META-LEVEL TRACKING FOR GESTURAL INTENT RECOGNITION

Mustafa Fanaswala and Vikram Krishnamurthy

Department of Electrical and Computer Engineering
University of British Columbia
Vancouver, BC, Canada.

ABSTRACT

In this paper, a novel mode-driven switching state space approach is proposed for the joint tracking and recognition of gestural commands. Gestures are modeled as spatio-temporal patterns comprised of syntactic sub-units called gesturelets. These gesturelets are directional vectors modulating a switching state space model. Stochastic context-free grammars (SCFG) are used as generative models for command gestures which impart a scale-invariant modeling framework. This translates into a method that is user-independent and robust to the signing variation between and among users. In addition to the modeling framework, we also design a library of useful gestural patterns that cannot be represented by regular grammars (hidden Markov models). Our approach combines tracking and recognition in a single framework and is able to deal with a high perplexity dataset. We demonstrate the effectiveness of our approach by comparing SCFG models with HMM models on synthetic gesture trajectories.

Index Terms— gestural command recognition, stochastic context-free grammars, meta-level tracking

1. INTRODUCTION

A novel static gesture recognition technique is proposed in this paper utilizing a joint tracking and classification approach. A syntactic framework is used to define gestures as a sequence of geometric primitives called gesturelets. Stochastic context-free grammars (SCFGs) are used as a generative model for complex spatio-temporal patterns composed of gesturelets from a small alphabet set. The expressive power of SCFGs is able to capture long-term dependencies in the gesture. Moreover, the self-embedding property of SCFGs is used for scale-invariant recognition of each gesture. As a result, our approach is robust to user variation in signing speed. Our proposed approach is also agnostic to the sensor modality used because it primarily depends on movement patterns of the hand that can be obtained from vision-based algorithms, time-of-flight sensors and accelerometer-based devices.

A physics-based generative model is proposed for gestures utilizing a regime-switching state space model. The 3D coordinates of the hand/finger is the state variable of interest whose evolution is driven by the gesturelets composing a specific gesture. A perspective projection through a pin-hole camera is used as the sensing modality with an additional "detector" stage to convert the image measurement into a point measurement. The switching state space model presents a hybrid estimation problem because in addition to the continuous valued 3D coordinates, we are also interested in recovering the discrete-valued modal state (gesturelet) driving the state-space model. We propose a novel Rao-Blackwellised particle filter to perform hybrid state estimation. In addition, the inference of SCFG models is carried out using a modified Earley-Stolcke parser. The modeling framework in this paper is largely derived from our previous work in radar tracking [1], [2], [3] where SCFG models have been applied to anomalous trajectory identification.

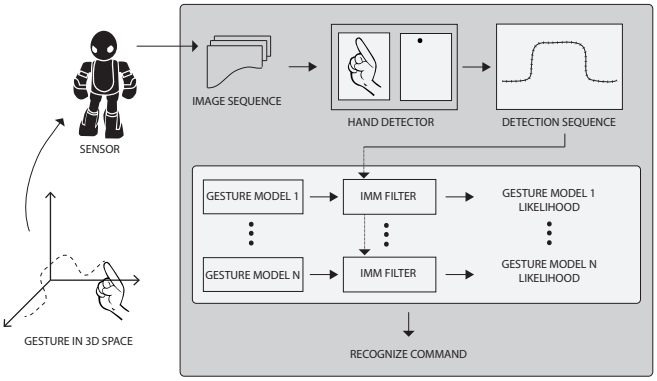

Fig. 1. The gesture recognition system diagram

Literature Survey: A gesture is treated as a space-time trajectory in [4] where the 3D coordinates of the hand is reduced to a 2D coordinate by a plane-fitting approach. The relative difference between 2D coordinates are then used as alphabets in a discrete hidden Markov model to perform recognition. Such an approach is only able to provide a discriminative model for gestures (as opposed to our physics-based generative model). Moreover, additional heuristics are required to account for temporal variation in gestures due to different signers. The

