Tracking humans from a moving platform

Larry Davis, Vasanth Philomin and Ramani Duraiswami
Computer Vision Laboratory
Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742, USA
{lsd, vasi, ramani}@cs.umd.edu

Abstract

Research at the Computer Vision Laboratory at the University of Maryland has focussed on developing algorithms and systems that can look at humans and recognize their activities in near real-time. Our earlier implementation (the W^4 system) while quite successful, was restricted to applications with a fixed camera. In this paper we present some recent work that removes this restriction. Such systems are required for machine vision from moving platforms such as robots, intelligent vehicles, and unattended large field of regard cameras with a small field of view. Our approach is based on the use of a deformable shape model for humans coupled with a novel variant of the Condensation algorithm that uses quasi-random sampling for efficiency. This allows the use of simple motion models which results in algorithm robustness, enabling us to handle unknown camera/human motion with unrestricted camera viewing angles. We present the details of our human tracking algorithms and some examples from pedestrian tracking and automated surveillance.

1 Introduction

The Computer Vision Laboratory at the University of Maryland has been investigating problems related to detection, tracking and analysis of human activities for almost ten years. Our earliest work focused on tracking of facial features in the context of recognizing human facial expressions from motion [11, 2]. The system described in [2], which involved robust flow estimation, image stabilization, and tracking of several facial features, required more than one minute of what then passed for CPU time per frame. Clearly, the system was far from real time, which limited both experimentation during development as well as performance evaluation over large data sets. Just seven years later at SIGGRAPH 2000 we demonstrated in collaboration with IBM Almaden a real time system for detection of faces and facial features, recognition of facial expressions and online mimicry of those facial expression on an electromechanical face.

Another early project was described in [5]. Here, multiperspective videos of humans in action were analyzed and 3D volumetric models with many degrees of freedom were fit to these images as the body was tracked through the sequence. This system, which was implemented using Khoros and ran on a rather underpowered UNIX workstation, took many minutes per frame to analyze. But at SIGGRAPH 1998 we were able to demonstrate, in collaboration with the M.I.T. Media Laboratory and the ATR Media and Communications Laboratory, a 3D motion capture system that utilized six cameras, eight PCs and was able to recover coarse body shape data at rates of 28 frames per second. That system used many of the processing elements integrated into our W^4 visual surveillance system, a PC based system that could detect and track people and their body parts at speeds of more than thirty frames per second [7]. The W^4 system, however, was designed with a stationary camera in mind (as shown by its heavy dependency on background subtraction techniques) and so takes a "Stop and Look" approach when dealing with moving camera platforms. More recently, in collaboration with Daimler-Chrysler Research, we addressed the problem of detection of humans from moving vehicles [4] using an efficient variant of a multifeature distance transform algorithm. The system described in [4] was able to achieve near real-time performance in natural environments as a result of an efficient organization of shape templates into a hierarchical data structure for matching (resulting in a matching strategy with logarithmic complexity rather than linear), a coarse-to-fine search over the transformation parameters and a SIMD (Single Instruction Multiple Data) implementation of the time-consuming steps of the algorithm.

We have been able to achieve these several order of magnitude increases in computational capability through a combination of better algorithm and data structure design, relentless increases in processing power of commodity com-

puting, and advances in both communication hardware and software for multiprocessor systems. In this paper we discuss recent research in our laboratory that addresses the related problem of tracking humans from moving camera platforms. In particular, we describe how efficient random sampling techniques can be employed in the Condensation algorithm [8] to improve its asymptotic complexity and robustness, especially for problems that involve high-dimensional state spaces.

This paper is organized as follows: Section 2 gives a brief introduction to the shape model used to model humans and explains how to learn the model automatically from segmented pedestrian contours. This gives a set of deformation parameters which along with the Euclidean parameters (translation, rotation and scaling) constitute the state space. Section 3 describes the tracking algorithm that addresses the issues of an unknown motion model, robustness to outliers, and use of quasi-random points for efficiency. In Section 4 we successfully apply this algorithm to real video sequences of pedestrians as well as automated surveillance sequences. Section 5 concludes the paper.

