On the Sustained Tracking of Human Motion

Yaser Ajmal Sheikh
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
yaser@cs.cmu.edu

Ankur Datta
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
ankurd@cs.cmu.edu

Takeo Kanade
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
tk@cs.cmu.edu

Abstract

In this paper, we propose an algorithm for sustained tracking of humans, where we combine frame-to-frame articulated motion estimation with a per-frame body detection algorithm. The proposed approach can automatically recover from tracking error and drift. The frame-to-frame motion estimation algorithm replaces traditional dynamic models within a filtering framework. Stable and accurate per-frame motion is estimated via an image-gradient based algorithm that solves a linear constrained least squares system. The per-frame detector learns appearance of different body parts and 'sketches' expected gradient maps to detect discriminant pose configurations in images. The resulting online algorithm is computationally efficient and has been widely tested on a large dataset of sequences of drivers in vehicles. It shows stability and sustained accuracy over thousands of frames.

1. Introduction

Applications such as monitoring of the elderly, surveillance, and human computer interfaces all stand to benefit greatly from a system that is able to determine what posture a person of interest is in and how their posture evolves over time. However, tracking humans across a video is among the most challenging problems in computer vision. In addition to inference from incomplete imaged data, which is common to all vision problems, humans are highly articulated, highly variant in appearance, and unpredictable in their behavior. This makes it difficult to develop useful, yet general, models of motion and appearance for use during motion estimation.

The traditional framework used for tracking is inherited from data association research in radar technologies, [1]. In these approaches, predictions under a dynamic model, e.g. constant motion, constant acceleration or nonparametric variants, are fused with measurements at each time instant to minimize the variance of the location estimate. Per-frame measurements of human body configurations have been an active area of research recently, e.g. [14], [5], [17], [22], [4]. On the other hand, although some attention has been paid to dynamic models of humans, e.g. [23], [16], [21] and [18], relatively less attention has been given to modeling generic human motion, because of the large number of complex influences on human behavior, [20]. By and large, dynamic models of humans are either nondescript, like constant velocity models, or highly domain specific, learnt from a corpus of training data.

Unlike point measurements generated from radar technology, video sequences provide far richer descriptions at each time instant. We argue that rather than using dynamic models that are difficult to quantify for humans, we should exploit the fact that appearance remains approximately constant *across* time instances to provide "predictions" at the incident of new time frames. Frame-to-frame motion estimation using the appearance constancy assumptions has generated a significant and successful body of work, e.g. [12], [3], [2]. The key proposition in this paper is that the framework of fusing measurements and prediction through dynamic models can be replaced by a framework for fusing per-frame detections and frame-to-frame motion estimation.

In this paper, we describe a novel approach that produces accurate estimates of an actor's posture at each time instance in a video. Estimates from a frame-to-frame articulated motion estimation algorithm are fused with a per-frame detection method that uses learnt appearance models. Key characteristics of the approach are that it is online and *sustainable*, i.e. tracking that can continue processing indefinitely and does not suffer from accumulated errors as the video sequence progresses. At each time instance, a measure of the covariance is maintained. The application domain used to test and demonstrate ideas in this paper is vehicle driver behavior.

978-1-4244-2154-1/08/$25.00

2. Related Work

A large body literature exists on 2D human pose detection and tracking, however, due to space constraints, we will review only the most relevant papers dealing with 2D pose detection and tracking of 2D configurations (with the exception of [2]). The interested reader is directed to surveys ([9], [13]) and a recent book by Forsyth *et al*. ([7]) for a more comprehensive overview of the area.

Human pose detection approaches can be divided into discriminative and generative approaches. Discriminative approaches attempt to a learn a direct mapping between image features, such as edges or image moments to the 2D human pose. Rosales *et al*. in [19] train several specialized mapping functions in a supervised setting to map from input silhouette moments to the 2D human pose. A major limitation of discriminative approaches is that their performance degrades significantly in images where it is hard to obtain reliable features, as is often the case in the cluttered scenes. Generative approaches on the other hand generate a number of plausible pose hypothesis which are then evaluated against the image for evidence. Felzenszwalb and Huttenlocher in [5] represented the 2D human pose as a collection of parts arranged in a deformable configuration, building upon the seminal work of Fischler and Elschlager in [6], who introduced the spring constraints for human pose modeling. The pictorial structures, as their model is called, represents each body part using a simple appearance model with deformable spring-like connections between pair of parts. Ramanan *et al*. in [17] built a person detector that localized limbs of people in lateral walking poses. These limb detections were then used to build an appearance model of the limbs which was used to detect limbs in successive frames. Sigal and Black in [22] described an algorithm to infer 2D human pose from a single image. Their algorithm integrated information from bottom-up body-part proposal processes using non-parametric belief propagation. In this paper, we describe a generative approach towards 2D human pose detection with a search algorithm that does not suffer from the limitations of high-dimensional search that is usually associated with generative approaches.

