Online Appearance Learning for 3D Articulated 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 human appearance modelling framework where colour distributions are associated with surface regions on an articulated body model is presented. In general, these distributions are unknown, multi-modal and changing in time. We therefore propose using recursively updated histograms to represent them. For a certain pose, a set of histograms may be collected and a likelihood constructed based on the histograms' similarity with the previously learned histograms. To ease histogram estimation and improve computational efficiency, a merging and splitting algorithm is derived which groups surface regions based upon histogram similarity and prior knowledge of clothing layout. An investigation of the behaviour of this likelihood shows it to be broad, smooth and peaked around the correct location, a good candidate for coarse sampling and gradient-based search methods. We show how conditioning the likelihood to maximise foreground usage reduces secondary maxima. Finally, we present results from tracking a challenging sequence.

1. Introduction

Tracking humans using computer vision techniques has drawn much attention recently. Not only does it present a challenging test-bed for evaluating tracking schemes, the resulting applications could revolutionize human-computer interaction. Tracking people is difficult not least because the visual appearance is complex and varies markedly. Human trackers often rely on a constrained appearance that restricts application. Learned appearance models have been investigated, for example by Sidenbladh *et al.* [4, 5]. However, they rely on off-line learning. This work's unique contribution is the specification of a computationally feasible appearance model based upon learning the colour distribution of points on the human body on-line.

2. Method

Using a probabilistic formulation, the tracking problem can be stated generally as that of estimating the probability density $p(\{\vec{\phi}_t\}|\{I_t\})$, where $\{I_t\}$ denotes the image sequence and $\{\vec{\phi}_t\}$ denotes the sequence of required pose parameters for time $t \in [0, T]$. Applying Bayes rule and assuming a Markovian relationship between frames, yields $p(\vec{\phi}_t|\{I_t\}) = p(I_t|\vec{\phi}_t) \int p(\vec{\phi}_t|\vec{\phi}_{t-1}) p(\vec{\phi}_{t-1}|\{I_{t-1}\}) d\vec{\phi}_{t-1}$. Each of the terms on the right-hand side has an intuitive meaning. The first represents the pose likelihood for the current image, the second represents the motion model and the third represents the previous a posteriori distribution. This manipulation effectively allows us to use a pose sampling method to update the previous distribution. Neither the likelihood nor motion terms can be specified analytically and must be modelled. Due to the high dimensionality and limited information available from a monocular view, a successful tracking framework will require strong likelihood and motion models. A recent review by Moeslund *et al.* [3] provides an introduction to human tracking. A likelihood model is now presented based upon the temporal consistency of colour distributions on the surface of an articulated body model.

2.1. Body Model

An object-based approach is used to model the surface and kinematics of the human body in which a set of rigid primitive parts are linked to form an hierarchical articulated structure. The pose parameter, $\vec{\phi}$, then becomes a global transformation and the relative orientation parameters for each primitive. Each body part, indexed by n, has a shape which is naturally described by some co-ordinate system, denoted here by $\vec{\omega}$. A cylinder, for example, is conveniently parameterised by a length and angle. To project points on the surface into the image we convert to Cartesian co-ordinates and chain homogeneous matrix transformations to convert up the body hierarchy, into scene co-ordinates and then into the image plane using a camera matrix.

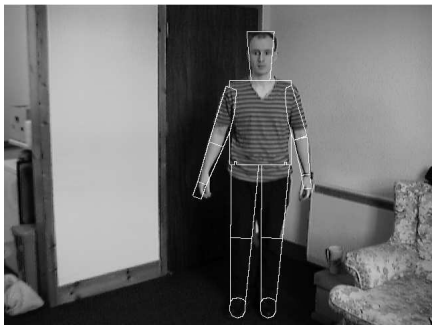

Figure 1. An example frame taken from a waving sequence with the model contour overlayed.

Figure 2. Probability map for a stationary lower arm histogram for images two seconds apart. It can be seen that the distribution changes.

