Office Activity Recognition using Hand Posture Cues

Brandon Paulson
Sketch Recognition Lab
3112 TAMU
College Station, TX 77843
bpaulson@cs.tamu.edu

Tracy Hammond
Sketch Recognition Lab
3112 TAMU
College Station, TX 77843
hammond@cs.tamu.edu

ABSTRACT
Activity recognition plays a key role in providing information for context-aware applications. When attempting to model activities, some researchers have looked towards Activity Theory, which theorizes that activities have objectives and are accomplished through tools and objects. The goal of this paper is to determine if hand posture can be used as a cue to determine the types of interactions a user has with objects in a desk/office environment. Furthermore, we wish to determine if hand posture is user-independent across all users when interacting with the same objects in a natural manner. Our initial experiments indicate that a) hand posture can be used to determine object interaction, with accuracy rates above 94% for a user-dependent system, and b) hand posture is dependent upon the individual user when users are allowed to interact with objects as they would naturally.

Categories and Subject Descriptors
I.5.4 [**Pattern Recognition**]: Applications.

General Terms
Algorithms, Performance, Experimentation, Human Factors.

Keywords
Activity recognition, CyberGlove, hand posture, office environment, context-aware, wearable computing.

1. INTRODUCTION
As the future of computing heads toward ubiquity, wearable computing, coupled with context-aware applications like Starner's proposed Remembrance Agent [25], will become more prevalent. According to Dey and Abowd, the most important types of context are location, identity, activity, and time [6]. Of these forms of context, activity is one of the hardest to capture and is used less frequently by many context-aware applications [7]. However, we believe activity-based context can play a significant role in applications, particularly those involving wearable and pervasive computers.

The goal of activity recognition is to aid context-aware applications by providing information to help explain what a user is currently doing (i.e. what activity the user is engaged in). An issue activity recognition researchers face, however, is how to define what an activity is and how to determine when it is taking place. One answer may lie in "Activity Theory" [13,20]. According to this theory, activities have objectives and are accomplished via tools and objects. Therefore, one can assume that if we can determine the object that a user is interacting with, then we may be able to imply something regarding the activity that the user is currently engaged in. Some frameworks have been created to model activities in this manner, but were implemented in virtual environments in which interaction with objects is assumed to be given by some form of sensor values [27,28]. When applying such frameworks to a real-world domain, we still face the issue of determining when an appropriate interaction is taking place. In order to achieve full contextual-awareness, one must address the category of contextual sensing as it is the lowest, most basic part of context-aware applications [22].

The two most common approaches to tracking haptic activity and interaction in a real-world setting are through cameras or wearable sensors. Vision-based techniques require cameras that are either placed within a room [19], or that are wearable [17]. Stationary cameras placed in a room have the advantage of being less obtrusive to the user, but makes the context-capturing system static to that one location. Wearable cameras allow for context-capturing systems to become mobile, but still have the problem which most vision-based approaches experience when dealing with the interaction of objects: occlusions (which typically occur because of the object itself).

Because our interests lay more with interaction and less with writing vision-based algorithms to handle occlusions, we decided to use glove-based sensor input provided by Immersion's 22-sensor CyberGlove II. The primary goal of our work to determine if hand posture can be used a cue to help determine the objects a user interacts with, thus providing some form of activity-related context. To give some real-world practicality to our problem, we chose to perform our experiments in an office domain, a setting we believe could benefit from context-aware applications.

A secondary goal of our work is to determine the variability of hand postures between different users who interact with the same objects or are asked to perform the same gestures. In other words, when users are allowed interact with objects or perform gestures as they would naturally, will they use similar or dissimilar hand postures? This motivation of creating systems which allow natural interaction rather than forcing the user to learn specified behaviors is shared with our previous work in sketch recognition [9].

