

TRACKING THE PATH SHAPE QUALITIES OF HUMAN MOTION

Kai Tu, Harvey Thornburg, Matthew Fulmer, and Andreas Spanias

Arts, Media and Engineering
Arizona State University
699 S. Mill Ave. Tempe, AZ 85287

ABSTRACT

We propose a probabilistic generative model for extracting intended path shape qualities of an object moving under human control in real time. At each instant, we decide whether the object is moving in a straight, curved, or random path, or whether it has stopped moving. Our model incorporates sensor noise as well as human imperfections in the intended motion. As well as tracking the object’s position, velocity, and motion direction, we compute the posterior probability of each shape quality hypothesis given all sensed-data in the horizon $[t - N + 1, t]$; the hypothesis maximizing this posterior is taken as the decision. The posterior is computed using the unscented Kalman filter (UKF), as our model is inherently nonlinear. The path-shape quality tracking is successfully embedded in a hybrid physical-digital interface where the position of an illuminated ball, sensed by a low-cost video camera array, triggers multimodal feedback in a mediated learning environment. We show successful results on a variety of real-world motion paths where the participant is given only verbal descriptions of how to move. Our generative model is further validated by user studies involving a simple color-based interaction, where participants discover shape quality controls as they interact.

Index Terms— Activity Analysis, Computer Vision, Human-Computer Interaction, Human Motion Analysis, Hybrid Physical-Digital Environment, Multimedia Signal Processing, Natural Information Interface, Unscented Kalman Filter, Video Sensing

1. INTRODUCTION

Natural information interfaces and hybrid physical-digital environments – where human movement in real physical spaces triggers software-based interaction – have received much attention in the interactive media and gaming worlds [3, 11, 9, 2, 1]. Applications range from the exploration of complex data sets [3], interactive dance performance [9], and stroke patient rehabilitation [5]. These systems prove advantageous as they eschew the physical encumbrances of traditional mouse-keyboard interaction, engage bodily-kinesthetic intelligence [4] and foster collaborative and social modes of interaction [1].

One such interface is SMALLab [1], a physically situated multimedia learning environment consisting of a 15’x15’x12’ open physical space surrounded by a loudspeaker array, a ceiling-mounted floor-projection system for multimodal feedback, and a video camera array for visual sensing. Students interact by guiding illuminated objects through the interior of the physical space. Object locations are sensed by the camera array at eight frames/sec, triggering multimodal (audio and visual) feedback.

A key challenge with SMALLab and related interfaces is that the familiar location-based controls of mouse-keyboard interaction may not be well-adapted to large physical spaces, especially where

portions of the visual feedback domain lie outside the participant’s field of vision. This situation presents a high cognitive load, as the participant must form an internal visual-spatial representation of the space as a whole [8, 10] while simultaneously attending to his/her movement and the multimodal feedback. As such, we desire control strategies which are more closely coupled with “felt” physical movement, for instance gesture-based control [11, 9, 2]. Unfortunately, human gesture recognition in SMALLab and other environments is currently infeasible due to adverse lighting conditions, the costs of high-quality video sensing, and the large number of participants. Instead, we propose controls which are closely allied with “felt” human movement, and observable from illuminated object motion. Examples include *path shape qualities* (straight, curved, random) or *dynamics qualities* (fast, slow). Since the meaning of “fast” or “slow” is highly context-dependent, we focus on shape qualities.

At each time t , we detect whether the object’s motion in $[t - N + 1, t]$ follows a straight, curved, or random path, or whether the object is not moving. We pursue a probabilistic approach based on a generative model for each “straight”, “curved”, “random”, or “stop” hypothesis. This model is developed in Section 2.1. Based on the generative model, we compute the posterior probability for each hypothesis as well as approximate minimum mean-square error (MMSE) “tracking” estimates of object location and motion direction, speed, and curvature, given sensed-data in $[t - N + 1, t]$ (Section 2.2). Tracking estimates may control secondary features of the interaction; e.g., average speed as a rough measure of the physical activity level of the participant can adapt the cognitive load of the multimodal feedback or trigger specific events which stimulate further activity. Section 3 discusses experimental results and user studies validating the generative model proposed in Section 2.1.