2 Learning a linear human model

The Point Distribution Model (PDM) [3] has proven to be a useful method for building a compact linear shape model from training examples of a class of shapes. The conventional PDM requires manual labelling of a set of points called the "landmark" points in each training image. These points are concatenated to form a shape vector and the shape vectors resulting from all the training images are aligned using Procrustes analysis [6]. A mean shape and a set of modes of variation are then generated using Principal Component Analysis (PCA). A method for automatically extracting the human silhouettes from a training set of images and building a linear shape model is described in [1]. First, the silhouettes are extracted using background subtraction followed by morphological operations and then tracing the boundary points of the resulting foreground regions to form edge chains. A uniform B-spline with the control points placed at approximately uniformly spaced intervals along the contour is produced efficiently from each of these silhouettes. The control points of the B-spline are then used as the landmark points in the PDM.

We use techniques similar to that described in [1] and [3] with some improvements to build a linear human model. One improvement is in the parameterization of the B-spline curve that is fitted to each extracted contour. Suppose that the set of points in a single human contour is $\{Q_k\}, k = 0, \ldots, m$, and we want to approximate these points with a p^{th} degree B-spline. Suppose that the values for the parameters $\bar{u}_k$ and the knot vector $U = u_0, ..., u_r$ are precomputed and known. We then set up and solve the (unique) linear least squares problem for the unknown control points P_i. Assume that $p \geq 1$, $m > n$ and $n \geq p$. We seek a p^{th} degree nonrational curve

$$C(u) = \sum_{i=0}^{n} N_{i,p}(u) P_i \quad u \in [0,1]$$

satisfying:

- $Q_0 = C(0)$ and $Q_m = C(1)$;
- the remaining Q_k are approximated in the least squares sense, i.e.

$$\sum_{k=1}^{m-1} |Q_k - C(\bar{u}_k)|^2$$

is a minimum with respect to the $n+1$ variables, P_i; the $\{\bar{u}_k\}$ are the precomputed parameter values and $N_{i,p}$ are the p^{th} degree B-spline basis functions. The resulting curve generally does not pass precisely through Q_k, and $C(\bar{u}_k)$ is not the closest point on $C(u)$ to Q_k.

The choice of $\bar{u}_k$ and U affects the shape and parameterization of the curve. The most common method for choosing $\bar{u}_k$ is the chord length parameterization, which is the one used in [1]. Here, if d is the total chord length given by

$$d = \sum_{k=1}^{m} |Q_k - Q_{k-1}|$$

then $\bar{u}_0 = 0$, $\bar{u}_m = 1$ and

$$\bar{u}_k = \bar{u}_{k-1} + \frac{|Q_k - Q_{k-1}|}{d} \quad k = 1, \ldots, m-1$$

This gives a good parameterization of the curve in the sense that it approximates a uniform parameterization. However, when the data takes very sharp turns such as in the case of human shapes, the chord length method does not perform well. We use the centripetal method ([9]) that gives better results with such data, where if

$$d = \sum_{k=1}^{m} \sqrt{|Q_k - Q_{k-1}|}$$

then $\bar{u}_0 = 0$, $\bar{u}_m = 1$ and

$$\bar{u}_k = \bar{u}_{k-1} + \frac{\sqrt{|Q_k - Q_{k-1}|}}{d} \quad k = 1, \ldots, m-1$$

The placement of the knots should reflect the distribution of the $\{\bar{u}_k\}$ and we choose the knot vector U as follows. Let $c = \frac{m+1}{n-p+1}$, then the internal knots are given by

$$i = \lfloor jc \rfloor \qquad \alpha = jc - i$$
$$u_{p+j} = (1-\alpha)\bar{u}_{i-1} + \alpha \bar{u}_i \quad j = 1, \ldots, n-p \qquad (1)$$

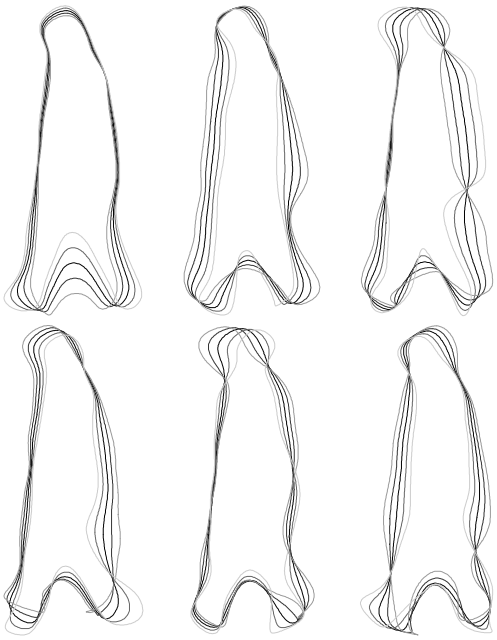

Figure 1. Modes of variation of pedestrian shapes

Equation (1) guarantees that every knot span contains at least one $\bar{u}_k$, and under this condition the matrix in the least squares formulation is positive definite and well-conditioned.