2. RELATED WORK
The term "activity recognition" does not solely include the recognition of activity through objects, as defined by Activity Theory, but also includes the recognition of activities a user performs with her own body. These "ambulatory" activities typically include standing, walking, running, sitting, ascending/descending stairs, and other common body-related

movements. Previous works have attempted to recognize movement-related activities using vision-based approaches [26] or wearable accelerometers [2,4,8,12,14]. Some approaches include other inputs like ambient light, barometric pressure, humidity, and temperature [15]. Minnen et al. use accelerometers and audio in order to capture "interesting" behavior in a journal which could be used to help treat people with behavioral syndromes like autism [18]. In [29], Ward et al. also use a combination of accelerometers and sound in order to determine activity in a workshop setting.

Other works have also shared a common motivation of activity recognition through objects; however, their approaches have been different than ours. In [23,24], objects are tagged with radio-frequency identification (RFID) sensors. The user then wears a glove with a built-in RFID tag reader which enables the system to determine the objects being interacted with. While this approach is not prone to much error, it constrains the user to only interacting with the tagged objects in his environment.

Some works have shared a similar domain as ours (an office setting). In [1], a vision-based system is used but requires prior knowledge about the layout of the room. This system is mainly used for security purposes, in order to determine if unauthorized peoples are performing unauthorized activities. Oliver et al. propose a system that includes audio, video, and computer interaction input [21]. However, their work focused primarily on providing an environmental context (i.e. determining if a conversation is taking place or a presentation is being given etc.) rather than determining the objects a user is interacting with.

The idea of using hand posture to recognize types of grasps has been proposed in previous works using both vision-based optical markers [5] and glove-based input [3]. While these works are similar to what we have done, we should make the distinction that our goal is to recognize objects rather than grasp types. Many times objects are interacted with using similar grasp types. Our goal is to determine if hand posture can yield a fine enough resolution to determine the object a user is currently interacting with, even if that object is grasped with the same grasp type as another object in the domain.

3. EXPERIMENT

For our experiments, we utilized a single, right-handed, wireless, 22-sensor CyberGlove II device developed by Immersion Corporation, sampled at 10 readings a second. The glove provides three flexion sensors per finger, four abduction (finger spread) sensors, a palm-arch sensor, and sensors that measure wrist flexion and abduction.

We had a total of 8 users participate in our data collection. All of the users were graduate students, and about half had previous experience interacting through the CyberGlove, mainly for sign language recognition tasks. We asked users to interact with objects that may be typically found around the desk in an office. Table 1 lists the 12 types of interactions we collected. Each user performed each interaction 5 different times, during which the user's hand posture was capture through the CyberGlove. For each interaction, we averaged the values of each sensor to create an input vector for our classifier.

Because we are focused primarily on determining how much information hand posture provides us, as well as, determining whether or not this information is common across all users, we decided to use simple 1-nearest neighbor classifier that could give us decent baseline results.

Table 1. The types of interactions we collected for our experiment.

Drinking from a cup	Dialing a telephone
Picking up earphones	Typing on a keyboard
Using a mouse	Drinking from a mug
Opening a drawer	Reading a piece of paper
Writing with a pen	Stapling papers
Answering a telephone	Waving to an office mate

4. RESULTS

We performed a set of tests to determine how well hand posture could be used for recognition on both a user-independent and user-dependent system. To determine the accuracy of a user-independent system, we performed leave-one-out cross-validation across all 8 of our users. This form of testing mimics a system which is trained offline and then used by a brand new user. The average accuracy of this type of system across all users was 62.5% with minimum accuracy of 41.7% and a maximum accuracy of 81.7%.

To simulate the effect of a user-dependent system, we trained the classifier using only data from a given test user. We tried different combinations of training and testing. In the first test, a single, random example of each interaction was used for training while the remaining examples were used for testing. In the next experiment, two samples of each interaction were randomly selected to use for training and three remaining samples were used to test, and so on. With a single user-specified training example for each interaction, the system outperforms the user-independent system with an average accuracy of 78.9%. When the system is trained with 4 examples of each interaction, the system achieves an accuracy of 94.2%.

5. DISCUSSION

Overall, it can be concluded that when users are allowed to interact with objects in a natural, unconstrained manner, a user-dependent system will likely be more suitable for recognition

Figure 1: Examples of variations in interaction with the same objects.