The model presented here uses super-quadrics to represent body parts. The pose vector has twenty-two components: six global transformation parameters and four Euler angles for each limb. The system does not model independent head, hand or foot motion. Currently, the camera is specified using orthographic projection since the scenes contain little perspective effect. The extension to perspective projection in straightforward. Figure 1 provides an example, showing the contour aligned to a frame from a sequence of a waving gesture. Results from processing this sequence are used throughout the report to illustrate ideas.

2.2. Region Features

A point on the surface of the articulated body model is specified by the body part, n, and co-ordinates, $\vec{\omega}$. Due to clothing motion, an inaccurate surface model, illumination changes and noise, the colour, $\vec{q}$, at a surface point must be represented by a distribution. For some regions, such as the face and hands these distributions can be estimated a priori. However, due to the varied nature of clothing and illumination this is not true in general and the distribution must be found on-line. In addition, as Figure 2 shows, this distribution changes over time, sometimes quickly. Since clothing is often textured the distributions can be multi-modal. Two key problems with this approach are density estimation and computational expense. In § 2.5 a histogram merging and splitting algorithm is defined that makes such an approach feasible.

We proceed by associating colour histograms with surface points on the body surface and denote these by $H_{n,\vec{\omega}}$. To model the likelihood of a hypothesized pose, $\vec{\phi'}$, we first project the model into the image. The set of visible pixels, denoted by $\{V\}$, is tagged with the corresponding body part number, n, and co-ordinate, $\vec{\omega}$. Using this set, hypothesized histograms, $H'_{n,\vec{\omega}}$, are built. Each of these hypothesized histograms is then compared to the corresponding learned histogram. Due to its theoretical properties and previous success for tracking colour distributions [1], the Bhattacharyya measure, Equation (1), was used to compare distributions. The likelihood, Equation (2), is formed by averaging this similarity measure over the set of visible pixels.

$$B_{n,\vec{\omega}} = \sum_{q} \sqrt{H_{n,\vec{\omega}}(q) H'_{n,\vec{\omega}}(q)} \tag{1}$$

$$p_R(I_t|\vec{\phi'}) = \frac{\sum_{\{V\}} B_{n,\vec{\omega}}}{|V|} \tag{2}$$

The likelihood response found from varying the lower arm pose in Figure 1 is graphed in Figure 3. From this graph two observations are made. First, the response is smooth and broad. Second, that the response remains constant as the model foreshortens in depth against the background. Without further knowledge of the scene we are not able to uniquely specify the pose. Furthermore, the likelihood has multiple maxima corresponding to occluded parts. Therefore, in this system, rather than propagating multiple solutions, the likelihood is conditioned such that it maximises foreground usage.

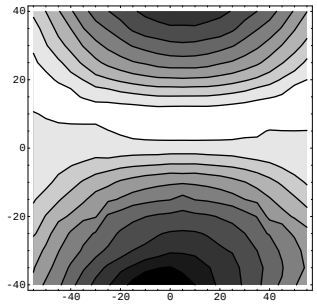

Figure 3. Contour plot of the region likelihood for the lower arm as it makes out-of-plane (abscissa) and in-plane (ordinate) rotations.

2.3. Background Model

A statistical background subtraction scheme is used to condition the likelihood to maximise foreground usage. The background is considered to be changing slowly with respect to the foreground. Background colour is assumed to be normally distributed, with means and diagonal covariance matrices specified for each pixel. To reduce the effect of shadows, colour is specified in chromaticity space $\vec{q} = [I = R+G+B, r = R/I, g = G/I]$ and the variance in the intensity channel is scaled to reduce its effect. The system is initially supplied with mean and covariance matrix estimates and these are recursively updated using the equations used by McKenna *et al.* [2]. The foreground likelihood, Equation (3), is the fraction of foreground usage. The multiple cue likelihood becomes $p_M(I_t|\vec{\phi}) = p_F(I_t|\vec{\phi})p_R(I_t|\vec{\phi})$, and is illustrated in Figure 4.

