Multi-Signal Gesture Recognition Using Temporal Smoothing Hidden Conditional Random Fields

Yale Song, David Demirdjian, and Randall Davis
MIT Computer Science and Artificial Intelligence Laboratory
32 Vassar Street, Cambridge, MA 02139
{yalesong,demirdj,davis}@csail.mit.edu

***Abstract*— We present a new approach to multi-signal gesture recognition that attends to simultaneous body and hand movements. The system examines temporal sequences of dual-channel input signals obtained via statistical inference that indicate 3D body pose and hand pose. Learning gesture patterns from these signals can be quite challenging due to the existence of long-range temporal-dependencies and low signal-to-noise ratio (SNR). We incorporate a Gaussian temporal-smoothing kernel into the inference framework, capturing long-range temporal-dependencies and increasing the SNR efficiently. An extensive set of experiments was performed, allowing us to (1) show that combining body and hand signals significantly improves the recognition accuracy; (2) report on which features of body and hands are most informative; and (3) show that using a Gaussian temporal-smoothing significantly improves gesture recognition accuracy.**

I. INTRODUCTION

Human communication is inherently both multimodal and multi-signal. Spoken language is often accompanied by non-verbal cues, such as body and/or hand gestures, eye gaze, head nod, or facial expressions that can be essential to understanding. Gestures in turn are often multi-signal, e.g., using both body and hand poses simultaneously, with both necessary for gesture understanding. Successful gesture recognition thus needs to be able to process multi-signal data seamlessly. Most current gesture recognition systems, however, concentrate on dealing with only a single signal.

We developed a multi-signal gesture recognition system that attends to body and hands, allowing a richer gesture vocabulary and more natural human-computer interaction. In this paper, we present the *signal understanding* part of the system, i.e., learning to recognize patterns of multi-signal gestures. The *signal processing* part (i.e., obtaining a temporal sequence of body and hand features) is described in a companion paper [16].

Discriminative hidden-state learning approaches (e.g., HCRF [14]) have recently shown promising results in many pattern recognition tasks. The main advantage of discriminative approaches compared over generative approaches (e.g., HMM [15]) is that they do not make the conditional independence assumption, which is often both too restrictive and unrealistic. It has been shown that when conditional independence does not hold, the asymptotic accuracy of discriminative models is higher than generative models [10].

In our task, the input signal patterns tend to exhibit long-range temporal-dependencies (e.g., body parts move coherently as time proceeds, hand poses are articulated in relation to body poses in a time-sequence, etc.). Thus, although a gesture label is given, individual observations may not be independent of each other; observations rather seem to be important clues to distinguish similar patterns of gestures.

Also, in our task body and hand pose signals are obtained by performing statistical estimation and classification, which themselves are not perfectly accurate; thus the input signal patterns exhibit high-frequency fluctuations in a time series, with a low SNR.

Previous work on HCRFs for gesture recognition [18] tried to resolve the first issue, capturing long-range dependencies among observation by defining a temporal window and concatenating signals within the window, creating a single large input feature. We take a slightly different approach. Instead of concatenating signals (which increases the dimensionality of the input feature vectors), we use a Gaussian temporal smoothing kernel, capturing long-range dependencies and making our framework less sensitive to the noise. We show that this improves upon the performance of previous work on HCRFs for gesture recognition [18], while at the same time keeping the same computational complexity of the original HCRF model [14].

The main contribution of this paper lies in this incorporation of a Gaussian temporal-smoothing kernel into the HCRF framework. Based on the results from an extensive set of experiments using 10 body-and-hand gestures from the NATOPS database [16], we (1) show that combining body and hand poses significantly improves the recognition accuracy; (2) report on which body and hand features are most informative for this recognition task; and (3) show that temporal smoothing improves system performance.

Section II describes related work on multi-signal gesture recognition and inference framework, Section III gives an overview of our gesture recognition system, Section IV describes our HCRFs with temporal-smoothing in more detail, and Section V describes experiments and results. Section VI concludes with our contributions and suggests directions for future work.

II. RELATED WORK

Gesture recognition is a broad area of research that is increasingly used in natural human-computer interaction. Gestures can range from dynamic human body motion through pointing device gestures to sign language. In this work, we are concerned primarily with multi-signal gestures

involving dynamic body movements and static or dynamic hand pose configurations. Here we review some of the recent efforts to the similar goal. For a more comprehensive review of gesture recognition, see [1], [11].