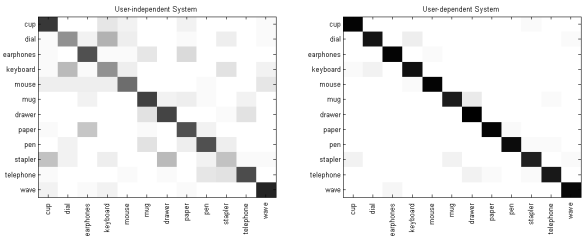

Figure 2: Confusion matrix for the user-independent system (left) and user-dependent system (right) for (top to bottom, left to right): cup, dial, earphones, keyboard, mouse, mug, drawer, paper, pen, stapler, telephone, and wave.

rather than a user-independent system. The reason for such a poorly performing user-independent system was due to a high degree of variation in the way in which users interacted with the same objects. Figure 1 shows some examples of these interaction variations.

When looking at the confusion matrix (Figure 2) for the user-independent system, it is easier to see the examples of interactions that had a high degree of variation across multiple users. In Figure 2, dark areas indicate high confusion with lighter areas indicating little confusion.

According to user-independent confusion matrix, the three actions that contained the most variance across all users (lightest colors along the diagonal) were dialing the telephone, typing on the keyboard, and stapling papers together. We can also see that there was a lot of confusion between dialing the telephone and typing on a keyboard. We believe most of this confusion can be attributed to the users who dialed the telephone with an open palm, the posture typically used when typing. Other areas of high confusion included: stapling and drinking from a cup (when both were interacted with using a circular grip), picking up earphones and picking up a piece of paper (both of which use pinching postures), and opening a drawer and using a stapler (both of which can be done with a circular grip).

The confusion matrix for the user-dependent system gives us idea of the interactions that varied most across an individual user. Obviously, this matrix contains much less confusion overall that the user-independent system; however, there are still some areas of confusion. Most notably, is the confusion that still occurs between typing on the keyboard and dialing the telephone. As with the user-independent system, there is still some confusion between interactions that can occur with objects held with a circular grip: cup and stapler, mug and stapler, drawer and telephone, and drawer and mug. There was also a small amount of confusion between picking up earphones and waving. We believe this was due to a single user who waved by bending the four non-thumb fingers towards the palm, rather than waving the hand with all fingers extended. Because of this type of wave, there were occasional examples of confusion with the pinching associated with holding a piece of paper.

6. FUTURE WORK

In this work we have seen that hand posture can be used as a cue in performing object-based activity recognition. However, we have also seen that some objects are interacted with using similar postures, even in a user-dependent system. For future work, we hope to combine hand posture with other forms of input that measure hand movement. These inputs could come in the form of accelerometers or 3-D position trackers. We believe that this extra information could be used to disambiguate interactions like using a stapler and drinking from a cup, both of which could potentially use a similar posture but would likely have different movements associated with them. This would essentially make our approach activity recognition based on hand gestures rather than hand postures.

The other obvious area for future work deals with segmentation and noise detection. In our experiments, data was recorded on an isolated interaction basis. For this approach to be beneficial to a real-world system, we would need to develop ways of designating when an interaction starts and stops. We also need to be able to detect instances when no interaction is taking place at all. The issue of segmenting hand postures and gestures is still an ongoing research effort [10,16].

In addition to these issues, we also plan to try our experiments with more sophisticated classifiers. We have showed that reasonable results can be given with a simple 1-nearest neighbor algorithm. In the near future we hope to implement and test other algorithms like neural networks, support vector machines, and hidden Markov models (HMMs). Using these more advanced classifiers will likely lead to higher accuracy, and may also provide an extra advantage of providing automatic segmentation. For example, Iba et al. were able to successfully recognize and segment hand gestures used to control mobile robots by introducing a ``wait state" into their HMM [11]. It will also be beneficial to analyze dimensionality reduction techniques, as the CyberGlove contains many sensors which could be producing extra noise during hand posture recognition.

