MLESAC-Based Tracking with 2D Revolute-Prismatic Articulated Models

G. McAllister, S.J. McKenna, I.W. Ricketts
Department of Applied Computing
University of Dundee, Dundee DD1 4HN, United Kingdom

(gmcallis, stephen, ricketts)@computing.dundee.ac.uk

Abstract

A model for tracking articulated objects is proposed using a novel 2D revolute-prismatic joint. An extension of the RANSAC and MLESAC algorithms incorporating feature weights is used to perform robust tracking. The models are suitable for tracking certain human body structures. Limbs are modelled as constrained planar patches. A patch can rotate about a joint point that is displaced relative to the previous patch. A scenario in which the forearm is tracked is used to illustrate the method.

1. Introduction

Two-dimensional articulated models have previously been used to track human motion. For example, connected planar patches were used by Ju *et al.* [2] to track human limbs based on optical flow. Here we propose articulated models consisting of connected planar patches connected in a quite different manner using what we refer to as *2D prismatic-revolute joints*. These allow a patch to rotate in the image plane about a joint point which is displaced some distance, s, in a direction orthogonal to the major axis of the previous patch. We note in passing that, despite the similar terminology, the models described here are quite different from the scaled prismatic models proposed by Morris and Rehg for articulated object tracking [4]. In order to track human motion using articulated 2D patch models, robust fitting is required to cope with noise, clutter and tracking ambiguities. This paper presents an extension of MLESAC, originally proposed by Torr and Zisserman [5] as a modification of RANSAC [1] for stereo matching. This extension incorporates measurement likelihoods to provide a more robust estimation. The approach is illustrated here using a scenario in which a vehicle driver's arm is tracked through a cluttered scene.

2. Articulated Model

An articulated model of the type illustrated in Figure 1 is in general an acyclic graph of planar patches. In each chain of patches, $\mathbf{m}_i$ is connected to $\mathbf{m}_{i-1}$. Here we constrain the motion of patches to image plane translations, rotations and scaling. The patches are connected using 2D prismatic-revolute joints.

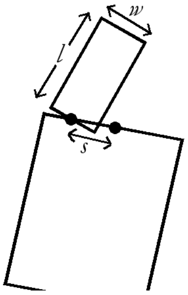

Figure 1. Articulated model

A patch $\mathbf{m}_i$ can rotate relative to $\mathbf{m}_{i-1}$ about a joint point $\mathbf{x}_i^J$ which is located at the midpoint of the bottom of the patch. A patch, $\mathbf{m}_i$, is defined in a frame of reference, $\mathcal{R}_i$, calculated from $\mathbf{m}_{i-1}$ and is parameterised as

$$\mathbf{m}_i = [w_i, l_i, s_i, \theta_i] \tag{1}$$

where w_i and l_i are the width and length of the planar patch respectively, s_i denotes the x-axis displacement of $\mathbf{x}_i^J$ in $\mathcal{R}_i$ and θ_i is the orientation of the patch about $\mathbf{x}_i^J$ with $\theta = 0°$ defined parallel to the y-axis of $\mathcal{R}_i$. $\mathcal{R}_i$ has as its origin the top side midpoint of $\mathbf{m}_{i-1}$ and its x-axis is co-linear with this top side. Having an x-displacement as the only positional parameter constrains $\mathbf{x}_i^J$ to lie on a line co-linear with the top side of $\mathbf{m}_{i-1}$, hence a *prismatic* joint. The joint is also *revolute* since θ_i controls the orientation of the planar patch about $\mathbf{x}_i^J$. The top, left and right sides of a patch will be denoted T, L and R respectively. For convenience, T, L

and R will also be used to denote the sets of pixels on the respective sides when they are rendered as line segments.