$$p_F(I|\vec{\phi}) = \frac{\sum_{\{V\}} p_B(I_{(x,y)})}{\sum_{(x,y)} p_B(I_{(x,y)})} \tag{3}$$

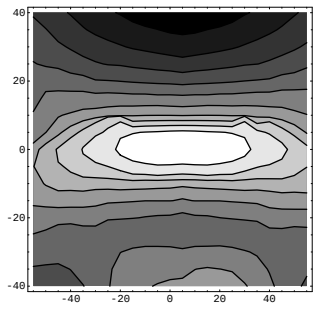

Figure 4. Contour plot of the multiple cue likelihood for the same conditions.

2.4. Histogram Re-estimation

Since the histograms are changing over time, the histograms are recursively updated as in Equation (4). In the event of a tracking error, we can increase the chance of recovery by only updating using pixels that are sufficiently different from the background. In the current system the update rate and update condition thresholds are constant and chosen empirically. They are 0.2 and 0.1 respectively for the sequences shows here.

$$H^t_{n,\vec{\omega}} = (1-k)H^{t-1}_{n,\vec{\omega}} + kH'_{n,\vec{\omega}} \tag{4}$$

2.5. Region Merging and Splitting

Assigned distinct colour distributions to each surface point is unfeasible for two reasons: one theoretical and one practical. Firstly, one cannot reliably estimate the histogram from a small number of samples. Secondly, storing and comparing such a large number of histograms is computationally infeasible.

These problems can be overcome by observing that people dress in a similar, structured manner and that the number of unique pieces of clothing is usually small (of the order of 10 compared to an order of 10,000 visible foreground pixels in the images under consideration). Therefore, we propose that points be grouped together into regions and single histograms associated with them.

To justify our region splitting and merging scheme we begin by considering the problem of estimating an unknown probability for a colour $\vec{q'}$ in a poorly represented histogram, using all the other surface histograms. If we assume that each histogram's contribution is conditionally independent of the others this value can be expressed as a weighted sum over all other histograms.

$$H_{n',\vec{\omega}'}(\vec{q}') = \sum_{n,\vec{\omega}} H_{n,\vec{\omega}}(\vec{q}')B(H_{n,\vec{\omega}}, H_{n',\vec{\omega}'})p_{(n,\vec{\omega}),(n',\vec{\omega}')} \tag{5}$$

The second term, the similarity between the distributions, is the Bhattacharyya measure for the known data. The third term, the prior, encodes the structure of the way people dress and can be learnt off-line.

Direct use of this sum would however, be computationally infeasible. Region merging is an approximation founded on the observation that large contributions to the sum are all similar to the distribution, and hence similar to one another. Equation (6) is a practical pairwise region merging condition based upon a fixed threshold. The threshold, K, can be changed to suit the application and controls how much local detail the system maintains and thus affects its speed. For large values, the system behaves like a template tracker and for small values more like a blob tracker.

$$B(H_{n,\vec{\omega}}, H_{n',\vec{\omega}'}) > \frac{K}{p_{(n,\vec{\omega}),(n',\vec{\omega}')}} \tag{6}$$

Once merged the histogram becomes the sum of its children, either of which may have previously been merged. The prior can either be learnt from a representative training set or modelled. In our experiments we learned a conservative prior, corresponding to the most general appearance by using the minimum observed similarity measure. As one would expect, this prior encodes that the limbs are usually rotationally symmetric and that opposing limbs are also similar. Since the distributions are changing, and regions could erroneously merge, regions can also split. Currently this is done using a threshold on the bin lookups. When split, the distribution is re-initialised.

