A Hierarchical Approach to Continuous Gesture Analysis for Natural Multi-modal Interaction

Ying Yin
Massachusetts Institute of Technology
32 Vassar St, Cambridge MA, 02139 USA
yingyin@csail.mit.edu

ABSTRACT

I propose a systematic hierarchical approach to continuous gesture analysis using a unifying framework based on abstract hidden Markov models (AHMMs). With this framework, I will develop a gesture-based interactive interface that allows users to do both manipulative and communicative gestures without artificial restrictions, and hence enabling natural interaction.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*input devices and strategies, interaction styles*

Keywords

Multi-modal interaction; natural human computer interaction; fingertip tracking; hand tracking

1. INTRODUCTION

Consider the following scenario:

> *An earthquake has hit a populous area. You are the incident commander at the crisis management center coordinating the search and rescue teams working in the field. There is a large tabletop display in front of you, showing the map of the site with live information coming from the field. A report about a big explosion at a chemical plant comes in and you move the map around, zoom in and rotate it to get a good view of the plant. Then you select a group of unmanned vehicles on the map with your hand, and speak to the interface "Go nearer to the explosion site to gather more information," while tracing the route the vehicles should take to avoid obstacles. After that you instruct a rescue team to evacuate the residents in the surrounding buildings by going under a bridge because the surface of the bridge is blocked. You gesture with one hand as the bridge and the other hand moving underneath it to emphasize this.*

The scenario above is an example in the Urban Search and Rescue (USAR) domain. It shows an application of a multimodal interface to a real-world problem. Gestures play an important part in this scenario, providing key information about location, method and timing of movements, and about spatial relationship among the objects being described [13].

Recent trends in user interfaces have brought on a new wave of interaction techniques that depart from the traditional mouse and keyboard. These include multi-touch interfaces such as the iPhone, the iPad and the Microsoft Surface® as well as camera-based systems such as the Microsoft Kinect and the Nintendo® Wii. Most of these devices gained instant popularity among consumers because they make interacting with computation more natural and effortless. Users feel more natural to directly manipulate the virtual objects by hands and/or body gestures as this is how we interact with our environment in everyday life.

My goal is to take this aspiration to the next level by developing an intelligent multi-modal interface for natural interaction. By *natural interaction*, I mean the kind of cognitively transparent, effortless multi-modal communication that can happen between people; I want to make this possible in human-computer interaction such that the computer interface understands what the user is saying and doing, and the user can simply behave.

Gesture plays an important part in multi-modal interaction, especially for conveying spatial information. The focus of my doctoral research is developing a hierarchical approach for continuous gesture analysis that can be easily applied in different domains and applications. Specifically, I focus on gestures made with hands.

I propose to develop a gesture-based interface that allows users to perform manipulative and communicative gestures continuously with no arbitrary restrictions. The key elements of the approach are the hierarchical framework for gestural analysis based on abstract hidden Markov models (AHMMs) and the accurate real-time detection of the onset of natural gestures that convey information intended by the users, while filtering out movements that do not convey information.

2. BACKGROUND

In a computer controlled environment people use hands to mimic both the natural use of the hand as a manipulator, and its use in human-machine communication [8]. Hence, a

ICMI'12, In Proceedings of the 14th ACM International Conference on Multimodal Interaction, pages 357–360, 2012.

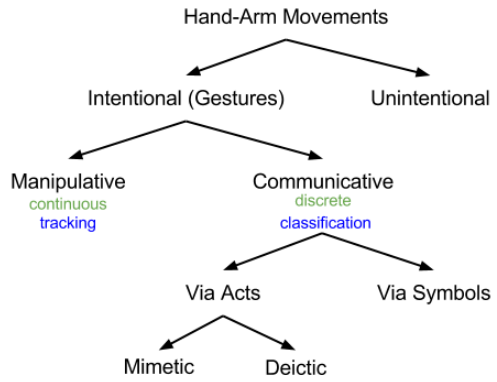

Figure 1: Gesture taxonomy

natural interface should handle both manipulative and communicative gestures.

Gestures originate as a mental concept, and are expressed through the motion of arms and hands [8]. The mental concept represents the intention of the gesturer. In this dimension, hand movements can be divided into different categories, each has its own characteristics and requires different responses from the computer interface. The gesture taxonomy in Figure 1, modified from [8], is the basis of my hierarchical approach to gesture interpretation.

Gestures, as *intentional* movements, should be distinguished from *unintentional* hand movements (like beats). By *unintentional* movements, I mean movements that are not intended to convey information. The distinction is important for a natural interface because there should not be any restriction on how people should place or move their hands when they are not doing any meaningful gestures.

