Continuous Body and Hand Gesture Recognition for Natural Human-Computer Interaction

YALE SONG, DAVID DEMIRDJIAN, and RANDALL DAVIS,
Massachusetts Institute of Technology

Intelligent gesture recognition systems open a new era of natural human-computer interaction: Gesturing is instinctive and a skill we all have, so it requires little or no thought, leaving the focus on the task itself, as it should be, not on the interaction modality. We present a new approach to gesture recognition that attends to both body and hands, and interprets gestures continuously from an unsegmented and unbounded input stream. This article describes the whole procedure of continuous body and hand gesture recognition, from the signal acquisition to processing, to the interpretation of the processed signals.

Our system takes a vision-based approach, tracking body and hands using a single stereo camera. Body postures are reconstructed in 3D space using a generative model-based approach with a particle filter, combining both static and dynamic attributes of motion as the input feature to make tracking robust to self-occlusion. The reconstructed body postures guide searching for hands. Hand shapes are classified into one of several canonical hand shapes using an appearance-based approach with a multiclass support vector machine. Finally, the extracted body and hand features are combined and used as the input feature for gesture recognition. We consider our task as an online sequence labeling and segmentation problem. A latent-dynamic conditional random field is used with a temporal sliding window to perform the task continuously. We augment this with a novel technique called multilayered filtering, which performs filtering both on the input layer and the prediction layer. Filtering on the input layer allows capturing long-range temporal dependencies and reducing input signal noise; filtering on the prediction layer allows taking weighted votes of multiple overlapping prediction results as well as reducing estimation noise.

We tested our system in a scenario of real-world gestural interaction using the NATOPS dataset, an official vocabulary of aircraft handling gestures. Our experimental results show that: (1) the use of both static and dynamic attributes of motion in body tracking allows statistically significant improvement of the recognition performance over using static attributes of motion alone; and (2) the multilayered filtering statistically significantly improves recognition performance over the nonfiltering method. We also show that, on a set of twenty-four NATOPS gestures, our system achieves a recognition accuracy of 75.37%.

Categories and Subject Descriptors: I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Motion*; I.5.5 [**Pattern Recognition**]: Implementation—*Interactive systems*

General Terms: Algorithms, Design, Experimentation

Additional Key Words and Phrases: Pose tracking, gesture recognition, human-computer interaction, online sequence labeling and segmentation, conditional random fields, multilayered filtering

ACM Reference Format:
Song, Y., Demirdjian, D., and Davis, R. 2012. Continuous body and hand gesture recognition for natural human-computer interaction. ACM Trans. Interact. Intell. Syst. 2, 1, Article 5 (March 2012), 28 pages.
DOI = 10.1145/2133366.2133371 http://doi.acm.org/10.1145/2133366.2133371

This work was funded in part by the Office of Naval Research Science of Autonomy program, contract no. N000140910625, and in part by the National Science Foundation grant no. IIS-1018055.
Authors' addresses: Y. Song (corresponding author), D. Demirdjian, and R. Davis, Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar St., Cambridge, MA 02139; email: yalesong@csail.mit.edu.

DOI 10.1145/2133366.2133371 http://doi.acm.org/10.1145/2133366.2133371

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. Evidence suggests that gesture-based interaction is the wave of the future, with considerable attention from both the research community (see recent survey articles by Mitra and Acharya [2007] and by Weinland et al. [2011]) and from the industry and public media (e.g., Microsoft Kinect). Evidence can also be found in a wide range of potential application areas, such as medical devices, video gaming, robotics, video surveillance, and natural human-computer interaction.

Gestural interaction has a number of clear advantages. First, it uses equipment we always have on hand: there is nothing extra to carry, misplace, or leave behind. Second, it can be designed to work from actions that are natural and intuitive, so there is little or nothing to learn about the interface. Third, it lowers cognitive overhead, a key principle in human-computer interaction: Gesturing is instinctive and a skill we all have, so it requires little or no thought, leaving the focus on the task itself, as it should be, not on the interaction modality.

