Adaptive Learning of Statistical Appearance Models for 3D Human Tracking

Timothy J. Roberts, Stephen J. McKenna, Ian W. Ricketts
Department of Applied Computing
University of Dundee, Scotland, DD1 4HN
troberts@computing.dundee.ac.uk

Abstract

A likelihood formulation for human tracking is presented based upon matching feature statistics on the surface of an articulated 3D body model. A benefit of such a formulation over current techniques is that it provides a dense, object-based cue. Multi-dimensional histograms are used to represent feature distributions and different histogram similarity measures are evaluated. An on-line region grouping algorithm, driven by prior knowledge of clothing structure, is derived that enables better histogram estimation and greatly increases computational efficiency. Finally, we demonstrate that the smooth, broad likelihood response allows efficient inference using coarse sampling and local optimisation. Results from tracking real world sequences are presented.

1 Introduction

Human tracking is a difficult and interesting problem and the recent surge in research interest is due to both the solutions' possible applications and the problems' challenging nature. Tracking people is difficult not least because of their complex appearance. This paper presents an appearance model that brings tracking in poorly constrained, real world scenes a step closer.

Much human tracking research adopts a state-based, probabilistic, Bayesian, analysis-by-synthesis framework. Furthermore, much of the research progress can be broadly categorised into either modelling advancements or inference advancements. We briefly discuss this framework and the state of the art in each area to provide a background to illustrate where our method contributes.

The aim of the tracking system is to find the state of the target $X_t = \{x_0, ..., x_t\}$, given the observations $Y_t = \{y_0, ..., y_t\}$ and a body of prior knowledge $p(X_t)$ (where $x \in \Re^n$ and $t \in [0, T]$). However, due to noise, model inaccuracies and loss of information there are usually genuine ambiguities and we must represent our knowledge using a belief distribution $p(X_t|Y_t)$ instead. In general, distributions at particular times can be found using Bayesian filtering:

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

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

The first term on the right hand side of Equation (2) represents the likelihood model and the second represents a prior over paths through state space. It is important to realise that the posterior probability distribution is induced by the chosen likelihood and prior. Accurate modelling of these terms allows for easier and more accurate estimation. For the case of modelling for human tracking there are many difficulties: the clothing can be loose, textured, changing over time and not known *a priori* and the inter-frame motion can be large and varied. The topic of this paper, which builds upon our previous work [9], is the derivation and evaluation of a computationally feasible likelihood model that allows tracking in real world environments. Many state-of-the-art systems rely upon likelihood models that assume restrictive scene conditions such as tight, high contrast, un-textured clothing or a static, known background. The system presented here is less restrictive in that it copes with more normal, textured and loose clothing.

There has been some previous work on learning detailed models of human appearance. Sidenbladh and Black [10] learned the distribution of edge, ridge and optical flow responses from images. They observed that edge and ridge cues provide sparse information about limb appearance. Sidenbladh *et al.* [11] used a modified PCA algorithm to account for viewpoint to find the principal components of texture and use these in a generative tracking framework. However, these systems had the disadvantage of requiring the appearance to be learned off-line.

Once a model is established, an algorithm is required to estimate the posterior distribution. Estimation in the case of human tracking is difficult because of the high dimensionality, occlusion, clutter, loss of depth and the nature of the kinematic structure of the body. These often result in a non-Gaussian, multi-modal distribution for which there is no analytic solution to the filtering problem [1, 5, 13]. Therefore, there has been much interest recently in sophisticated particle filtering techniques that can represent and efficiently infer the posterior distribution, including particle annealing [3], genetic cross-over and state space partitioning [4], local optimisation [2, 12] and lattice sampling [7]. A further benefit of the formulation presented in this paper is that estimation is eased due to the smooth, broad properties of the likelihood response.