The traditional framework used for tracking is a legacy from the data association research [1]. In these approaches, predictions from a dynamic model are fused with measurements to obtain the location estimates. Recent research under such a framework have involved application of complex dynamical models to track human motion in images. Isard and Blake in [11] describe the use of 'factored sampling' along with learnt dynamical models to propagate distributions over position and shape over time. Pavlovic *et al*. in [15] use the human motion capture data to learn a switching linear dynamic system model for tracking of human walking. A major difficulty with the wide-spread application of dynamical models to human tracking under general settings is the problem of constructing or learning an effective model of human dynamics. In this paper, we argue that the framework employing complex dynamical model for human tracking can be replaced by one employing frame-to-frame motion estimation. In other words, the framework of fusing predictions from dynamical models with measurements can instead be supplanted by a framework fusing the per-frame detections with frame-to-frame motion estimation. In this paper, we show encouraging results showing the potential of this framework for sustained human pose tracking across thousands of frames.

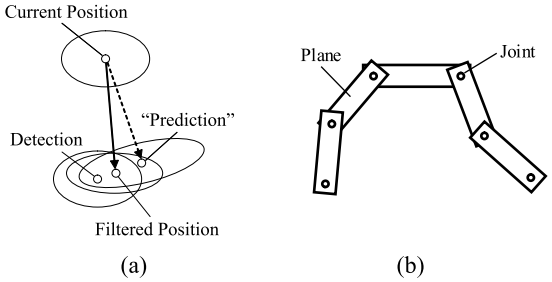


Figure 1. Estimation Framework. (a) The frame-to-frame motion estimation "predicts" a new location for the anatomical landmark and the per-frame detection provides a measurement for the new location. The mean and covariances of these two estimates are fused using the Kalman update equation. (b) The body model used in this paper is an articulated chain.

3. Data Model and Estimation Framework

The problem tackled in this paper is to locate, in an online manner, the configuration $\mathbf{X}_t \in \mathrm{R}^{2p}$ of p human anatomical landmarks (see Figure 1(b)) in gray-scale images $\mathbf{I}_t$ from a sequence ordered by $t = 1, \cdots, F$. The algorithm we propose is near realtime (5 FPS) and operates online (as opposed to batch mode algorithms). One common approach to solving this problem, is to adopt an inductive framework which separates the problem into initialization and tracking. Initialization is defined as a base case, "*given* $\mathbf{I}_0$*, locate* $\mathbf{X}_0$" and the tracking problem is defined inductively, "*given* $\mathbf{I}_t$ *and* $\mathbf{X}_t$ *and the new image* $\mathbf{I}_{t+1}$*, locate* $\mathbf{X}_{t+1}$." This approach suffers both from sudden failure and gradual error accumulation. In other words, if a configuration is incorrectly located at $t = T$, then for all $t > T$ failure is virtually guaranteed — there is no mechanism to recover from sudden failure. Additionally, if each configuration $\mathbf{X}_t$ is located with error $\epsilon \sim \mathcal{N}(\mathbf{0}, \Sigma)$, then at time $t = T$ the error would have accumulated to $\epsilon \sim \mathcal{N}(\mathbf{0}, (T - t_0)\Sigma)$.

In contrast to this inductive framework, we propose to use filtering, [1], to combine the *frame-to-frame* estimates of the configuration with the *per-frame* estimates. Unlike filtering algorithms, we do not describe a dynamic model (such as constant velocity or acceleration) or a distribution to sample from, but instead use the frame-to-frame motion

Objective

Given the current image, $\mathbf{I}_t$ and the triple $(\mathbf{I}_{t-1}, \mathbf{X}^+_{t-1}, \Sigma^+_{t-1})$, estimate $(\mathbf{X}^+_t, \Sigma^+_t)$.

Algorithm

1. **Per-Frame Detection**: Find the most likely configuration and its covariance, $(\bar{\mathbf{X}}_t, \bar{\Sigma}_t)$, given the image $\mathbf{I}_t$ (Section 4).
2. **Frame-to-Frame Motion Estimation**: Find the most likely configuration and its covariance $(\mathbf{X}^-_t, \Sigma^-_t)$ given the images $\mathbf{I}_t$ and $\mathbf{I}_{t-1}$ (Section 5).
3. **Update Configuration and Covariance**: Fuse $(\bar{\mathbf{X}}_t, \bar{\Sigma}_t)$ and $(\mathbf{X}^-_t, \Sigma^-_t)$ to get corrected locations $(\mathbf{X}^+_t, \Sigma^+_t)$. (Section 6.2)

Figure 2. Sustained Tracking of Human Motion

estimate as our "prediction" at time t. Detection at each time instance serve as "measurements". Figure 1(a) illustrates the concept. Thus, given $(\mathbf{X}^-_t, \Sigma^-_t)$ from frame-to-frame motion estimation, and the per-frame estimates of the same, $(\bar{\mathbf{X}}_t, \bar{\Sigma}_t)$, we fuse both to estimate the corrected configuration $(\mathbf{X}^+_t, \Sigma^+_t)$. The complete algorithm is described in Figure 2.

