Classifying Social Actions with a Single Accelerometer

Hayley Hung*
T.U. Delft
The Netherlands
h.hung@tudelft.nl

Gwenn Englebienne*
University of Amsterdam
The Netherlands
Englebienne@gmail.com

Jeroen Kools
University of Amsterdam
The Netherlands

ABSTRACT

In this paper, we estimate different types of social actions from a single body-worn accelerometer in a crowded social setting. Accelerometers have many advantages in such settings: they are impervious to environmental noise, unobtrusive, cheap, low-powered, and their readings are specific to a single person. Our experiments show that they are surprisingly informative of different types of social actions. The social actions we address in this paper are whether a person is speaking, laughing, gesturing, drinking, or stepping. To our knowledge, this is the first work to carry out experiments on estimating social actions from conversational behavior using only a wearable accelerometer. The ability to estimate such actions using just the acceleration opens up the potential for analyzing more about social aspects of people's interactions without explicitly recording what they are saying.

Author Keywords

Human behavior, social actions, wearable sensors

ACM Classification Keywords

H.1.2 Models and Principles: User/Machine Systems—*Human Information Processing*

General Terms

Algorithms; Experimentation; Human Factors; Measurement; Performance

INTRODUCTION

In this paper, we propose to analyze face-to-face interactive behavior automatically in dense crowded social gatherings such as that shown in Fig. 1 . In this context, we consider dense crowded social gatherings to have many people who are limited to no more than 1m^2 per person — although the methods proposed here would also work in less crowded situations. The analysis of such densely crowded social situations is appealing since such social events are organized to bring people together to meet, socialize, influence each other, forge new relationships or foster existing ones. And yet, to our knowledge, such scenarios have not been systematically studied in terms of whether wearable sensors, and particularly accelerometers, can provide sufficient information to classify social behavior reliably during conversations. In this paper, we present work where we organized, recorded, and labeled data from a crowded social gathering from which systematic experimentation could be carried out.

Figure 1. Examples of the type of crowded scenario this work could address. Note that the camera only captures a small proportion of the 270 attendees of this event.

Analyzing densely crowded scenes in this way is highly desirable because it enables a dense sampling of behavior in a space with small, cheap, disposable wearable devices. It is difficult for conventional sensors such as cameras and microphones to cope with the challenges that crowded social environments bring. That is, computer vision techniques cannot track large numbers of people at an event due to the many occlusions and also complexity of the data association problem with keeping track of many people at once. Likewise, speech processing cannot analyze people's speech or even speaking status robustly due to the high auditory noise of such events. In our experiments, we have even found that radio-based sensors such as ultra-wideband (UWB) localization systems struggle to deal with crowded environments because of interference, reflection and attenuation of the radio waves. A single accelerometer, in comparison, which could for example be worn as a conference badge, does not suffer these flaws and is easy to wear and use. Restricting the number of available sensors in each device is also appealing in terms of low battery consumption. Practically speaking, the hardware set up we use could scale to hundreds if not thousands of users and this is the type of scenario we target.

The novel contribution of this paper is in systematically investigating the challenging task of estimating social actions associated with conversing using just a single-body worn accelerometer. Specifically, the actions that we target relate directly to behaviors that occur during conversations: speaking, laughing, gesturing, drinking, and stepping. The reasons for targeting these behaviors are (i) they provide some indication of how socially active someone is, (ii) some behaviors such

*H. Hung and G. Englebienne contributed equally to this work.

UbiComp'13, September 8–12, 2013, Zurich, Switzerland.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-1770-2/13/09...$15.00.
http://dx.doi.org/10.1145/2493432.2493513

as stepping and drinking have been reported to be correlated across subjects who are talking together [7], and (iii) identifying when people speak is a good indication of group hierarchy such as dominance [11].