We also use a weighted least squares method to align two shapes in the Procrustes analysis, where the weights are chosen so that more significance is given to the more stable landmark points i.e. the points which vary their position the least over the entire training set. As a result, emphasis is given to aligning the stable parts of the object rather than the unstable parts during shape alignment. Figure 1 shows some of the significant modes of variation of the human shapes in the training set of pedestrian contours.

3 Tracking algorithm

The Condensation algorithm [8] has attracted much interest in the active vision area as it offers a framework for dynamic state estimation where the underlying probability density functions (pdfs) need not be Gaussian. The algorithm is based on a Monte Carlo or sampling approach, where the pdf is represented by a set of random samples. As new information becomes available, the posterior distribution of the state variables is updated by recursively propagating these samples (using a motion model as a predictor) and resampling. An accurate dynamical model is essential for robust tracking and for achieving real-time performance. This is due to the fact that the process noise of the model has to be made artificially high in order to track objects that deviate significantly from the learned dynamics, thereby increasing the extent of each predicted cluster in state space. One would then have to increase the sample size to populate these large clusters with enough samples. A high-dimensional state space (required for tracking complex shapes such as pedestrians) only makes matters worse. Even when one uses a "perfect" pseudo-random sequence for generating N sample points, the sampling error will only decrease as $O(N^{-1/2})$ as opposed to $O(N^{-1})$ for another class of sequences known as quasi-random sequences which have low discrepancy. We introduced quasi-random sampling in the context of the Condensation algorithm in [10] and showed that even in low dimensions, a significantly fewer amount of sample points were needed to achieve the same sampling error when compared to pseudo-random sampling. For reasons of brevity, the details are not discussed here; the readers are referred to [10]. In typical implementations of the Condensation algorithm, a "perfect" pseudo-random number generator is almost never used and a linear congruential generator (such as the system supplied `rand()` function) is used instead. These generators, although very fast, have an additional inherent weakness that they are not free of sequential correlation on successive calls, i.e. if k random numbers at a time are used to generate points in k-dimensional space, the points will lie on $(k-1)$-dimensional planes and will not fill up the k-dimensional space.

Since we do not want to make any assumptions about how the vehicle and the pedestrian are moving or about the viewing angle, we propose using a zero-order motion model with large process noise high enough to account for the greatest expected change in shape and motion. In other words, we need to concentrate our samples in large regions around highly probable locations from the previous time step. These high-dimensional regions which correspond to the large process noise can now be efficiently sampled using quasi-random sampling as described below.

Given the sample set $\{(\mathbf{s}_{t-1}^{(n)}, \pi_{t-1}^{(n)})\}$ at the previous time step, $\pi_{t-1}^{(n)}$ being the associated probabilities, we first choose a base sample $\mathbf{s}_{t-1}^{(i)}$ with probability $\pi_{t-1}^{(i)}$. This yields a small number of highly probable locations, say M, the neighborhoods of which we must sample more densely. If there were just one region requiring a dense concentration, an invertible mapping from a uniform space to the space of equal importance could be constructed, as given below in Equation (3) for the case of a multi-dimensional Gaussian. Since we have M regions, the importance function cannot be constructed in closed form. One therefore needs an alternative strategy for generating from the quasi-random distribution, a set of points that samples important regions densely.

We have devised a simple yet effective strategy that achieves these objectives. Let the M locations have centers $\mu^{(j)}$ and variances $\sigma^{(j)}$ based on the process noise, where these quantities are k-dimensional vectors. We then overlay $M+1$ distributions of quasi-random points over the space, with the first M distributions made Gaussian, centered at $\mu^{(j)}$ and with diagonal variance $\sigma^{(j)}$ (3). Finally, we also overlay a $(M+1)$th distribution that is spread uniformly over the entire state space. This provides robustness against sudden changes in shape and motion. The total number of points used is N, where

$$N = N_1 + N_2 + \ldots + N_{M+1}, \tag{2}$$

the sample size in the Condensation algorithm. We have in effect chosen $\mathbf{s}_t^{(n)}$ by sampling from $p(\mathbf{X}_t/\mathbf{X}_{t-1} = \mathbf{s}_{t-1}^{(i)})$.

