STOCHASTIC MODELING OF MOTION TRACKING FAILURES

Shiloh L. Dockstader[†‡], *Nikita S. Imennov*[#], *and A. Murat Tekalp*[‡*]

[‡]Dept. of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627
[*]College of Engineering, Koç University, Istanbul, Turkey
[#]Dept. of Biomedical Engineering, University of Rochester, Rochester, NY 14627
URL: http://www.ece.rochester.edu/~dockstad/research/

ABSTRACT

This research introduces a new and effective method of predicting motion tracking failures and demonstrates its application towards the analysis of gait and human motion. We define a tracking failure as an event and describe its temporal characteristics using a hidden Markov model (HMM). This stochastic model is trained using previous examples of tracking failures and is applied to the Kalman-based tracking of a parametric, structural model of the human body. With an observation sequence derived from the noise covariance matrices of the structural model parameters, we show a causal relationship between the conditional output probability of the HMM and imminent tracking failures. Results are demonstrated on a variety of multi-view sequences of complex human motion.

1. INTRODUCTION

Numerous applications in video processing require the accurate and robust tracking of various objects, features, or models [1]. We focus on the application of gait analysis in which a number of relevant gait variables must be extracted from a moving structural model of the human body [2]. The end objective is to track the model of the body long enough to extract useful gait variables while ensuring sufficient accuracy in the extracted information. We limit the scope of this contribution to the latter issue, which in turn requires an understanding of tracking failures, but refer the reader to [3] for a detailed presentation of an appropriately accurate and robust tracking algorithm.

Although few algorithms can boast unlimited tracking duration and accuracy, little attention has focused on the corresponding detection, prediction, and analysis of terminal failures. Pasqual *et al*. [4] introduce an algorithm that explicitly addresses the uncertainty of tracking. They suggest a method of feature substitution using optical flow, texture, and implicit depth and then switch modalities as needed to prolong and enhance tracking performance. Darrell *et al*. [5] present an interesting approach to tracking using depth estimation, color segmentation, and intensity pattern classification. The method effectively increases tracking robustness using multiple modalities, but does not explicitly address the detection or prediction of tracking failures. In an attempt to leverage the knowledge of tracking failures, Shearer *et al*. [6] employ complementary region- and edge-based algorithms for tracking objects. The approach includes simultaneous monitoring for tracking confidence and uses this information to share data between the two algorithms. The method of detecting failures, however, is quite fundamental, not entirely robust, and does not lend itself to prediction.

The majority of previous research in the field of failure detection and prediction has occurred not in the vision community, but elsewhere [7]. Dobra and Festila [8] develop a technique for detecting failures based on coefficient changes and statistical decision methods. Mehra *et al*. [9] use an Interacting Multiple Model Extended Kalman Filter (IMM-EKF) to detect and identify failure modes. The method represents each failure mode by a model and combines the outputs of the models to detect failures. With applications to control theory, Doraiswami *et al*. [10] propose a three-stage process for failure detection and isolation. The method isolates faults by computing the maximum correlation between residual measurements and estimates of the residual generated using a number of failed hypotheses.

In contrast to earlier research, this work introduces an original method for the accurate prediction of model-based tracking failures using a stochastic, temporal model. The approach assumes a structural model in which the parameters are being tracked with some inherent confidence, as would be the case with a Kalman-based tracker. Using features derived from the parameters' noise covariance matrices, and assuming a Kalman infrastructure, we build a time-varying observation vector that we hypothesize will correlate with the occurrence of tracking failures. This observation vector provides the input to a carefully constructed hidden Markov model [11][12] that is initially trained on and later used to identify sequences of frames immediately preceding various tracking failures.

2. THEORY

The proposed method of modeling tracking failures is built upon an occlusion-adaptive, multi-view algorithm for feature tracking [3]. The tracking is applied to a 3-D structural model of the human body and characterized by stochastic kinematic constraints that limit the variability and improve the accuracy of the underlying body configuration estimates [2].