Gestures are then further divided into manipulative and communicative categories. For manipulative gestures, the system needs to respond frame by frame while the user is doing certain direct manipulations. This can be achieved by tracking the hand state in real-time. For communicative gestures, the system needs to respond to the meaning of the gesture when the user finishes the gesture. This part is accomplished by gesture recognition.

3. RESEARCH QUESTIONS & HYPOTHESES

This thesis focuses on several research questions that aim to solve some of the key challenges in developing a multimodal interface with natural gesture input.

- **Fine-grained hand pose estimation.** The hand model for gestural input can range from a very simplistic one (one point) to a very comprehensive one (a 26 degrees of freedom (DOF) skeleton). There is a trade off between the accuracy of the model and computational complexity. The question here is what hand model is sufficient for the HCI tasks for both manipulative and communicative gestures.

  My hypothesis is that a simplified 3-D skeletal hand model with fingertip positions is useful for manipulative gestures because we need to know exactly where the fingertips are. For communicative gestures, we only need to know the meaning of the gesture instead of the exact spatial parameters. Hence example-based template models would be more suitable which also require less computation.

- **Online gesture spotting.** A real-time gesture-driven HCI application requires gesture spotting with minimum delay. The questions are how to accurately determine the start and the end of a gesture in real-time, and how to filter out unintentional hand movement without arbitrary restrictions.

  My hypothesis is that the characteristics in the movement of hands (such as speed, repetitive patterns, and hand poses) and the context of the applications can be used to differentiate gestures from non-gestures, and manipulative gestures from communicative ones.

- **Unified hierarchical model for gestural analysis.** I propose a unified framework based on AHMMs that distinguishes unintentional movements, manipulative gestures and communicative gestures in real-time and provides appropriate responses to the user. The research questions are how to first differentiate the broader categories of gestures, and then further classify the gestures into more specific categories, and whether this improves the recognition accuracy.

  My hypothesis is that by dividing the hand movements into broader categories first, then using different models for different categories can improve the performance of the system because the features that are relevant for each category are different. For manipulative gestures, we need to track the positions and orientation of fingertips and hands so that the position and the orientation of the virtual object being manipulated can be updated. For deictic gestures, we need to know only the general pointing direction. For symbolic gestures, both the static hand poses and the dynamic hand movements need to be considered as features in the models.

4. RELATED WORK

There is a large body of active research on using the hand as an input modality for human-computer interaction.

As [1], I use the Kinect sensor to capture hand movements and configurations. However, in addition to detecting touch-based gestures, I also detect above-surface 3D gestures. Tracking fingertips is usually sufficient for manipulating objects on the 2D surface [8] as exemplified in [7, 1, 4]. I developed my own finger tracking method which makes fewer assumptions about hand positions as compared to [1].

To distinguish meaningful gestures from unintentional movements, i.e., *gesture spotting*, Peng et al. [9] train several nongesture hidden Markov models (HMMs) by identifying nongestures from the training data. But in real life the possible unintentional movements are limitless, and it is not clear how this approach will scale.

For dynamic gesture recognition, HMM is a commonly used technique [11]. Wang et al. [12] introduced hidden conditional random fields (HCRF) for gesture recognition, but they do not handle continuous input. Morency et al. [5] present a latent-dynamic conditional random field (LDCRF) model that is able to perform sequence labeling and segmentation simultaneously for bounded input sequences. There is not much prior work that distinguishes manipulative gestures and communicative gestures. Oka et al. [7]

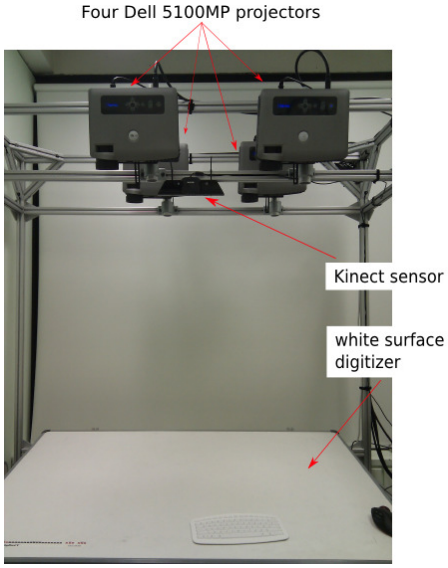

Figure 2: System setup.

developed a system that allows both direct manipulation and symbolic gestures. They regard gestures with an extended thumb as direct manipulation and those with a bent thumb as symbolic gestures. But this seems to be arbitrary and "unnatural".

