omega_b)} \quad (9)$$

The clothing structure prior, $p(b', \omega'_b | (b, \omega_b))$, in (9) is learned from example images of differently clothed people by manually aligning the model to the image and performing an exhaustive pairwise comparison. The prior is set to the average of the observed similarities. Examples are shown in Table 1. Subjective priors were assigned based on an informal analysis of example images. Learning more complex clothing structure priors from large data sets is deferred to future work.

Merging is an $O(u^2)$ operation, where u is the number of unique regions. However, this cost is greatly outweighed by the improvement in computational efficiency due to reductions in the number of region comparisons and storage overhead. Fig. 4 illustrates a body part model with region grouping leading to shared feature distributions.

Since regions can erroneously merge we also introduce a splitting operation. However, performing this in a similar manner to merging requires unique histograms to be stored for every atomic region, resulting in a large storage overhead. Therefore, we currently use an heuristic splitting criterion based upon a threshold on the sum of histogram lookups in an atomic region from the current image. This rule is particularly efficient which is important since it is performed for every atomic region in every frame.

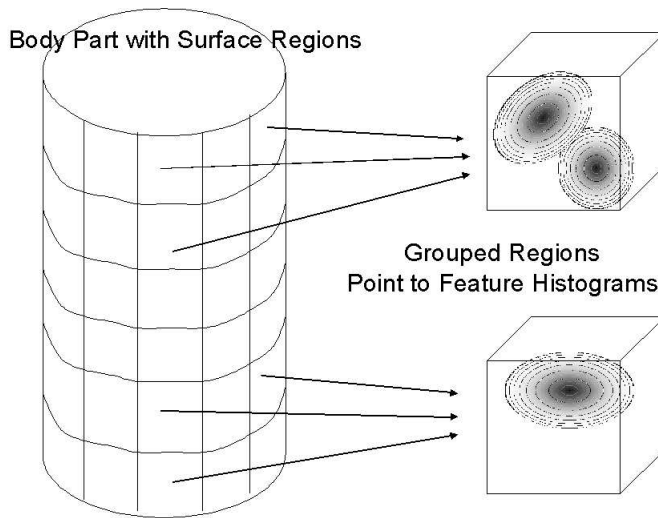


Fig. 4. A body part where grouped regions have associated feature distributions.

3.2.2 Region Comparison Techniques

Many histogram similarity measures have been proposed. These include inter-bin measures such as the Bhattacharyya, Jeffrey, Minkowski, Intersection, χ^2 , and Kullback Leibler, and intra-bin measures such as QBIC and the Earth Movers distance, see e.g. [40]. Inter-bin measures are favoured here because of their lower computational cost. Sum of histogram back-projections, which is quicker to calculate online, can also be used but allows less discriminatory power since it uses each of the measurements independently, ignoring how these are distributed.

3.2.3 Background Model

The distribution induced by the likelihood model described so far cannot be used to disambiguate certain poses. For example, consider the waving sequence, where the lower arm, which is uniformly coloured, rotates in depth. As the arm foreshortens the hypothesised histograms will remain approximately constant and therefore so will the likelihood. This problem is illustrated in Fig. 5. To overcome this problem multiple solutions could be propagated using a semi-parametric or non-parametric density representation [12,41]. However, this approach is particularly expensive when propagating an appearance estimate and only delays decision making to a higher-level stage. For these reasons the approach taken in this work was to condition the likelihood to maximise the foreground usage as determined by a statistical background model. The background is modelled using a multi-variate Gaussian in chromaticity-intensity space for each pixel i . These Gaussian densities are recursively updated as described by McKenna *et al.* [42]. The resulting modified likelihood