Two-dimensional prismatic-revolute joints can be used for tracking human body structures such as the arm-torso and leg-torso joints of an upright human. Rotation in depth due to a person turning around is handled by the prismatic joint parameter, s. In order to illustrate the use of this novel kinematic model for tracking, consider a forearm-hand and outstretched finger as in Figure 1. In this case we have two patches: $\mathbf{m}_h$ is used to track the forearm-hand and $\mathbf{m}_f$ is used to track the outstretched finger. The origin and y-axis of $\mathcal{R}_h$ are defined as the image origin $(0, 0)$ and the image y-axis respectively. $\mathcal{R}_f$ is defined relative to $\mathbf{m}_h$. While it is a reasonable model of an outstretched finger, a planar patch is only a crude approximation to the combined hand and forearm: width changes and clothing are not accurately modelled. Nonetheless, a planar patch is a useful representation for tracking provided a robust fitting method is used. In this scenario, each patch is fitted based on image evidence for T, L and R. The bottom sides of patches do not correspond to useful local image evidence. L and R are the longest sides and evidence should be available along their entire length. However T is relatively short and in the case of $\mathbf{m}_h$ only partial, noisy image evidence will be available locally. A decision was therefore taken to decompose fitting for this scenario. The parallel left and right sides were used to fit the w, s, and θ parameters while the top sides were used to fit the l parameters. The next section describes the measurement and robust fitting processes used to fit this articulated model.

3. Robust fitting

The articulated model is fitted hierarchically to measurements made on a foreground probability image, I, which is calculated using a statistical colour modelling approach described elsewhere [3]. Since feature point extraction and fitting methods are the same for each patch in the model, the patch subscript will be dropped. Search is centred on a predicted patch $\mathbf{m}^* = [w^*, l^*, s^*, \theta^*]$.

Pairs of points are sampled at uniform intervals along the parallel sides $\mathbf{L}^*$ and $\mathbf{R}^*$. More formally, N pairs $(\mathbf{l}_n, \mathbf{r}_n) \in (\mathbf{L}^*, \mathbf{R}^*)$ are determined, where $n = 1, \ldots, N$. The points $\mathbf{l}_n$ and $\mathbf{r}_n$ are positioned at a distance $\frac{l^*}{N}(n - 0.5)$ from $\mathbf{T}^*$. A similar scheme is used along the top side except that single points rather than pairs are sampled. M points are selected on $\mathbf{T}^*$ where the m^{th} point is at a distance $\frac{w^*}{M}(m - 0.5)$ from the end of $\mathbf{T}^*$.

For each pair of points, $(\mathbf{l}_n, \mathbf{r}_n)$, two co-linear search line segments of length λ are defined. These are centred on $\mathbf{l}_n$ and $\mathbf{r}_n$ and are orthogonal to $\mathbf{L}^*$ and $\mathbf{R}^*$. Let $\mathbf{s^L}_n$ and $\mathbf{s^R}_n$ denote the sets of all pixels on these two rendered line segments. A likelihood is computed for each pair of pixels $(\mathbf{x^L}, \mathbf{x^R})$ such that $\mathbf{x^L} \in \mathbf{s^L}_n$ and $\mathbf{x^R} \in \mathbf{s^R}_n$. This likelihood is then maximised over these pairs of search points.

In the experiments described here, the likelihood of a pair was calculated by combining steered, first-derivative filter responses at $\mathbf{x^L}$ and $\mathbf{x^R}$ together with the foreground probability mass between them. The response of a Sobel filter steered to orientation θ^* at pixel $\mathbf{x}$ in image I is denoted $h(\theta^*, \mathbf{x})$. The foreground probability mass between a co-ordinate pair $(\mathbf{x^L}, \mathbf{x^R})$ is denoted by $f(\mathbf{x^L}, \mathbf{x^R})$ and is the average value of the pixels in I on the line segment between $\mathbf{x^L}$ and $\mathbf{x^R}$. Therefore, the likelihood can be written as:

$$p(I|\mathbf{x^L}, \mathbf{x^R}) = \left(h(\theta^*, \mathbf{x^L}) - h(\theta^*, \mathbf{x^R})\right) f(\mathbf{x^L}, \mathbf{x^R}) \quad (2)$$