7. CONCLUSION

In this paper we have shown that hand posture can be used as a cue to perform object-based activity recognition, which can provide important context to context-aware applications. Furthermore, we have determined that when users are allowed to interact with objects as they would naturally, a user-dependent system is preferable over a user-independent system because of the high variation between users interacting with the same objects. We have shown that such a user-dependent system is capable of producing up to 94% accuracy using a simple nearest neighbor algorithm along with the raw sensor values from the CyberGlove II.

8. ACKNOWLEDGMENTS

Funding provided in part by NSF IIS Creative IT Grant #0757557 and NSF IIS HCC Grant #0744150.

9. REFERENCES

[1] Ayers, D. and Shah, M. Monitoring human behavior from video taken in an office environment. *Image and Vision Computing 19*, 12 (2001), 833-846.

[2] Bao, L. and Intille, S.S. Activity recognition from user-annotated acceleration data. *Pervasive Computing 3001*, Springer, (2004), 1-17.

[3] Bernardin, K., Ogawara, K., Ikeuchi, K., and Dillmann, R. A sensor fusion approach for recognizing continuous human grasping sequences using hidden Markov models. *IEEE Transactions on Robotics 21*, 1 (Feb. 2005), 45-57.

[4] Bussmann, J.B.J., Martens, W.L.J., Tulen, J.H.M., Schasfoot, F.C., van den Berg-Emons, H.J.G., and Stam, H.J. Measuring daily behavior using ambulatory accelerometry: the activity monitor. *Behavior Research*

Methods, Instruments, & Computers 33, 3 (2001), 349-356.

[5] Chang, L.Y., Pollard, N.S., Mitchell, T.M., and Xing, E.P. Feature selection for grasp recognition from optical markers. In *Proceedings IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '07)* (San Diego, California, Oct. 29-Nov. 2, 2007), 2007, 2944-2950.

[6] Dey, A.K. and Abowd G.D. Towards a better understanding of context and context-awareness. In *Workshop on the What, Who, Where, When, and How of Context-Awareness at the 2000 Conference on Human Factors in Computing Systems (CHI '00*) (The Hague, The Netherlands, April 1-6, 2000). ACM Press, New York, NY, 2000.

[7] Dey, A.K., Salber, D., Abowd, G.D., and Futakawa, M. The conference assistant: combining context-awareness with wearable computing. In *Third International Symposium on Wearable Computers (ISWC '99)* (San Francisco, CA, USA, Oct. 18-19, 1999). 1999, 21-28.

[8] Foerster, F., Smeja, M., and Fahrenberg, J. Detection of posture and motion by accelerometry: a validation study in ambulatory monitoring. *Computers in Human Behavior 15*, 5 (1999), 571-583.

[9] Hammond, T., Eoff, B., Paulson, B., Wolin, A., Dahmen, K., Johnston, J., and Rajan, P. Free-sketch recognition: putting the chi in sketching. In *Extended Abstracts on Human Factors in Computing Systems (CHI '08)* (Florence, Italy, April 5-10, 2008), ACM Press, New York, NY, 2008, 3027-3032.

[10] Harling, P.A. and Edwards, A.D.N. Hand tension as a gesture segmentation cue. In *Proceedings of Gesture Workshop on Progress in Gestural Interaction*, Springer-Verlag, London, UK, 1997, 75-88.

[11] Iba, S., Weghe, J.M.V., Paredis, C.J.J., and Khosla, P.K. An architecture for gesture-based control of mobile robots. In *Proceedings IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '99)* (Oct. 17-21, 1999), 1999, 851-857 vol.2.

[12] Kern, N., Schiele, B., and Schmidt, A. Multi-sensor activity context detection for wearable computing. *Ambient Intelligence 2875*, Springer, (2003), 220-232.

[13] Kuutti, K. Activity theory as a potential framework for human-computer interaction research. *Context and Consciousness: Activity Theory and Human-computer Interaction*. MIT Press, Cambridge, MA, 1995.

