To appear, *International Joint Conference on Artificial Intelligence (IJCAI) 2015.*

Continuous Body and Hand Gesture Recognition for Natural Human-Computer Interaction: Extended Abstract*

Yale Song[1†] **and Randall Davis**[2]
[1]Yahoo Labs [2]Massachusetts Instutute of Technology
[1]yalesong@yahoo-inc.com [2]davis@csail.mit.edu

Abstract

We present a new approach to gesture recognition that tracks body and hands simultaneously and recognizes gestures continuously from an unsegmented and unbounded input stream. Our system estimates 3D coordinates of upper body joints and classifies the appearance of hands into a set of canonical shapes. A novel multi-layered filtering technique with a temporal sliding window is developed to enable online sequence labeling and segmentation. Experimental results on the NATOPS dataset show the effectiveness of the approach. We also report on our recent work on multimodal gesture recognition and deep-hierarchical sequence representation learning that achieve the state-of-the-art performances on several real-world datasets.

1 Introduction

For more than 40 years, human-computer interaction has been focused on the keyboard and mouse. Although this has been successful, as computation becomes increasingly mobile, embedded and ubiquitous, it is far too constraining as a model of interaction. Gestural interaction has a number of advantages over the conventional interaction model. It uses equipment we always have on hand; there is nothing extra to carry, misplace, or leave behind. It also can be designed to work from natual and intuitive actions; gesturing is instinctive and a skill we all have, so it requires little thought, leaving the focus on the task itself, as it should be, not on the interaction modality. The goal of this work is to enable natural gesture-based interaction using a camera sensor in a non-intrusive way .

We present a new vision-based approach to gesture recognition that tracks body and hands simultaneously and recognizes gestures continuously from an unsegmented and unbounded camera input stream. Most current systems focus on one source of input, *e.g.*, body pose [Shotton *et al.*, 2011]. Yet human gesture is most naturally expressed with both body and hands. Our system tracks both body and hands, allowing a richer gesture vocabulary and more natural interaction.

Gesture recognition can be viewed as the task of sequence labeling and segmentation from an unsegmented and unbounded input stream. Morency *et al.* [2007] presented Latent-Dynamic Conditional Random Fields (LDCRF) for offline sequence labeling and segmentation, but it assumed a bounded input sequence, limiting its practical use for online gesture recognition. Our work extends LDCRF to an online domain by incorporating our novel multi-layered filtering technique with a temporal sliding window. We demonstrate our system on the NATOPS aircraft handling signals dataset [Song *et al.*, 2011b].

This extended abstract summarizes the material presented in [Song *et al.*, 2012a]. In addition, we briefly introduce our recent work on multimodal gesture recognition [Song *et al.*, 2012b] and deep-hierarchical sequence representation learning [Song *et al.*, 2013] that achieve the state-of-the-art results on a number of real-world datasets.

2 Body and Hand Tracking

As our motivating scenario, we selected an official gesture vocabulary used for communication between aircraft marshallers and pilots on a carrier deck. The vocabulary is defined in the Naval Air Training and Operating Procedures Standardization (NATOPS), a standard manual for general flight and operating procedures used by the US Navy. Our NATOPS dataset [Song *et al.*, 2011b] contains 24 gesture classes, with each gesture performed by 20 participants 20 times, resulting in 9,600 gesture instances in total. See [Song *et al.*, 2011b] for more details about the dataset.

Our system starts by receiving RGBD images from a single stereo camera.[1] As we receive the images, we perform background subtraction using a combination of a codebook approach [Kim *et al.*, 2005] and a depth-cut method, *i.e.*, once obtaining a probable region-of-interest (foreground object), we filter out pixels whose distance is further away then the foreground object. We then track body and hands from foreground RGBD images, described below.

2.1 3D Upper Body Pose Estimation

We represent an upper body using a 3D skeleton model consisting of 6 body parts (head, torso, upper and lower arms) and 9 joints (head, chest, navel, shoulders, elbows, wrists). This model is paramaterized by a 14D vector with 8 local

*This paper is an extended abstract of our ACM TiiS article [Song *et al.*, 2012a].

†The work was done when Y. Song was at MIT CSAIL.