The most likely pair can be used to specify hypotheses $(\tilde{w}, \tilde{s}, \tilde{\theta})$ for (w, s, θ). The hypothesis $\tilde{w}$ is just the Euclidean distance between the pixels in the most likely pair. Let $\tilde{\mathbf{c}}_n$, denote the midpoint of the most likely pair. Probabilities for the N sample pairs are thus obtained:

$$p_n^{\mathbf{C}} = p(I|\mathbf{c}_n, \tilde{w}_n) = p(I|\mathbf{x^L}, \mathbf{x^R}) \quad (3)$$

A similar process is performed for each of the M sample points along $\mathbf{T}^*$. For each such point, a search line segment of length μ is centred on the point, orthogonal to $\mathbf{T}^*$. A maximum likelihood pixel $\tilde{\mathbf{t}}_m$ is found on each of the M search line segments. The likelihood again combines edge strength with the local foreground probability mass within the patch.

$$p(I|\mathbf{x^T}) = h(\theta^*, \mathbf{x^T}) f(\mathbf{x_0}, \mathbf{x^T}) \quad (4)$$

where $\mathbf{x_0}$ is the search line endpoint inside the planar patch. Let $p_m^{\mathbf{T}}$ denote the likelihood at the most likely pixel on the m^{th} search line. The measurements $\tilde{\mathbf{c}}_n$, $\tilde{w}_n$ and $\tilde{\mathbf{t}}_m$ and the associated probabilities, $p_n^{\mathbf{C}}$ and $p_m^{\mathbf{T}}$, are used in the robust fitting process. We will now describe the RANSAC and MLESAC fitting methods before discussing our extension.

3.1. RANSAC and MLESAC

Random sample consensus (RANSAC), maximum-likelihood consensus (MLESAC) and our extension all have at heart the idea proposed in the original RANSAC paper [1] of fitting a model to the best points in a data set by iteratively sampling and fitting to random subsets of the data. Subsets are chosen of the minimum size required to instantiate the model. The model is scored on how well it fits the data points. A stopping criterion is specified and once this criterion has been satisfied the model which optimises the fitting score in some respect is chosen as the new model estimate. For our purposes, the sample set consists of indices into the arrays of measurements. A random model, $\mathbf{m}'$, is instantiated by generating uniform random numbers $p, q \in [1, N]$ and $r \in [1, M]$. The model parameters s' and θ' are determined by two centre points, $\tilde{\mathbf{c}}_p$ and $\tilde{\mathbf{c}}_q$. The width is simply $w' = \tilde{w}_p$, and l' is determined as the distance to the top edge point, $\tilde{\mathbf{t}}_r$, given s' and θ'.

The proposed extensions to RANSAC have involved changing the objective function, C, used to evaluate model goodness of fit. In each scheme the underlying sampling algorithm is the same. In RANSAC itself those feature points whose distance from the model are below a threshold score zero while the outliers score a constant penalty. Good models minimise this score although if the threshold is set too high then all models will score zero. Formally, RANSAC defines an objective function based on some error measure, e, as

$$C = \sum_j \rho(e_j) \tag{5}$$

where

$$\rho(e) = \begin{cases} 0 & e < \text{threshold} \\ \text{constant} & e \geq \text{threshold} \end{cases} \tag{6}$$

MLESAC is a probabilistic version of RANSAC where the distance of data points from the model is assumed distributed according to a mixture of a Gaussian and a uniform distribution. The distribution over distance of a data point from the model is

$$Pr(e) = \alpha \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{e}{2\sigma^2}\right) + (1-\alpha)\frac{1}{v} \tag{7}$$

where the mixing parameter α is calculated using expectation-maximisation and v is the size of the search. This gives a more robust measure of error than least-square methods. The objective function that MLESAC aims to minimise is

$$C = -\log\left(\prod_j Pr(e_j)\right) \tag{8}$$

3.2. Extension of MLESAC

The MLESAC scheme weights all measurements equally in the fitting process, regardless of their likelihood. The scheme can be made more robust by incorporating these measurement likelihoods into the fitting process. MLESAC can also be extended into a tracking framework with the inclusion of a temporal prior $p(\mathbf{m}^{\tau-1})$ on the model parameters.