[14] Lee, S.-W. and Mase, K. Activity and location recognition using wearable sensors. *IEEE Pervasive Computing 1*, 3 (2002), IEEE Educational Activities Department, 24-32.

[15] Lester, J., Choudhury, T., Kern, N., Borriello, G., and Hannaford, B. A hybrid discriminative/generative approach for modeling human activities. In *International Joint Conferences on Artificial Intelligence (IJCAI '05)*, 2005, 766-772.

[16] Li, C. and Prabhakaran, B. A similarity measure for motion stream segmentation and recognition. In *Proceedings of the 6th International Workshop on Multimedia Data Mining (MDM '05)* (Chicago, Illinois, Aug. 21, 2005), ACM Press, New York, NY, 2005, 89-94.

[17] Mayol, W.W. and Murray, D.W. Wearable hand activity recognition for event summarization. In *Proceedings of the Ninth IEEE International Symposium on Wearable Computers (ISWC '05)* (Oct. 18-21, 2005), IEEE Computer Society, 2005, 122-129.

[18] Minnen, D., Starner, T., Ward, J.A., Lukowicz, P., and Troster, G. Recognizing and discovering human actions from on-body sensor data. In *IEEE International Conference on Multimedia and Expo (ICME '05)* (July 6, 2005), 2005, 1545-1548.

[19] Moore, D.J., Essa, I.A., Hayes, M.H., III. Exploiting human actions and object context for recognition tasks. In *Proceedings of the Seventh IEEE International Conference on Computer Vision (ICCV '99)*, 1999, 80-86 vol.1.

[20] Nardi, B.A. Studying context: a comparison of activity theory, situated action models, and distributed cognition. *Context and Consciousness: Activity Theory and Human-computer Interaction*. MIT Press, Cambridge, MA, 1995.

[21] Oliver, N., Garg, A., and Horvitz, E. Layered representations for learning and inferring office activity from multiple sensory channels. *Computer Vision and Image Understanding 96*, 2 (2004), 163-180.

[22] Pascoe, J. Adding generic contextual capabilities to wearable computers. In *Second International Symposium on Wearable Computers (ISWC '98)* (Oct. 19-20, 1998), 1998, 92-99.

[23] Patterson, D.J., Fox, D., Kautx, H., and Philipose, M. Fine-grained activity recognition by aggregating abstract object usage. In *Proceedings of the Ninth IEEE International Symposium on Wearable Computers (ISWC '05)* (Oct. 18-21, 2005), IEEE Computer Society, 2005, 44-51.

[24] Philipose, M., Fishkin, K.P., Perkowitz, M., Patterson, D.J., Fox, D., Kautz, H., and Hahnel, D. Inferring activities from interactions with objects. *IEEE Pervasive Computing 3*, 4 (2004), 50-57.

[25] Starner, T., Mann, S., Rhodes, B., Levine, J., Healey, J., Kirsch, D., Picard, R.W., Pentland, A. Augmented reality through wearable computing. *Presence: Teleoperators and Virtual Environments 6*, 4 (1997), 386-398.

[26] Sun, X., Chen, C.-W., and Manjunath, B.S. Probabilistic motion parameter models for human activity recognition. In *Proceedings of the 16th International Conference on Pattern Recognition (ICPR '02)*, 2002, 443-446 vol.1.

[27] Surie, D., Lagriffoul, F., Pederson, T., and Sjolie, D. Activity recognition based on intra and extra manipulation of everyday objects. *Ubiquitous Computing Systems 4836*, (2007), Springer, 196-210.

[28] Surie, D., Pederson, T., Lagriffoul, F., Janlert, L.-E., and Sjolie, D. Activity recognition using an egocentric perspective of everyday objects. *Ubiquitous Intelligence and Computing 4611*, (2007), Springer, 246-257.

[29] Ward, J.A., Lukowicz, P., Troster, G., and Starner, T.E. Activity recognition of assembly tasks using body-worn microphones and accelerometers. *IEEE Transactions on Pattern Analysis and Machine Intelligence 28*, 10 (2006), 1553-1567.