[1]Our work predates the now-popular Kinect-based gesture recognition systems; we used a single Bumblebee2 stereo camera.

variables (3D ball-and-socket joints for shoulders, 1D revolute joints for elbows) and 6 global variables (3D translation and rotation). From the 14 parameters, we can reconstruct 3D coordinates of the 9 joints by solving the forward kinematics problem [Denavit and Hartenbert, 1955].

We formulate pose estimation as a sequential Bayesian filtering problem. Let $\mathbf{x}_t$ be a parameterization of an upper body pose and $\mathbf{I}_t$ a vector representation of an RGBD image at time t. Our goal is to estimate a posterior state density $p(\mathbf{x}_t \mid \mathbf{I}_{1:t})$ having observed images $\mathbf{I}_{1:t} = [\mathbf{I}_1 \cdots \mathbf{I}_t]$ and knowing the prior state density $p(\mathbf{x}_{t-1})$. We solve this problem using a Particle Filter [Isard and Blake, 1998], which represents $p(\mathbf{x}_t \mid \mathbf{I}_{1:t})$ as a multinomial non-Gaussian distribution with a set of N weighted particles $\{(\mathbf{s}_t^1, \pi_t^1) \cdots (\mathbf{s}_t^N, \pi_t^N)\}$. Each sample $\mathbf{s}_t^i$ represents a pose configuration, and the weights $\pi_t^{1:N}$ are obtained by computing the likelihood $\pi_t^i = p(\mathbf{I}_t \mid \mathbf{x}_t = \mathbf{s}_t^i)$ and normalizing them so that $\sum_N \pi_t^i = 1$.

Our likelihood function is defined as $p(\mathbf{I}_t \mid \mathbf{x}_t = \mathbf{s}_t^i) = \exp\{\epsilon(\mathbf{I}_t,\ \mathbf{s}_t^i)\}^{-1}$, where the fitting error $\epsilon(\mathbf{I}_t,\ \mathbf{s}_t^i)$ is a weighted sum of three error terms : 3D visible-surface point clouds, 3D contour point clouds, and a Motion History Image (MHI) [Bobick and Davis, 2001]. The first two features capture spatial discrepancies in static poses; the third captures discrepancies in the temporal dynamics of motion. We chose the weights for each error term empirically.

We initialized the Particle Filter with 500 random pose samples, and computed the prior $p(\mathbf{x}_0)$ by assuming the "T-pose" (arms stretched to the side) and fitting the model to the image with exhaustive search. Our C++ implementation based on streaming SIMD extensions 2 (SSE2) takes no more than 0.3 seconds for each iteration on an Intel Xeon 2.4 GHz machine. The pose estimate at time t is then obtained by computing the weighted sum, $\mathbb{E}[\mathbf{x}_t] = \sum_N \mathbf{s}_t^i\, \pi_t^i$.

The final body features include 3D joint velocities for elbows and wrists. To obtain this, we reconstruct a normalized skeletal model with the estimated joint angles and fixed-length limbs, so that all generated models have the same set of limb lengths across participants. This reduces cross-participant variances resulting from different body measures.

2.2 Hand Shape Classification

Our approach to hand tracking is different from body tracking in that, while we estimate 3D coordinates of body joints, we classify the appearance of a hand into one of four canonical shapes (thumb up and down, palm open and close), selected based on the types of gestures used in the NATOPS scenario.

We define two small search regions (56×56 pixels) around the estimated wrist positions, and slide a 32×32 pixel window in each region with a step size of 8 pixels, computing the HOG features [Dalal and Triggs, 2005]. We classify the hand shape using a pre-trained SVM [Chang and Lin, 2011]. The classifier is defined with 5 classes (4 positive, 1 negative) and trained on a set of manually labeled examples subsampled from the NATOPS dataset. After we classify all windows in each search region, we compute the mean probability estimate from the classification results and drop the negative class to obtain the final 8D hand feature representation.

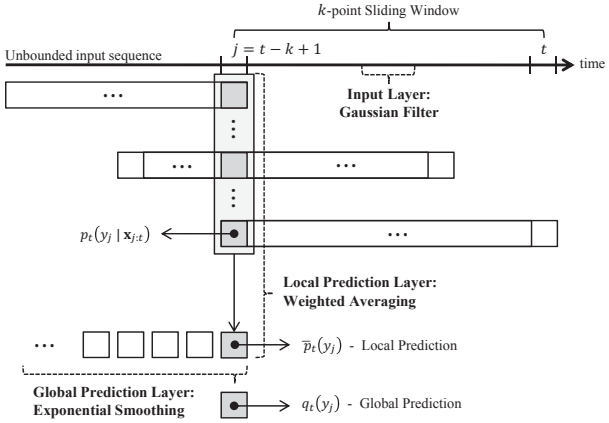

Figure 1: An illustration of multi-layered filtering.