RELATED WORK

The work in this paper is motivated by a number of social psychological studies that have shown that (i) speakers tend to move more than listeners [12], (ii) laughter and joking is correlated with sudden bursts of motion [8], and (iii) drinking and stepping often occur almost synchronously during conversations [8, 7]. More synchronous behavior during conversation is also correlated with group cohesion [16] and influence [14]. Moreover, people who are getting along well during conversations mimic each other's behavior [3]. Also, knowing when someone is speaking provides information about the speaking turns of a conversation, which can then be used to automatically identify the social relationships between people such as their cohesion [5], or who is dominant in the group [6].

Most of the related literature on estimating human activities from accelerometer data have focused on identifying activities that are relevant to health such as fall detection [4, 18], ordinary daily activities including walking, running, sitting, climbing the stairs [9], daily household activities including eating or drinking, vacuuming or scrubbing, lying down [1], or to identify modes of transport taken [15]. For these types of activities, excellent classification accuracy is possible, even when only a single body worn accelerometer is used. Cattuto et al. [2] have used wearable sensors to analyze social interactions in crowded social settings, but without having an explicit evaluation of the precision and recall of their interaction estimation method.

The use of wearable sensors for measuring social interactions has also been addressed extensively, but often in situations where there is much less crowding and in conjunction with relatively clean audio data [17, 13, 10]. Techniques have ranged from using a classifier with simple derived features of the raw signal such as its mean or variance, to spectral feature extraction or multi-scale feature extraction. However, for more subtle behavior during conversations in highly social settings, we are not aware of any existing literature that has addressed this task.

REPRESENTING SOCIAL BEHAVIOR

We suggest that by using accelerometer readings from a single body-worn sensor (hung around the neck) in crowded social settings, it is possible to automatically detect social actions during conversations. The link between body motion and social behavior has been well-documented by social psychologists [12, 8, 3]. Specifically, existing research in social psychology cites a strong correlation between speech and the body gestures of both speaker and listener [12, 8].

DATA

We tested our approach on a newly collected dataset. To obtain natural behavior and the crowd density that we are interested in, we organized a social event from which the appropriate behaviors could be observed. A total of 32 student volunteers from different universities took part in the data collection. The volunteers were briefed that the aim of the event was to play a quiz game in teams, where the quiz was designed to span a wide variety of topics so that only diverse teams could be competitive. In order to form competitive teams, the volunteers had to (i) meet new people from different backgrounds, and (ii) form teams of four people to play the quiz. To increase motivation, prizes (personal music players and book vouchers) were awarded to the top 3 winning teams. Each participant was fitted with a sensor pack consisting of a tri-axial accelerometer (measuring acceleration along the X, Y and Z axes), an UWB indoor positioning device and a proximity sensor (the position and proximity data was not used here). 12 wireless microphones were also given to a random sample of the participants and three overhead fish eye cameras were mounted over the 5m × 6m area, which was marked on the floor and where the experiment took place; all the participants were requested to stay within the marked space during the recording. An example snapshot of the scene is shown in Fig. 2 .

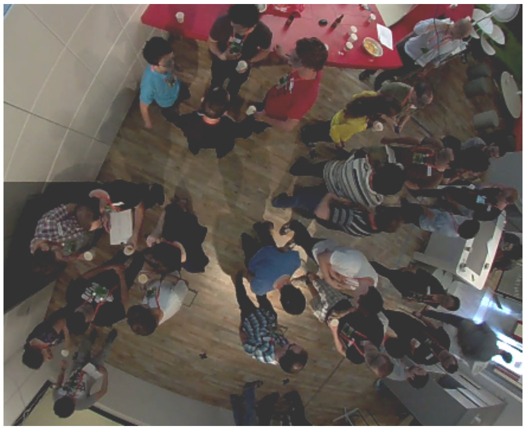

Figure 2. Snapshot of our data.

Due to the complex nature of the setup, not all sensors performed as expected. A small number of accelerometers failed due to a firmware bug, and the UWB positioning data is sparse and imprecise due to the crowded setup of the experiment. For the purposes of this work, we used the data from 9 of the subjects, which had both microphone and valid accelerometer data. We annotated the behavior of these subjects every 2s for the actions: *speaking*, *laughing*, *gesturing* (either hand or head), *stepping* (or walking) and *drinking*.

