RECOGNIZING AND DISCOVERING HUMAN ACTIONS FROM ON-BODY SENSOR DATA

D. Minnen, T. Starner

Georgia Institute of Technology
College of Computing
Atlanta, GA 30332-0280 USA

J. A. Ward, P. Lukowicz, G. Tröster

ETH - Swiss Federal Institute of Technology
Wearable Computing Lab
8092 Zürich, CH

ABSTRACT

We describe our initial efforts to learn high level human behaviors from low level gestures observed using on-body sensors. Such an activity discovery system could be used to index captured journals of a person's life automatically. In a medical context, an annotated journal could assist therapists in helping to describe and treat symptoms characteristic to behavioral syndromes such as autism. We review our current work on user-independent activity recognition from continuous data where we identify "interesting" user gestures through a combination of acceleration and audio sensors placed on the user's wrists and elbows. We examine an algorithm that can take advantage of such a sensor framework to automatically discover and label recurring behaviors, and we suggest future work where correlations of these low level gestures may indicate higher level activities.

1. INTRODUCTION

On-body sensors provide a unique perspective into the actions and behaviors of their owner. A first-person view of the environment and a direct, mobile relationship with the wearer frees recognition systems built around on-body sensors from many of the traditional perceptual problems that have plagued sensor systems based in the environment. Perceptual difficulties such as occlusion, scale, and interference can be greatly diminished, while practical problems of deployment and coverage can also be alleviated.

We believe that the reduction of perceptual problems is also accompanied by a new set of physical constraints that, rather than hinder analysis of gestural and behavioral data, can be used to make simplifying assumptions that aid recognition systems. In this paper, we present motivating research that shows that careful placement of on-body sensors can lead to simple, yet highly effective, detection and segmentation methods for context dependent, "interesting" behaviors. Our goal is to use such techniques, along with traditional data-driven analysis and machine learning methods, to build an end-to-end system that processes raw data from on-body sensors and learns contextual "primitive actions." Ultimately, we plan to extend this learning to higher level scripts composed of temporal relationships between these primitives.

Automatically learning such scripts and primitive actions provides several benefits. Traditional arguments for unsupervised learning, such as reducing cost by precluding the need to manually label data, aiding adaptation to non-stationary patterns, and providing early exploratory tools, continue to apply in this setting. Furthermore, such systems have the ability to learn useful patterns of behavior that may be unknown or simply transparent to human observers, possibly including the user himself (for example, bouncing a knee nervously when talking on the phone).

One practical application of activity discovery may be in long term medical or behavioral treatments. For example, austistic children often exhibit individualized self-stimulation ("stimming") behaviors such as rocking back and forth, walking on toes, or sudden flapping of the hands. If a system can be created that can discover such behaviors and correlate them to each other and to environmental factors, an automatic, annotated journal of the child's behavior can be maintained. Such a journal may become a valuable tool for therapists in describing children's behavior and in analyzing the effect of various treatments on that behavior.

2. RELATED WORK

Many other researchers have explored activity recognition using on-body accelerometers, but most have worked with comparatively simple activities such as walking, running, sitting, and shaking hands [1, 2, 3]. We are unaware of previous work that has explored the discovery of recurring activities from accelerometer data. In the audio domain, situation analysis was investigated by Peltonen et al. [4], and work by Clarkson et al. explored context awareness from wearable sensors [5].

Ashbrook and Starner discovered "significant locations" through unsupervised analysis of large GPS datasets [6], but the discovery relied on the spatial nature of the data and not on temporal patterns. Much literature in the data mining community is concerned with clustering time series data (see [7] for a review) but that work typically assumes a fixed

0-7803-9332-5/05/$20.00 ©2005 IEEE

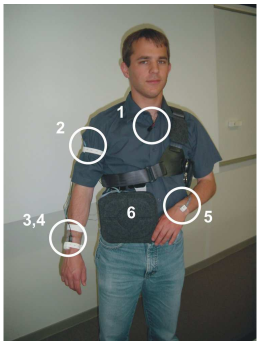

Fig. 1. Sensor locations: microphones (1, 2, 3), accelerometers (4 and 5), and a wearable computer (6).

pattern length and univariate data, two unrealistic assumptions with regard to human actions. Systems such as MEME in the bio-informatics field discovers noisy nucleotide patterns [8], and recent innovations allow for variable length patterns. The weak notion of time and naturally discrete alphabet, however, lead to several simplifications that are unrealistic for accelerometer or audio data representing human activities.