3 Continuous Gesture Recognition

An LDCRF [Morency *et al.*, 2007] is a discriminative model with latent variables that performs sequence labeling and segmentation from a bounded input sequence. Given a pair of input sequence $\mathbf{x} = [\mathbf{x}_1 \cdots \mathbf{x}_T]$ and a label sequence $\mathbf{y} = [y_1 \cdots y_T]$ (assumed discrete), the model computes a conditional probability distribution $p(\mathbf{y}|\mathbf{x})$ by learning within-class and between-class dynamics with a sequence of latent variables $\mathbf{h} = [h_1 \cdots h_T]$. Given a test sequence $\mathbf{x}$, the model computes a label sequence by solving $\mathbf{y}^* = \arg\max_y p(\mathbf{y}|\mathbf{x})$ and obtains sequence boundaries by looking for discontinuities in the predicted label sequence. Unfortunately, the forward-backward inference (belief propagation) makes the model inapplicable to an unbounded input scenario. Various filtering techniques have been proposed to remedy this (*e.g.*, forward-only inference [Huang *et al.*, 2011]), but these often resort to an approximate inference scheme [Murphy, 2002].

3.1 Multi-Layered Filtering

We present a multi-layered filtering technique with a temporal sliding window for online sequence labeling and segmentaion, which can be used with an LDCRF while still maintaining the advantages of exact inference via belief propagation. Specifically, we define three layers of filters: a Gaussian temporal-smoothing filter in the input layer, a weighted-average filter in the local prediction layer, and an exponentiated-smoothing filter in the global prediction layer. Figure 1 illustrates our approach.

We define a k-point temporal sliding window (this provides locally-exact inference results). At each time t, a k-point window slides forward and evaluates a sequence of k frames $\mathbf{x}_{j:t} = [\mathbf{x}_{j=t-k+1} \cdots \mathbf{x}_t]$ to compute $p_t(\mathbf{y}_{j:t} \mid \mathbf{x}_{j:t})$ using an LDCRF. The result is a Y-by-k matrix, where each column vector $p_t(y_i \mid \mathbf{x}_{j:t})$ is a probability estimate of Y class labels for the i-th frame ($t - k + 1 \leq i \leq t$).

Input Layer: Temporal patterns of gestures exhibit long-range temporal dependencies, *e.g.*, body parts move smoothly and coherently as time proceeds. Because our body and hand signals are obtained via statistical estimation, the signals also exhibit high-frequency fluctuations (*i.e.*, noise). To capture long-range dependencies, previous work proposed a tech-

nique that concatenates neighboring signals within a small temporal window, creating a single large input feature vector for each time frame [Quattoni *et al.*, 2007]. This, however, increases the model complexity and does not address the noisy input problem explicitly [Song *et al.*, 2011a].

We define a Gaussian temporal-smoothing filter with a normalized ω-point weighted kernel $g(\omega)$ and perform a convolution of the input signals with $g(\omega)$. Each element of the kernel is computed as $g(\omega)[i] = \exp\{-1/2(\alpha \cdot 2i/\omega)^2\}$ for $-(\omega-1)/2 \leq i \leq (\omega-1)/2$, with α set to be inversely proportional to the standard deviation of a Gaussian distribution. Intuitively, the filter computes a weighted mean of neighboring input signals, capturing long-range dependencies and providing robustness to noise. This approach has the advantage of keeping the dimensionality of input feature vectors the same, keeping the model complexity unchanged.

Local Prediciton Layer: As the k-point window slides forward and computes $p_t(\mathbf{y}_{j:t} \mid \mathbf{x}_{j:t})$, we obtain k prediction results per frame. We make a local prediction for the first frame $\bar{p}_t(y_j)$ within the current window (*i.e.*, the tail edge) by computing a weighted average of k previous prediction results $\bar{p}_t(y_j) = \sum_{i=1}^{k} \gamma_i p_{t-i+1}(y_j \mid \mathbf{x}_{j-i+1:t-i+1})$ where $\gamma_{1:k}$ is a normalized uniform weight vector ($\gamma_i = 1/k$).