For the purposes of this work, 10 minutes of the mingling part of the event was manually labeled with the corresponding actions, which resulted in 813 sequences and a total of 2:04 hours' worth of readings. This involved manually associating each person in the video with their corresponding sensor readings and the labelling the social actions appropriately. We selected a part of the event that was highly social to maximize on the amount of examples of each class type. The actions that we are interested in can overlap, so that a single sequence may be associated with multiple labels. For example, a person may regularly be both speaking and gesturing, and we even observed three instances of a person "*speaking*", "*laughing*", "*gesturing*" and "*stepping*" at the same time in our dataset. Our approach is, therefore, to create a different classifier per action, discriminating between the action and its absence: speech vs. non-speech, gesturing vs. non-gesturing, *etc.*

CLASSIFYING SOCIAL ACTIONS

In practice, the different data sequences have varying lengths and the corresponding data has a complex structure. For example, the acceleration measured just before somebody starts speaking may be different from the acceleration when the person is in mid-flow, or when the person concludes his turn. To capture this, we trained two different Hidden Markov Models (HMMs) for each action: one trained on data annotated with the action (positive HMM) and one trained with a random sample of sequences not associated with the action (negative HMM). We used the standard algorithms for HMM training and inference, working with log-probabilities to avoid numerical underflows, and fixed the number of states per HMM to five. Both the positive and the negative set contain sequences that are associated with multiple labels. During inference, labels are then assigned to the observed sequence by comparing the likelihood of the data under the positive and negative HMM, disregarding the HMMs associated with the other actions.

Some model selection was done beforehand on data collected at a different event, with different people. This included selecting the number of states per HMM and the parameters of the feature extraction: the window length used to discretize time and the number of frequency bins per time window. The distribution parameters of the HMMs were optimized by maximum likelihood on left-out data from the same event as the test data.

EXPERIMENTAL RESULTS

Our aim is to evaluate the extent to which a single accelerometer can distinguish between the socially relevant actions of gesturing, speaking, drinking, laughing and stepping. To test this, we selected sequences from the recorded accelerometer data which were labelled with the corresponding actions from nine people. We collected the 3-dimensional readings with a sampling frequency of 20Hz. This data was transformed by a discrete Fourier transform and the resulting frequency data was windowed with a Hamming window of 10 seconds. For each time window, we binned the frequency components using 8 bins, resulting in 24 feature dimensions per time window. The bins were logarithmically spaced from 0 to 8Hz, to increase the resolution at low frequencies while keeping the dimensionality of the data acceptably low.

We performed 10-fold cross-validation on a random permutation of the sequences; our model is not tied to the behavioural idiosyncrasies of any single person and can be applied on a large scale. The recognition results reported here are averaged over the ten runs of crossvalidation.

Table 1 show the recognition results for the actions that we annotated. It should be noted that our classes are heavily unbalanced, since for most actions people spend the vast majority of the time not performing the action. Speech is an exception, as some people can spend more time speaking than listening, while others will speak very little. Therefore, we chose the more informative measures of precision, recall and F1 measures, to handle the class imbalance.

	gesture	**step**	**drink**	**laugh**	**speech**
Precision	0.59	1.00	1.00	1.00	0.64
Recall	0.24	0.21	0.21	0.38	0.82
F1	0.34	0.35	0.35	0.56	0.72

Table 1. Precision, recall and F-measure for the different action categories in our dataset

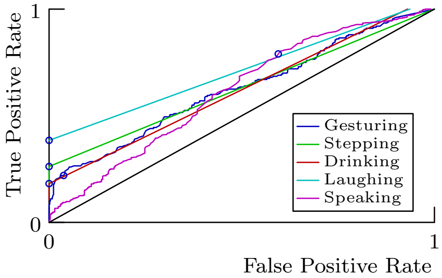

Figure 3. Receiver Operating Curve of the different classifiers on our dataset. The points on the curve corresponding to the precision and recall figures shown in Table 1 are indicated by a blue circle.