2.1 Structural Modeling

The suggested structural model employs fifteen parameters ($p_1 \ldots p_{15}$) that are measured in three-dimensional, body-centered coordinates, as indicated in Figure 1. The origin of the coordinate system, p_0, corresponds to a fixed position on the 3-D model.

[†] Corresponding Author: dockstad@ieee.org

The time-varying coordinate axes are uniquely determined at each frame, k, by interpreting the velocity of the origin. We assume that during a normal gait cycle, the body moves forward, tangential to the transverse (x-y) and saggital (x-z) planes and orthogonal to the coronal plane (y-z).

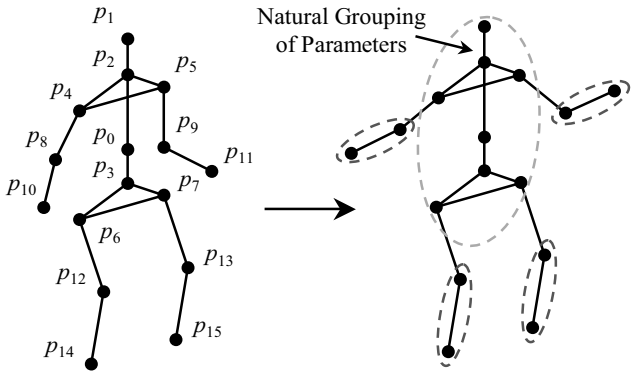


Figure 1. Structural model of the human body showing the natural grouping of model parameters.

The underlying tracking algorithm is implemented using a Kalman filter due to its convenient application of dynamics via linear systems theory. Any number of techniques might also be considered, however [13]. We introduce a time-varying state vector as

$$\boldsymbol{\sigma}[k] \equiv \left[\boldsymbol{\sigma}_1[k]\ \boldsymbol{\sigma}_2[k]\ \cdots\ \boldsymbol{\sigma}_m[k]\ \cdots\ \boldsymbol{\sigma}_{2N+1}[k]\right]^T . \quad (1)$$

Here, $\boldsymbol{\sigma}_m[k]$, $m \le N$ denotes the 3-D position of the m^{th} parameter in our body-centered coordinate system, while

$$\boldsymbol{\sigma}_{m+N+1}[k] = \left.\frac{\partial \boldsymbol{\sigma}_m[k]}{\partial k}\right|_{m \le N} \quad (2)$$

indicates an approximation of the true velocity of the m^{th} parameter. The corresponding state equation is given by

$$\hat{\boldsymbol{\sigma}}[k] = \boldsymbol{\Psi}[k]\hat{\boldsymbol{\sigma}}[k-1] . \quad (3)$$

We develop an error covariance matrix, $\boldsymbol{\Gamma}[k \mid k-1]$, that depicts our confidence in the predictions of the state estimates. The update equation is indicated by

$$\boldsymbol{\Gamma}[k \mid k-1] = \boldsymbol{\Psi}[k]\boldsymbol{\Gamma}[k-1]\boldsymbol{\Psi}^T[k] + \mathbf{Q}[k] , \quad (4)$$

where $\mathbf{Q}[k]$ represents a Gaussian noise covariance matrix which is iteratively modified over time to account for the deviations between the predictions and corrections of the state estimates. The system then defines a Kalman gain matrix, $\mathbf{D}[k]$, according to

$$\mathbf{D}[k] = \boldsymbol{\Gamma}[k \mid k-1]\boldsymbol{\Phi}^T[k]\left(\boldsymbol{\Theta}[k] + \boldsymbol{\Phi}[k]\boldsymbol{\Gamma}[k \mid k-1]\boldsymbol{\Phi}^T[k]\right)^{-1} , \quad (5)$$

where $\boldsymbol{\Phi}[k] = \left[\mathbf{I} \vdots \mathbf{0}\right]_{N \times (2N+1)}$ indicates the linear observation matrix and $\boldsymbol{\Theta}[k]$ is a recursively updated observation noise covariance matrix. The remaining steps of the trajectory filtering include