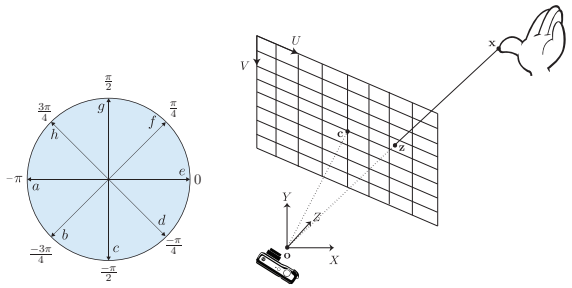

Fig. 2. (a) shows the gesturelets and the radial angular directions that they represent. (b) shows the projection of a point in 3D space onto the camera sensor array.

SCFG framework is able to perform scale-invariant recognition due to the expressive power of its self-embedding rules. When the Markovian assumption is invalid, certain variations like the coupled HMM[5] and the hierarchical HMM[6] have also been used for complex interactions in gestures. However, such techniques cannot model unbounded long-range dependencies which are conveniently captured by the hierarchical branching structure of SCFG models.

2. SWITCHING STATE SPACE MODELS FOR GESTURAL COMMANDS

In this section, the switching state space model relating the temporal evolution of a gesture in 3D to the sensor observation is described. A users hand is assumed to move in one of $|\mathcal{Q}|$ modes, where each mode q is associated with a particular state dynamical model. The modes refer to gesturelets which are geometric primitives of gestural patterns. They correspond to unit vectors in 8 quantized radial directions as shown in Fig. 2a. The primary assumption in this paper is that the mode dynamics evolve according to a SCFG process such that $q_k \sim \mathbf{P}\{q_k|q_{1:k-1}\}$.

2.1. Mode-Dependent State Dynamics

The 3D position and velocity of the hand with respect to a world coordinate system is represented by the state variable $\mathbf{x}_k = [x_k, y_k, z_k, 1, \dot{x}_k, \dot{y}_k, \dot{z}_k, 0]^\intercal$ in homogeneous coordinates [7]. The evolution of the state can be modeled by a mode-driven constant velocity state space model

$$\mathbf{x}_k = F\mathbf{x}_{k-1} + G\mathbf{w}_k(q_k), \quad (1)$$

$$\mathbf{z}_k = H\mathbf{x}_k + \mathbf{v}_k, \quad (2)$$

where F, G are standard for a constant velocity model [8]. The sampling time is represented by $\delta\tau$. Only the state dynamics are mode-dependent through a modulated process noise $\mathbf{w}_k(q_k) \sim \mathcal{N}\{\mathbf{0}, Q(q_k)\}$ such that the process noise in each mode is normally distributed with a mode-dependent covariance $Q(q_k)$. For the radial directions shown in Fig. 2a, we use the process noise covariance

$$Q(q) = \rho \begin{bmatrix} \sigma_o^2 & 0 \\ 0 & \sigma_a^2 \end{bmatrix} \rho^T,$$

$$\rho(q) = \begin{bmatrix} \sin(q) & \cos(q) \\ -\cos(q) & \sin(q) \end{bmatrix},$$

where σ_o^2 is the variance orthogonal to the mode direction represented by q and σ_a^2 is the variance along the direction of mode q. The rotation matrix ρ is used for proper orientation to the mode q. The process noise is a 2-dimensional random variable restricted to small accelerations (nearly constant velocity model) only in the x and y directions. The simplifying approach in our modeling approach is that the gesture occurs in the $x-y$ plane and that there is negligible movement in the z direction.