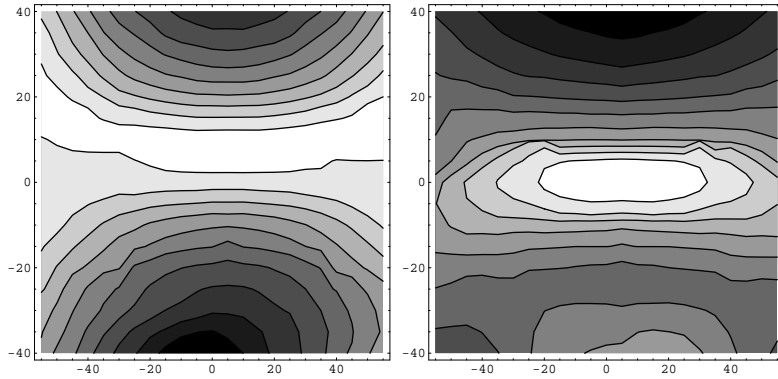


Fig. 5. Visualisation of the likelihood whilst rotating the lower right model arm against frame 10 of the waving sequence. The solution is centralised. Abscissa: out of plane rotation, ordinate: in plane rotation. The central ridge in the first plot has a large likelihood and illustrates the inability of the model to resolve out of plane rotations. The second plot illustrates how conditioning the likelihood to maximise foreground usage results in a single solution.

is illustrated in Fig. 5 and is defined as:

$$p(y_t|x_t) = p_{dynamic}p_{background} \quad (10)$$

$$p_{background}(y_t|x_t) = \frac{\sum_{i \in V} p_f(i)}{\sum_{i \in I} p_f(i)} \quad (11)$$

where $p_f(i)$ denotes the foreground probability density at pixel i , I is the set of all pixels in the image, and V is the set of pixels corresponding to the body. The effects of shadow could be reduced by using a model similar to McKenna *et al* [42].

3.2.4 Appearance Update

Fig. 6 shows a cropped region from the back-projection of the arm histogram onto two frames, two seconds apart. It can be seen that the foreground appearance changes over time, sometimes quite quickly. The histograms are recursively updated to account for such changes using Equation (12). The rate of adaptation is controlled by a constant, c . To reduce the chance of the tracker drifting the appearance histograms are updated using only those pixels that are sufficiently different from the background. Specifically, the histogram H_t used for estimation at time t is formed from a weighted average of the histogram at time $t-1$ and a histogram, H'_t , formed from those pixels on the best pose estimate that are different from the background (based upon a threshold).

$$H_t = cH'_t + (1 - c)H_{t-1} \quad (12)$$

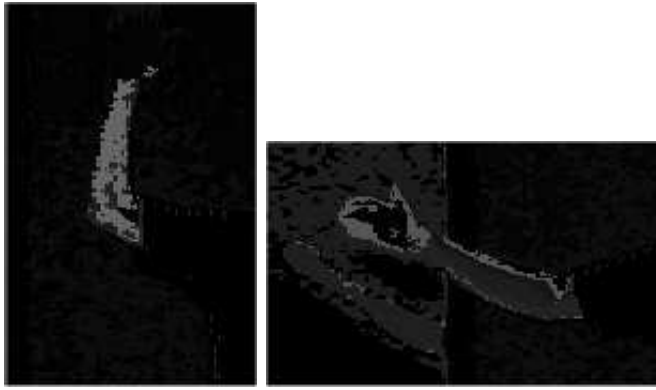


Fig. 6. Probability map for a lower arm histogram for frames from the waving sequence two seconds apart. It can be seen that the distribution has changed. The background has also changed.

3.3 Inference Method

The focus of this paper is not a new inference technique. However, the advantages of the described dynamic likelihood over existing likelihoods, such as its broad, smooth likelihood profile, can allow for easier and more accurate inference. In the implementation described here, an iterative maximum likelihood optimization scheme was employed. A key reason for choosing maximum likelihood over a Bayesian filtering scheme, such as particle filtering, is that the enlarged state space requires much more memory.