$$\hat{\boldsymbol{\sigma}}[k] = \hat{\boldsymbol{\sigma}}[k \mid k-1] + \mathbf{D}[k]\left(\hat{\mathbf{y}}[k] - \boldsymbol{\Phi}[k]\hat{\boldsymbol{\sigma}}[k \mid k-1]\right) \quad (6)$$

and

$$\boldsymbol{\Gamma}[k] = \left(\mathbf{I} - \mathbf{D}[k]\boldsymbol{\Phi}[k]\right)\boldsymbol{\Gamma}[k \mid k-1] , \quad (7)$$

where $\hat{\mathbf{y}}[k]$ is a vector of three-dimensional, image-derived observations. It is this final noise covariance at each frame that lays the foundation for the prediction of tracking failures. For a more thorough treatment of this theory, including an extension to gait variable extraction, we refer the reader to [2] and [3].

2.2 Tracking Failure Definition

A terminal tracking failure is defined as an immediate and sustained loss in tracking accuracy at one or more structural model parameters. The acceptable magnitude of such a loss is application dependent and, in the case of this research, driven by the ultimate accuracy required in gait variable extraction. We quantify an immediate loss via the distance between a parameter's estimated and ground-truth values and a sustained loss via the RMS error between the same measures taken over a period of T frames. We designate for the m^{th} model parameter p'_m and p''_m as the error thresholds for immediate and sustained losses in tracking accuracy, respectively.

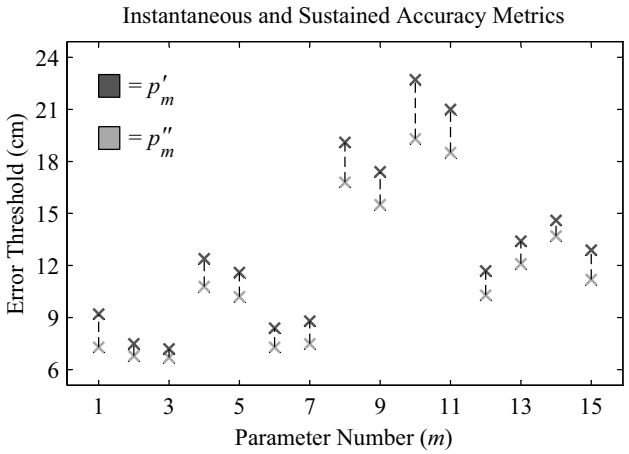


Figure 2. Experimentally derived parameters for instantaneous and sustained tracking accuracy thresholds.

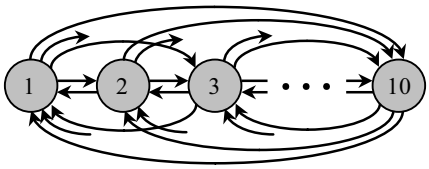
An analysis of typical image sequences used for tracking and interpreting human motion shows that most failures are preceded by $T \equiv 30$ or fewer frames of relevant data. The tracking accuracy parameters, p'_m and p''_m, corresponding to such failures are empirically derived and summarized in Figure 2.

2.3 Markov Modeling

Tracking failures are not easily characterized by the changing positions of model parameters over the course of time, but are correlated, however, with temporal changes in noise covariance measurements. The matrix $\boldsymbol{\Gamma}[k]$ is a $(2N+1)\times(2N+1)$ block diagonal matrix in which the 3×3 matrices along the diagonal of the first $N \times N$ quadrant represent the 3-D noise distributions for each of the m model parameters. Let the determinant of the m^{th} matrix at frame k be denoted by $o_m[k] = \left|\boldsymbol{\Gamma}_m[k]\right|$, and let

$$\mathbf{o}_k \equiv \left[o_1[k]\ o_2[k]\ \cdots\ o_m[k]\ \cdots\ o_N[k]\right]^T \quad (8)$$

indicate the vector observation for the entire structural model at frame k. A corresponding observation sequence extracted over T adjacent frames is denoted by $\mathbf{O} = \left(\mathbf{o}_1\ \mathbf{o}_2\ \cdots\ \mathbf{o}_T\right)$.