The sensor observation is modeled as the perspective projection of a pinhole camera model. A 3D point in real world homogeneous coordinates is represented as $\mathbf{s} = [x, y, z, 1]$, where $\mathbf{s}$ is the positional subset of the target state $\mathbf{x}$. The projection operation is a non-linear operation represented by the camera projection matrix $P = K\,[M|\mathbf{t}]$, where K represents the intrinsic parameters of the camera, M is the orientation of the camera-centered frame with reference to the world coordinate frame and $\mathbf{t}$ is the translation of the camera-centered frame from the world origin. The projected 2D homogeneous coordinates $\tilde{\mathbf{u}} = [\tilde{u}, \tilde{v}, \tilde{w}]^\intercal$ are obtained by the action of the camera projection matrix $\tilde{\mathbf{u}} = P\mathbf{s}$. The image coordinates $\mathbf{z} = [u, v]^\intercal$ are obtained by the normalization to the $z = 1$ plane such that $u = \frac{\tilde{u}}{\tilde{w}}$ and $v = \frac{\tilde{v}}{\tilde{w}}$. The perspective projection operation is shown in Fig. 2b.

The non-linear operation involved in the perspective projection is not amenable towards use with a Kalman filter. However, inspired by the projective Kalman filter [9], an adaptive measurement matrix $H_k = \alpha_k P$ can be designed to incorporate the effects of the non-linear operation such that

$$\alpha_k = \frac{1}{P^3 \cdot \hat{\mathbf{x}}_{k|k-1}} \mathbf{I}_{4\times 4}. \quad (3)$$

The notation P^3 refers to the 3rd row of the projection matrix P and $\hat{\mathbf{x}}_{k|k-1}$ is the predicted state using the dynamics in (1). The perspective projection model projects a point in world coordinates into pixel coordinates on a image. We assume that a detection operator D operates on each image captured by the camera and outputs the centroid of the hand (or tip of the finger) as a point measurement.

2.2. Gesture Models

In this section, we describe SCFG gesture models for the gestural patterns shown in Fig. 3.

1. *Right-angular patterns:* The right-angular gestural models are characterized by sentences of the form

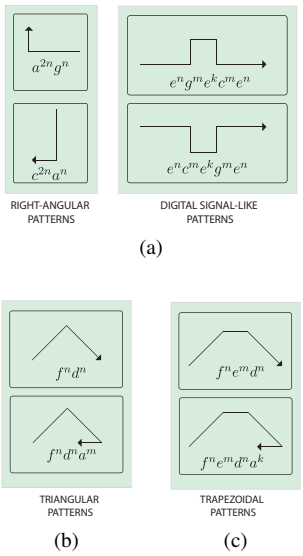

Fig. 3. (a) shows patterns having a right angular part. These can be used as directional commands signifying operations like "next" or "previous". Also depicted are patterns looking like a digital signal. (b) shows triangular patterns which are also called arcs. (c) shows trapezoidal patterns which are closely related to arcs.

$a^{2n}g^n$. They can be used to denote directional commands such as "left", "right", "next", "previous" etc. The primary SCFG rule for such patterns is of the form $X \rightarrow A\,A\,X\,G$ with X being a self-embedding rule. Repeated applications of such a production rule generates 2 a's for every g.

2. *Digital-signal like patterns:* Digital-signal like patterns are represented by rules of the form $e^n g^m e^k c^m e^n$. These are intuitive patterns that can be used for positional commands like "top", "bottom", "front", "back" etc depending on the orientation of the hump. The most characteristic feature of this gesture is an equal number of movements in opposite directions represented by the gesturelet directions g and c. These can be captured by a self-embedding rule of the form $X \rightarrow G\,X\,C$ to generate equal number of g's and c's. In addition, an equal of e's can be generated by a different self-embedding rule of the form $S \rightarrow E\,S\,E$.

3. *Triangular and Trapezoidal patterns:* Triangular gestural models are characterized by sentences of the form $f^n d^n$ or $f^n d^n a^m$ with two neighboring sides of equal length. Such patterns are differentiated from trapezoidal patterns of the form $f^n e^m d^n$ or $f^n e^m d^n a^k$. A recurring theme in SCFG models of such patterns is the self-embedding rule $X \rightarrow F\,X\,D$ to ensure equal lengths for the gesturelets f and d.

3. RAO-BLACKWELLIZED SYNTACTIC TRACKING FOR CLASSIFICATION

In this section, a novel Rao-Blackwellised multiple model particle filter that uses the one-step prediction probability of an SCFG model as a mode proposal density is derived. The gesture recognition task is viewed as a model classification problem using likelihoods computed from the Earley-Stolcke parser.