Recently, many efforts have been made to build multimodal gesture recognizers. In [2], Althoff *et al.* used trajectories of head and hands to recognize gestures for in-car control systems. Two different recognizers were developed, rule-based and HMM-based. When tested with 5 common in-car control gestures (left, right, forward, backward, and wipe) using either head or hand gestures, the two recognizers achieved similar recognition accuracy (90%). In [9], Li *et al.* presented multi-signal pointing-direction estimation in a human-robot interaction scenario, using a combination of head orientation, body pose, and hand pose information. Head orientation was determined by tracking eye-gaze using FaceLAB; body pose was estimated using a particle filter; and hand pose was classified using a multi-resolution image querying method. These three signals were then used to determine the pointing direction. However, all of these efforts were either tested on fairly simple tasks (i.e., recognizing single-signal or static gestures) or used statistical inference frameworks that were not particularly well suited to these tasks (i.e., unable to capture complex long-range dependencies in the input signals).

There have also been active efforts to build a robust inference framework for pattern analysis tasks based on discriminative learning. In [8], Lafferty *et al.* introduced CRFs, a discriminative learning approach that does not make conditional independence assumptions. In [14], Quattoni *et al.* introduced HCRFs, an extension to CRFs that incorporates hidden variables. Many other variants of HCRFs have been introduced since then [18], [12], [5], but most of these were tested only on single-signal pattern recognition tasks (e.g., POS tagging [8], object recognition [14], body gesture recognition [18], [12], and phone classification [5]) and paid less attention to dealing with noisy input signals.

In this work, we demonstrate that discriminative hidden-state learning approaches are well suited to multi-signal gesture recognition tasks, and that significant improvements in recognition accuracy can be achieved by incorporating Gaussian temporal-smoothing into the inference framework.

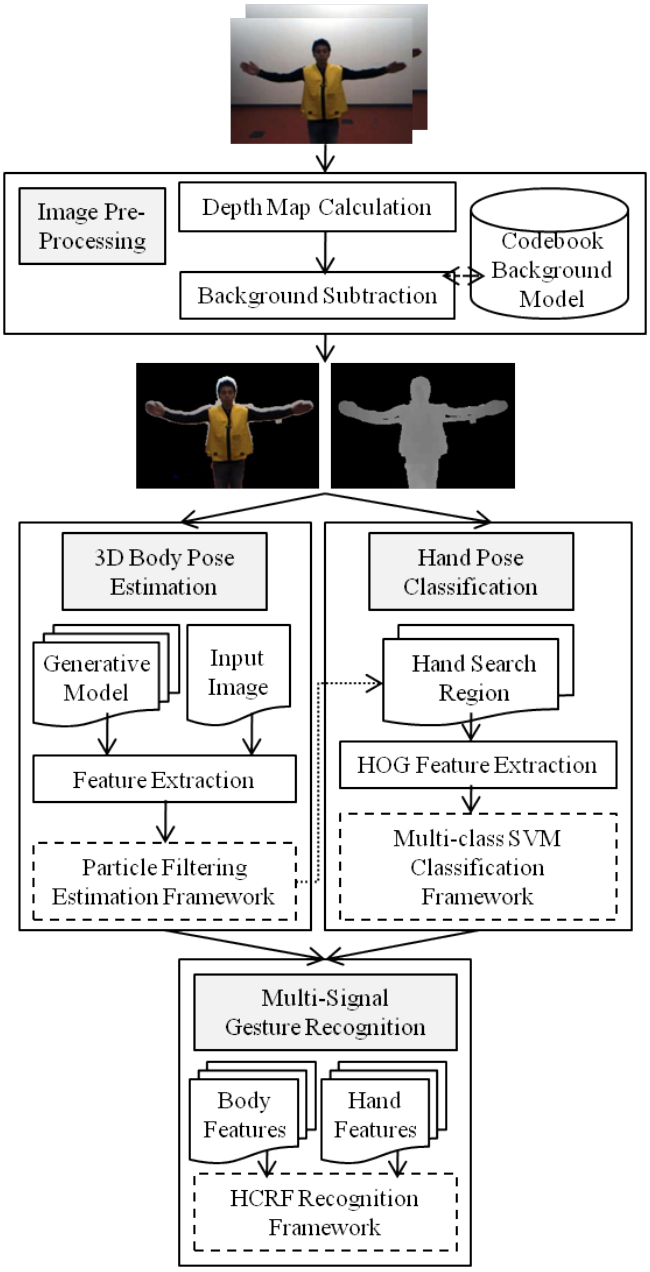

Fig. 1. Multi-signal gesture recognition framework.