Number of States, $R = 10$; Observation Length, $T = 30$

Discrete Alphabet [Codebook] Size, $M = 2^{10} = 1024$

Tracking Failure Likelihood, $\Pr\left[(\mathbf{o}_{k-T+1}, \mathbf{o}_{k-T+2}, \cdots, \mathbf{o}_k) \mid \lambda_S\right]$

Figure 3. Hidden Markov model.

We introduce a nearly ergodic HMM, $\lambda_S = (A, B, \zeta)$, to describe the stochastic properties associated with tracking failures. Using sequences of length $T = 30$ and an observation vector of dimension $N = 15$, one finds in practice that the vector observations are conveniently clustered into $M = 1024$ discrete symbols, thus yielding a 10-bit HMM codebook. The number of states, R, used by the model is motivated by the articulated structure of the underlying body model. In the majority of training sequences, failures show a dependency on the confidence of the torso parameters as well as the number of accompanying body limbs being successfully tracked. This is in contrast to a dependency on the actual parameters of the various body limbs. Thus, an appropriate number of states is the minimum best describing this phenomenon, or $R = 2 \cdot 5 = 10$ (the torso + zero to four limbs). This grouping of body parameters and the corresponding HMM topological description are illustrated in Figures 1 and 3, respectively.

Estimating the remaining parameters of λ_S is performed using the Baum-Welch method. The only restriction on the topology of the HMM is that, where $A = \{a_{ij}\}$,

$$a_{ij} = 0, \; i > 5, j < 6. \tag{9}$$

This constraint allows for the progression of certain noise covariance changes from which the algorithm cannot recover (e.g., a tracking degradation within the parameters of the torso). With this restriction, the model is estimated according to

$$\hat{\lambda}_S = \arg\max_{\lambda_S} \left(\prod_i \Pr\left[\mathbf{O}^{(i)} \mid \lambda_S\right] \right), \tag{10}$$

where $\mathbf{O}^{(i)}$ is the i^{th} observation training sequence. The optimization procedure supports the initial choices of $R = 10$, $M = 1024$, and $T = 30$ by producing, at least, a local maximum at these values. The initial state distribution, ζ, the state-transition probabilities, A, and the observation symbol probabilities, B, are then calculated accordingly. Initial values for these parameter estimates are based on a uniform distribution.

2.4 Tracking Failure Prediction

For all sequences introduced after the initial training set, the measure $\Pr\left[(\mathbf{o}_{k-T+1}, \cdots, \mathbf{o}_k) \mid \lambda_S\right]$ or, alternatively, $\Pr\left[\mathbf{O} \mid \lambda_S\right]$ may be used to test the likelihood that the observation, in which $T = 30$, was produced by λ_S. A greater likelihood suggests a more confident measure that, in turn, implies a greater correlation between the observed sequence and those known to correspond with imminent tracking failures. A simple threshold placed on this output probability is a sufficient mechanism for detecting or predicting such events. Thus, where a single fixed-state sequence is denoted by $\mathbf{r} = (r_1 \; r_2 \; \cdots \; r_T)$, we have

$$\Pr\left[\mathbf{O} \mid \lambda_S\right] \equiv \sum_{\text{all } \mathbf{r}} \Pr\left[\mathbf{O} \mid \mathbf{r}, \lambda_S\right] \cdot \Pr\left[\mathbf{r} \mid \lambda_S\right] \underset{\text{No Failure}}{\overset{\text{Failure}}{\gtrless}} \lambda'_S. \tag{11}$$

The above output probability is estimated using the well-known forward estimation procedure.

3. EXPERIMENTAL RESULTS

To test the contribution we collect a number of video sequences of complex human motion, captured at $\Delta t = 1/30$ sec using three or four views of a single scene. Training of both the HMM and the vector quantization scheme is based on approximately 45000 frames of relevant video data, while testing is based on approximately 15000 frames of data. Ground truth measurements are based on a number of methods, including the use of markers placed on the body as well as manual interpretation of feature points. Figure 4 illustrates tracking results from several views immediately preceding a correctly detected failure. The failure is shown in the last row of the figure; the upper rows present earlier frames, thus demonstrating the progression of the failure.