4. Per-Frame Detection

The per-frame detection approach described in this section takes an image $\mathbf{I}_t$ and estimates the configuration $(\bar{\mathbf{X}}_t, \bar{\Sigma}_t)$ for that image. The primary requirement for detectors is that the *false positives* be minimized, so as to reliably correct the configurations provided by the frame-to-frame motion estimation in Section 5. Like [17], we focus on detecting specific discriminate configurations. The set of configurations that can be detected may be increased at the expense of processing time.

From a corpus of labeled training data, we learn correlated appearances of each body part in isolation. The labeled data are aligned using procrustes analysis and the frequency of observing gradients over normalized (x, y) coordinate is learnt for each body part. This is recorded in a two dimensional histogram of frequencies, $\mathcal{A}_i$, each bin corresponding to an (x, y) location. Figure 3 shows the distribution for each body part, and distinctive shapes (e.g. parallel edges of lower arms) can be observed. The gradient information generated from the clothing of particular individuals in the corpus are damped out and the correlated gradients, such as the parallel edges of the lower arm, are amplified. With these appearance models for individual body parts, $\{\mathcal{A}\} = [\mathcal{A}_1, \cdots, \mathcal{A}_N]$, we can evaluate individual proposal configurations, by "sketching" the expected gradient map of the proposal. The likelihood of a given configuration is estimated by,

$$f(\mathbf{X}|\{\mathcal{A}\}; \Delta\mathbf{I}_t) = \prod_{(x,y)} e^{\Delta\mathbf{I}_t(x,y) \times \mathcal{A}_\mathbf{X}(x,y)}, \qquad (1)$$

where $\Delta\mathbf{I}_t$ is the gradient magnitude map of $\mathbf{I}_t$, and $\mathcal{A}_\mathbf{X}$ is the expected sketch of the configuration $\mathbf{X}$. The sketch $\mathcal{A}_\mathbf{X}$ takes the histograms of individual body part frequencies and transforms them to the location of the body part defined by $\mathbf{X}$. Figure 3(b) illustrates the concept, where sketches are generated for three different configurations. Note that the gradient information due to the background is suppressed before evaluation.

Another challenge in detection is efficient search of the $2P$ space of solutions. A naïve search of the space is likely to suffer from local minima and will be computationally expensive. Instead, we learn a low dimensional space corresponding to the specific body pose we are detecting. A compact linear subspace is learnt, $\mathcal{X}$, from the training set of configurations,$\{\mathbf{X}\}$, corresponding to the pose to be detected. Principal component analysis is used to find a low dimensional linear subspace.

$$\mathbf{X}^T = \begin{bmatrix} x^1_1 & \cdots & x^1_p \\ x^2_1 & \cdots & x^2_p \\ & \vdots & \\ x^n_1 & \cdots & x^n_p \end{bmatrix}, \qquad (2)$$

where each row of $\mathbf{X}^T$ represents $\mathbf{X}_t \in \mathrm{R}^{2p}$ of p human anatomical landmarks in one of the n training examples.

Taking the Singular Value Decomposition (SVD) of (mean compensated) $\mathbf{X}$, we get $\mathbf{X} = W\ D\ V^T$. We can then project the data-matrix $\mathbf{X}$ to a lower-dimensional subspace of k dimensions by retaining the first k singular values and setting the remaining to zero. Therefore, the compact linear space $\mathcal{X}$ is spanned by,

$$V_k = \begin{bmatrix} | & | & & | \\ v_1 & v_2 & \cdots & v_k \\ | & | & & | \end{bmatrix}. \qquad (3)$$

We can then construct the k-dimensional approximation of $\mathbf{X}_t \in \mathrm{R}^{2p}$ of p human anatomical landmarks,

$$\bar{\mathbf{X}}_t = \sum_{i=1}^{k} \bar{c}_i v_i. \qquad (4)$$

In experiments, we have found that a low dimensional (2 in our experiments) subspace suffices in describing the variation for a single pose. This variation appears closely related to the anthropometry of the actor (and does *not* correspond to an isotropic scaling). The most likely detection is then determined by bounded minimization (determined by anthropometric limits in the training set),

$$\bar{\mathbf{X}} = \arg\max_{\hat{\mathbf{X}} \in \mathcal{X}} f(\hat{\mathbf{X}}|\{\mathcal{A}\}). \qquad (5)$$