2. PROPOSED METHOD

Let $Y_{t-N+1:t}$ denote sensed-data location observations in the horizon $[t - N + 1, t]$ ¹. Each $Y_t \in \mathbb{R}^2$ consists of horizontal planar coordinates. Let \mathcal{H}_t denote one of four shape quality hypotheses: $\mathcal{H}_t = \text{‘S’}$ “straight”, ‘C’ “curved”, ‘R’ “random”, and $\mathcal{H}_t = 0$ “stop”. We compute \mathcal{H}_t^* as the minimum-error, or Bayes decision:

$$\mathcal{H}_t^* = \underset{\mathcal{H}_t}{\operatorname{argmax}} P(\mathcal{H}_t)P(Y_{t-N+1:t}|\mathcal{H}_t) \quad (1)$$

Modeling $P(\mathcal{H}_t)$ as uniform, the Bayes decision requires only the likelihood $P(Y_{t-N+1:t}|\mathcal{H}_t)$. A generative model for this likelihood, developed in Section 2.1, encodes information about motion dynamics and sources of uncertainty for each hypothesis. Should any of

¹Here we adopt the conventions: $i : j$ denotes the set of consecutive integers between i and j inclusive; $X_{i:j}$ denotes the sequence of values X_i, X_{i+1}, \dots, X_j .

this information change, the error-optimal solution (1) automatically adjusts.

2.1. Probabilistic generative motion model

Straight and curved hypotheses are modeled as special cases of *directional motion*, where random and stop hypotheses are modeled as Brownian motion. We model directional motion as follows. Let $l_t \in \mathbb{R}^2$ be the inherent object location. Observed location equals inherent location plus Gaussian noise:

$$Y_t \sim \mathcal{N}(l_t, \lambda_Y I) \quad (2)$$

Now, let $\theta_t \in [-\pi, \pi]$ be the motion direction, $v_t \in \mathbb{R}$ be the speed, and $\omega_t \in \mathbb{R}$ be the instantaneous path curvature. Curvature is the rate of direction change per arc length, and speed the rate of arc length increase per unit time; hence

$$\begin{aligned} \theta_t &= \theta_{t-1} + v_t \omega_t \\ l_t &\sim \mathcal{N}\left(l_{t-1} + v_t \begin{bmatrix} \cos \theta_t \\ \sin \theta_t \end{bmatrix}, v_t^2 \lambda_L^{(DIR)} I\right) \end{aligned} \quad (3)$$

The Gaussian uncertainty in (3) accounts for spurious movements; we expect the magnitude of such movements to be proportional to the speed, which accounts for the variance term $v_t^2 \lambda_L^{(DIR)}$.

The difference between straight and curved motion is solely a function of curvature, as opposed to speed. For both hypotheses, we expect speed to vary continuously in a manner proportional to its value; i.e.:

$$\log v_t \sim \mathcal{N}(\log v_{t-1}, \lambda_v) \quad (4)$$

Hence, the straight/curved discrimination reduces to the probabilistic specification of the process $\{\omega_t\}$, including the initial value ω_0 . For straight motion $\omega_t \approx 0$; for curved ω_t differs noticeably from zero. Furthermore, we expect $\{\omega_t\}$ to vary continuously. We propose three additional considerations as a result of preliminary user studies. First, absent state and observation noise, the straight/curved discrimination should be *scale-invariant*. Second, path scale should be governed solely by speed. The first two conditions mean: consider “primed” and “unprimed” realizations (2, 3, 4) with inputs $\{v'_t, \omega'_t\}$ vs. $\{v_t, \omega_t\}$, $\lambda_L^{(DIR)} = \lambda_Y = 0$, $v'_t = K v_t$, and all other parameters identical. Then ω'_t and ω_t should relate in a way such that $Y'_t = K Y_t$, and the straight/curved discrimination should be unaffected. Third, by physical constraints curvature cannot become arbitrarily large. Hence “snaked” motions, where the variance of $\{\omega_t\}$ is bounded, (Fig. 1) should be favored.