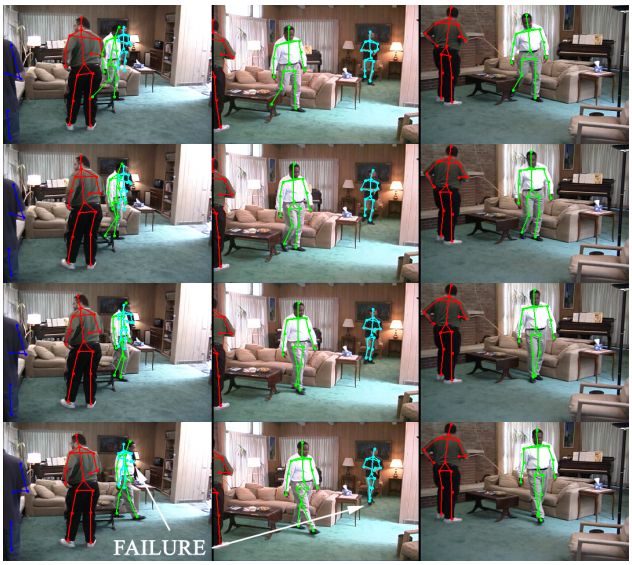


Figure 4. Tracking failure progression.

The output probability, $\Pr\left[\mathbf{O} \mid \lambda_S\right]$, of the HMM for a particular sequence is shown as a function of time in Figure 5. The vertical dashed lines show known tracking failures, according to our earlier definition, while the gray curve shows the HMM output. There is little correlation between the failures and the output taken more than 15-30 frames in advance. However, as the tracking failure draws closer, the characteristic changes in the Kalman noise covariance measurements drive the output probability to a more deterministic and correlated state.

To demonstrate the existence of a temporal correlation in the noise covariance features and, ultimately, the utility of the Markov assumption, we construct a more fundamental metric for

comparison. In particular, we consider $\|\mathbf{o}_k\|$ at each frame k. In both cases, we develop a simple threshold, λ_S', one for $\|\mathbf{o}_k\|$ and another for $\Pr[\mathbf{O}|\lambda_S]$, that maximizes detection accuracy given a specificity of 99% or greater. The goal here, of course, is to minimize the number of false negatives (i.e., missing actual failures). Using such a threshold for both cases yields the results shown in Table 1. The proposed Markov scheme produces a maximum accuracy of nearly 98%, while the alternate metric generates only 88%. In the instance of the alternate metric, $\|\mathbf{o}_k\|$, the cost of maintaining a specificity of 99% is a lower threshold yielding far too many false positives.

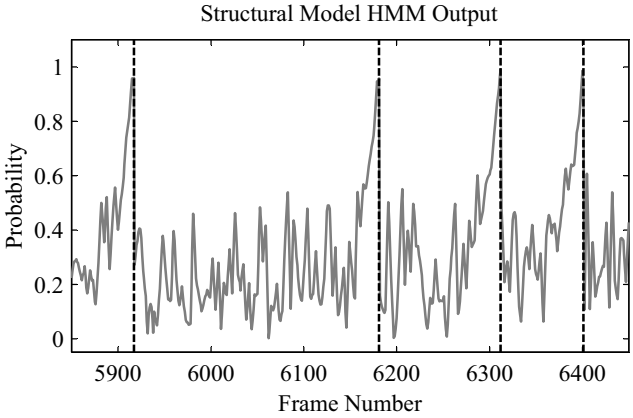


Figure 5. HMM conditional output probability taken as a function of frame number.

A secondary, and somewhat surprising, result is that in the case of the HMM, the chosen threshold flags a detection 7-8 frames before the failure nearly 40% of the time. A closer inspection shows a failure detection within 5-10 frames approximately 80% of the time. This suggests that the HMM could be used to not only detect failures, but to predict them as well.