3.1. SCFG-based Multiple Model Particle Filter

The filtering density $\mathbf{P}\{q_{1:k-1}, \mathbf{x}_{1:k-1}|\mathbf{z}_{1:k-1}\}$ involving both a discrete-mode and continuous state can be numerically approximated by the random measure $\{(q_{1:k}^{(i)}, \mathbf{x}_{1:k}^{(i)}), w_k^{(i)}\}_{i=1}^{N_p}$ consisting of N_p particles and weights $w_k^{(i)}$.

Denote the conditional probability distribution of the mode as the modal probability $\chi_{1:k} = \mathbf{P}\{q_{1:k}|\mathbf{z}_{1:k}\}$. We observe that conditioned on the mode sequence $q_{1:k}$, the density $\mathbf{P}\{\mathbf{x}_{1:k-1}|q_{1:k-1}, \mathbf{z}_{1:k-1}\}$ is Gaussian and can be computed analytically using the optimal Kalman filter if the marginal posterior density $\mathbf{P}\{q_{1:k}|\mathbf{z}_{1:k}\}$ is known. The modal density satisfies the alternative recursion

$$\mathbf{P}\{q_{1:k}|\mathbf{z}_{1:k}\} = \frac{\mathbf{P}\{\mathbf{z}_k|q_{1:k-1}, \mathbf{z}_{1:k-1}\}\mathbf{P}\{q_k|q_{1:k-1}\}}{\mathbf{P}\{\mathbf{z}_k|\mathbf{z}_{1:k-1}\}} \times \mathbf{P}\{q_{1:k-1}|\mathbf{z}_{1:k-1}\}, \quad (4)$$

where the term $\mathbf{P}\{\mathbf{z}_k|q_{1:k-1}, \mathbf{z}_{1:k-1}\}$ is implicitly dependent on past base state values $\mathbf{x}_{1:k}$. Instead of approximating the entire filtering density, a weighted set of samples $\{q_{1:k}^{(i)}, w_k^{(i)}\}_{i=1}^{N_p}$ is used to only represent the marginal posterior distribution $\chi_{1:k}$. The marginal density of the base $\mathbf{x}_{1:k}$ is a Gaussian mixture

$$\mathbf{P}\{\mathbf{x}_{1:k}|\mathbf{z}_{1:k}\} = \int \mathbf{P}\{\mathbf{x}_{1:k}|q_{1:k}, \mathbf{z}_{1:k}\} \sum_{i=1}^{N_p} w_k^{(i)} \delta_{q_{1:k}^{(i)}}(q_{1:k})$$

$$= \sum_{i=1}^{N_p} w_k^{(i)} \mathbf{P}\{\mathbf{x}_{1:k}|q_{1:k}^{(i)}, \mathbf{z}_{1:k}\} \quad (5)$$

that can be computed efficiently with a Kalman filter bank. The Rao-Blackwellised particle filter samples the modal state $q_k \sim \zeta_{k|1:k-1}$ from the one-step prediction probability in (8). We sample $q_k^{(i)}$ and then propagate the mean $\hat{\mathbf{x}}_k^{(i)}$ and covariance $\Sigma_k^{(i)}$ of $\mathbf{x}_k$ with a Kalman filter [8]. The conditional density of the discrete-valued mode history $q_{1:k-1}$ is approximated by a set of N_p weighted random particles as the empirical random measure $\{q_{1:k-1}^{(i)}, w_{k-1}^{(i)}\}_{i=1}^{N_p}$.