III. System Overview

Fig. 1 shows an overview of our gesture recognition system. The system starts by receiving pairs of time-synchronized images recorded from a Bumblebee2 stereo camera[1], producing 320 x 240 pixel resolution images at 20 FPS.

For the first part in the pipeline, image pre-processing, depth maps are calculated, and the images are background subtracted using depth information and a codebook background model that is trained off-line. For the second part, 3D body pose estimation, we construct a generative model of the human upper body, and compare various features extracted from the model to corresponding features extracted from input image. We estimate body pose in a multi-hypothesis Bayesian inference framework with a particle filter [7]. For the third part, hand pose classification, we define two small search regions around estimated wrist joint positions and slide a window within each region to search for hands. A multi-class SVM classifier [17] is trained off-line based on HOG descriptors [4] extracted from manually-segmented images of hands, and is used to classify hand poses. In the last part, multi-signal gesture recognition, we perform recognition with a combination of body and hand pose information. An HCRF with a Gaussian temporal-smoothing kernel is trained off-line using a supervised gesture data set,

[1]http://www.ptgrey.com

and is used to perform gesture recognition.

The system builds on our previous work [16]; this paper reports on the gesture recognition part of the system. A detailed description of the 3D body pose estimation and hand pose classification part of the system is in a companion paper [16].

IV. Multi-Signal Gesture Recognition

The goal here is to learn a classifier $p(y \mid \mathbf{x})$ that predicts a gesture label $y \in \mathcal{Y}$ given a temporal sequence of input images $\mathbf{x} = \{\mathbf{x}_1, \cdots, \mathbf{x}_T\}$. For each image $\mathbf{x}_t$, we extract body pose features $\phi(\mathbf{x}_t^1) \in \mathcal{R}^{N_1}$ and hand pose features $\phi(\mathbf{x}_t^2) \in \mathcal{R}^{N_2}$; each $\mathbf{x}_t$ is represented as a multi-signal feature-vector

$$\phi(\mathbf{x}_t) = \left(\phi(\mathbf{x}_t^1)\, \phi(\mathbf{x}_t^2)\right)^T. \tag{1}$$

We briefly review HCRF to set the context for our work, and describe the formulation of our model in detail.

A. HCRFs: A Review

An HCRF [14] is a discriminative framework for building probabilistic models to label segmented sequential data (i.e., data that has been divided at signal boundaries, such as gesture start and end). The framework extends CRF models [8], which assumes a tree-structured undirected graph G, by incorporating hidden state variables into the graphical structure. The framework is designed to capture complex dependencies in observations efficiently, without attempting to specify exact conditional dependencies. The goal is to learn a mapping function of observations $\mathbf{x}$ to class labels $y \in \mathcal{Y}$, by introducing hidden state variables $h \in \mathcal{H}$ to compactly represent the distribution of observations. The conditional probability distribution $p(y|\mathbf{x};\theta)$ of a class label y given a set of observation $\mathbf{x}$ with parameter vector θ is constructed as

$$p(y \mid \mathbf{x};\, \theta) = \sum_{\mathbf{h}} p(y, \mathbf{h} \mid \mathbf{x};\, \theta) = \frac{1}{Z} \sum_{\mathbf{h}} e^{\Psi(y, \mathbf{h}, \mathbf{x}; \theta)} \tag{2}$$

where Z is a *partition function* defined as

$$Z = \sum_{y \in \mathcal{Y}} \sum_{\mathbf{h}} p(y, \mathbf{h} \mid \mathbf{x};\, \theta) \tag{3}$$

and $\Psi(y, \mathbf{h}, \mathbf{x}; \theta)$ is a *potential function* defined as

$$\begin{aligned} \Psi(y, \mathbf{h}, \mathbf{x}; \theta) &= \sum_{v \in V} \theta_V \cdot f\left(v, \mathbf{h}|_v, y, \mathbf{x}\right) \\ &+ \sum_{(i,j) \in E} \theta_E \cdot f\left((i,j), \mathbf{h}|_{(i,j)}, y, \mathbf{x}\right). \end{aligned} \tag{4}$$

The potential function models dependencies in the graphical structure, where θ_V and θ_E are parameters that determine dependencies within $\mathbf{h}|_S$, a set of components of $\mathbf{h}$ associated with the vertices and edges in subgraph S of G. Therefore, it is crucial to design the potential function carefully. We describe our potential function below.