3. Tracking Results

Using the multiple cue likelihood a walking sequence containing a subject with a highly textured foreground, in a scene with nonuniform lighting and background clutter was tracked. A constant velocity motion model was used to provide a starting point for a gradient search of the likelihood. Due to the high dimensionality the search space was decomposed hierarchically into torso and head, and individual limbs. The system quickly converges to representing five distinct colour distributions. It can be seen that the tracking is successful, although the alignment is sometimes approximate. This is due to the limited tracking scheme that was employed. The implementation was coded in C++ and takes approximately 10 seconds to process one frame. The sequence can be found on-line at www.computing.dundee.ac.uk/staff/troberts.

4. Conclusions and Future Work

A colour-based appearance modelling framework was presented. A likelihood model was constructed based upon the similarity of colour distributions on the surface of an articulated model. This likelihood response was shown to be strong, smooth and broad. A region merging and splitting scheme was presented which makes such an approach feasible. We also showed how to use foreground usage to condition the likelihood.

Three key areas will be explored in future work. Firstly, we plan to construct an importance sampling function using region back-projections. Secondly, after fitting regions the edge direction field will be used to improve alignment in a manner similar to the work of Wachter and Nagel [6]. Finally, we plan to make a quantitative comparison of different region features, such as edge strength.

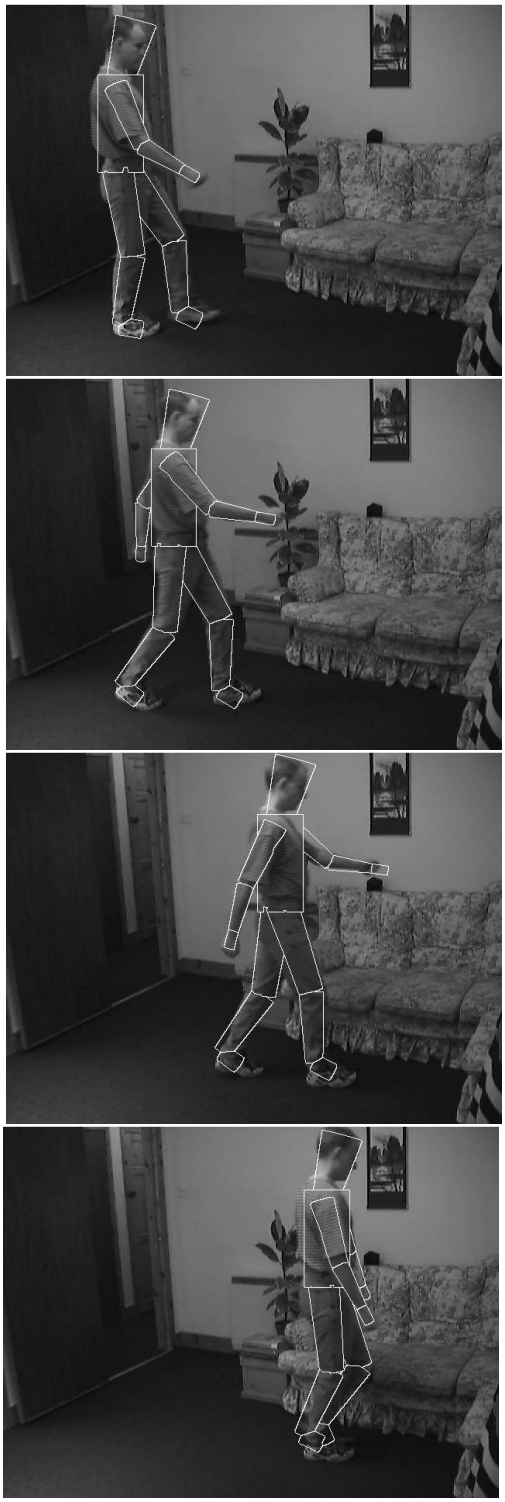

Figure 5. Tracking a walking sequence

References

[1] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of nonrigid objects using mean shift. *Computer Vision and Pattern Recognition*, pages 673–678, 2000.

[2] S. 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.

[3] T. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231–268, March 2001.

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

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

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