The prediction for the random measure $\{q_{1:k-1}^{(i)}, w_{k-1}^{(i)}\}_{i=1}^{N_p}$ is performed using a suitable proposal $\pi_{\text{RBPF}}\{q_k^{(i)}|q_{1:k-1}^{(i)}, \mathbf{z}_{1:k-1}\}$. The bootstrap proposal is chosen such that prediction of the modal state is given by

$$\pi_{\text{RBPF}}\{q_k^{(i)} = v|q_{1:k-1}^{(i)}\} = \zeta_{k|k-1}^{(i)}(v), \quad (6)$$

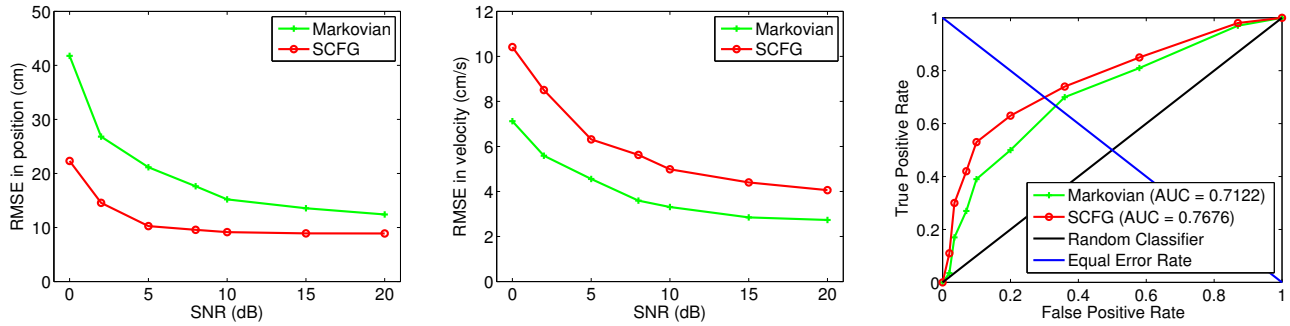

Fig. 4. (a) shows the RMSE in position (x, y only) (b) shows the RMSE in velocity (x,y only) and (c) shows the receiver operating characteristics for the triangular gesture models. The area under the curve (AUC) is a metric between 0 and 1, where the best classifier has an AUC closest to 1.

where $\zeta^{(i)}_{k|k-1}(v)$ is the SCFG one-step prediction probability in (8) for $v \in \mathcal{V}$ and $i = 1, \ldots, N_p$. At the end of each cycle of the sequential importance sampling step, the base state marginal is computed using (5) such that $\hat{\mathbf{x}}_{k|k} = \mathbf{E}\{\mathbf{x}_k|\mathbf{z}_{1:k-1}\}$ and $\Sigma_{k|k} = cov(\mathbf{x}_k|\mathbf{z}_{1:k})$.

The measurement update for the modal state is performed by the incremental importance weight update for the particles. In the case of using the bootstrap proposal, the importance weights reduce to the mode likelihood

$$\mathbf{P}\{\mathbf{z}_k|q_{1:k-1}, \mathbf{z}_{1:k-1}\} = \mathcal{N}\{\mathbf{z}_k; \mathbf{z}_{k|k-1}, S_k\}, \qquad (7)$$

where S_k is the covariance after measurement update in the Kalman filter[8]. The one-step ahead prediction $\mathbf{P}\{q_k|q_{1:k-1}\}$ for the SCFG can be computed from a left-right pass over an observed sequence. The one-step prediction utilizes the probabilistic rules of the SCFG model to predict the next mode in the sequence and is used as a proposal density for the particle filter in Sec. 3.1. The one-step prediction probability

$$\mathbf{P}\{q_k = v|q_{1:k-1}\} = \frac{\mathbf{P}\{q_{1:k-1}, q_k = v\}}{\mathbf{P}\{q_{1:k-1}\}} = \frac{\zeta_k(v)}{\zeta_{k-1}}, \qquad (8)$$

where $v \in \mathcal{Q}$ is an element of the finite mode set $\mathcal{Q}$ and ζ_k is the prefix probability of a partially observed string. The Earley-Stolcke parser [10] provides an efficient algorithm to compute the prefix probability and the related one-step prediction probability in (8). For the sake conciseness, the reader is referred to previous work [1] for the expressions required to compute these quantities.