Following previous work on CRFs [8], parameter optimization is performed using:

$$L(\theta) = \sum_{i=1}^{N} \log p(y_i \mid \mathbf{x}_i;\, \theta) - \frac{1}{2\sigma^2} ||\theta||^2 \tag{5}$$

where the second term, the *regularization factor*, is introduced to prevent overfitting of the training data. The optimal parameter values are obtained by solving the maximum log-likelihood function $\theta^* = \arg\max_\theta L(\theta)$ using belief propagation [13]. Finally, a class label for a new observation is determined as

$$y^* = \arg\max_{y \in \mathcal{Y}} p(y \mid \mathbf{x};\, \theta). \tag{6}$$

Similar to [8], we assume that the underlying graph satisfies the first-order Markov property, forming a tree-structured chain. Therefore, belief propagation [13] can be used for efficient parameter estimation and inference.

B. HCRFs with Gaussian Temporal-Smoothing

Our potential function is defined as

$$\begin{aligned} \Psi(y, \mathbf{h}, \mathbf{x}; \theta) &= \sum_t K\left(\phi(\mathbf{x}), g(\omega), t\right) \cdot \theta(h_t) \\ &+ \sum_t \theta(y, h_t) + \sum_{t-1,t} \theta(y, h_{t-1}, h_t) \end{aligned} \tag{7}$$

where $K\left(\phi(\mathbf{x}), g(\omega), t\right)$ is a Gaussian temporal-smoothing kernel, which performs a convolution of the input feature vector $\phi(\mathbf{x})$ and the ω-point Gaussian window $g(\omega)$. The Gaussian window is computed as

$$g(\omega)[n] = e^{-\frac{1}{2}\left(\alpha \frac{n}{\omega/2}\right)^2} \tag{8}$$

where $-\frac{\omega-1}{2} \leq n \leq \frac{\omega-1}{2}$, and α is inversely proportional to the standard deviation of a Gaussian random variable. [2] The Gaussian window $g(\omega)$ is normalized so that $\sum_n g(\omega)[n] = 1$. Intuitively, the kernel computes for each time frame a weighted mean of ω neighboring feature vectors with a Gaussian filter, centering the filter at the current time frame. This process produces a feature vector at each time frame that both incorporates observations some time distance away from the current frame, and reduces signal noise.

The first term in Eq. 7 captures dependencies between the temporal smoothed input feature vectors and hidden state variables; the second term captures dependencies between class labels and hidden states variables; and the last term captures dependencies among class labels and two time-consecutive hidden state variables.

V. Experiment And Result

We conducted an extensive set of experiments using our gesture recognition system with a body-and-hand gesture dataset [16]. We briefly describe the dataset, and (1) show that combining body and hand poses significantly improves the recognition accuracy; (2) describe which body and hand features are most informative for this recognition task; and (3) show that temporal smoothing significanly improves performance.

[2] Following [6], we set α=2.5.

