Detecting Conversing Groups with a Single Worn Accelerometer

Hayley Hung*
TU Delft
Delft, The Netherlands
h.hung@tudelft.nl

Gwenn Englebienne*
University of Amsterdam
Amsterdam, The Netherlands
g.englebienne@uva.nl

Laura Cabrera-Quirós
TU Delft
Delft, The Netherlands
l.c.cabreraquiros@tudelft.nl

ABSTRACT

In this paper we propose the novel task of detecting groups of conversing people using only a single body-worn accelerometer per person. Our approach estimates each individual's social actions and uses the co-ordination of these social actions between pairs to identify group membership. The aim of such an approach is to be deployed in dense crowded environments. Our work differs significantly from previous approaches, which have tended to rely on audio and/or proximity sensing, often in much less crowded scenarios, for estimating whether people are talking together or who is speaking. Ultimately, we are interested in detecting who is speaking, who is conversing with whom, and from that, to infer socially relevant information about the interaction such as whether people are enjoying themselves, or the quality of their relationship in these extremely dense crowded scenarios. Striving towards this long-term goal, this paper presents a systematic study to understand how to detect groups of people who are conversing together in this setting, where we achieve a 64% classification accuracy using a fully automated system.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems—*Human Information Processing*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing Methods*; G.3 [**Probability and Statistics**]: Time Series Analysis

Keywords

Human behavior; human factors; wearable sensors; data mining

1. INTRODUCTION

In this paper, we propose to detect conversing groups automatically in dense crowds during social gatherings using only a single worn accelerometer per person. Our long term goal with such a set-up is to be able to analyse socially relevant behaviour in such gatherings to understand if people enjoyed themselves, the type of mood of the event, whether people would come again, *etc.* This requires us to try to get a deeper understanding of what is happening in the crowd in terms of the social interactions and relationships between the people. By taking inspiration from findings in social psychology, we aim to address such problems while maintaining privacy and minimising additional effort of the user.

To that end, in this paper, we go significantly beyond prior work where the state of the art addressed whether socially relevant actions during conversations, such as speaking, could be estimated from just a single body worn accelerometer [13]. Our aim is to try to understand deeply the problems and opportunities of automatically analysing social behaviour using such an approach, as a starting point for other systems that may wish to embellish with more sensors, or indeed to strip down to as few sensors as possible.

Importantly for research in estimating social attributes such as dominance[15], leadership [16], or cohesion [14], one of the most informative features is typically based on the management of turns (turn-taking patterns) within a group. Therefore, the ability to automatically detect where a conversing group is, and when someone is speaking, are vital action units from which semantically higher level social concepts can be inferred.

Specifically in this paper, we address the task of estimating whether two people are conversing in dense crowded social gatherings by just measuring the movement via a body worn accelerometer. It is worth considering this modality as in such scenarios, relying on proximity alone to make robust detections is challenging as the detections are bursty. Audio data would be contaminated by significant background noise from all conversations happening at the gathering making robust audio processing too challenging. Moreover, the use of audio can be considered privacy invasive to wearers.

Prior work has used infrared (IR) sensors to detect the proximity of others as a proxy for interaction [4, 11] in much less crowded scenarios which showed poor accuracies when accumulated over short periods (5 minutes) [4]. The lack of reliability of IR sensors is partially caused by how people choose to place themselves in a space which can depend on the shape of a room or its furniture for example [22]. Under such scenarios audio sensors have been used to detect conversational interactions [5, 4, 11, 20] robustly.

The Scenario Challenges.

As already mentioned, in this paper, we address the problem of addressing socially relevant behaviour in dense crowded social gatherings such as the conference pictured in Fig. 1. Note that this image captures just a small corner of the entire space of the event, which in this case consisted of 300 participants. In such social settings, groups can tend to form, merge, and split, but still remain together spatially as different social networks with common acquaintances

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

ICMI'14, November 12–16, 2014, Istanbul, Turkey.

http://dx.doi.org/10.1145/2663204.2663228.

Figure 1: The complexity of crowded social encounters.

are found in the same physical space. For large groups (> 5 people) sustaining a single informal conversation is not possible [8]. However, perhaps due to the familiarity of the group members with each other, they may tend to orientate themselves towards the entire group, while gazing at the few people they are actually talking to. An example of this is shown in the crowded social gathering in Fig. 1 where the highlighted group are all acquainted with each other. Although there are 3 different conversing groups (shown by the red lines), the people still try to remain quite open to starting a conversation or including one of the others in their conversation at any moment. Therefore relying on physical cues such as just directed proximity and body orientation also does not address the problem sufficiently well. Using radio-based sensors to detect proximity allows for slightly less direction-specific detections. However, when events are densely crowded, it is likely that more people will be detected as being proximate than are actually participating in a conversation, or indeed could potentially converse together.