In the first frame, the pose was manually initialised and initial surface feature distributions extracted. Then for each subsequent frame a two stage search was first performed. Firstly, the state space was coarsely sampled around an estimate given by the constant velocity motion model. The number and spacing of samples was chosen empirically using the likelihood response. For example, in the case of the upper arm with three degrees of freedom, sampling at four half limb widths in all directions at two half limb width intervals requires 64 samples. Then the best results of the hierarchical coarse sampling were used to seed a local gradient-based search. Including the hierarchical sampling reduces the chance of getting trapped in local maxima, thereby allowing larger inter-frame motion. It is particularly useful when self-occlusion of the human body causes gradient information to be lost.

4 Results

As previously mentioned, zeroth-order chromaticity-intensity statistics were used as features. A good histogram size was found empirically to be $12 \times 12 \times 8$ bins. No prior colour information was used. The system was implemented in

C++. Preprocessing to find the histogram bins, an efficient model projection implementation and loop unrolling resulted in efficient likelihood calculations, the main computational burden for most trackers. The system required around $100MB$ to store the appearance model and made of the order of 10,000 samples per frame at around $10ms$ per sample on a $2Ghz$ PC. Whilst the computation time for an edge based likelihood (e.g. [21]) is 1 to 2 orders of magnitude less, many more samples will be needed for reliable tracking and the final result will be less discriminatory, particularly in scenes with large amounts of clutter.

4.1 Likelihood Investigation

Fig. 7 shows different similarity measures as the model upper right arm undergoes image-plane rotation. It can be seen that grouping has a large effect on the profile of the response. A large amount of grouping causes local detail to be lost and localization suffers. In the case of too little grouping, the histograms are poorly estimated and the response is less smooth and has significant secondary maxima. It can also be seen that some similarity measures produce smoother responses and are less sensitive to the amount of grouping. In particular, the Bhattacharyya coefficient was found to work well for tracking and grouping. The back-projection worked well when all background pixels were sufficiently different from the foreground.

4.2 Grouping Results

Fig. 8 illustrates the result of applying the grouping scheme to the first frame in the waving sequence. The number of unique regions is plotted against time. It can be seen that the system quickly converges to a stable region representation. It can be seen that occasionally split and merges are performed after the system has reached a steady state, this is primarily due to new regions becoming visible.

4.3 Tracking Results

In this section we present the results of tracking for qualitative evaluation. Comparing the results to ground truth is deferred to future work. Fig. 9 shows the result from successfully tracking in an everyday indoor scene. The subject is wearing loose-fitting clothes with both textured and plain regions. The background contains a significant amount of clutter, similarly coloured objects and uneven, natural lighting. The sequence was captured at 12 frames per second