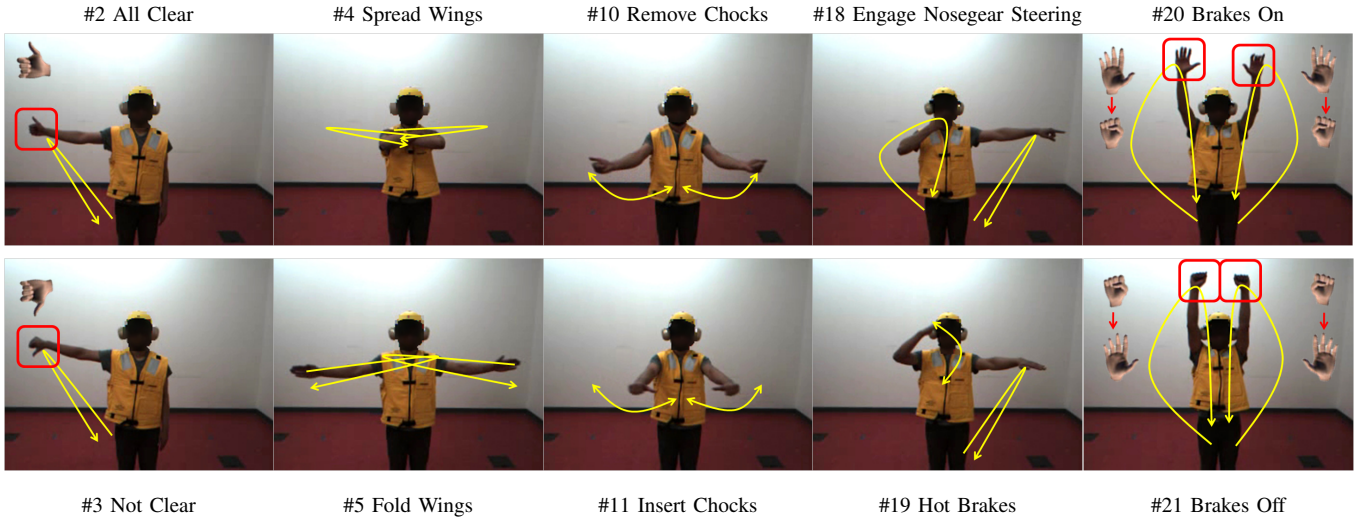

Fig. 2. Ten NATOPS aircraft handling signal gestures [16]. Body movements are illustrated in yellow arrows, and hand poses are illustrated with synthesized images of hands. Red rectangles indicate hand poses are important in distinguishing the gesture pair.

A. NATOPS Aircraft Handling Signal Dataset

We used the NATOPS dataset [16], a body-and-hand gesture dataset containing an official gesture vocabulary used for communication between carrier deck personnel and Navy pilots (e.g., yes/no signs, taxing signs, fueling signs, etc.). The dataset contains 24 gestures, with each gesture performed by 20 subjects 20 times, resulting in 400 samples per gesture. Each sample had a unique duration; the average length of all samples was 2.34 sec (σ^2=0.62).

We selected five pairs of gestures (see Fig. 2) that are particularly interesting because in general the gestures in each pair are very similar, and in fact two pairs (#2 & #3 and #20 & #21) are indistinguishable in the absence of knowledge of hand pose. For example, gestures #20 ("brakes on") and #21 ("brakes off") are performed by raising both hands, with either open palms that are closed ("brakes off"), or vice versa ("brakes on"). Here, the role of hand pose is crucial to distinguishing two very similar gestures with opposite meanings. As a more subtle case, gestures #10 ("insert chocks") and #11 ("remove chocks") are performed with both arms down and waving them in/outward. The only difference is the position of thumbs: inward ("insert chocks") and outward ("remove chocks").

Experiments were conducted using combinations of body and hand features extracted in our previous work [16]. There were 4 body features and 2 hand features. The four body features were joint angles (T), angular velocities (dT), joint coordinates (P), and the corresponding velocities (dP).

The joint angle features (T and dT) are 8 DOF vectors (3 for shoulder and 1 for elbow, for each arm), and the joint coordinate features (P and dP) are 12 DOF vectors (3D coordinates of elbows and wrists for both arms). The uniform-length relative joints are obtained by configuring a generative model with the estimated joint angles with uniform limb lengths (so that their joint coordinates have less variance), and recording joint coordinates relative to the chest point.

The two hand features were a "soft decision" and "hard decision." The soft decision (S) is an 8 DOF vector with probability estimates obtained from the SVM (4 hand poses for each hand), while the hard decision (H) is a 2 DOF vector obtained by selecting the highest probability estimate for each hand. Intuitively, S has richer information about the shape of hands, while H has a lower degree of freedom, which can reduce the computational cost in an estimation step.

All experiments were conducted with n-fold cross validation (n-CV), allowing us to perform a cross-subject analysis, i.e., train the model with a dataset that does not include gesture examples performed by subjects in a test dataset, resulting in more accurate measurement of performances. We measured accuracy with an F1 score ($\text{F1}=2*\frac{\text{precision}\cdot\text{recall}}{\text{precision + recall}}$). In all tests, we set the regularization factor in Eq. 5 at 1,000 which, based on our preliminary experiments, helps prevent overfitting.