4. NUMERICAL EXAMPLES

Simulations are carried out using synthetic data assuming an ideal pin-hole camera without any lens distortion effects. The characteristics of a cell-phone camera are assumed with focal length of 4mm and a camera sensor format of $\frac{1}{3.2}$". A trajectory is simulated in 3D coordinates in the plane at $z = 1m$ perpendicular to the optical axis following different gesture shapes. Perspective projection is then applied to the location component of the target state contaminated with additive noise producing sensor measurements that act as input to the proposed filter. The additive noise models the efficacy of a hand/finger localization algorithm used in segmenting the hand from the background of the image. It is typically impulsive in nature, but a simplifying Gaussian assumption can be made analogous to [9]. A filter bank is maintained with each filter tuned to a particular gesture model. The model with the maximum likelihood at the end of the static gesture is chosen as the correct model and the filtered state estimate from that model is used for performance metrics. The root mean square error in position and velocity is used as a tracking metric and model mismatch rates are used as a classification metric. The results are shown in Fig. 4. The SCFG model have lower RMSE than the Markov chain based models. Only the receiver operating curve (ROC) for the triangular gesture models is shown in Fig. 4c. Other models have similar ROC curves in which the SCFG models have a larger area under the curve. The SCFG models are not learned, but are initialized using system-theoretic constraints derived in [1]. A simulated dataset is used to calculate average lengths of the gestures. Fully connected Markov models are learned from this dataset as the competing framework.

5. CONCLUSION

A novel physics-based generative model utilizing an SCFG modulated state-space is formulated for various gestural commands. The joint tracking and classification of gestures is carried out using a novel Rao-Blackwellised SCFG particle filter employing a projective Kalman filter for the continuous state variable. From numerical simulations, it can be observed that the SCFG models outperform Markovian models. In particular, at low SNRs (when the hand/finger detection algorithm is performing poorly), the long-range dependencies in the spatial pattern inform better tracking and classification.

6. REFERENCES

[1] M. Fanaswala and V. Krishnamurthy, "Detection of anomalous trajectory patterns in target tracking via stochastic context-free grammars and reciprocal process models," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 1, pp. 76–90, 2013.

[2] M. Fanaswala and V. Krishnamurthy, "Syntactic models for trajectory constrained track-before-detect," *IEEE Transactions on Signal Processing*, vol. 62, no. 23, pp. 6130–6142, Dec 2014.

[3] M. Fanaswala and V. Krishnamurthy, "Spatiotemporal trajectory models for meta-level target tracking," *IEEE Aerospace and Electronic Systems Magazine*, vol. 30, no. 1, pp. 16–31, Jan 2015.

[4] Yanghee Nam, K Wohn, et al., "Recognition of space-time hand-gestures using hidden Markov model," in *ACM symposium on Virtual reality software and technology*, 1996, pp. 51–58.

[5] Matthew Brand, Nuria Oliver, and Alex Pentland, "Coupled hidden Markov models for complex action recognition," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1997*. IEEE, 1997, pp. 994–999.

[6] Graeme S Chambers, Svetha Venkatesh, Geoff AW West, and Hung Hai Bui, "Hierarchical recognition of intentional human gestures for sports video annotation," in *16th International Conference on Pattern Recognition*. IEEE, 2002, vol. 2, pp. 1082–1085.

[7] R. Szeliski, *Computer Vision: Algorithms and Applications*, Texts in computer science. Springer, 2010.

[8] Y. Bar-Shalom, T. Kirubarajan, and X. Li, *Estimation with Applications to Tracking and Navigation*, John Wiley & Sons, Inc., New York, NY, USA, 2002.

[9] Cristian Canton-Ferrer, Josep R Casas, A Murat Tekalp, and Montse Pardas, "Projective Kalman filter: Multiocular tracking of 3d locations towards scene understanding," in *Machine Learning for Multimodal Interaction*, pp. 250–261. Springer, 2006.

[10] Andreas Stolcke, "An efficient probabilistic context-free parsing algorithm that computes prefix probabilities," *Computational Linguistics*, vol. 21, no. 2, pp. 165–201, 1995.