The conversion from a uniform quasi-random distribution to a Gaussian quasi-random distribution is achieved using the mapping along the lth dimension

$$y_{jl} = \mu_l^{(j)} + \sqrt{2}\sigma_l^{(j)}\ \mathrm{erf}^{-1}\left((2\xi_l - 1)\right), \tag{3}$$

where erf^{-1} is the inverse of the error function given by

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt,$$

and ξ_l represents the quasi-randomly distributed points in $[0, 1]$.

Finally, we measure and compute the probabilities $\pi_t^{(n)} = p(\mathbf{Z}_t/\mathbf{X}_t = \mathbf{s}_t^{(n)})$ for these new sample positions in terms of the image data $\mathbf{Z}_t$. We use a measurement density based on the multi-feature distance transform algorithm (see [4] for details) that has been successfully used for detecting pedestrians from static images. Therefore

$$\log p(\mathbf{Z}_t/\mathbf{X}_t) = \log p\left(\mathbf{Z}/\mathbf{X}\right) \propto \left\{ -\frac{1}{M} \sum_{i=1}^{M} d_{typed}^2(z_i, I) \right\},$$

where the z_i's are measurement points along the contour, I is the image data, and $d_{typed}(z_i, I)$ denotes the distance between z_i and the closest feature of the same type in I. We use oriented edges discretized into eight bins as the features in all our experiments.

4 Results

We now present some results on tracking pedestrians from a moving vehicle (Figure 2) and humans from an overhead surveillance camera that pans from side to side (Figures 3 and 4). First, a linear shape model was built from automatically segmented human contours using the techniques described in Section 2 and the dimensionality was reduced using PCA to find an eight-dimensional space of deformations. We used $N = 2000$ samples in the Condensation algorithm and introduced 10% of random samples at every iteration to account for sudden changes in shape and motion. Figures 2, 3 and 4 show the tracker output as contours corresponding to the modal (highest probability) state and the mean state. The tracker was able to recover very quickly from failures due to sudden changes in shape or motion and track people through partial occlusion. Figure 3 shows a specific example where the person being tracked is temporarily occluded by a pole between Frames 38 and 50.

5 Conclusions

In this paper, we have developed a framework for tracking humans from moving camera platforms. Our approach used the Condensation tracker and extended it to high-dimensional problems by incorporating quasi-Monte Carlo methods into the conventional algorithm. Specifically, we overlaid layers of quasi-random Gaussian grids over the state space which allowed for efficient sampling. As a result, we could handle general situations where there are no restrictions on the dynamics of the camera or the human being tracked and there are no assumptions on the viewing angle.

Acknowledgements

We gratefully acknowledge the partial support of ONR contract N000149510521 and Department of Justice contract JUST1999LTVXK019.

References

[1] A. Baumberg and D. C. Hogg. Learning flexible models from image sequences. In *Proc. European Conference on Computer Vision*, 1994.

[2] M. J. Black and Y. Yacoob. Recognizing facial expressions in image sequences using local parameterized models of image motion. *International Journal of Computer Vision*, 25(1):23–48, 1997.

[3] T. F. Cootes, C. J. Taylor, A. Lanitis, D. H. Cooper, and J. Graham. Building and using flexible models incorporating grey-level information. In *Proc. IEEE International Conference on Computer Vision*, pages 242–246, 1993.

[4] D. Gavrila and V. Philomin. Real-time object detection for "smart" vehicles. In *Proc. IEEE International Conference on Computer Vision*, volume 1, pages 87–93, Kerkyra, Greece, 1999.

[5] D. M. Gavrila and L. Davis. Towards 3-D model-based tracking and recognition of human movement: a multi-view approach. In *Int. Workshop on Face and Gesture Recognition*, Zurich, Switzerland, 1995.

Figure 2. Tracking results for Daimler-Chrysler pedestrian sequence using quasi-random sampling. Dark - Modal state estimate; Light - Mean state estimate.