In the general case, we define some measure, e, of the distance between measurement points and the model as before. The variance in Equation (7) is set inversely proportional to the feature probability, p_j. This penalises models which fit poorly to high probability feature points. The objective function to be maximised is the posterior probability at time τ:

$$C = p(\mathbf{m}^\tau | \mathbf{m}^{\tau-1}, I) = \mathcal{L}p(\mathbf{m}^{\tau-1}) \tag{9}$$

3.3. Model-specific fitting

Now we discuss the fitting for this model in particular. The image measurements are not made in the same parameter space as the model so fitting can not directly be achieved in $[w, l, s, \theta]$ space. The fitting problem is subdivided into fitting to w', s' and θ' in $[x, y, w]$ space and fitting to l' in l space. w', s' and θ' define a line on a subspace in $[x, y, w]$ space since w' is constant along the length of the planar patch. The fitting error, e_i^C, for each data point, $(\mathbf{c}_i, w_i)$, is calculated by measuring the 3D perpendicular distance of the point from the line. The fitting error for l' is calculated by first converting all measurements $\mathbf{t}_j$ into length values by using s' and θ'. Then the fitting error for each length, e_j^T, is the distance between l' and each length data point. The objective function for this model is

$$C = p(\mathbf{m}^\tau | \mathbf{m}^{\tau-1}, I) = \mathcal{L}^C \mathcal{L}^T p(\mathbf{m}^{\tau-1}) \tag{10}$$

where $\mathcal{L}^C$ and $\mathcal{L}^T$ are $\mathcal{L}$ obtained by setting $e = e^C, p = p^C$ and $e = e^T, p = p^T$ respectively.

4. Results

Two sets of results are presented here for qualitative evaluation. The nature of the model makes defining an objective measure of performance problematic. The model is only a crude approximation to the arm and as such a range of parameters may be deemed acceptable in any given frame. However, these results clearly show the benefits of incorporating feature weights into MLESAC for model fitting. The first results, given in Figure 2, demonstrate the need for a robust fitting technique. The top row shows the model estimate from a least-squares (LS) fit in each frame and the bottom row shows the ‘Extended MLESAC’ estimate. The advantages of a robust technique are immediately apparent. The LS fitting completely loses track of the arm while Extended MLESAC tracks reliably throughout the sequence. The second result is shown in Figure 3. In order to compare the robustness of ‘standard’ and ‘extended’ MLESAC a deliberate error was introduced by manually instantiating the model at time $\tau = 50$. Standard MLESAC in the top row fits to the measurements regarding them equally valid and as a result fails to lock back onto the hand. Extended MLESAC on the bottom row prefers a model which includes the high probability measurements inside the hand and locks back on immediately, demonstrating the advantages of incorporating measurement probabilities into the fitting process over regarding each measurement equally.