Global Prediction Layer: A sequence of local prediction results alone may still provide noisy output, resulting in highly fluctuating prediction labels. To smooth out the local prediction results, a global prediction $q_t(y_j)$ is made over the local prediction results using exponential smoothing. Intuitively, the smoothing rate should be adaptive to how confident the local prediction is, putting less weight to the past if the current prediction is highly confident. Therefore, we set the smoothing rate adaptively to the maximum probability of the local prediction $q_t(y_j) = \alpha \cdot \bar{p}_t(y_j) + (1-\alpha) \cdot q_{t-1}(y_{j-1})$ where $\alpha = \max \bar{p}_t(y_j)$ so that the more confident a local prediction is, the less smoothing performed.

4 Experiments

We evaluated various aspects of our system using the NATOPS dataset [Song *et al.*, 2011b]; here, we briefly summarize the highlights of the results, see [Song *et al.*, 2012a] for the complete results. In all our experiments, we chose the optimal parameter values via 5-fold cross validation. For parameter optimization, we selected three pairs of gestures from the NATOPS dataset (*all/not clear, insert/remove chocks, breaks on/off*) that are difficult to distinguish from each other within a pair without knowing both body and hand poses. We then evaluated our approach on all 24 gestures in the dataset.

Input layer: We evaluated our Gaussian temporal-smoothing (GTS) filter on the input layer, varying the window size ω from 0 (no smoothing) to 9 (19 frames; roughly one second in the NATOPS dataset). We compared our approach to [Quattoni *et al.*, 2007], which used temporal concatenating window (TCW) to capture long-range dependencies.

Figure 2 shows our approach outperforming TCW in terms of both the mean F1 score and the training time. The best mean F1 scores from both approaches were 0.8573 for our GTS ($\omega = 3$) and 0.8487 for the TCW ($\omega = 1$); the baseline score ($\omega = 0$) was 0.8268. In case of TCW, the mean F1 score has sharply decreased at $\omega > 1$, suggesting the model has started to overfit due to the increased model complexity, caused by the concatenation of neighboring input features. The linear increase in training time of TCW is another consequence of the increased model complexity. Our approach performs convolution of neighboring signals with a Gaussian filter and keeps the model complexity unchanged, allowing us to capture long-range temporal dependencies more effectively than the competing approach.

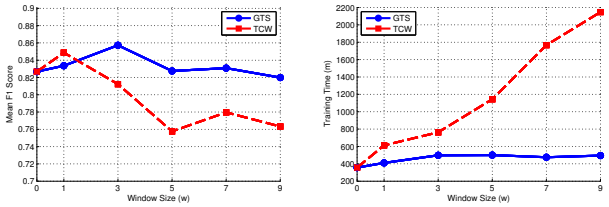

Figure 2: Mean F1 scores and training time comparisons of our Gaussian temporal-smoothing (GTS) to the temporal concatenating window (TCW) of Quattoni *et al.* [2007].

Method	Train	Validation	Test
Offline	.9514 (.03)	.8785 (.04)	.8645 (.05)
Online	.8332 (.02)	.7783 (.06)	.7668 (.05)
Online (P)	.9587 (.01)	.8907 (.04)	.8758 (.05)
Online (IP)	**.9650 (.02)**	**.8912 (.04)**	**.8977 (.05)**

Table 1: F1 scores of various approachs (means and standard deviations from 5-fold cross validation). We evaluated three verions of our online approach, enabling filtering methods on the input layer (I) and/or on the prediction layer (P).

Prediction layer: We evaluated our filters on the local and global prediction layers, varying the window size k from 20 to 80 frames (*i.e.*, one to four seconds in the NATOPS dataset). We compared our method to two baselines. The *offline LDCRF* assumes bounded input and makes predictions in batch; hence the prediction delay is infinite with unbounded input. The *online LDCRF without filtering* is similar to ours without the filtering step: it assumes unbounded input and makes predictions continuously as the observation is made, sliding a temporal window and selecting the prediction result as the last frame within the window; hence the prediction delay is zero. Our approach can be considered as an extension of the online LDCRF with filtering on the prediction layer; the prediction delay is k because of the filtering step.