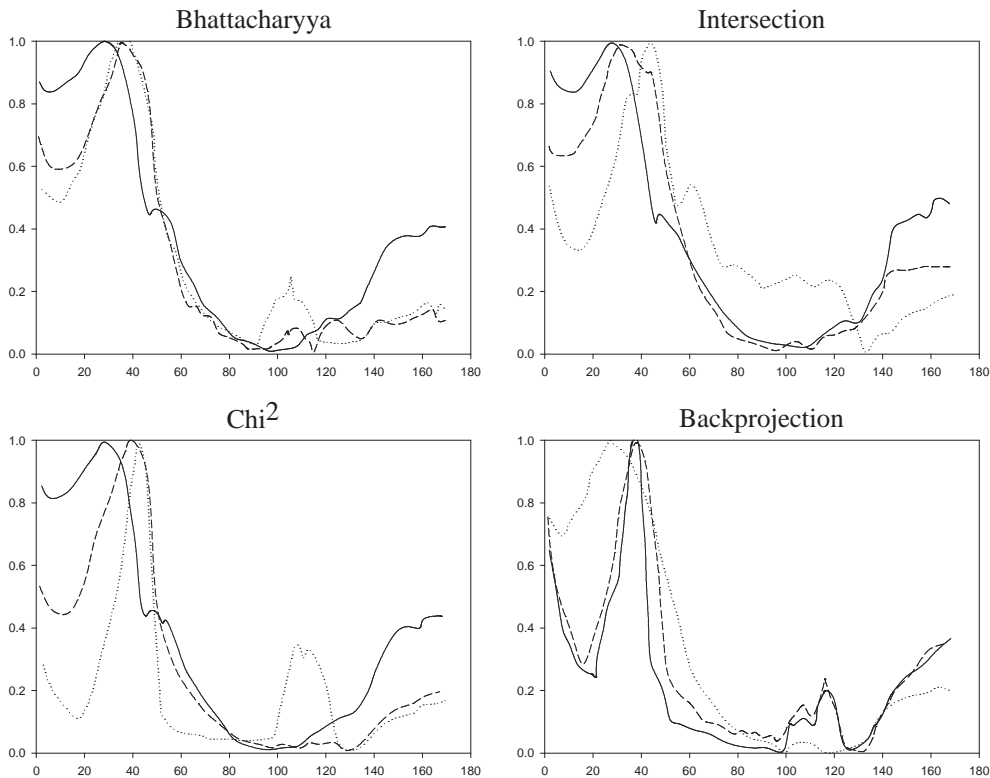


Fig. 7. Investigating the effects of region grouping and different similarity measures. Plots show the dynamic likelihood, $p_{dynamic}$ (scaled to the unit interval), versus upper arm rotation (degrees) for frame 10 of the waving sequence for three levels of grouping: solid= 5 regions, dashed= 20 regions, dotted= 120 regions. The true rotation angle was 39° .

and at a resolution of 640×480 pixels. The frame-rate was lower than is usual, making tracking more difficult. The sequence contains 72 frames (6 seconds) which compares favourably to the length of sequences used in related published results. The update constant c in Equation (12) was set to 0.2. The reader is referred to the full video sequence at www.computing.dundee.ac.uk/staff/troberts/.

It can be seen that the tracker maintains lock throughout the sequence including during the frequent self-occlusions. The hierarchical sampling along with a low frame rate makes the result a little jumpy. Furthermore, in some frames the upper arm is localised only approximately due to constraints on the model deformations. A strength of this region-based formulation is that the tracker degrades gracefully under such conditions. The most significant error is that it switches the legs half-way through the sequence. This is due to the low frame rate and highly symmetric appearance and could be overcome by using a better motion model. In the final frames, tracking of a foot is inaccurate because of the heavy shadowing and the similarity of the background and foreground distribution. The effects of shadowing have not been investigated further. The tracker is prone to error when applied to scenes with complex