5. PROPOSED WORK AND PROCEDURE

Our custom tabletop interface includes four 1280×1024 pixel projectors that provide a 2560×2048 pixel resolution display. The display is projected onto a flat white surface digitizer. The projected displays were mechanically aligned to produce a single seamless large display area. One Microsoft Kinect motion sensor is placed above the center of the tabletop among the projectors (Figure 2).

5.1 Hand Tracking

I use the averaging background method for background subtraction, which learns the average and average difference of each pixel in the depth image. Then I use 1 iteration of morphological opening to clear out small pixel noise. I find connected components by finding all contours that are greater than the minimum length of a hand. These components are considered to be the forelimbs and are approximated with convex hulls and bounding boxes. The hand region is at either end of the bounding box depending on the position of the arm relative to the table.

I base the estimation of the hand model on geometric properties of the hand. First, I compute the convexity defects from the convex hull of the forelimb. For an extended finger, it has one convexity defect on each side, and the two adjacent sides of the defects form an acute angle. Then, I iterate through the adjacent convexity defects, and mark the intersection of those sides that form an angle smaller than a threshold value as the potential fingertip. I further refine the fingertip position by searching in the direction of the finger and finding the edge where the gradient of the depth value is greater than 0.05.

5.2 Hierarchical Gestural Analysis

I propose a systematic approach to enabling gestural interaction according to the hierarchical taxonomy of gestures. I will explore the differences between unintentional movements, manipulative gestures and communicative gestures. Unintentional movements, like "beats", tend to be fast, repetitive, and close to the body; manipulative gestures tend to be slower, and have definite hand poses. I can also use the context constraints to narrow down the possibilities of gestures. For example, if the hand is at the position of a movable object, the probability it is a manipulative gesture is higher. I can combine the hand movement features and the context constraints to build a probabilistic model to differentiate the broader categories of the hand movements.

5.2.1 Abstract Hidden Markov Model

I will combine different levels of recognition using the abstract hidden Markov model (AHMM), which is a probabilistic model used to explain the interaction between behaviors at different levels of abstraction [2]. I can map the levels in the gesture taxonomy into the levels of abstraction in an AHMM. The higher abstract levels represent the mental intention of the gesturer. I need to determine what is the optimal number of levels to use in AHMM. One possibility is using the number of levels in the hierarchy of the taxonomy. But the greater the number, the greater the computational complexity. A flatter model may be more reasonable and may not affect accuracy very much. The bottom level in the AHMM consists of observations and states in a typical HMM. An HMM is suitable for modeling sequential data such as time series, and has been used widely for dynamic gesture recognition with reasonable success. Kalman filter model is widely used for tracking objects. It can be viewed as a special case of HMM where state transition probabilities and emission probabilities are all Gaussian. As a result, I propose to combine these two together under a unified framework of AHMM where a Kalman filter is used for tracking hands for manipulative gestures and HMMs are used for dynamic gesture recognition. Static gestures are treated as a special case of dynamic gestures with only one state.

I will start with a simple 1-level AHMMs [6] (Figure 3). G_t represents the mental concept of the gesture that the gesturer currently has. It includes unintentional movements (U), manipulative gestures (M), and various communicative gestures (C_i). S_t is the hidden state of the hand pose and movement, which is essentially a vector quantization of the actual, observed (but noisy) feature vector X_t. F_t^G is a binary indicator variable that is "on" (has value 1) if the lower level HMM at time t has just "finished" (i.e., is about to enter an end state), otherwise it is "off" (has value 0).

AHMM is a generative model. I will use discriminative training methods to further improve the discriminative power of the model.

5.2.2 Gesture Spotting

I will detect the start and the end of a hand movement by the amount of motion. This forms a segment which is a candidate hand movement. Instead of using hard coded threshold values for detecting candidate start and end frame as in [3], I will use machine learning methods to make the result more robust. I propose to use a support vector machine (SVM) to classify each frame into either a candidate start, candidate end or neither. One important question

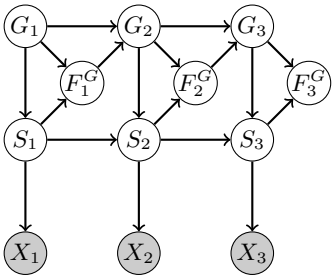

Figure 3: A 1-level AHMM.

here is what features I should use for the feature vector. I will experiment with the criteria used in [3] which are motion energy, static poses, and curvature, and will also explore other possible features such as positions relative to the body or table, and accompanying speech etc.