A detailed look at Table 1 shows some interesting patterns. *Gesturing*, a label which contains head and hand gestures, is not recognised sufficiently well to be useful. The precision is barely better than random and the recall is worse than random, so that the resulting F1 measure is itself worse than random. Just looking at the F1 measure, one might be tempted to say the same of *stepping* and *drinking*, but a closer look reveals that although we miss many instances of these actions (low recall,) when we do detect them we can be very confident of the detection (high precision). This is not surprising from the way these actions are performed. For example, one can take a sip of a bottle with very little motion of the upper body, while emptying a bottle may lead to very distinctive tilting back of the head and torso, so that very distinctive motion patterns are associated with these actions. *Laughing* is similarly distinctive, and also easier to detect.

Probably the most interesting result concerns *speaking*, however. This action is by far the most prominent in the dataset, and we detect it with both high precision and recall from the accelerometer data. We consider this a remarkable result since, after all, the accelerometer has too low a sampling frequency to directly measure sound vibrations in the torso. Instead, the social interaction behaviour of the listener changes when he becomes a speaker, and this is measured by the accelerometer. This indirect detection of speech is surprisingly good, and informal tests show that it can be improved even more with careful tuning of the model parameters. We choose not to do so for this analysis to avoid the risk of overfitting on our dataset.[1]

[1] Incidentaly, we should highlight that each action detector is independent of the others, and that there is no reason why they should be constrained to have the same features or model complexity.

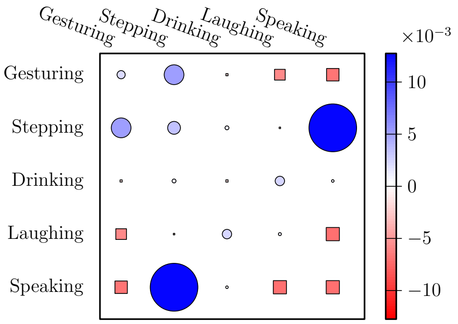

Figure 4. Probability of action co-occurence in a 10s window when two people are not conversing, subtracted from when they are in the same group. Positive values:blue circles; Negative values: red squares.

Our classifier can be tuned to trade off precision for recall, or vice versa. The actual performance of the classifier is therefore better represented as a Reciever Operating Characteristic (ROC) curve, as depicted in Fig. 3 . The black line indicates the performance of a classifier that would return a random label to each query, while the other curves indicate the actual performance of our classifiers. All classifiers perform better than random and they all attain a non-zero true positive rate for a zero false positive rate.

DISCUSSION

Our results highlighted some interesting findings about which social actions were more or less difficult to automatically detect. In particular, we believe that the detection of these actions could be used to then estimate who could be talking with whom. To understand better how these classes of behavior could be useful for detecting social interactions, we analysed the co-occurrence of the labelled actions from two different people when they belonged to the same conversing group (in-group co-occurence probability), and when they were not (out-group co-occurrence probability). Figure 4 shows the out-group probalitiy matrix subtracted from the in-group probability matrix. We observe clear differences in the in-group behavior. For example, we see that stepping while someone else is speaking is more likely to occur with people in the same group, compared to when they are in different groups. Simultaneous speaking is also less probable within the same conversation. This analysis suggests that the social actions that had high precision but low recall, could stil be very informative for detecting conversing groups.

CONCLUSION

In this paper, we have demonstrated the feasibility of using accelerometers to automatically recognised socially relevant actions. Automatically detecting *speaking* was achieved very good performance, while *stepping*, *drinking*, and *Laughing* had very high precision. Further analysis of the co-occurrence between social actions between conversing and not conversing people showed that differences in behaviour matched the social actions that obtained high performing estimates. We plan to further improve the classifier performance by investigating better feature representation models and to investigate how to estimate when people are speaking together, using the estimated social actions proposed here.