It is easily shown that for any input sequence $\{\delta_t\}$, the following satisfies the scale-invariant conditions

$$\omega_t = \omega_{t-1} + v_t^{-1} \delta_t \quad (5)$$

For the condition regarding snaked motion, we let $\{\delta_t\}$ be the result of passing zero-mean Gaussian white noise $\{\eta_t\}$ through a bandpass filter with transfer function

$$H(z) = \frac{(1 - p_L)(1 - p_H)(1 - z^{-2})}{4(1 - p_L z^{-1})(1 - p_H z^{-1})} \quad (6)$$

This filter has zeros at DC ($z = 1$) and Nyquist ($z = -1$), and poles at $z = p_L, p_H$ where $p_L = F(\phi_L)$; $p_H = F(\phi_H)$, $F(\phi) = (1 - \sin \phi)/(\cos \phi)$, and ϕ_L and ϕ_H are the lower and higher cutoff frequencies of the bandpass characteristic.

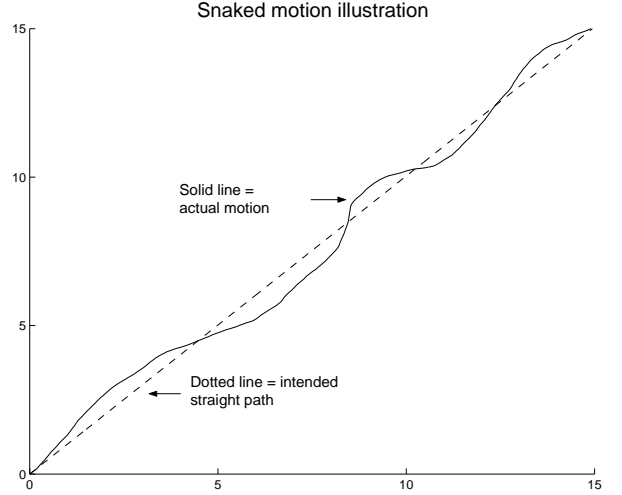


Fig. 1. Example of snaked motion. The dotted line represents the intended path, the solid line the actual path.

As we will show, the zero at DC guarantees that $\{\omega_t\}$ has bounded variance, at least when $\{v_t\}$ is constant. If $v_t = v_0 \forall t$, (5) and (6) imply that $\{\omega_t\}$ is Gaussian white noise $\{\eta_t\}$ processed by a filter with transfer function

$$H_\omega(z) = \frac{(1 - p_L)(1 + p_H)(1 + z^{-1})}{4v_0(1 - p_L z^{-1})(1 - p_H z^{-1})} \quad (7)$$

If both $|p_L|$ and $|p_H|$ are strictly less than 1 (guaranteed by the mapping $F(\cdot)$, when $\phi_L, \phi_H \in (0, \pi)$) it follows that $\{\omega_t\}$ is a Gauss-Markov process with finite power spectral density [7]. Hence, the variance of $\{\omega_t\}$ is bounded.

Thus, we distinguish straight vs. curved hypotheses by choices of λ_η (the variance of the white Gaussian process $\{\eta_t\}$) as well as initial curvature $P(\omega_0)$. By scale invariance, the same choices of λ_η hold in a variety of situations – for instance, when the user stands still and guides the object by hand around his/her body, or when the user carries the object through the space.

The upper half of Figure 2 displays sample straight vs. curved process realizations using $\phi_L = 0.1$ and $\phi_H = 0.5$ rad/frame, where the frame rate is eight frames/sec. Here

$$\lambda_\eta = \begin{cases} 30.0, & \mathcal{H} = \text{'C' } \\ 0.5, & \mathcal{H} = \text{'S' } \end{cases} \quad (8)$$

and $P(\omega_0) = \mathcal{N}(0, v_0^{-1} \lambda_\eta)$ in either case.

By contrast, random and stop hypotheses are modeled as Brownian:

$$l_t \sim \mathcal{N}\left(l_{t-1}, \lambda_L^{(B)} I\right) \quad (9)$$

where $\lambda_L^{(B)}$ is much greater for random motion ($\mathcal{H} = \text{'R'}$) than for stopped ($\mathcal{H} = 0$). Unlike straight/curved discrimination, random/stop discrimination is highly scale-dependent. Thus, $\lambda_L^{(B)}$ must be set roughly to the square of the expected per-frame drift $\|l_t - l_{t-1}\|$. The lower half of Figure 2 shows sample process realizations from the generative models for each hypothesis. Here $\lambda_L^{(B)} = 0.25$ for random motion and 0.5 for stop motion.