Current gesture recognition is, however, still sharply limited. Most current systems concentrate on one source of input signal, for example, body or hand. Yet human gesture is most naturally expressed with both body and hands: Examples range from the simple gestures we use in everyday conversations, to the more elaborate gestures used by baseball coaches giving signals to players, soldiers gesturing for tactical tasks, and police giving signals to drivers. Considering only one source of signal (e.g., body or hand) severely restricts the expressiveness of the gesture vocabulary and makes interaction far less natural.

Gesture recognition can be viewed as a task of statistical sequence modeling: Given example observation sequences, the task is to learn a model that captures spatio-temporal patterns in the sequences, so that the model can perform sequence labeling and segmentation on new observations. One of the main challenges here is the task of online sequence segmentation. Most current systems assume that signal boundaries and/or the length of the whole sequence are known a priori. However, interactive gesture understanding should be able to process continuous input seamlessly, that is, with no need for awkward transitions, interruptions, or indications of boundaries between gestures. We use the terms *unsegmented* and *unbounded* to clarify what we mean by continuous input. Continuous input is unsegmented, that is, there is no indication of signal boundaries, such as the gesture start and end. Continuous input is also unbounded, that is, the beginning and the end of the whole sequence are unknown, regardless of whether the sequence contains a single gesture or multiple gestures. This is unlike work in most other areas with continuous input. In speech recognition, for example, most systems rely on having signal segmentation (e.g., by assuming that silence of a certain length indicates the end of a sentence) and deal with bounded conversations (e.g., making an airline reservation). Interactive gesture understanding from input that is continuous (both unsegmented and unbounded) requires that sequence labeling and segmentation be done simultaneously with new observations being made.

This article presents a new approach to gesture recognition that tracks both body and hands, and combines the two signals to perform online gesture interpretation and segmentation continuously, allowing richer gesture vocabulary and more natural human-computer interaction. Our main contributions are threefold: a unified framework for continuous body and hand gesture recognition; a new error measure, based on Motion History Image (MHI) [Bobick and Davis 2001], for body tracking that captures dynamic attributes of motion; and a novel technique called *multilayered filtering* for robust online sequence labeling and segmentation.

We demonstrate our system on the NATOPS body and hand gesture dataset [Song et al. 2011b]. Our extensive experimental results show that examining both static and dynamic attributes of motion improves the quality of estimated body features, which in turn improves gesture recognition performance by 6.3%. We also show that our multilayered filtering significantly improves recognition performance by 15.78% when added to the existing latent-dynamic conditional random field model. As we show in Section 4, these improvements are statistically significant. We also show that our continuous gesture recognition system achieves a recognition accuracy of 75.37% on a set of twenty-four NATOPS gestures.

Section 1.1 gives an overview of our system; Section 1.2 reviews some of the most related work in pose tracking and gesture recognition, making distinctions to our work; Section 2 describes body and hand tracking; Section 3 describes continuous gesture recognition; and Section 4 shows experimental results. Section 5 concludes with a summary of contributions and suggesting directions for future work.

Some of the material presented in this article has appeared in earlier conference proceedings [Song et al. 2011a, 2011b]. Song et al. [2011a] described gesture recognition of segmented input. This article extends our previous work to the continuous input domain and presents a new approach to performing online gesture interpretation and segmentation simultaneously (Section 3.2). Body and hand tracking was described in Song et al. [2011b]. Here, we include a deeper analysis of the body tracking, evaluating the performance of an MHI-based error measure we introduced in Song et al. [2011b] (Section 4.4). None of the experimental results reported in this article has appeared in any of our earlier work. Song et al. [2011b] also introduced a body and hand gesture dataset; here we give an experiment protocol on a set of all twenty-four gestures in the NATOPS dataset, and report a recognition accuracy of 75.37% (Section 4.7).

1.1. System Overview

Figure 1 shows an overview of our system. The three main components are a 3D upper-body posture estimator, a hand shape classifier, and a continuous gesture recognizer.