Our approach outperformed the online LDCRF without filtering by a large margin, achieving a mean F1 score of 0.8758 compared to 0.7668; the difference was statistically significant ($p < 0.001$). The performance of an offline version can be considered as an empirical upper bound because it performs infernce over the entire input signal; the online versions have to make predictions based on local observations. The offline version achieved a mean F1 score of 0.8645, comparable to our filtering method (the difference was not statistically significant). Figure 3 shows a qualitative comparison of sequence segmentaion results. Online LDCRF without our multi-layered filters (middle) fluctuates over time, resulting

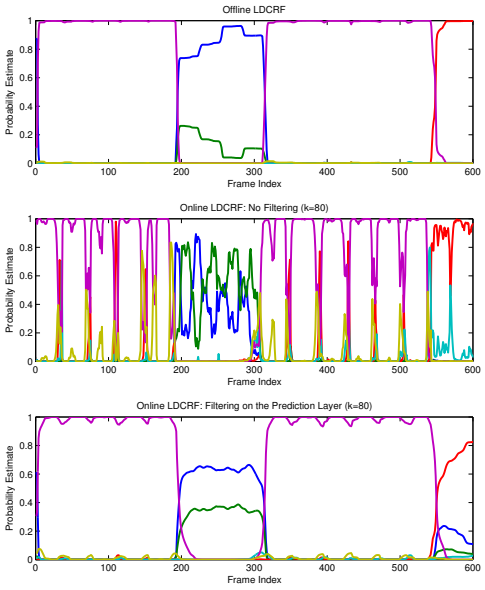

Figure 3: Qualitative evaluation of sequence segmentation. Shown here are probability estimate plots of three competing approaches. Ground-truth sequence boundaries are roughly at the 200th, the 310th, and the 550th frames.

in inaccurate sequence segmentation results. In contrast, our approach (bottom) provides more accurate results, similar to the offline version (top).

Multi-layered filtering: Table 1 shows mean F1 scores on all four approaches we tested. Our multi-layered filtering approach, Online (IP), which combines filtering on both the input and the prediction layer, provides the best performance, with an F1 score of 0.8977 on a set of 6 gestures. Figure 4 shows a confusion matrix obtained from the full 24 gestures, setting the number of latent states to 7, the LDCRF regularization factor to 10^3, the Gaussian window size $\omega = 5$, and the temporal window size $k = 60$. The overall recognition accuracy was 75.37% and the mean F1 score was 0.7349.

5 Recent Work

5.1 Multimodal Gesture Recognition

Many real-world gesture recognition tasks involve data obtained from multiple views, or *modalities*, including body postures, hand shapes, facial expressions, voice, etc. These modalities often interact with each other over time, providing important cues to understanding the behavior, *e.g.*, an angry gesture manifested by loud voice with exaggerated gestures.

In [Song *et al.*, 2012b], we presented *multi-view latent variable discriminative models* that jointly learn both modality-shared and modality-specific sub-structures to capture the interaction between modalities. Knowledge about the underlying structure of the data is formulated as a multi-chain structured latent conditional model, explicitly learning the interaction between modalities using disjoint sets of latent variables, one set per modality. The chains are tied using a pre-determined topology that repeats over time; we presented three topologies – linked, coupled, and linked-coupled – that differ in the type of interactions that they model. We showed that our model achieves the best performance on the NATOPS dataset by learning the relationship between body and hand signals. In [Song *et al.*, 2012c], we extended this model with Kernal Canonical Correlation Analysis [Hardoon *et al.*, 2004] and showed that it achieves the best performance on an audio-visual agreement-disagreement recognition task [Bousmalis *et al.*, 2011] by learning correlation and interaction between audio-visual signals.

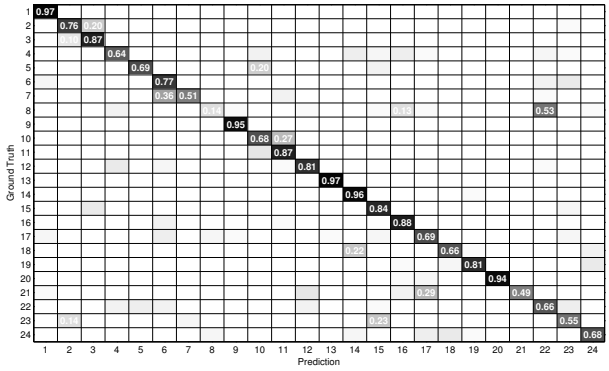

Figure 4: Confusion matrix over 24 gesture classes.