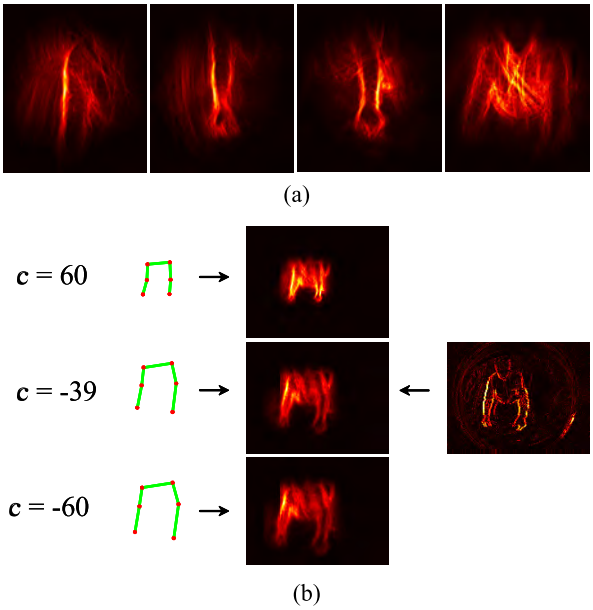


Figure 3. Configuration Sketching. (a) Appearance models for different body parts (left to right): left upper arm, left forearm, right lower arm, and torso. (b) Sketching the gradient map for three different configurations in the one-dimensional configuration space. The configuration best corresponding to the image gradient is for c = -39.

The covariance of the detection $\bar{\mathbf{X}}$ is then computed as follows,

$$\bar{\Sigma} = V_k V_k^T. \tag{6}$$

5. Frame-to-Frame Motion

To estimate frame-to-frame motion we use the algorithm for computing articulated motion in [3]. This approach adopts the cardboard model, proposed by Ju *et al.* in [12], where each rigid part of an articulated object as a plane connected at different joints. The motion of each plane is approximated by the six parameters of a 2D affine transformation. The set of affine transformations relating $\mathbf{I}_t$ and $\mathbf{I}_{t+1}$ are denoted as $\mathbf{A}_t \in \mathbf{R}^{6N}$. In the model used in this paper there are five planes related by four joints, i.e. $N = 5$ and $j = 4$. Articulation induces a set of linear constraints that $\mathbf{A}_t$ must satisfy, i.e. $\mathbf{\Theta}(\mathbf{X}_t)\mathbf{A}_t = \mathbf{0}$ (we follow the constraint described in [3]).

The transformations $\mathbf{A}_t$ are estimated by minimizing the sum of squared difference,

$$g(\mathbf{A}_t|\mathbf{X}_{t-1}, \mathbf{I}_{t-1}, \mathbf{I}_t) = \|w(\mathbf{I}_{t-1}; \mathbf{A}_t) - \mathbf{I}_t\|_2, \tag{7}$$

where $w(\cdot)$ is a warping function that transforms an image according to the affine transformations in $\mathbf{A}_t$[1]. Gauss-Newton minimization of this function yields an algorithm that iteratively solves a linear least squares system, $\mathbf{\Gamma}\hat{\mathbf{A}}_t = \mathbf{b}$ subject to $\mathbf{\Theta}(\mathbf{X}_t)\hat{\mathbf{A}}_t = \mathbf{0}$, where $\hat{\mathbf{A}}_t$ is the current estimate of $\mathbf{A}_t$. The matrix $\mathbf{\Gamma}$ is a block diagonal matrix where each block contains the gradient information to solve for the affine coefficient of a single body part. The 'interaction' between the motion of body parts is captured in the linear constraints. To solve the linearly constrained least squares problem, at each iteration a Karush-Kuhn-Tucker (KKT) system is solved,

$$\begin{bmatrix} \mathbf{\Gamma}^T\mathbf{\Gamma} & \mathbf{\Theta}^T \\ \mathbf{\Theta} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{A}}_t \\ \lambda \end{bmatrix} = \begin{bmatrix} \mathbf{\Gamma}^T\mathbf{b} \\ \mathbf{0} \end{bmatrix}. \tag{8}$$

6. Correction

In this section, we describe how to propagate the covariance across frames and how to obtain the corrected localization and covariance from the per-frame detection and frame-to-frame motion estimates.