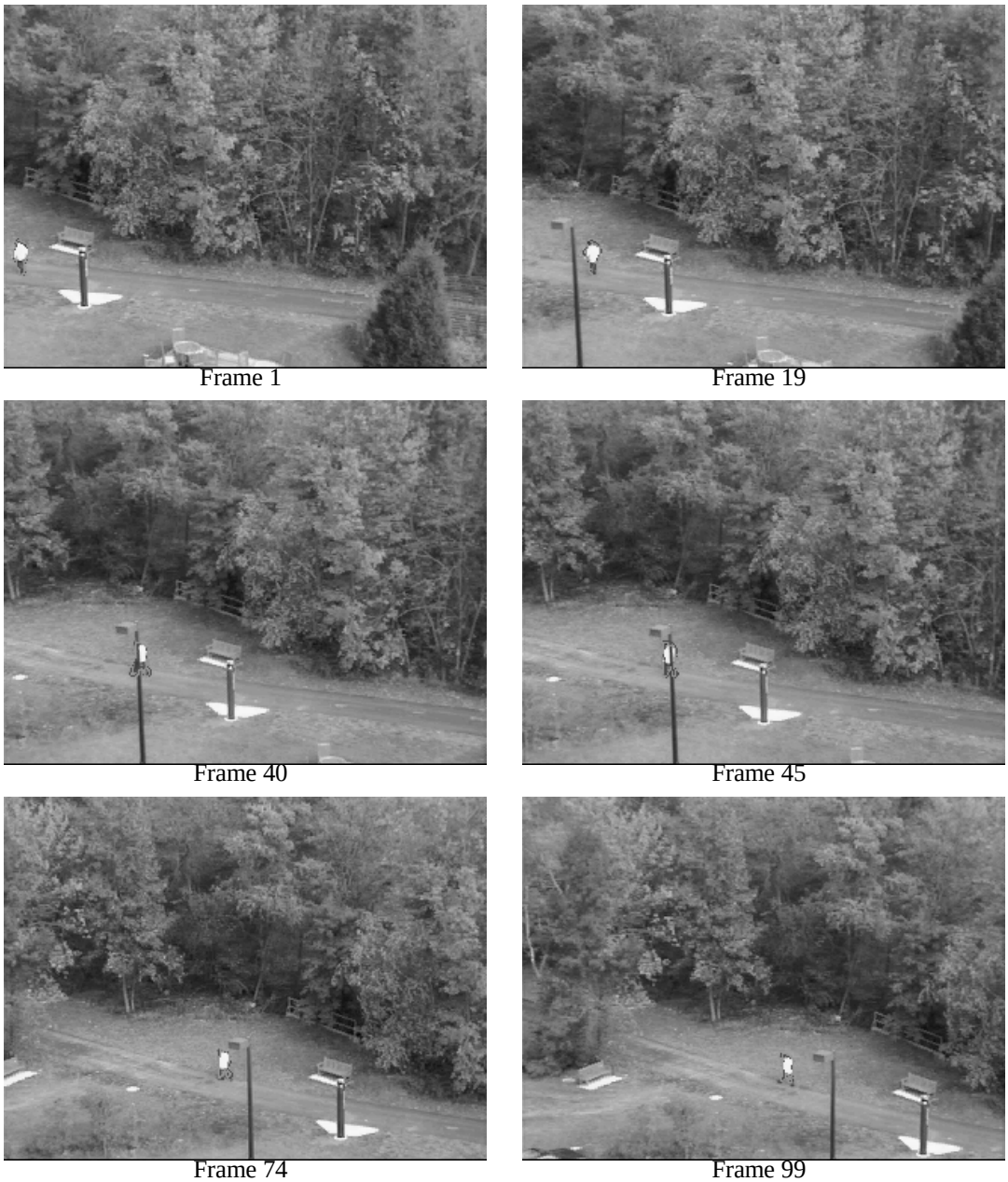

Figure 3. Tracking results for a surveillance sequence with occlusion using quasi-random sampling. Dark - Modal state estimate; Light - Mean state estimate.