3. ON-BODY SENSORS

The on-body sensors used to support our research are based on the ETH PadNET sensor network equipped with two 3-axis accelerometers and two Sony microphones [9]. The accelerometers were placed on the user's wrists to capture the gross motion of the hands, while the microphones were attached at the wrist and upper forearm of the right arm (see Figure 1).

The choice of placement of the microphones proved essential to segmenting "interesting" behaviors. Because the distance between the microphones is constant due to the fixed length of the user's forearm, the intensity difference between the microphones can be used as a rough indication of the proximity of an audio source to the user's hand. We expect sounds caused by the hand interacting with an object to be very loud in the wrist-mounted microphone relative to the forearm. In contrast, sounds that are far from the user should have roughly equal intensity at both microphones (see Figure 2). Previous experiments placed the second microphone at the user's chest (microphone position 1 in 1), but this led to significantly worse segmentation performance due to the high variation in distance between the microphones as the user moved his arm, thus breaking the correlation between intensity difference and proximity.

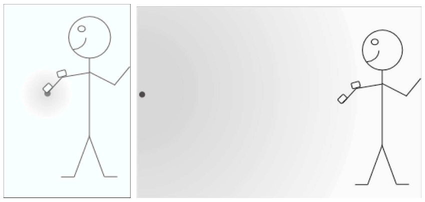

Fig. 2. When an interaction at the hand creates a sound, the sound intensity at the hand is much greater than that at the elbow. When a sound occurs in the environment, the sound intensity at the hand and elbow are approximately equal.

4. MODELING PRIMITIVE ACTIONS

Given the ability to roughly segment actions that are caused by the user, it is still necessary to recognize known behaviors and to cluster new behaviors to facilitate learning and adaptation.

For audio classification, the raw 44.1kHz signal is downsampled to 2kHz before a 100ms sliding window is used to compute FFT coefficients at 25ms increments. Linear discriminant analysis (LDA) is used to reduce the dimensionality of the resulting FFT coefficients. Finally, classification is achieved by projecting each test point into the LDA space and finding the nearest class centroid using the L_2-norm.

Hidden Markov models (HMMs) were used to classify each segment according to the observed accelerometer data. First the data is transformed into a feature vector consisting of the number of peaks within the segment considering all three axes, the mean amplitude of these peaks, and the raw x-axis data. A HMM was trained for each class using a mixture of Gaussian distributions for each observation node. The number of nodes and number of mixture elements in each model was manually specified for each class. Finally, a classification decision is made by selecting the class corresponding to the model with the highest likelihood for the test data.

After both the audio and accelerometer-based classification process, the results are fused using a simple heuristic. Whenever the two modalities yield the same classification, this class is accepted. If the classifiers disagree, however, the segment is labeled as not belonging to any class.

Figure 3 depicts some of the accelerometer data, the synchronized audio intensity difference signal, and the segment labels from our first experiment. In this domain, subjects equipped with the PadNET sensors used tools typical for a wood workshop. Five subjects followed a twenty-one step assembly script which involved nine activities: hammering, sawing, filling, drilling, sanding, grinding, tightening a screw, tightening a vise, and opening a drawer. The system correctly labeled each frame 70%, 66.1%, and 60.5% of the time for the user dependent, user adapted, and user independent cases, respectively. While such results may ap-