6.1. Covariance Propagation

To propagate the covariance matrix at each time instance, we have to consider both the uncertainty introduced by solving the KKT system and the transformation induced by the affine motion. From [10] we have,

$$\Sigma'_{\mathbf{x}} = \mathrm{J}_{\mathbf{x}}\Sigma_{\mathbf{x}}\mathrm{J}_{\mathbf{x}}^{\top} + \mathrm{J}_{\mathbf{A}}\Sigma_{\mathbf{A}}\mathrm{J}_{\mathbf{A}}^{\top} \tag{9}$$

where

$$\mathrm{J}_{\mathbf{A}} = \begin{bmatrix} \frac{\partial u}{\partial a_1} & \cdots & \frac{\partial u}{\partial a_6} \\ \frac{\partial v}{\partial a_1} & \cdots & \frac{\partial v}{\partial a_6} \end{bmatrix} \tag{10}$$

$$= \begin{bmatrix} x_1 & y_1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & x_1 & y_1 & 1 \end{bmatrix}. \tag{11}$$

and

$$\mathrm{J}_{\mathbf{x}} = \begin{bmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{bmatrix} \tag{12}$$

$$= \begin{bmatrix} a_1 - 1 & a_2 \\ a_4 & a_5 - 1 \end{bmatrix}. \tag{13}$$

The matrix $\Sigma_{\mathbf{x}}$ is the covariance at the previous time instance and $\Sigma_{\mathbf{A}}$ is the inverse of the information matrix of the least squares system in Equation 8, i.e. $(\mathbf{\Omega}^{\top}\mathbf{\Omega})^{-1}$, where

$$\mathbf{\Omega} = \begin{bmatrix} \mathbf{\Gamma}^T\mathbf{\Gamma} \\ \mathbf{\Theta} \end{bmatrix}. \tag{14}$$

6.2. Update

Given the per-frame detection, $(\bar{\mathbf{X}}_t, \bar{\Sigma}_t)$, and the frame-to-frame prediction, $(\mathbf{X}_t^-, \Sigma_t^-)$, the update formulae are

[1] This requires the support of each affine transformation to be defined. This is done by specifying a rectangle around each pair of points.

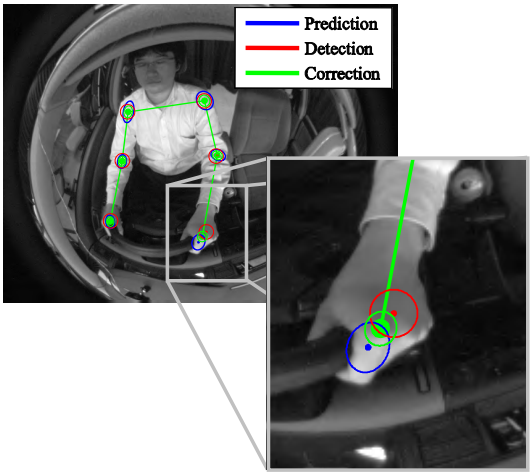


Figure 4. Update of configuration and covariance. The red dot and ellipse represent the per-frame detection, $(\bar{\mathbf{X}}_t, \bar{\Sigma}_t)$, the blue dot and ellipse represent the frame-to-frame prediction $(\mathbf{X}_t^-, \Sigma_t^-)$ and the green dot and ellipse represent the correct configuration and covariance $(\mathbf{X}_t^+, \Sigma_t^+)$.

used to obtain the corrected configuration and covariance $(\mathbf{X}_t^+, \Sigma_t^+)$. They may be computed as, [8],

$$\mathbf{X}_t^+ = \mathbf{X}_t^- + \Sigma_t^-(\Sigma_t^- + \bar{\Sigma}_t)^{-1}(\bar{\mathbf{X}}_t - \mathbf{X}_t^-), \quad (15)$$

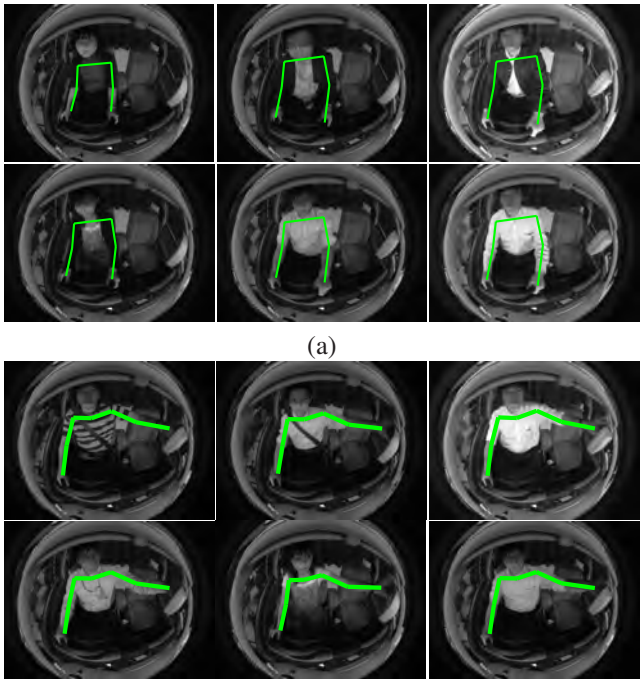
$$\Sigma_t^+ = (\mathbf{I} - \Sigma_t^-(\Sigma_t^- + \bar{\Sigma}_t)^{-1})\Sigma_t^-, \quad (16)$$