2 Method

To begin we ask the question: what are the properties of a good likelihood formulation $p(Y_t|X_t)$? A good generative model will potentially utilise all the input data and will be able to accurately re-synthesise an appropriate representation of the input data given the solution. Furthermore, a good model should have characteristics which allow for easy and accurate estimation, such as a strong, broad response around the solution and a large discriminatory power to reduce secondary maxima. However, the dimensionality of the integral in the Bayesian filtering equation (2) grows with time as we consider more and more information and direct evaluation becomes prohibitively complex. Therefore, it is

usual to consider the state evolution to be a Markov process and the distributions can then be found recursively using:

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

A difficulty in the case of visual human tracking, where y_t is an image, is that, due to the wide variety in a subject's appearance, the single frame likelihood model, $p(y_t|x_t)$, is not known *a priori*. Such models may be constructed by making assumptions on the scene conditions but the assumptions do not allow for accurate estimation in other environments. Therefore, we conclude that in real world scenarios, where such conditions are not known, accurate inference, under the Markov assumption, requires the state, x, to include appearance information. This is the basic idea surrounding on-line, adaptive background models, such as in [6].

The central problem is that of estimating a complex changing appearance given the limited data available. The problem is hopelessly under constrained without prior knowledge. Furthermore, there is the question of computational efficiency. In the remainder of this paper we present and evaluate a scheme to model the appearance of humans using a computationally efficient, on-line estimation technique. The synthesis of views requires knowledge of both the body's structure and surface properties and therefore the state description is partitioned into geometric and appearance components, $x = (x_g, x_a)$.

2.1 Geometric Model

The body is highly deformable and exact modelling of its form is infeasible in this context. Its important properties can be captured using an articulated body model. In this system a 3D geometric model is used since the body's surface can, in general, vary with viewpoint. A 3D articulated body model is a collection of 3D geometric primitives connected in a hierarchical fashion to form a kinematic tree. The state space, x_g, then becomes the relative position and orientation of the primitives and their shapes and sizes. The advantages of this description are that it has a lower dimensionality, is computationally efficient, captures the kinematic structure of the body, allows for easy encoding of prior knowledge such as joint limits and automatically handles self occlusion. For a discussion of more advanced geometric models see, for example, [8].

Each of the geometric primitives, indexed by b, has a surface that is naturally described using some co-ordinate system, denoted in this paper by ω_b. For example, the surface of a cylinder is conveniently described by a length and angle. A point on the subject is then specified by the pair (b, ω_b). To project a surface point into the image plane the co-ordinates are converted to Cartesian form. Homogeneous, relative transformations are chained together to project up the kinematic tree into world co-ordinates and finally, using a camera model, into the image plane.

In the current implementation we represent the body using truncated, elliptic cross-section cylinders with constant, known size and shape. The camera is modelled using an orthographic projection since the sequences under consideration do not contain strong perspective effects. However, the extension to perspective projection is straightforward. The system currently allows no independent head, hand or foot motion leaving a total of 22 degrees of freedom, encoded as 3 root translations and 19 Euler angles (four for each

limb and three for the trunk). Prior knowledge on joint angles is encoded using a ramp function at boundaries.

2.2 Appearance Model

Due to body model inaccuracies, discretisation and noise, a feature such as colour or a local filter response, at a point on the surface of a body model will have a probability distribution rather than a single value. The appearance component of the state is the set of feature distributions for the whole body, or equivalently over all points, Ω_b, on each body part b:

$$x_a = \{p_{(b,\omega_b)} : 0 < b \leq B, \omega_b \in \Omega_b\}. \tag{4}$$

In this paper we consider colour distributions only. Since clothing is also often textured these distributions can be multi-modal. We therefore propose using normalised multi-dimensional histograms to represent these distributions and denote them by $H_{(b,\omega_b)}$. In general these distributions are not known *a priori*. Some distributions, such as skin, can be estimated off-line and this framework also allows us to incorporate such prior knowledge in a principled fashion. This helps with automating initialisation, for example.