In the first part of the pipeline, image preprocessing (Section 2.1), depth maps are calculated using images captured from a stereo camera, and the images are background subtracted using a combination of an offline trained codebook background model [Kim et al. 2005] and a "depth-cut" method.

For 3D body posture estimation (Section 2.2), we construct a generative model of the human upper-body, and fit the model to observations by comparing various features extracted from the model to corresponding features extracted from observations. In order to deal with body posture ambiguities that arise from self-occlusion, we examine both static and dynamic attributes of motion. The static attributes (i.e., body posture features) are extracted from depth images, while the dynamic attributes are extracted from MHI [Bobick and Davis 2001]. Poses are then estimated using a particle filter [Isard and Blake 1998].

For hand shape classification (Section 2.3), we use information from body posture estimation to make the hand tracking task efficient: Two small search regions are defined around estimated wrist joints, and our system searches for hands in only these regions. A multiclass SVM classifier [Vapnik 1995] is trained offline using manually-segmented images of hands. HOG features [Freeman et al. 1998; Dalal and Triggs 2005] are extracted from the images and used as an image descriptor.

In the last part, continuous gesture recognition (Section 3), we form the input feature by combining body and hand information. A Latent-Dynamic Conditional Random Field (LDCRF) [Morency et al. 2007] is trained offline using a supervised body and hand gesture dataset. The LDCRF with a temporal sliding window is used to perform online sequence labeling and segmentation simultaneously. We augment this with our