ACKNOWLEDGEMENTS

This research was partially supported by a Marie Curie Research Training Network fellowship in the project “AnaSID” (PIEF- GA-2009-255609) and the Amsterdam “zorg en ict” project healthlab. We thank Matthew Dobson, Claudio Martella, and Maarten van Steen (VU University of Amsterdam) and Ben Kröse (University of Amsterdam) for their assistance during the data collection.

REFERENCES

1. L. Bao and S. Intille. Activity recognition from user-annotated acceleration data. *Pervasive Computing*, pages 1–17, 2004.
2. C. Cattuto, W. Van den Broeck, A. Barrat, V. Colizza, J. Pinton, and A. Vespignani. Dynamics of Person-to-Person Interactions from Distributed RFID Sensor Networks. *PLOS ONE*, 5(7):e11596, 07 2010.
3. T. L. Chartrand and J. A. Bargh. The chameleon effect: the perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, 76(6):893–910, 1999.
4. C. Doukas, I. Maglogiannis, P. Tragas, D. Liapis, and G. Yovanof. Patient fall detection using support vector machines. *Artificial Intelligence and Innovations 2007: from Theory to Applications*, pages 147–156, 2007.
5. H. Hung and D. Gatica-Perez. Estimating cohesion in small groups using audio-visual nonverbal behavior. *Multimedia, IEEE Transactions on*, 12(6):563–575, 2010.
6. D. Jayagopi, H. Hung, C. Yeo, and D. GaticaPerez. Modeling dominance in group conversations from non-verbal activity cues. *IEEE Transactions on Audio, Speech and Language Processing*, 2008.
7. A. Kendon. Movement coordination in social interaction: Some examples described. *Acta Psychologica*, 32:101–125, 1970.
8. A. Kendon. *Conducting Interaction: Patterns of Behavior in Focused Encounters*. Cambridge University Press, 1990.
9. J. R. Kwapisz, G. M. Weiss, and S. A. Moore. Activity recognition using cell phone accelerometers. *ACM SIGKDD Explorations Newsletter*, 12(2):74–82, 2011.
10. B. Lepri, J. Staiano, G. Rigato, K. Kalimeri, A. Finnerty, F. Pianesi, N. Sebe, and A. Pentland. The SocioMetric Badges Corpus: A Multilevel Behavioral Dataset for Social Behavior in Complex Organizations. In *SocialCom/PASSAT*, pages 623–628. IEEE, 2012.
11. M. S. Mast. Dominance as Expressed and Inferred Through Speaking Time. *Human Communication Research*, (3):420–450, July 2002.
12. D. McNeill. *Language and Gesture*. Cambridge University Press New York, 2000.
13. D. O. Olguin, B. N. Waber, T. Kim, A. Mohan, K. Ara, and A. Pentland. Sensible Organizations: Technology and Methodology for Automatically Measuring Organizational Behavior. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 39(1), 2009.
14. K. O’Quin and J. Aronoff. Humor as a Technique of Social Influence. *Social Psychology Quarterly*, 44(4):349–357, Dec. 1981.
15. S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava. Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks (TOSN)*, 6(2):13, 2010.
16. E. J. Romero and K. W. Cruthirds. The Use of Humor in the Workplace. *Academy of Management Perspectives*, 20(2):58–69, 2006.
17. D. Wyatt, T. Choudhury, J. Bilmes, and J. A. Kitts. Inferring colocation and conversation networks from privacy-sensitive audio with implications for computational social science. *ACM Trans. Intell. Syst. Technol.*, 2(1):7:1–7:41, Jan. 2011.
18. T. Zhang, J. Wang, P. Liu, and J. Hou. Fall detection by embedding an accelerometer in cellphone and using KFD algorithm. *International Journal of Computer Science and Network Security*, 6(10):277–284, 2006.