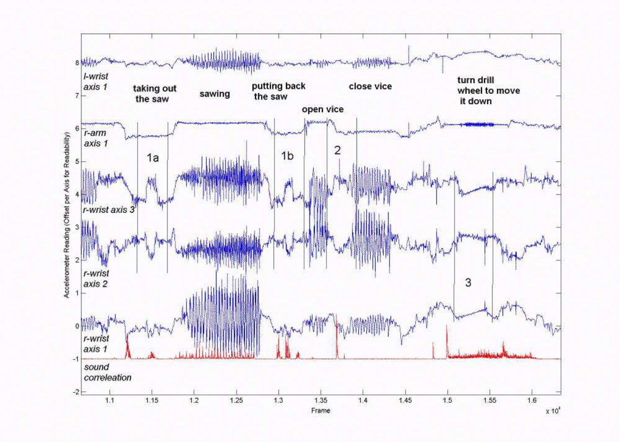

Fig. 3. Labeled audio signal from wood workshop experiment

pear low, note that this is a raw, frame–level metric. Evaluating the system at the action–level based on which frames represented "serious" errors like misrecognizing an activity, inserting an activity where there was none, or failing to recognize an activity was much more promising, yielding 8.2%, 8.6%, and 12.7% error rates for the three conditions. One of the most encouraging results is that the system could use the sound features to avoid insertion errors – as little as 0.6% of the frames in the difficult user independent case (for more results, see [10]). Such low false positive rates inspired us to investigate a system that could automatically discover the classes of activity that a user performs.

Thus, we are currently developing methods of learning primitive actions without prior models or labeled training data. Building on PERUSE, an unsupervised algorithm that analyzes time series to find recurring patterns [11], we are designing a system for discovering primitive actions by analyzing accelerometer data. Our method relies on the statistical similarity between occurrences of a particular action and the relative dissimilarity between different actions.

The problem of discovery in time series data is particularly difficult relative to isolated and continuous recognition because neither the location of each occurrence nor a model of each class is known. Thus, both the location and model parameters must be simultaneously estimated from the data using only the expected pattern length as guidance. To address this problem, our system first slides a fixed–length window across the time series to generate a set of potential occurrences. Then, for each occurrence, a 10 state left–right HMM is initialized using the segmental k-means algorithm and then adapted via Baum-Welch reestimation [12]. During this process, the covariance matrix of each Gaussian observation distribution is held fixed at $k\mathbf{I}$, where k is set proportional to the total data variance for the corresponding dimension. After initialization, the HMM is used to find the n best matches in the dataset, where n is a parameter of the system. The matches are located by using Viterbi alignment to build a trellis for each sequence and then scanning the last row (corresponding to the last state of the HMM) for a maximum. This procedure is repeated for each possible starting position, but since the length of the trellis is constrained to be proportional to the expected pattern length, the complexity only grows linearly with the data.

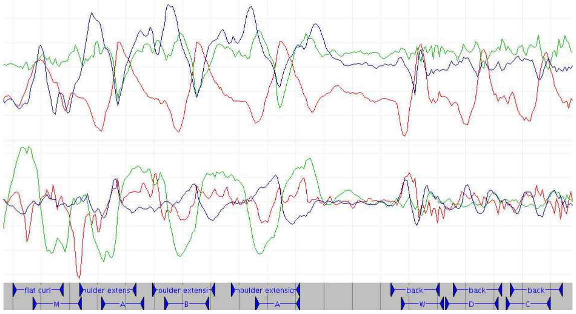

Fig. 4. Pattern discovery results on the exercise dataset. The top row of labels are ground truth, while the bottom row was automatically discovered. The segment labeled "W" is the only major error.

Finally, each potential occurrence is scored by the sum of the data log–likelihoods of its n best matches, and the best occurrence is selected. Importantly, the log–likelihood is normalized by the length of the matching sequence, yielding the average likelihood per frame. The n occurrences are then removed from the time series, and the algorithm iterates to find the next pattern.

Figure 4 shows results of discovered patterns in a weight lifting domain. A 3–axis accelerometer mounted on the wrist recorded data at 100Hz while the subject performed several repetitions of six kinds of dumbbell exercises including a shoulder press, tricep extension, and bicep curl. Eight sequences, each roughly one minute long, were analyzed after each sensor reading was transformed by $x' = sign(x) \cdot ln(|x| + 1)$. The system successfully discovered all six actions in the dataset, though several discovered patterns refered to the same action. Of the occurrences labeled, 68% corresponded to manually labeled ground truth. While preliminary, these results suggest the feasibility of activity discovery using on-body sensing.

5. FUTURE WORK: REPRESENTING AND LEARNING BEHAVIORAL SCRIPTS

We wish to progress from the ability to recognize an individual's activities from continuous data, to discovering individual actions, to learning higher level behavioral scripts. Such scripts can supply background information to aid future data interpretation and allow prediction of future events and event sequences based on current observations [13, 6]. For example, consider a typical daily occurrence of a professor entering her office. In general, it is difficult to describe the exact steps that she may take or to specify a probability distribution over a range of possible actions before observing the idiosyncrasies of the particular professor. She may unlock her office door, hang her jacket, and then grab a cup of coffee in the office kitchen. She could also check her email, skim the front page of the day's newspaper, or rush off to a class that she teaches each morning.