B. Does Combining Body and hand Pose Really Help?

The first question was whether combining body and hand poses helps to improve recognition performance. To determine this, we compared recognition performance under two conditions: body feature only (BO) and body and hand features (BH), i.e., BO contained only body features, while BH contained body and hand features. Since there were two hand features (S and H), we averaged the two test results for the BH condition. For each test, we performed 4-CV analysis, varying the number of hidden states from 3 to 4 and taking an average. Since a 4-CV analysis performs four repetitive tests, we get variances in the results; we performed independent samples T-tests to see if the differences between two conditions (BO and BH) were statistically significant.

Table I shows means and standard deviations for overall recognition accuracy rates averaged over 10 gestures, as

TABLE I

BODYONLY VS. BODYHAND

Body Feature	Hand Feature		T-test result
	BO, $\mu(\sigma^2)$	BH, $\mu(\sigma^2)$	
T	20.09 (3.57)	27.02 (3.83)	$t(22)$=1.00, p=.326
P	23.26 (11.07)	32.73 (20.57)	$t(22)$=1.21, p=.240
dT	62.47 (7.21)	76.23 (8.10)	$t(22)$=4.06, p=.001
dP	70.94 (6.73)	80.65 (5.30)	$t(22)$=3.82, p=.001

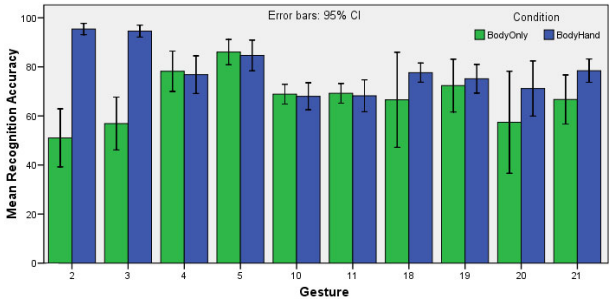

Fig. 3. Per-gesture comparisons of BodyOnly and BodyHand.

TABLE II

VARIOUS COMBINATIONS OF BODY AND HAND FEATURES

Body Feature	Hand Feature		
	H, $\mu(\sigma^2)$	S, $\mu(\sigma^2)$	Average
dT	78.02 (10.97)	82.27 (10.42)	80.15 (10.82)
dP	80.72 (9.85)	86.02 (8.32)	83.37 (9.37)
dTdP	80.08 (8.21)	80.86 (9.51)	80.47 (8.82)
Average	79.61 (9.67)	83.05 (9.60)	·

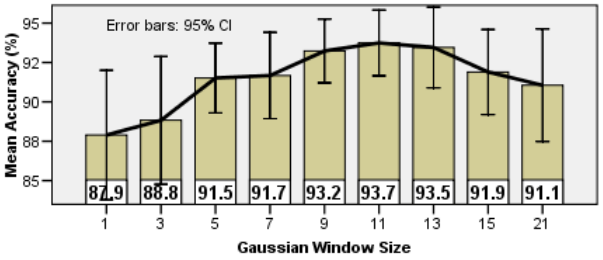

Fig. 4. Recognition accuracy for different window sizes.

well as the results from independent samples T-tests. In all our test cases, using body and hand pose together resulted in higher recognition accuracy rates. For two of the body pose features (dT and dP) the differences were statistically significant (p=.001).

Fig. 3 shows per gesture comparisons of the two conditions (BO and BH). Note that the graph used only the higher performing body features dT and dP. As expected, the performance difference was significant for the 4 gestures (#2, #3, #20, and #21) where the hand pose plays an important role in defining the gesture (see Fig. 3). The difference is especially obvious for gesture pair #2 and #3, where recognition without knowing hand pose was no better than random. Our result indicated that using body and hand pose together on these 4 gestures achieved on average 27.5% higher accuracy; for the other 6 gestures there were slight differences, but none were significant.

C. Which features are most informative?

Various types of body or hand features have been explored in gesture recognition research, but there is no clear sense as to which features are most informative. In response, we compared the system's recognition accuracy using various combinations of three body features (dT, dP, and dTdP) and two hand features (S and H). For each test case we performed 10-CV analysis, varying the number of hidden states from 3 to 5 and taking an average.