where $\mathbf{I}$ is an identity matrix. Figure 4 shows the fusion of location and covariance of the detection (red) and motion (blue) estimates, along with the corrected configuration (green).

7. Results

We have tested extensively on videos of drivers inside vehicles. The data set contained seven different subjects wearing (on average) four different outfits each. Each sequence contained actors engaged in seven typical driver behaviors (e.g. adjusting the rear view mirror, opening the glove compartment and changing gears) and the duration of each sequence was 3566 frames on average captured at 60 FPS at a resolution of 320×240. This algorithm was used to track the upper torso of person from video. Six points were tracked of which four were joints between five body parts. A least squares system of 30 unknowns and 8 linear constraint equation was solved. On a QuadCore 2.66 Ghz, with 4GB RAM, a C++ implementation of the system runs at 58 frames per second (at a resolution of 160×120).

We trained detectors for two poses, the neutral pose of the driver with both hands on the steering wheel and 'adjusting the rear view mirror' pose. For neutral poses, the detection was trained on a relatively small training set of twenty five labeled images (containing five of the seven subjects, each in three different outfits). The pose space was also



(a)

(b)

Figure 5. Detections for various individuals wearing a variety of different clothes in the neutral pose (a) and reaching for the rear view mirror (b).

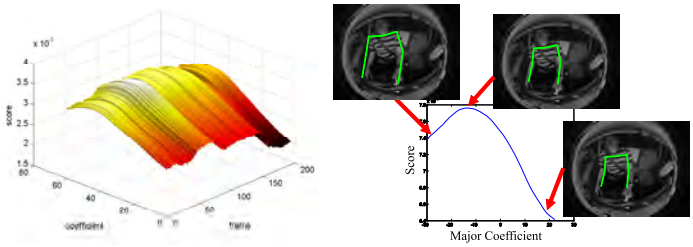


Figure 6. Detection score against various coefficient values across multiple frames (left). The function is well-behaved usually displaying a distinct unique maxima. The maxima corresponds to the correct configuration (right).

learnt from this small data set. Figure 5 shows the result on six images outside of the training set for the neutral pose and for the rear-view mirror pose. Figure 6 shows the cost function varying across time for different poses (in terms of the major projection coefficient). The cost function is smooth with a distinct maxima that corresponds to the correct configuration.

The frame-to-frame motion estimation algorithm is able to accurately track across hundreds of frames with only accumulated drift error. Figure 7 shows several sequences where the body is tracked in the absence of detections for hundreds of frames. Each row corresponds to frames from a single sequence. As the sequence progresses without detections, the uncertainty of the location accumulates (cor-

responding to drift error). The ability of the approach to recover from drift and localization errors is shown in Figure 8. The first row shows actors reaching for the rear view mirror and accumulating significant drift errors along the way. As the actors return to the neutral configuration, the detector is fired and this reduces the covariance. In the second and third rows, significant errors are introduced because of self-occlusion of the face with the shoulder. Once the actors return to neutral poses, the error in configurations is quickly corrected and covariance is controlled again.

8. Summary

In this paper, we propose an algorithm that fuses per-frame detections with frame-to-frame motion estimates for sustainable tracking of articulated objects, like humans. The detection algorithm learns part based appearance models of gradient information from training data. During detection, "sketches" of candidate configurations are compared against (background suppressed) gradient maps of the current image and the configuration corresponding to the most likely sketch is selected. During frame-to-frame motion estimation, an articulated motion algorithm is used that iteratively solves a linearly constrained least squares system. Frame-to-frame location and covariance estimates are updated (when detections are available) to produce corrected estimates of body configuration. This framework keeps estimation errors introduced by drift, occlusion or appearance variations under control. A C++ implementation of the algorithm runs at 58 frames per second and has been tested on over 30 sequences, each containing seven distinct operations performed over thousands of frames.

Acknowledgements

The research described in this paper was supported by the DENSO Corporation, Japan.

References

[1] Y. Bar-Shalom. Tracking and data association. *Academic Press Professional*, 1987.