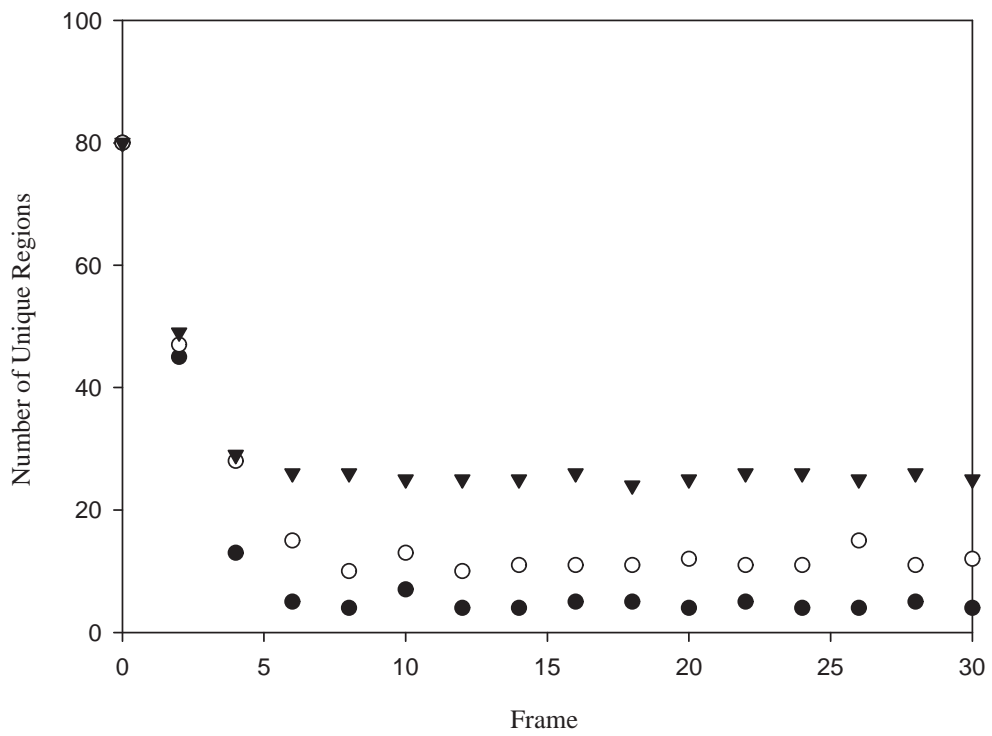
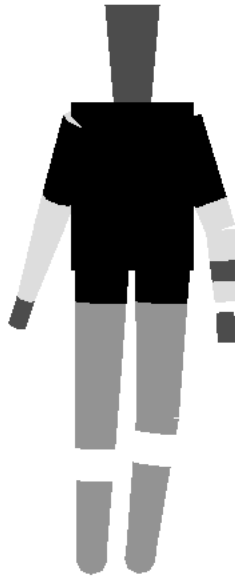


Fig. 8. Results from applying region merging to the first frame of the waving sequence. Top: visualisation of the largest grouped regions. Bottom: plot showing the behaviour of the grouping algorithm for three different merge thresholds.



Fig. 9. Tracking a highly textured subject through a walking cycle containing self-occlusion and motion blur in a cluttered indoor scene at low frame rate.

motions or large appearance changes. The use of a motion prior and a static likelihood cue would reduce such errors.

5 Conclusions and Future Work

A likelihood formulation was presented to allow for detailed, accurate pose estimation in unknown scenes. The model was based upon estimating the feature statistics of regions on the surface of a 3D articulated body model. Two problems with this approach are computational efficiency (in terms of both memory and computation time) and density estimation. A region grouping algorithm was presented to overcome these difficulties and its benefits were illustrated.

The tracker worked well in some real world scenes of moderate complexity and, due to the properties of the likelihood model, inference was efficient and straightforward. However, in scenes with more complex motions or significant appearance changes the tracker would be expected to diverge. We believe that combining dynamic and static cues and incorporating a pose prior and motion model would significantly reduce this problem and is the most pressing issue to address in future work. In this regard the method should be seen as a first step towards addressing how to incorporate a model of estimated appearance into a human tracking system. In addition several other possible directions for future research can be identified:

Shape Model The body model used is somewhat restrictive and sometimes only allows approximate registration. This can be seen when the shoulder moves relative to the torso, for example. Spring-type joints could be used to improve this situation. However, more detailed modelling increases the dimensionality of the search space. It would also be interesting to model other object occlusion to allow for tracking in more realistic environments.

Likelihood Model The method proposed here is extensible to other feature statistics such as texture descriptors. We plan to investigate their use in this context. The model should also be combined with a static likelihood component to allow for automatic (re)initialisation.

Inference We plan to use the feature histograms to construct a body part importance sampling function which will allow for a greater range of movements and recovery from error. A method for propagating multiple state estimates (particles) when the state incorporates a dynamic appearance component is also needed.