Table II shows comparisons of the resulting performance. Hand feature S performed significantly better than H ($t(178)$=2.24, p=.018), achieving on average 3.44% higher accuracy rate. For body pose, dP performed the best, while the performances obtained using dT and dTdP were similar. We found no statistical significant in body feature differences.

For the features we used, a combination of dP (uniform-length relative body joint velocity) and S (probability estimates of a hand pose) was the most informative feature for this task.

D. Does Gaussian Temporal-Smoothing Help?

The third experiment aimed to measure the advantage of a Gaussian temporal smoothing HCRF. Based on the previous results, we selected dPS as a feature combination (joint velocities for body and soft decision for hands). All tests were performed with 10-CV analysis, fixing the number of hidden states at 5, and varying the Gaussian window size from 1 to 21 (using only odd numbers).

As can be seen in Fig. 4, Gaussian temporal-smoothing significantly improved the performance: when compared to non-smoothing (ω=1, 12.1% error), a half-second sized Gaussian window (ω=11, 6.3% error) was able to reduce 48% of remaining errors. The performance dropped as the window size increased beyond ω=11, indicating that it started losing some important local/high-frequency gesture information when the Gaussian window size was larger than half a second. Fig. 6. shows confusion matrices comparing ω=1 and ω=11 (best performing setting). We can see that both false positives and false negatives were decreased for all individual classes, with the highest gain achieved for gesture #10 (22% improvement).

Fig. 5 shows distributions of hidden states for each gesture class, when the dPS feature combination was used with $|\mathcal{H}|$=5 and ω=11. Here we can see that the hidden states are roughly evenly distributed over the gesture classes, suggesting that the number of hidden states was appropriate.

One important thing to notice is that temporal-smoothing not only improves recognition accuracy significantly (by considering long-range input features and increasing SNR), but

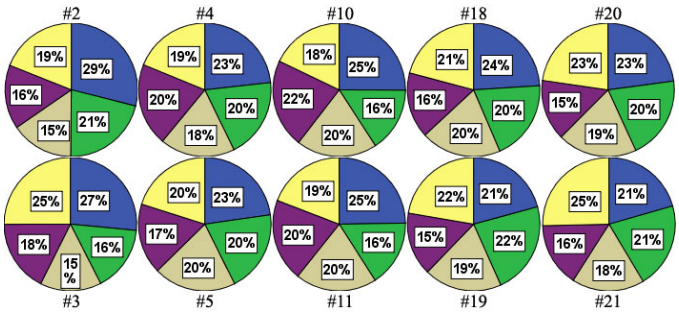

Fig. 5. Distributions of assigned hidden states ($|\mathcal{H}|$=5, ω=11). The numbers enclosed in each area indicates the hidden state assignments.

TABLE III

CONFUSION MATRICES COMPARING ω=1 AND ω=11.

	#2	#3	#4	#5	#10	#11	#18	#19	#20	#21
#2	**1.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
#3	0.00	**0.98**	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00
#4	0.00	0.00	**0.79**	0.03	0.08	0.01	0.01	0.01	0.00	0.01
#5	0.00	0.00	0.06	**0.92**	0.01	0.01	0.01	0.01	0.00	0.00
#10	0.00	0.00	0.06	0.01	**0.73**	0.11	0.00	0.00	0.00	0.00
#11	0.00	0.01	0.03	0.02	0.14	**0.86**	0.01	0.00	0.00	0.01
#18	0.00	0.01	0.01	0.00	0.01	0.00	**0.90**	0.08	0.01	0.04
#19	0.00	0.00	0.02	0.01	0.00	0.01	0.07	**0.88**	0.03	0.03
#20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	**0.87**	0.06
#21	0.00	0.00	0.03	0.01	0.00	0.00	0.00	0.01	0.09	**0.85**

No Temporal-Smoothing ($|\mathcal{H}|$=5, ω=1)

	#2	#3	#4	#5	#10	#11	#18	#19	#20	#21
#2	**1.00**	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
#3	0.00	**0.99**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
#4	0.00	0.00	**0.87**	0.01	0.01	0.00	0.02	0.00	0.00	0.01
#5	0.00	0.00	0.03	**0.98**	0.00	0.01	0.00	0.00	0.01	0.00
#10	0.00	0.00	0.03	0.00	**0.95**	0.09	0.00	0.00	0.00	0.00
#11	0.00	0.00	0.01	0.01	0.03	**0.89**	0.00	0.00	0.01	0.01
#18	0.00	0.00	0.02	0.00	0.01	0.01	**0.95**	0.07	0.00	0.02
#19	0.00	0.00	0.00	0.00	0.00	0.00	0.03	**0.93**	0.01	0.00
#20	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	**0.92**	0.07
#21	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.05	**0.88**