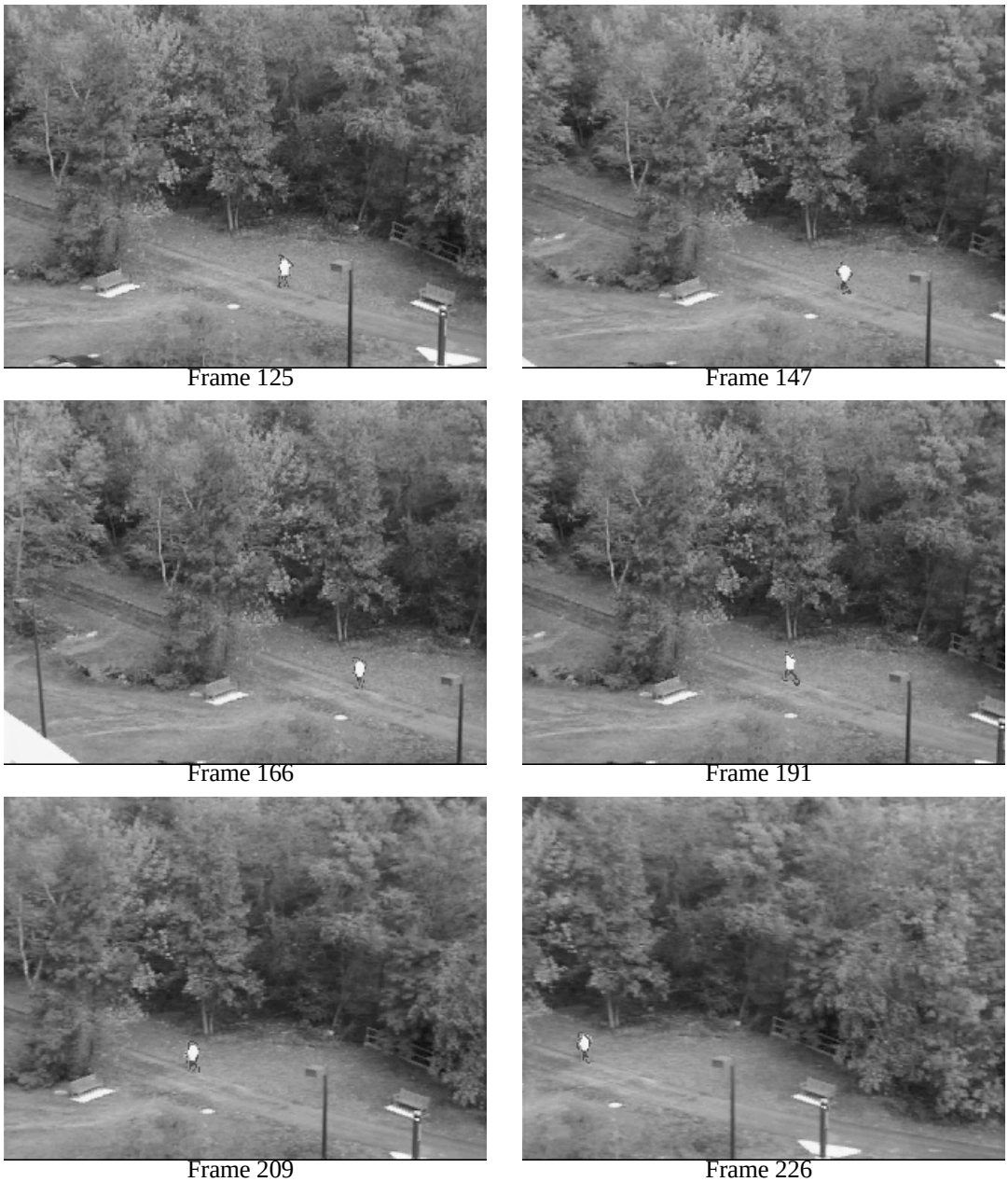

Figure 4. Tracking results for a surveillance sequence with occlusion using quasi-random sampling (cont'd from Figure 3). Dark - Modal state estimate; Light - Mean state estimate.

[6] C. Goodall. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society B*, 53(2):285–339, 1991.

[7] I. Haritaoglu, D. Harwood, and L. Davis. W4: Who, when, where, what: A real time system for detecting and tracking people. In *Face and Gesture Recognition Conference*, pages 222–227, Japan, 1998.

[8] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *Proc. European Conference on Computer Vision*, pages 343–356, Freiburg, Germany, 1996.

[9] E. T. Y. Lee. Choosing nodes in parametric curve interpolation. In *CAD*, volume 21, pages 363–370, 1989.

[10] V. Philomin, R. Duraiswami, and L. S. Davis. Quasi-random sampling for condensation. In *Proc. European Conference on Computer Vision*, 2000.

[11] Y. Yacoob and L. S. Davis. Recognizing human facial expressions from long image sequences using optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(6):636–642, 1996.