As described, our system could potentially discover recurring actions such as unlocking a door, unfolding a newspaper, sitting down at a desk, or hanging a jacket. The next goal then, is to extend this ability by learning temporal, causal, and statistical relationships between such actions. In isolation, the ability to discover and subsequently recognize an action only provides the ability to annotate a behavior. When coupled with behavioral scripts, however, such a system gains the ability to make predictions about future events, thereby enabling proactive aid, better planning, and more sophisticated models of normal behavior that can be used to enhance recognition rates [14].

Several representations for higher level scripts are possible. Due to the uncertainty inherent in human actions and in the accuracy of the recognition system, probabilistic models are likely to prove more useful than logical representations. Past research has looked at the use of bigrams and trigrams as an effective yet computationally inexpensive representation [6], while others have turned to more sophisticated models such as stochastic grammars, which provide greater representational power but are more computational expensive and more difficult to accurately learn [14, 15]. Both static and dynamic Bayesian networks represent other viable models for which a variety of powerful learning and inference algorithms have been developed [16].

6. CONCLUSIONS

We have discussed research–in–progress for discovering behavioral scripts using body–worn microphones and accelerometers. Preliminary results show that activity discovery is possible using such sensors, but that much research is needed to improve the accuracy and efficiency of the discovery process.

7. REFERENCES

[1] J. Mantyjarvi, J. Himberg, and T. Seppanen, "Recognizing human motion with multiple acceleration sensors," in *International Conference on Systems, Man and Cybernetics*, 2001, vol. 3494, pp. 747–752.

[2] C. Randell and H. Muller, "Context awareness by analysing accelerometer data," in *ISWC*, 2000, pp. 175–176.

[3] K. Van-Laerhoven and O. Cakmakci, "What shall we teach or pants?," in *ISWC*, 2000, pp. 77–83.

[4] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa, "Computational auditory scene recognition," in *International Conference on Acoustics, Speec, and Signal Processing*, May 2002, vol. 2, pp. 1941–1944.

[5] B. Clarkson, N. Sawhney, and A. Pentland, "Auditory context awwareness in wearable computing," in *Workshop on Perceptual User Interfaces*, November 1998.

[6] D. Ashbrook and T. Starner, "Using GPS to learn significant locations and predict movement across multiple users," in *Personal and Ubiquitous Computin*, October 2003, pp. 275–286.

[7] E. Keogh, J. Lin, and W. Truppel, "Clustering of time series subsequences is meaningless: Implications for past and future research," in *ICDM*, 2003, pp. 115–122.

[8] T.L. Bailey and C. Elkan, "Unsupservised learning of multiple motifs in biopolymers using expectation maximization," *Machine Learning*, vol. 21, pp. 51–83, 1995.

[9] H. Junker, P. Lukowicz, and G. Tröster, "PadNET: Wearable physical activity detection network," in *ISWC*, October 2003, pp. 244–245.

[10] J. A. Ward, P. Lukowicz, G. Tröester, and T. Starner, "Human activity recognition using body worn microphones and accelerometers for assembly tasks," *Submitted to PAMI*, 2005.

[11] T. Oates, "PERUSE: An unsupervised algorithm for finding recurring patterns in time series," in *ICDM*, 2002, pp. 330–337.

[12] L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Readings in speech recognition*, pp. 267–296, 1990.

[13] R. Schank and R. Abelson, *Scripts, Plans, Goals, and Understanding*, Lawrence Erlbaum Associates, 1977.

[14] D. Minnen, I. Essa, and T. Starner, "Expectation grammars: Leveraging high-level expectations for activity recognition," in *CVPR 2003*, June 2003.

[15] A. Stolcke and S. Omohundro, "Inducing probabilistic grammars by bayesian model merging," in *Grammatical Inference and Applications*, 1994, pp. 106–118.

[16] K. Murphy, *Dynamic Bayesian Networks: Representation, Inference and Learning*, Ph.D. thesis, UC Berkeley, July 2002.