[2] C. Bregler and J. Malik. Tracking people with twists and exponential maps. *IEEE International Conference on Computer Vision and Pattern Recognition*, 1998.

[3] A. Datta, Y. Sheikh, and T. Kanade. Linear motion estimation for systems of articulated planes. *IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.

[4] A. Fathi and G. Mori. Human pose estimation using motion exemplars. *IEEE International Conference on Computer Vision*, 2007.

[5] P. Felzenszwalb and D. Huttenlocher. Efficient matching of pictorial structures. *IEEE International Conference on Computer Vision and Pattern Recogition*, 2000.

[6] M. Fischler and R. Elschlager. The representation and matching of pictorial images. *IEEE Transactions on Computers*, 1973.

[7] D. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien, and D. Ramanan. Computational studies of human motion: Part 1, tracking and motion synthesis. *Foundations and Trends in Computer Graphics and Vision*, 2006.

[8] D. Forsyth and J. Ponce. Computer vision – a modern approach. *Prentice Hall*, 2003.

[9] D. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 1999.

[10] R. Hartley and A. Zisserman. Multiple view geometry in computer vision. *Cambridge University Press*, 2000.

[11] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. *European Conference on Computer Vision*, 1996.

[12] S. Ju, M. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated image motion. *Automatic Face and Gesture Recognition*, 1996.

[13] T. Moeslund, A. Hilton, and V. Krger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 2006.

[14] G. Mori, X. Ren, A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. *IEEE International Conference on Computer Vision and Pattern Recognition*, 2004.

[15] V. Pavlovic, james Rehg, and J. MacCormick. Learning switching linear models of human motion. *Neural Information Processing Systems*, 2000.

[16] V. Pavolvić, J. Rehg, T.-J. Cham, and K. Murphy. A dynamic bayesian network approach to figure tracking using learned dynamic models. *IEEE International Conference on Computer Vision*, 1999.

[17] D. Ramanan, D. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. *IEEE International Conference on Computer Vision and Pattern Recognition*, 2005.

[18] L. Ren, A. Patrick, A. Efros, J. Hodgins, and J. Rehg. A data-driven approach to quantifying natural human motion. *ACM Transactions on Graphics*, 2005.

[19] R. Rosales and S. Sclaroff. Learning body pose via specialized maps. *Neural Information Processing Systems*, 2002.

[20] Y. Sheikh, M. Sheikh, and M. Shah. Exploring the space of human motions. *IEEE International Conference on Computer Vision*, 2005.

[21] H. Sidenbladh, M. Black, and L. Sigal. Implicit probabilistic models of human motion for synthesis and tracking. *European Conference on Computer Vision*, 2002.

[22] L. Sigal and M. Black. Predicting 3d people from 2d pictures. *Conference on Articulated Motion and Deformable Objects*, 2006.

[23] C. Wren and A. Pentland. Dynamic models of human motion. *IEEE Proceedings of Automatic Face and Gesture Recognition*, 1998.