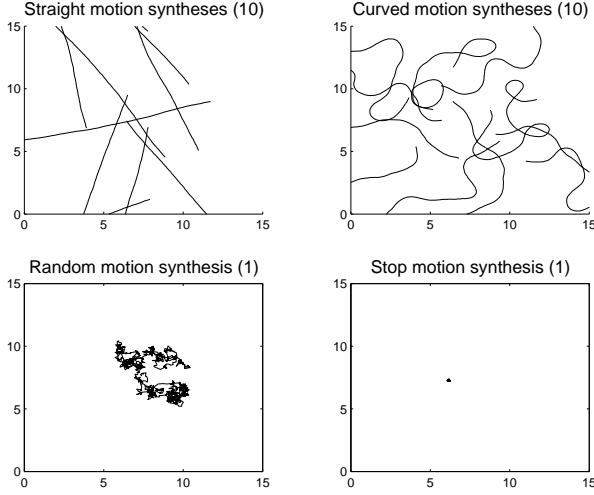


Fig. 2. Sample process realizations for straight, curved, random, and stop shape quality hypotheses.

2.2. Inference methodology

Via (1), shape quality determination requires computation of $P(Y_{t-N+1:t}|\mathcal{H}_t)$. Let

$$S_t \triangleq \{l_t, \theta_t, v_t, \omega_t, \delta_t, \delta_t^{(1)}, \eta_t^{(0:1)}\} \quad (10)$$

where $\delta_t^{(1)} \in \mathbb{R}$, $\eta_t^{(0:1)} \in \mathbb{R}^2$ are auxiliary states necessary to model $\{\delta_t\}$ as a first-order Gauss-Markov dependence.

Let $t_0 \triangleq t - N + 1$. We begin by factoring

$$P(Y_{t_0:t}|\mathcal{H}_t) = P(Y_{t_0}|\mathcal{H}_t) \times \prod_{s=t-N+2}^t P(Y_s|Y_{t_0:s}, \mathcal{H}_t) \quad (11)$$

From (2, 3, 5, 4, 8, 9, 10), one may factor

$$P(Y_{t_0:t}, S_{t_0:t}|\mathcal{H}_t) = P(S_{t_0}|\mathcal{H}_t)P(Y_{t_0}|S_{t_0}) \times \prod_{s=t-N+2}^t P(S_s|S_{s-1}, \mathcal{H}_t)P(Y_s|S_s) \quad (12)$$

Then, conditional independences indicated by (12) imply

$$P(Y_s|Y_{t_0:s}, \mathcal{H}_t) = \int P(Y_s|S_s)P(S_s|Y_{t_0:s}, \mathcal{H}_t)dS_s \quad (13)$$

If the filtered posterior, $P(S_s|Y_{t_0:s}, \mathcal{H}_t)$, is Gaussian:

$$P(S_s|Y_{t_0:s}, \mathcal{H}_t) \sim \mathcal{N}(\hat{S}_s, P_s) \quad (14)$$

then using (2) it is easily shown

$$P(Y_s|Y_{t_0:s}, \mathcal{H}_t) \sim \mathcal{N}(H\hat{S}_s, HP_sH^T + \lambda_Y I) \quad (15)$$

where H is the observation matrix: $HS_t = l_t$. Because of nonlinearities in $P(S_t|S_{t-1}, \mathcal{H}_t)$, the filtered posterior may not be Gaussian. Nonetheless, as we can write $S_t = g(S_{t-1}, z_t)$, where z_t is Gaussian, the UKF [6] can be used to approximate the filtered posterior as Gaussian. The UKF is initialized at time t_0 and propagates to time t .