	TP	*TN*	*FP*	*FN*	$\frac{TP}{TP+FP}$ *(Sens)*	$\frac{TN}{TN+FN}$ *(Spec)*	$\frac{TP+TN}{TP+FP+TN+FN}$ *(Accuracy)*
λ_S	1192	12710	204	102	85.4%	99.2%	97.8%
$\|\mathbf{o}_k\|$	1185	11323	1591	109	42.7%	99.0%	88.0%

Table 1. Quantitative results comparing proposed HMM architecture to a more fundamental approach.

4. CONCLUSIONS

This research introduces a new method of detecting and predicting motion tracking failures using hidden Markov modeling. The approach defines a failure as an event and uses the output probability of a trained HMM to detect, and even predict, such events. The vector observations for the model are derived from the time-varying noise covariance matrices of a Kalman filter that tracks the parameters of a structural model of the human body. The results clearly show the correlation between the proposed Markov metric and subsequent tracking failures as well as the utility of the Markov model over a more fundamental approach. The proposed theory is demonstrated on several multi-view sequences of complex human motion in support of various applications in gait and motion disorder analysis.

5. ACKNOWLEGMENTS

This research was supported in part by grants from the Center for Future Health, CEIS, Eastman Kodak Company, and the RUBI program (NSF-REU# EEC-0097470). The authors also thank Dr. Michel J. Berg for his insightful comments and suggestions.

6. REFERENCES

[1] J. Davis and S. Taylor, "Analysis and recognition of walking movements," *Proc. of the Int. Conf. on Pattern Recognition*, Québec City, Canada, 11-15 August 2002, pp. 315-318.

[2] S. L. Dockstader, K. A. Bergkessel, and A. M. Tekalp, "Feature extraction for the analysis of gait and human motion," *Proc. of the Int. Conf. on Pattern Recognition*, Québec City, Canada, 11-15 August 2002, pp. 1441-1455.

[3] S. L. Dockstader and A. M. Tekalp, "Multiple Camera Tracking of Interacting and Occluded Human Motion," *Proceedings of the IEEE*, vol. 89, no. 10, pp. 1441-1455, October 2001.

[4] A. A. Pasqual, K. Aizawa, and M. Hatori, "Use of multiple visual features for object tracking," *Proc. of SPIE*, San Jose, CA, 25-27 January 1999, vol. 3653, pp. 946-955.

[5] T. Darrell, G. Gordon, M. Harville, and J. Woodfill, "Integrated Person Tracking Using Stereo, Color, and Pattern Detection," *Int. J. of Computer Vision*, vol. 37, no. 2, pp. 175-185, June 2000.

[6] K. Shearer, K. D. Wong, and S. Venkatesh, "Combining Multiple Tracking Algorithms for Improved General Performance," *Pattern Recognition*, vol. 34, no. 6, pp. 1257-1269, June 2001.

[7] V. B. Vagapov, "Determination of Tracking Failure Probability in Multidimensional Radio Tracking Systems," *Radiotekhnika*, vol. 37, no. 12, pp. 40-42, December 1982.

[8] P. Dobra and C. Festila, "Fault detection and diagnosis for continuous time process," *Proc. of the IFAC Workshop on Intelligent Manufacturing Systems*, Bucharest, Romania, 24-26 October 1995, pp. 389-393.

[9] R. Mehra, C. Rago, and S. Seereeram, "Autonomous failure detection, identification and fault-tolerant estimation with aerospace applications," *Proc. of the IEEE Aerospace Conf.*, Aspen, CO, 21-28 March 1998, vol. 2, pp. 133-138.

[10] R. Doraiswami, C. P. Diduch, and J. Kuehner, "Failure detection and isolation: A new paradigm," *Proc. of the American Control Conf.*, Arlington, VA, 25-27 June 2001, vol. 1, pp. 470-475.

[11] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-286, February 1989.

[12] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden Markov model," *Proc. of the Conf. on Computer Vision and Pattern Recognition*, Champaign, IL, 15-18 June 1992, pp. 379-385.

[13] M. Isard and A. Blake, "Condensation - Conditional Density Propagation for Visual Tracking," *Int. J. of Computer Vision*, vol. 29, no. 1, pp. 5-28, 1998.