Gaussian Temporal-Smoothing ($|\mathcal{H}|$=5, ω=11)

it does so without increasing the computational complexity of inference. Previous work on HCRF for gesture recognition [18] defined a window to concatenate neighboring input features, thus increasing the dimensionality. Our approach computes a weighted mean of neighboring input features, thus it does not increase the dimensionality, and there is no added complexity to the original HCRF model (additions and multiplications in the kernel operation can be negligible compared to the complexity of the inference algorithm).

VI. CONCLUSION AND FUTURE WORK

We presented a Gaussian temporal-smoothing HCRF capable of capturing long-range dependencies, increasing SNR, and improving performance, while at the same time keeping the same computational complexity of the original HCRF model [14]. Through an extensive set of experiments, we (1) showed that combining body and hand signals significantly improves the recognition accuracy; (2) reported on which features of body and hands are most informative; and (3) showed that using a Gaussian temporal-smoothing HCRF significantly improves the performance.

Our current system can be improved in a number of ways. Of the most interest is allowing non-segmented continuous time-series input. In [12], Morency *et al*. presented an LDCRF that does not require its input sequence to be segmented, and showed that it is suitable for a number of gesture recognition tasks. However, the experiments were conducted with binary classification tasks only (e.g., head nod or eye gaze-aversion). Our gesture dataset includes 10 body-and-hand gesture classes, and exhibit many similar subpatterns during gesticulation; tasks that are not clear to work well with non-segmented input stream. We plan to implement this for our future work.

VII. ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their helpful comments and suggestions. This work was funded by the Office of Naval Research Science of Autonomy program, Contract #N000140910625, and by NSF grant #IIS-1018055.

REFERENCES

[1] U. V. Agris, J. Zieren, U. Canzler, B. Bauer, and K.-F. Kraiss. Recent developments in visual sign language recognition. *Univ Access Inf Soc* 6:323–362, 2008.

[2] F. Althoff, R. Lindl, and L. Walchshaeusl. Robust multimodal hand- and head gesture recognition for controlling automotive infotainment systems. In *VDI-Tagung: Der Fahrer im 21*. Jahrhundert, Braunschweig, Germany, Nov 2005.

[3] G. Castellano, L. Kessous, and G. Caridakis. Emotion recognition through multiple modalities: Face, body gesture, speech. In *Affect and Emotion in Human-Computer Interaction*, pp.92–103, 2008.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pp.866–893, 2005.

[5] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt. Hidden conditional random fields for phone classification. In *INTERSPEECH*, 2005.

[6] F. Harris. On the use of windows for harmonic analysis with the discrete fourier transform. In *Proc. of the IEEE*, 66(1):51–83, 1978.

[7] M. Isard and A. Blake. CONDENSATION-conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.

[8] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pp.282–289, 2001.

[9] Z. Li and R. Jarvis. A multi-modal gesture recognition system in a human-robot interaction scenario. In *ROSE*, pp.41–46, 2009.

[10] T. M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

[11] S. Mitra and T. Acharya. Gesture Recognition: A Survey. *IEEE SMC-C* 37(3):311–324, 2007.

[12] L.-P. Morency, A. Quattoni, and T. Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In *CVPR*, pp.1–8, 2007.

[13] J. Pearl. Reverend bayes on inference engines: A distributed hierarchical approach. In *AAAI National Conference on AI*, pp.133–136, PA, 1982.

[14] A. Quattoni, M. Collins, and T. Darrell. Conditional random fields for object recognition. In *NIPS*, pp.1097–1104, 2004.

[15] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proc. of the IEEE* 77(2): 257–286, 1989.

[16] Y. Song, D. Demirdjian, and R. Davis. Tracking Body and Hands For Gesture Recognition: NATOPS Aircraft Handling Signals Database. In *FG*, 2011.

[17] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer, 2nd edition, Nov 1999.

[18] S. B. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell. Hidden conditional random fields for gesture recognition. In *CVPR*, pp.1521–1527, 2006.