Acknowledgments

T. J. Roberts was supported by an EPSRC studentship.

References

- [1] J. Deutscher, A. Blake, I. Reid, Articulated body motion capture by annealed particle filtering, in: *IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2, South Carolina, USA, 2000, pp. 126–133.
- [2] T. B. Moeslund, E. Granum, A survey of computer vision-based human motion capture, *Computer Vision and Image Understanding* 81 (3) (2001) 231–268.
- [3] T. Collins, A. Lipton, T. Kanade, Special section on video surveillance, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (8) (2000) 745–746.
- [4] N. Jovic, J. Gu, H. C. Shen, T. S. Huang, Computer modeling, analysis, and synthesis of dressed humans, in: *IEEE Conference on Computer Vision and Pattern Recognition*, Santa Barbara, CA, 1998, pp. 528–534.
- [5] D. M. Gavrilu, The visual analysis of human movement: A survey, *Computer Vision and Image Understanding* 73 (1) (1999) 82–98.
- [6] J. K. Aggarwal, Q. Cai, Human motion analysis: A review, *Computer Vision and Image Understanding* 73 (3) (1999) 428–440.
- [7] T. B. Moeslund, F. Bajers, Summaries of 107 computer vision-based human motion capture papers, Tech. Rep. LIA99-01, University of Aalborg (1999).
- [8] A. Baumberg, D. Hogg, Learning deformable models for tracking the human body, in: M. Shah, R. Jain (Eds.), *Motion-based Recognition*, Kluwer, 1997, pp. 39–60.
- [9] R. Bowden, T. A. Mitchell, M. Sarhadi, Non-linear statistical models for the 3D reconstruction of human pose and motion from monocular image sequences, *Image and Vision Computing* 18 (9) (2000) 729–737.
- [10] S. Gong, S. J. McKenna, A. Psarrou, *Dynamic Vision: From Images to Face Recognition*, Imperial College Press, 2000.
- [11] S. Ju, M. Black, Y. Yacoob, Cardboard people: a parameterized model of articulated image motion, in: *IEEE International Conference on Face and Gesture Recognition*, Killington, VT, USA, 1996, pp. 38–44.
- [12] T. J. Cham, J. M. Rehg, A multiple hypothesis approach to figure tracking, in: *IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2, Fort Collins, Colorado, USA, 1999, pp. 239–245.

- [13] D. Marr, K. H. Nishihara, Representation and recognition of the spatial organization of three dimensional structure, *Proceedings of the Royal Society of London* 200 (1978) 269–294.
- [14] D. Hogg, Model-based vision: A program to see a walking person, *Image and Vision Computing* 1 (1) (1983) 5–20.
- [15] I. A. Kakadiaris, D. Metaxas, Model-based estimation of 3D human motion with occlusion based on active multi-viewpoint selection, in: *IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, USA, 1996, pp. 81–87.
- [16] C. Bregler, J. Malik, Tracking people with twists and exponential maps, in: *IEEE Conference on Computer Vision and Pattern Recognition*, Santa Barbara, CA, 1998, pp. 8–15.
- [17] C. Sminchisescu, B. Triggs, Covariance scaled sampling for monocular 3D body tracking, in: *IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1, Hawaii, 2001, pp. 447–454.
- [18] I. Karaulova, P. Hall, A. Marshall, A hierarchical model of dynamics for tracking people with a single video camera, in: *British Machine Vision Conference*, Bristol, 2000, pp. 352–361.
- [19] R. Plankers, Human body modelling from image sequences, Ph.D. thesis, EPFL, Switzerland (2001).
- [20] D. M. Gavrila, L. S. Davis, 3D model-based tracking of humans in action: A multi-view approach, in: *IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, USA, 1996, pp. 73–80.
- [21] S. Wachter, H. H. Nagel, Tracking persons in monocular image sequences, *Computer Vision and Image Understanding* 74 (3) (1999) 174–192.
- [22] R. Ronfard, C. Schud, B. Triggs, Learning to parse pictures of people, in: *European Conference on Computer Vision*, Copenhagen, 2002, pp. 700–714.
- [23] A. Elgammal, R. Duraiswami, L. Davis, Efficient non-parametric adaptive color modeling using fast gauss transform, in: *IEEE Conference on Computer Vision and Pattern Recognition*, 2001, pp. 563–570.
- [24] D. Comaniciu, V. Ramesh, P. Meer, Real-time tracking of nonrigid objects using mean shift, *IEEE Conference on Computer Vision and Pattern Recognition* (2000) 673–678.
- [25] S. J. McKenna, Y. Raja, S. Gong, Tracking colour objects using adaptive mixture models, *Image and Vision Computing* 17 (3-4) (1999) 225–231.
- [26] C. R. Wren, A. Azarbayejani, T. J. Darrell, A. Pentland, Pfunder: Real-time tracking of the human body, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19 (7) (1997) 780–785.