At time t , we can obtain additional information from the filtered posterior (14). This information includes conditional approximate MMSE² tracking estimates of the current location l_t , speed v_t , motion direction θ_t , and path curvature ω_t , as these are all components of S_t . Let Z_t be the desired component. The conditional MMSE estimate, $Z_{t|t_0:t}^*$, is:

$$\begin{aligned} Z_{t|t_0:t}^* &= E(Z_t|Y_{t_0:t}) \\ &= \sum_{\mathcal{H}_t} P(\mathcal{H}_t|Y_{t_0:t})E(Z_t|Y_{t_0:t}, \mathcal{H}_t) \end{aligned} \quad (16)$$

where $E(Z_t|Y_{t_0:t}, \mathcal{H}_t)$ is the corresponding component of \hat{S}_s in (14) and $P(\mathcal{H}_t|Y_{t_0:t})$ is computed via (1, 11, 15). The final algorithm outputs are the posterior, $P(\mathcal{H}_t|Y_{t_0:t})$, the Bayes minimum-error decision (\mathcal{H}_t maximizing the posterior), and $Z_{t|t_0:t}^*$ evaluated for $Z_t \in \{l_t, v_t, \theta_t, \omega_t\}$.

3. EXPERIMENTAL RESULTS AND CONCLUSION

As discussed in Sec. 2, the path quality decision (1) approximately minimizes the decision error and the instantaneous location, direction, speed, and curvature estimates (16) are approximately MMSE given the generative model developed in Section 2.1. The only approximation lies in the use of the unscented transform to approximate the posterior $P(S_s|Y_{t_0:s}, \mathcal{H}_t)$ in (14). Hence, validating the overall approach rests on the validation of the generative model of Section 2.1 on sample paths and by conducting interaction-based user studies.

Sample paths and shape quality tracking results are shown in Fig. 3. These paths are generated by human participants without aid of multimodal feedback; they are told only to “move straight, move curved, move random, then stop”. All trials use a common observation rate of eight frames/sec, with a two-second horizon ($T = 16$ frames). Tracking results seem in each case correct, only with slight lags most noticeable during the straight \rightarrow curved transition due to the extent of the sliding window. In particular, the snaked motion (upper right Fig. 3) is correctly classified as straight.

We then conducted an informal user study involving seven adult participants. The shape quality tracking was embedded in a color-based interaction where the mixture of red, green, and blue correspond to suitably scaled versions of the log posterior (straight \leftrightarrow red; curved \leftrightarrow green; random \leftrightarrow blue). Participants’ tasks were to discover the shape quality controls and feedback mappings, to control feedback in a predictable manner, and to articulate their discoveries via questionnaire (Fig. 4). They were not told anything in advance about either controls or feedback. Users had little difficulty discovering and articulating the straight and curved controls (for each, five found these “easy” and two “difficult” to discover) but had considerably more difficulty with the blue \leftrightarrow random mapping. Informally, we observed that the majority of users also did not seem to grasp that only horizontal planar (2D) motion was tracked; those who did complained about the lack of 3D tracking, perhaps because they found the 2D-based control slightly cumbersome as it did not encourage full-body motion.

4. CONCLUSION

The success of our proposed generative model for tracking path shape qualities is, on the whole, validated by our preliminary user studies.

²The reason these estimates are approximately MMSE, rather than exactly so, is due by the Gaussian approximation of the posterior (14) by the UKF.

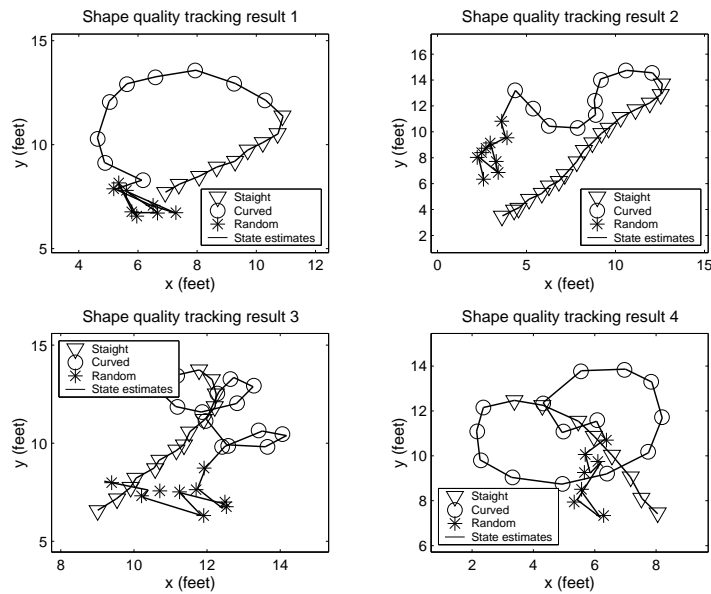


Fig. 3. Shape quality tracking results: triangles represent straight motion, circles curved, asterisks random.