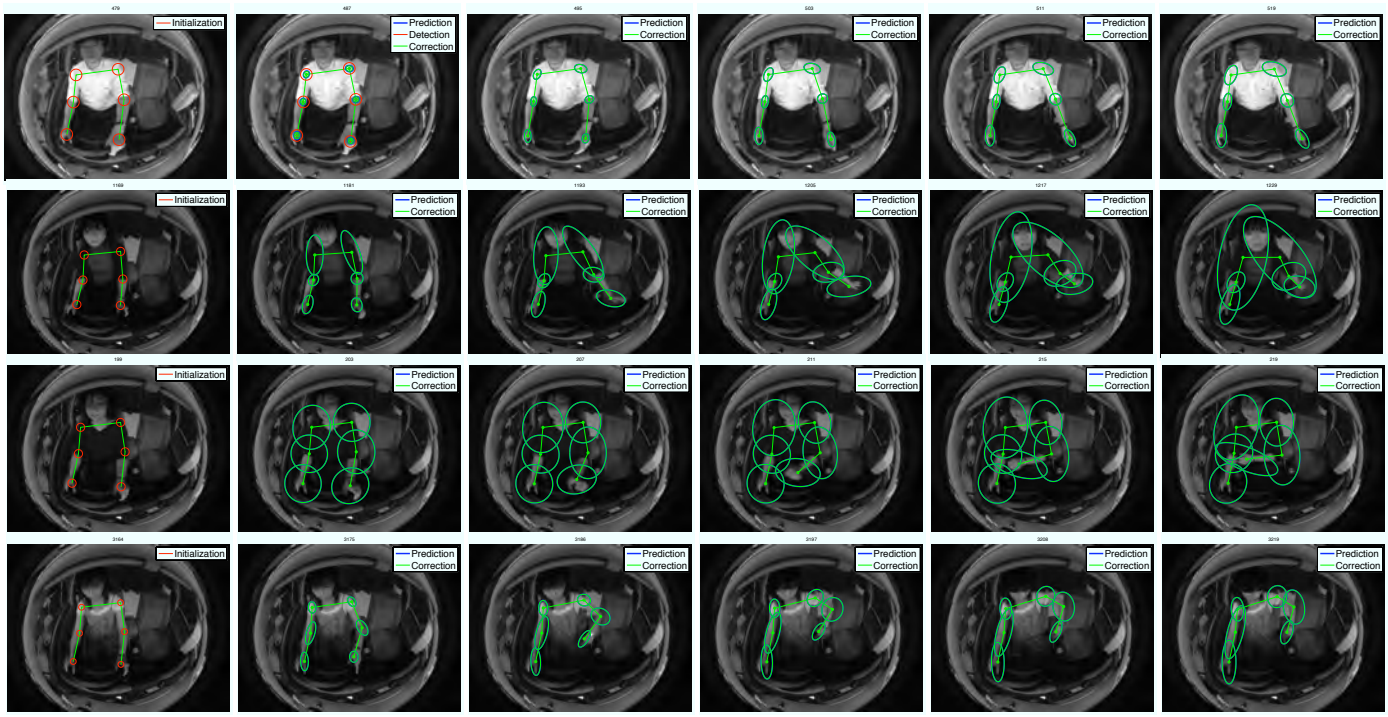


Figure 7. The detector automatically initializes the first frame in the neutral pose. As the actor leaves the neutral pose, the frame-to-frame motion estimate is the only cue for tracking. In the absence of detections, the covariance of the configuration steadily grows.

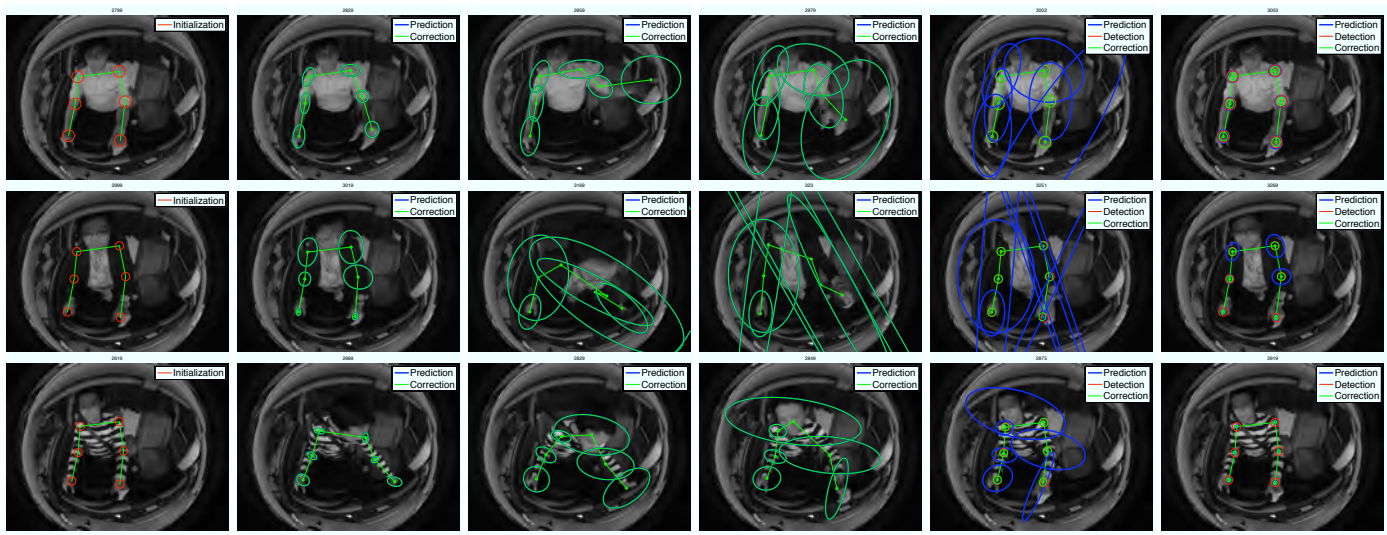


Figure 8. Recovery from drift and error. The first row shows the actor reaching for the rear view mirror and accumulating significant drift errors along the way. As the actor returns to the neutral configuration, the detector is fired and reduces the estimate covariance. In the second and third rows, significant errors are introduced because of self-occlusion of the face with the shoulder. Once the actors return to neutral poses, the error in configurations is quickly corrected and covariance is reduced.