5.2 Hierarchical Sequence Summarization

Recent progress has shown that learning from deep-hierarchical feature representations can lead to improvements in various computer vision tasks [Bengio, 2009].

Motivated by the observation that human activity data contains information at various temporal resolutions, in [Song *et al.*, 2013] we presented a novel approach to action recognition that learns multiple layers of discriminative feature representations at different temporal granularities. Our approach, dubbed *hierarchical sequence summarization*, builds up a deep-hierarchical feature representation dynamically and recursively, by alternating sequence learning and sequence summarization. For sequence learning, we use a CRF with latent variables to learn hidden spatio-temporal dynamics in human action. For sequence summarization, we group local observations that share certain similarities in the latent space. For each layer we learn an abstract feature representation through neural networks. This procedure is repeated to obtain a hierarchical sequence summary representation. We developed an efficient method to train our model and showed that its complexity grows only sub-linearly with the depth of the hierarchy. We also showed that our approach achieves a near perfect recognition accuracy (99.59%) on the ArmGesture dataset [Quattoni *et al.*, 2007].

6 Conclusions

We presented a new approach to gesture recognition that tracks body and hands simultaneously and recognizes gestures continuously from an unsegmented and unbounded input stream. We also introduced our recent work on multimodal gesture recognition and deep-hierarchical sequence representation learning that achieves the state-of-the-art performance on several real-world datasets.

References

[Bengio, 2009] Yoshua Bengio. Learning deep architectures for ai. *FTML*, 2(1), 2009.

[Bobick and Davis, 2001] Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *PAMI*, 23(3), 2001.

[Bousmalis *et al.*, 2011] Konstantinos Bousmalis, L Morency, and Maja Pantic. Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition. In *FG*, 2011.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *TIST*, 2(3):27, 2011.

[Dalal and Triggs, 2005] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[Denavit and Hartenbert, 1955] J. Denavit and R. S Hartenbert. A kinematic notation for lower-pair mechanisms based on matrices. *ASME J. Appl. Mechan.,*, 22, 1955.

[Hardoon *et al.*, 2004] David Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *NECO*, 16(12), 2004.

[Huang *et al.*, 2011] Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. Virtual rapport 2.0. In *IVA*, 2011.

[Isard and Blake, 1998] Michael Isard and Andrew Blake. Condensationconditional density propagation for visual tracking. *IJCV*, 29(1), 1998.

[Kim *et al.*, 2005] Kyungnam Kim, Thanarat H Chalidabhongse, David Harwood, and Larry Davis. Real-time foreground–background segmentation using codebook model. *Real-time imaging*, 11(3), 2005.

[Morency *et al.*, 2007] L Morency, Ariadna Quattoni, and Trevor Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In *CVPR*, 2007.

[Murphy, 2002] Kevin Murphy. *Dynamic bayesian networks: Representation, inference and learning*. PhD thesis, Ph. D. dissertation, UC Berkeley, 2002.

[Quattoni *et al.*, 2007] Ariadna Quattoni, Sybor Wang, Louis-Philippe Morency, Michael Collins, and Trevor Darrell. Hidden conditional random fields. *PAMI*, (10), 2007.

[Shotton *et al.*, 2011] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. *CVPR*, 2011.

[Song *et al.*, 2011a] Yale Song, David Demirdjian, and Randall Davis. Multi-signal gesture recognition using temporal smoothing hidden conditional random fields. In *FG*, 2011.

[Song *et al.*, 2011b] Yale Song, David Demirdjian, and Randall Davis. Tracking body and hands for gesture recognition: NATOPS aircraft handling signals database. In *FG*, 2011.

[Song *et al.*, 2012a] Yale Song, David Demirdjian, and Randall Davis. Continuous body and hand gesture recognition for natural human-computer interaction. *TiiS*, 2(1):5, 2012.

[Song *et al.*, 2012b] Yale Song, Louis-Philippe Morency, and Randall Davis. Multi-view latent variable discriminative models for action recognition. In *CVPR*, 2012.

[Song *et al.*, 2012c] Yale Song, Louis-Philippe Morency, and Randall Davis. Multimodal human behavior analysis: learning correlation and interaction across modalities. In *ICMI*, 2012.

[Song *et al.*, 2013] Yale Song, Louis-Philippe Morency, and Randall Davis. Action recognition by hierarchical sequence summarization. In *CVPR*, 2013.