1. Is blue color easy to achieve?
2. Is red color easy to achieve?
3. Is green color easy to achieve?
4. How would you evaluate the overall difficulty for you to establish those colors designed for different motions?
5. Is it too easy so that you lose you interest to explore very soon?
6. Is it too difficult so that you easily get bored and tired?
7. Could you share with me some of your findings when you are playing with the interactive system?
8. What are the interesting things when you are playing this game?
9. What gets you bored?
10. What do you think can improve the system so that it becomes more interesting for people to participate?

Fig. 4. Questionnaire for the user study.

Most importantly, naive participants were able to reproduce path segments correctly detected as straight, curved, or random given only verbal cues. An exploratory interaction study did highlight some needs for improvement, particularly in the ability of participants to discover mappings based on the random path shape quality, and in the lack of complete encouragement for full-body motion. Priorities for future research include extending the generative model to 3D motion to better facilitate full-body interactions, further optimizing the computational cost, and conducting more formal user studies involving cognitive load measurements and external assessments, especially involving the target K-12 population of SMALLab.

5. REFERENCES

- [1] D. Birchfield, T. Ciufu, G. Minyard, G. Qian, W. Savenye, H. Sundaram, H. Thornburg, and C. Todd, "SMALLab: a mediated platform for education", in *Proceedings of ACM SIGGRAPH*, Boston, 2006
- [2] T. Blaine, "The convergence of alternate controllers and musical interfaces in interactive entertainment", in *Proceedings of the 2005 International Conference on New Interfaces for Musical Expression (NIME05)*, Vancouver, BC 2005
- [3] C. Cruz-Neira, D. Sandin, T. Defanti, R. Kenyon, and J. Hart, "CAVE: audio visual experience automatic virtual environment", *Communications of the ACM* 35(6):64-72, 1992
- [4] H. Gardner, *Frames of Mind: The Theory of Multiple Intelligences*, Basic Books, New York, NY 1983
- [5] H. Huang, J. He, T. Rikakis, T. Ingalls, and L. Olson, "Design of biofeedback system to assist the robot-aided movement therapy for stroke rehabilitation", In *Proceedings of Society for Neuroscience 34th Annual Meeting*, San Diego, CA, 2004.
- [6] S. Julier and J. Uhlmann, "A new extension of the Kalman filter to nonlinear systems", in *Proceedings of the 11th International Symposium on Aerospace/Defense Sensing, Simulation, and Controls*, Orlando, FL 1997
- [7] B. Porat. *Digital Processing of Random Signals: Theory and Methods*, Englewood Cliffs, NJ: Prentice Hall, 1993
- [8] A. Siegel and S. White, "The development of spatial representations of large-scale environments". In H. Reese, ed. *Advances in Child Development and Behavior*, 10-45, Academic Press, New York, NY 1975.
- [9] G. Qian, F. Guo, T. Ingalls, L. Olson, J. James, and T. Rikakis. "A gesture-driven multimodal interactive dance system", in *Proceedings of IEEE International Conference on Multimedia and Expo*, Taipei, 2004.
- [10] D. Waller, E. Hunt, and D. Knapp, "The transfer of spatial knowledge in virtual environment training", *Presence: Teleoperators and Virtual Environments* 7(2):129-143, 1998
- [11] C. Wren, F. Sparacino, A. Azarbayejani, T. Darrell, T. Starner, A. Kotani, C. Chao, M. Hlavac, K. Russell, and A. Pentland *Perceptive Spaces for Performance and Entertainment: Untethered Interaction using Computer Vision and Audition*, MIT Media Laboratory PCS Tech. Report # 372, 1997