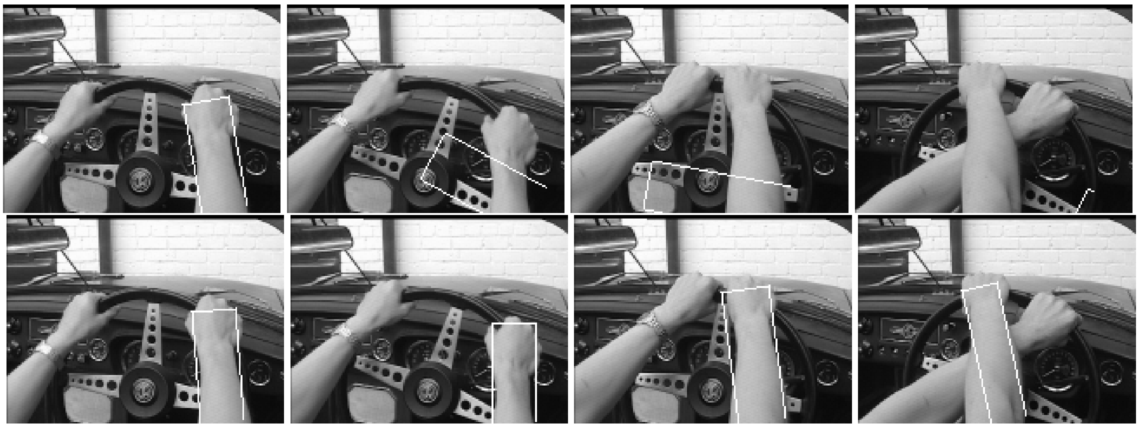

Figure 2. Comparison of fitting methods. Top row: Least-squares , bottom row: Extended MLESAC

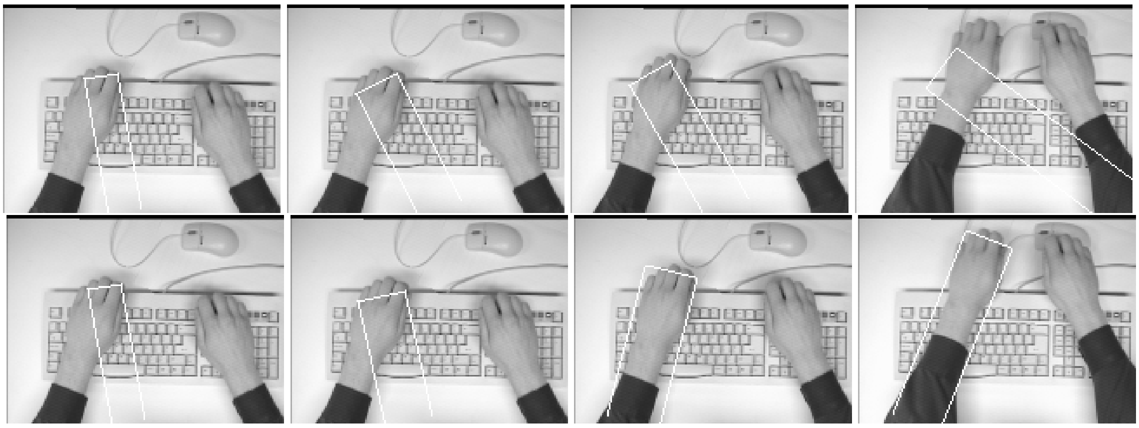

Figure 3. Key frames $\tau = 50, 51, 53, 67$ illustrating performance after tracking failure. Top row: Standard MLESAC, bottom row: Extended MLESAC

5. Discussion and Future Work

This paper has described a novel articulated model using planar patches connected by what we have named a *2D prismatic-revolute* joint. The joint can handle rotation in depth of certain structures with a 2D model. We have also described the robust fitting of this model using a novel extension of MLESAC which includes a temporal prior on the model parameters and possesses a greater degree of robustness over MLESAC by incorporating measurement probabilities into the fitting process. We have demonstrated this robustness in addition to illustrating the need for a robust fitting solution through the inability of least-squares fitting to cope with the inevitable outliers in the data.

References

[1] M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography. *Communications ACM*, 24:381–395, 1981.

[2] S. X. Ju, M. J. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated image motion. In *IEEE International Conference on Face & Gesture Recognition*, pages 38–44, Killington, VT, October 1996.

[3] G. McAllister, S. McKenna, and I. Ricketts. Hand tracking for behaviour understanding. *Image and Vision Computing*. Submitted.

[4] D. D. Morris and J. M. Rehg. Singularity analysis for articulated object tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 289–296, Santa Barbara, June 1998.

[5] P. Torr and A. Zisserman. MLESAC: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, 78(1-7):138–156, 2000.