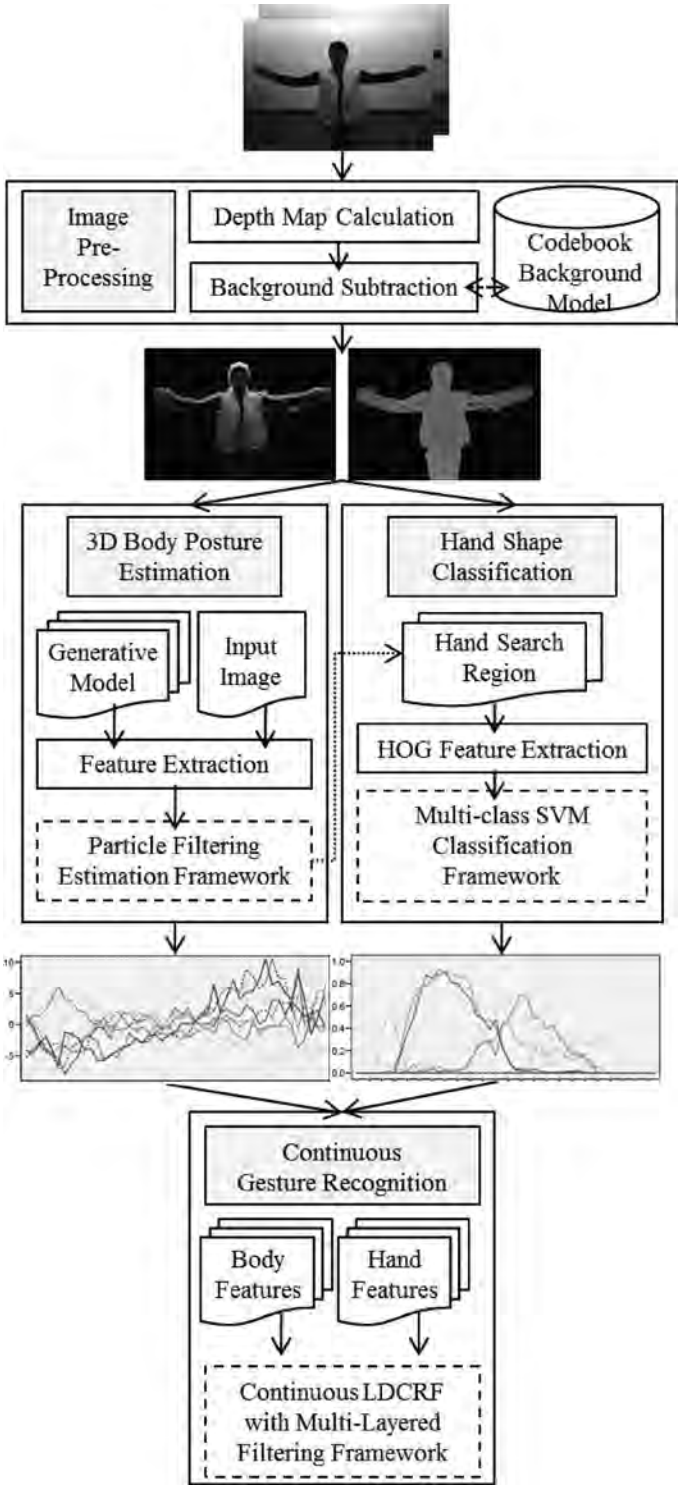

Fig. 1. A pipeline view of our unified framework for continuous body and hand gesture recognition.

multilayered filtering to make our task more robust. The multilayered filter acts both on the input layer and the prediction layer: On the input layer, a Gaussian temporal-smoothing filter [Harris 1978] is used to capture long-range temporal dependencies and make our system less sensitive to the noise from estimated time-series data, while not increasing the dimensionality of input feature vectors and keeping the model complexity the same. The prediction layer is further divided into local and global prediction layers, where we use a weighted-average filter and a moving-average filter, respectively, to take weighted votes of multiple overlapping prediction results as well as reduce noise in the prediction results.

1.2. Related Work

The topics covered in this article range broadly from body and hand tracking to gesture recognition with online sequence labeling and segmentation. This section reviews some of the most relevant work; comprehensive survey articles include Poppe [2007] for body tracking, Erol et al. [2007] for hand tracking, and Mitra and Acharya [2007] and Weinland et al. [2011] for gesture recognition.

Gesture-based interfaces typically require robust pose tracking. This is commonly done by wearing specially designed markers or devices (e.g., Vicon motion capture system or colored gloves [Yin and Davis 2010]). However, the most natural form of gestural interaction would not require additional markers or sensors attached to the body. We take a vision-based approach and perform motion tracking based on data from a single stereo camera, not using any special marker device attached to the body.

Several successful vision-based pose tracking approaches have been reported, falling generally into two categories: *model-based* methods, which try to reconstruct a pose model by fitting a kinematic model to the observed image [Deutscher et al. 2000; Sminchisescu and Triggs 2003; Lee and Cohen 2006]; and *appearance-based* methods, which assume a pose vocabulary and try to learn a direct mapping from features extracted from images to the vocabulary [Brand 1999; Shakhnarovich et al. 2003; Mori and Malik 2006]. Model-based methods are in general not affected by a camera viewpoint, do not require a training dataset, and are generally more robust in 3D pose estimation. Appearance-based methods require a large training dataset and in general are more sensitive to camera viewpoints, but once a mapping function is learned, classification is performed efficiently. Recent works take a hybrid approach, combining ideas from the two conventional methods and using advanced depth sensing cameras for 3D data acquisition (e.g., Time of Flight (ToF) [Gokturk et al. 2004] or structured light [Fofi et al. 2004]). Schwarz et al. [2011] use a ToF camera to obtain depth images. They detect anatomical landmarks to fit a skeleton body model, solving constrained inverse kinematics. A graph is constructed from the depth data, and geodesic distances between body parts are measured, making the 3D positions of anatomical landmarks invariant to pose. Similar to our work, they use optical flow between subsequent images to make tracking robust to self-occlusion. Shotton et al. [2011] obtain depth images from a structured light depth sensing camera (i.e., Microsoft Kinect). They take an object recognition approach: A per-pixel body part classifier is trained on an extensive training dataset. The results are reprojected onto 3D space, and local means are used to generate confidence-scored 3D proposals of body joints.

In this work, we take a model-based approach for body posture estimation, because reconstructing body posture in 3D space provides important information, such as pointing direction. Hand shapes, by contrast, are more categorical, that is, it is typically not crucial to distinguish fine-grained details of hand shape in order to understand a body and hand gesture. Therefore, we take an appearance-based approach to hand shape classification.

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.