- [27] D. A. Forsyth, M. M. Fleck, Body plans, in: IEEE Conference on Computer Vision and Pattern Recognition, Puerto Rico, 1997, pp. 678–683.
- [28] H. Sidenbladh, F. de la Torre, M. J. Black, A framework for modeling the appearance of 3D articulated figures, in: IEEE International Conference on Face and Gesture Recognition, Grenoble, 2000, pp. 368–375.
- [29] C. Cedras, M. Shah, Motion-based recognition: A survey, *Image and Vision Computing* 13 (2) (1995) 129–155.
- [30] M. J. Black, Y. Yacoob, S. Ju, Recognizing human motion using parameterized models of optical flow, in: M. Shah, R. Jain (Eds.), *Motion-based Recognition*, Kluwer, 1997, pp. 245–269.
- [31] H. Sidenbladh, M. J. Black, D. J. Fleet, Stochastic tracking of 3D human figures using 2D image motion, in: *European Conference on Computer Vision*, Dublin, 2000, pp. 702–718.
- [32] Y. Song, X. Feng, P. Perona, Towards detection of human motion, in: *IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1, Hawaii, 2000, pp. 810–817.
- [33] T. B. Moeslund, E. Granum, Multiple cues for model-based human motion capture, in: *IEEE International Conference on Face and Gesture Recognition*, Grenoble, 2000, pp. 362–367.
- [34] T. Zhao, T. Wang, H. Shum, Learning a highly structured motion model for 3D human tracking, in: *Asian Conference on Computer Vision*, Melbourne, 2002, p. 144149.
- [35] C. Sminchisescu, Consistency and coupling in human model likelihoods, in: *IEEE International Conference on Face and Gesture Recognition*, Washington, 2002, pp. 27–32.
- [36] T. Darrell, G. G. Gordon, M. Harville, J. Woodfill, Integrated person tracking using stereo, color, and pattern detection, *International Journal of Computer Vision* 37 (2) (2000) 175–185.
- [37] R. Okada, Y. Shirai, J. Miura, Tracking a person with 3D motion by integrating optical flow and depth, in: *IEEE International Conference on Face and Gesture Recognition*, Grenoble, 2000, pp. 336–341.
- [38] H. Sidenbladh, M. J. Black, Learning image statistics for Bayesian tracking, in: *IEEE International Conference on Computer Vision*, Vol. 2, Vancouver, 2001, pp. 709–716.
- [39] J. Park, O. Hwang-Seok, D. Chang, E. Lee, Human posture recognition using curve segments for image retrieval, in: *SPIE Conference on Storage and Retrieval for Media Databases*, Vol. 3972, 2000, pp. 2–11.
- [40] J. Puzicha, Y. Rubner, C. Tomasi, J. M. Buhmann, Empirical evaluation of dissimilarity measures for color and texture, *IEEE International Conference on Computer Vision* (1999) 1165–1173.

- [41] M. Isard, A. Blake, Contour tracking by stochastic propagation of conditional density, in: European Conference on Computer Vision, Vol. 1, Cambridge, 1996, pp. 343–356.
- [42] S. J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, H. Wechsler, Tracking groups of people, *Computer Vision and Image Understanding* 80 (1) (2000) 42–56.