Estimating the feature distributions from the limited data available is difficult, especially when it is considered that 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. We can therefore use the histograms from other points on the body to estimate an unknown bin q:

$$H_{(b,\omega_b)}(q) \approx \sum_{b'} \sum_{\omega'_b} H_{(b',\omega'_b)}(q) p(H_{(b',\omega'_b)} | H_{(b,\omega_b)}) \tag{5}$$

The conditional probability can be modelled in a Bayesian fashion using a likelihood determined from a similarity measure, S, on the known histogram bins and a prior determined from knowledge of clothing structure:

$$p(H_{(b',\omega')} | H_{(b,\omega_b)}) \approx S(H_{(b',\omega')}, H_{(b,\omega_b)}) P_{(b,\omega),(b',\omega')} \tag{6}$$

However, direct use of the sum in (5) is not computationally feasible since it involves summing over all points on the body for each unestimated histogram bin. Therefore, we propose to group regions based upon the observation that large contributions to the sum must be similar to the histogram in question and therefore similar to each other. Therefore, the sum is reasonably well approximated by the average bin value taken from the group of similar regions. Figure 1 illustrates our model.

To perform region merging, a threshold K is introduced. It controls the level of detail represented by the system and elegantly 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 resolution, there will be an optimal choice of threshold that allows the appearance to be well estimated without excessive loss of local structure. The merging decision criterion then becomes:

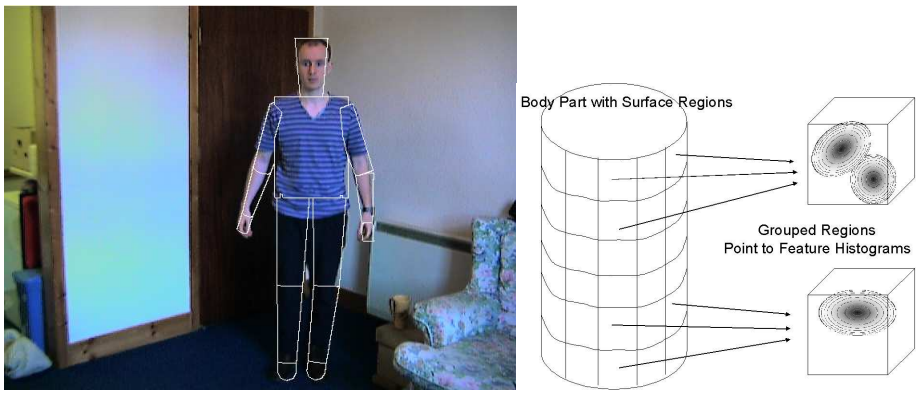

Figure 1: An articulated body model with feature distributions defined over surface patches. These patches are grouped based upon similarity and knowledge of clothing structure. The model is overlayed on a frame from a waving gesture sequence used throughout the paper to illustrate ideas.

$$S(H_{b,\omega_b}, H_{b',\omega'_b}) > \frac{K}{P_{(b,\omega_b),(b',\omega'_b)}} \tag{7}$$

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. Since regions can erroneously merge we also introduce a splitting operation. However, performing this in the same manner as merging requires unique histograms to be stored for every atomic region, resulting in a large storage overhead. Therefore, we currently use an *ad hoc* splitting criterion based upon a threshold on the sum of back-projections in an atomic region from the current image to split.

The clothing structure prior is learned from example images of differently clothed people by manually aligning the model to the image and performing exhaustive pairwise comparisons. The prior is set to the average of the observed similarities. Examples are shown in Table 1.