5.3 Multi-modal Interaction with Speech

I will use an off-the-shelf speech recognizer, and investigate how to combine speech and gesture to make the interaction more natural. I will explore the alignment of speech and gesture, mutual disambiguation of speech and gesture, and how speech and gesture complement each other.

In addition to using deictic gestures to provide spatial information as a complement to speech [10], I will explore the use of speech as a complement to manipulative gestures based on the finding from our user study [15]. I observed that manipulative gestures are at times accompanied by adjectives and adverbs that refine the actions.

5.4 Motivating Application and User Study

I am developing a browser-based game-like application that allows the use of hand gesture for interaction. So far, I have developed the infrastructure that can enable gesture input to a web application. Some of the gestural input functionalities I want to demonstrate are: moving units from one location to another; panning the map; command a unit to build different kinds of buildings; command units to attack the enemies by doing an attack gesture and pointing to the location of attack.

I will bootstrap the system with a set of natural gestures defined based on the user studies done by Yin et al. [15] and Wobbrock et al. [14] on the set of natural gestures people do for surface computing. I will next conduct a user study where the user can perform both manipulative and communicative gestures, aiming to evaluate the accuracy of hand tracking and gesture recognition.

6. CONTRIBUTIONS

To date, I have developed the finger tracking method described in Section 5.1 using a depth sensor. The average fingertip position error (Euclidean distance) is 5.3px, about 10mm on the tabletop. The method is non-obtrusive, easy to setup and relies on fewer assumptions about how users move their hands on the tabletop. I believe our method is general and accurate enough to allow manipulative interaction using hands on a tabletop display.

The expected contribution of this work is a unified framework for continuous gesture input that allows users to perform manipulative and communicative gestures seamlessly with no arbitrary restrictions. Part of the framework also includes a machine learning based probabilistic model for detecting the onset of natural gestures that convey information intended by the users, while filtering out movements that do not convey information.

7. ACKNOWLEDGEMENTS

I would like to thank my advisor Professor Randall Davis for his support and invaluable advice for my thesis work.

8. REFERENCES

[1] C. Harrison, H. Benko, and A. Wilson. Omnitouch: wearable multitouch interaction everywhere. In *UIST'11*, pages 441–450, 2011.

[2] J. Johns and S. Mahadevan. A variational learning algorithm for the abstract hidden markov model. In *Proceedings of the 20th national conference on Artificial intelligence - Volume 1*, pages 9–14, 2005.

[3] H. Kang, C. Woo Lee, and K. Jung. Recognition-based gesture spotting in video games. *Pattern Recognition Letters*, 25(15):1701–1714, 2004.

[4] E. Larson, G. Cohn, S. Gupta, X. Ren, B. Harrison, D. Fox, and S. Patel. Heatwave: thermal imaging for surface user interaction. In *CHI'11*, pages 2565–2574, 2011.

[5] L. Morency, A. Quattoni, and T. Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In *CVPR'07*, pages 1–8, 2007.

[6] K. Murphy. *Dynamic bayesian networks: representation, inference and learning*. PhD thesis, University of California, 2002.

[7] K. Oka, Y. Sato, and H. Koike. Real-time fingertip tracking and gesture recognition. *IEEE Computer Graphics and Applications*, 22(6):64–71, 2002.

[8] V. I. Pavlovic, R. Sharma, and T. S. Huang. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:677–695, 1997.

[9] B. Peng and G. Qian. Online gesture spotting from visual hull data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(6):1175–1188, 2011.

[10] I. Rauschert, P. Agrawal, I. R. Pyush, and R. Sharma. Designing a human-centered, multimodal GIS interface to support emergency management, 2002.

[11] R. Sharma, J. Cai, S. Chakravarthy, I. Poddar, and Y. Sethi. Exploiting speech/gesture co-occurrence for improving continuous gesture recognition in weather narration. In *FG'00*, pages 422–427. IEEE, 2000.

[12] S. Wang, A. Quattoni, L. Morency, D. Demirdjian, and T. Darrell. Hidden conditional random fields for gesture recognition. In *CVPR'06*, volume 2, pages 1521–1527. IEEE, 2006.

[13] A. Wexelblat. *A feature-based approach to continuous-gesture analysis*. PhD thesis, MIT, 1994.

[14] J. Wobbrock, M. Morris, and A. Wilson. User-defined gestures for surface computing. In *CHI'09*, pages 1083–1092, 2009.

[15] Y. Yin and R. Davis. Toward natural interaction in the real world: Real-time gesture recognition. In *ICMI'10*, November 2010.