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

2.3 Likelihood Formulation

To compare a hypothesised geometric configuration to the image, hypothesised feature histograms, H', are collected by casting a ray at each pixel into the world and determining

its point of intersection with the hypothesised body model. The hypothesised histograms are compared to the model histograms using a similarity measure. The likelihood as defined in Equation (8) is 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.

$$p(y_t|x_t) \propto \frac{\sum_{\{V\}} S(H'_{(b,\omega)}|H_{(b,\omega)})}{|V|} \tag{8}$$

2.3.1 Region Comparison Techniques

There are many histogram similarity measures, these include inter-bin measures such as the Bhattacharyya coefficient, the Jeffrey distance, the Minkowski distance, Intersection, χ^2, and the Kullback Leibler divergence and intra-bin measures such as QBIC and the Earth movers distance. Inter-bin measures are favoured here because of their lower computational cost. Sum of histogram back-projections, which is much quicker to calculate online, can also be used but allows less discriminatory power since it uses the measurements independently and ignores how these might be distributed.

Figure 2 shows different similarity measures as the model upper right arm undergoes image-plane rotation. It can be seen that some similarity measures produce smoother responses and are less sensitive to the amount of grouping. A lower number of regions tends to result in a smoother likelihood response. We found that tracking and grouping using the Bhattacharyya coefficient worked best. The back-projection works well when every background pixel is sufficiently different from the foreground.

The posterior distribution induced by this likelihood model is multi-modal and cannot be used to disambiguate certain poses. For example, consider the waving sequence where the lower arm, which is uniformly coloured, rotates in depth. The hypothesised histograms will remain approximately constant and therefore so will the likelihood. In addition, regions that are hidden do not contribute. To overcome this problem multiple solutions should be propagated. However, due to the difficulties of this approach with an adaptive appearance model we choose to condition the likelihood to maximise foreground usage, this is illustrated in Figure 3. The background is modelled pixel-wise using Gaussians in chromaticity-intensity space which are recursively updated using the equations from [6]. We stress that the likelihood formulation does not require a static background, rather a more advanced representation of the posterior distribution is required.

2.4 Inference

The focus of this paper is not a new inference technique. However, the advantages of this formulation over existing ones, such as its broad, smooth likelihood profile, can allow for easier and more accurate inference. In the first frame the geometric model is manually initialised. A hierarchical, best-first search is then performed by coarsely sampling the state space around an estimate given by a constant velocity motion model. The number and spacing of samples is chosen empirically using the similarity responses. For example, in the case of the upper arm with three degrees of freedom, sampling up to four half-widths in all directions at two half-width intervals requires 64 samples. This estimate is then used to seed a multi-dimensional gradient-based search and this has the effect of reducing the chance of getting trapped in local maxima and therefore allows more general motion. Hierarchical sampling is particularly useful in the case of the human

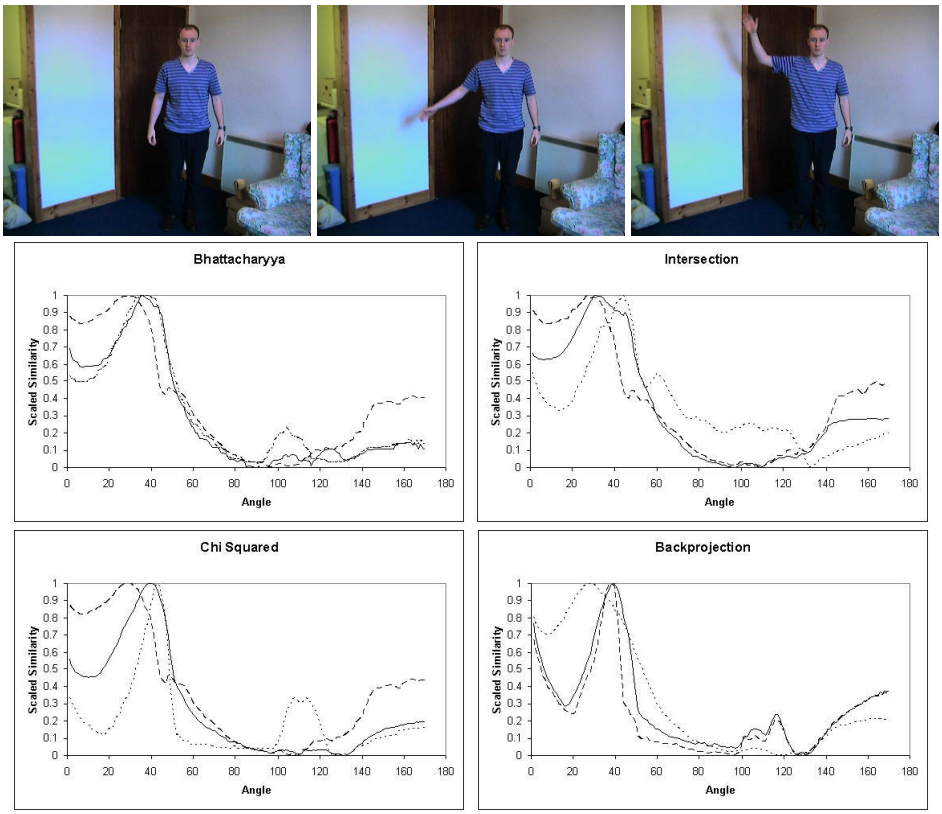

Figure 2: Plots of different similarity measures vs. upper arm rotation for three levels of grouping: dashed= 5 regions, solid= 20 regions, dotted= 120 regions.

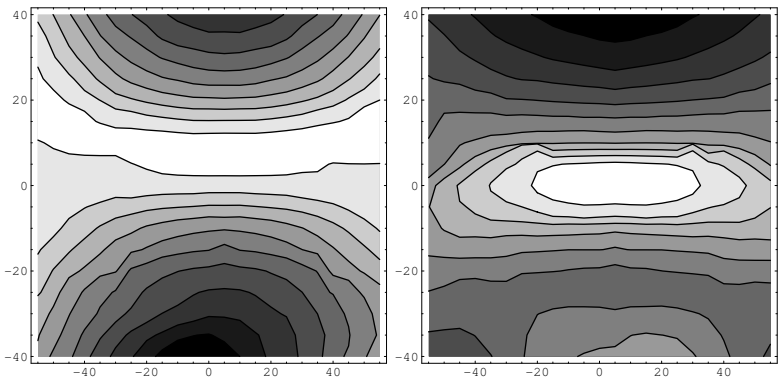

Figure 3: Abscissa: out of plane rotation, ordinate: in plane rotations. The central uniform ridge in the first plot has a large likelihood and illustrates the inability of the model to resolve out of plane rotations. The second plot illustrates how conditioning the likelihood to maximise foreground usage results in a single solution.

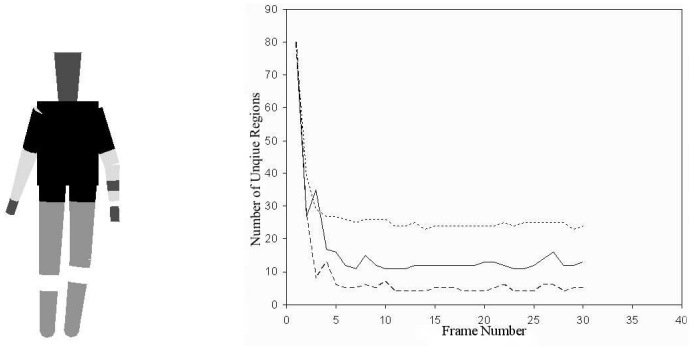

Figure 4: Results from region merging. Left: examples of the largest grouped regions. Right: plot showing the behaviour of the grouping algorithm for three different merge thresholds.

body where occlusion causes gradient information to be lost. Once the maximum is found the appearance model is updated recursively as in Equation (9). To make the tracker more robust and reduce the chance of the tracker diverging we update the appearance model using only those pixels that are sufficiently different from the background.

$$H_t = kH'_t + (1-k)H_{t-1} \tag{9}$$

3 Implementation and Results

Currently the system uses zeroth-order chromaticity-intensity statistics and histogram sizes are chosen using heuristics based upon the number of expected samples. The system has prior information on clothing structure, but no colour prior.

The system is implemented in C++. Use of caching techniques, preprocessing of histogram bins, efficient model projection and extensive loop unrolling result in very efficient likelihood calculations, the main computational burden for most trackers. The system requires around 100MB to store the appearance model and processes each frame in around 10 seconds.

Figure 4 shows how the region grouping algorithm behaves for the waving sequence. It can be seen that the system quickly converges to a stable region representation. Figure 5 shows the system successfully tracking in a cluttered indoor scene. The subject is wearing loose-fitting clothes with both textured and plain regions. The background contains many edges, and similarly coloured objects. The reader is referred to the video sequence at www.computing.dundee.ac.uk/staff/troberts/. The sequence was captured at 12 frames per second, at a resolution of 640×480. This is a smaller than usual frame-rate making tracking more difficult. In this sequence the appearance is stored in $12 \times 12 \times 6$ bin histograms.

Figure 5: Tracking a highly textured subject in a cluttered indoor scene.

4 Conclusions and Future Work

A likelihood formulation was presented that models the region statistics on the surface of an articulated body model to allow for detailed, accurate pose estimation in unknown scenes . Two problems with this approach are histogram density estimation and computational efficiency. An adaptive region grouping algorithm was derived to overcome these difficulties and its benefits where illustrated.

The method is very extensible. In the immediate future we will be testing different feature statistics. We also plan to use the feature histograms to construct an importance sampling function which we believe will allow for a much greater range of movements and recovery from error. Another possible inference extension is to build a motion model using PCA and coarsely sample along the eigenvectors. We plan to address the current problem of propagating multiple states and to use a more principled histogram update technique. In the more distant future we will investigate switching between different features online to improve tracker performance.

References

[1] T. J. Cham and J. M. Rehg. A multiple hypothesis approach to figure tracking. In *Computer Vision and Pattern Recognition*, volume 2, pages 239–245, 1999.

[2] K. Choo and D. J. Fleet. People tracking using hybrid Monte Carlo filtering. In *International Conference on Computer Vision*, pages 321–328, 2001.

[3] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *Computer Vision and Pattern Recognition*, volume 2, pages 126–133, 2000.

[4] J. Deutscher, A. Davison, and I. Reid. Automatic partitioning of high dimensional search spaces associated with articulated body motion capture. In *Computer Vision and Pattern Recognition*, volume 2, pages 669–676, 2001.

[5] J. Deutscher, B. North, B. Bascle, and A. Blake. Tracking through singularities and discontinuities by random sampling. In *International Conference on Computer Vision*, pages 1144–1149, September 1999.

[6] S. J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler. Tracking groups of people. *Computer Vision and Image Understanding*, 80(1):42–56, October 2000.

[7] D. Ormoneit, C. Lemieux, and D. J. Fleet. Lattice particle filters. *UAI*, 2001.

[8] R. Plankers. *Human Body Modelling From Image Sequences*. PhD thesis, EPFL, Switzerland, 2001.

[9] T. J. Roberts, S. J. McKenna, and I. W. Ricketts. Online appearance learning for 3D articulated human tracking. In *International Conference on Pattern Recogntion*, 2002.

[10] H. Sidenbladh and M. J. Black. Learning image statistics for Bayesian tracking. In *International Conference on Computer Vision*, volume 2, pages 709–716, 2001.

[11] H. Sidenbladh, F. de la Torre, and M. J. Black. A framework for modeling the appearance of 3D articulated figures. In *Automatic Face and Gesture Recognition*, pages 368–375, 2000.

[12] C. Sminchisescu and B. Triggs. Covariance scaled sampling for monocular 3D body tracking. In *Computer Vision and Pattern Recognition*, pages I:447–454, 2001.

[13] S. Wachter and H. H. Nagel. Tracking persons in monocular image sequences. *Computer Vision and Image Understanding*, 74(3):174–192, June 1999.