Despite these challenges, the automated analysis of interactive behaviour has many potential applications, of both scientific and commercial nature. For example, wearable sensors can be used for the analysis of individual and social behaviour in large crowds, for on-line evaluation of the success-rate of social events such as conferences or trade fairs, for the inclusion of social factors in handling people logistics in massive crowds such as at large sport events.

An important requirement for such a device is that wearers must not feel as if their privacy is invaded. Therefore, it must be unobtrusive and their raw speech should not be recorded.

Robust wireless transmission of real-time measurements is challenging in situations with many nodes in a restricted area. It is clear that additional sensors such as extra accelerometers on different parts of the body, gyroscopes or a magnetic compass would provide additional information, at the cost of increased bandwidth requirements and device price, and that this could lead to improved recognition rates. In this work, we focus on showing how surprisingly informative the lowly accelerometer is.

The practical appeal of such a system is that the use of a single sensor significantly reduces the power consumption. In addition, unlike mobile-based applications, designing for incentivised uptake (to compensate for increased battery usage and consent to the access of private personal data) is not necessary. One can imagine such a device could be easily attached to a conference badge, for example without further intervention from the wearer.

Theoretical Issues.

Experimentally, since our aim is to carry out a systematic study of how to detect conversing groups, we provide here a more formal definition by social psychologist Adam Kendon [17] of what we precisely mean by this. Kendon made clear distinctions between differing taxonomies of interacting groups. Co-located and co-ordinated group behaviour were named focused encounters when for example, people gathered together to perform in a marching band, play football, or watch a match. Within the set of all possible focused encounters are a specific type, called F-formations which define a small group of people who spatially and orientationally arrange themselves to facilitate conversation. An F-formation arises when

> "two or more individuals in close proximity orient their bodies in such a way that each of them has an easy, direct and equal access to every other participant's transactional segment, and when they maintain such an arrangement, they can be said to create an F-formation" [6](p.243).

Importantly, this definition makes explicit the fact that it is not necessary for someone to speak at all while being in an F-formation. However, other behaviours during participation in an F-formation, which for example correspond to listening, are still clearly important parts of being in a conversation. Moreover, findings in social psychology also suggest that other co-ordinated body motions also exist when people are in an F-formation, which are not necessarily directly related to the management of the turn-taking itself (*e.g.* simultaneous shifts of posture [17] or behavioural mimicry [3]). It is also important to highlight that while a conversing group might imply a cluster of multiple F-formations who are spatially close, here, the definition refers more specifically to one conversation happening for which all members of the F-formation are equally engaged as either the speaker(s) or listener(s).

Novel Contributions.

The contributions of this paper are: (i) we demonstrate method of using accelerometers alone to model and automatically detect instantaneous conversation-related social actions online, through a systematic analysis of streaming accelerometer readings, which out-performs the state of the art [13]; (ii) We show that the synchronicity as expressed through the mutual information between these social actions is indicative of whether people are part of the same F-formation, (iii) we show that even using imperfect recognition of such social actions results in good F-formation recognition, while direct computation of the mutual information between people's raw acceleration utterly fails to do the same.

2. RELATED WORK

2.1 Activity and Action Recognition

The majority of related work on human activity recognition using accelerometers have tended to concentrate on non-social activities such as fall detection [7, 33], ordinary daily activities including walking, running, sitting, climbing the stairs [19], daily household activities including eating or drinking, vacuuming or scrubbing, lying down [1], or to identify modes of transport taken [31]. Classifying these types of activities is possible with excellent performance. However, the aim of this paper is to measure behaviour for which the link between the activity and the behaviour is not as direct. That is, the movements associated with speaking for example, are physical manifestations of the cognitive process of speaking but do not directly produce the spoken behaviour [17].

Matic et al. also used acceleration to detect speaking status by strapping an accelerometer to the chest so that vibrations directly caused by speaking could be detected [23]. This essentially limited the possibility to interpret the verbal content of a conversation but would be much less practical to implement in a crowded social setting. In [13] an initial study was carried out to see if socially relevant actions during conversations such as speaking and laughing could be detected. In those experiments, the class data was pre-segmented and a fixed window size was used. We improved upon that work by demonstrating the feasibility of measuring the same behaviours from continuous accelerometer data and parameters trained on each class explicitly, leading to significant improvements in the performance.

2.2 Detecting Conversing Groups

There has been much prior research on analysing and estimating aspects of social behaviour using wearable sensors on both the small group scale with 2-6 people in a room [18, 10] and on larger scale, with 50 - 100 participants [9]. To our knowledge, most automated analysis of face-to-face scenarios have been carried out on small groups (2-6 people [18, 10]). Studies which have analysed larger groups of people (10-200) have used sensors such as bluetooth and/or infrared (IR) to detect when people are close together, as a proxy for interaction [21, 26, 9, 4, 11], or audio [32] However, these studies were often carried out over weeks, sometimes also exploiting other communication methods such as call frequency. [26, 9]. Moreover, it is expected that at any moment in time, the number of detected neighbours is in general far fewer than would occur in crowded situations. More systematic evaluation of conversation detection and the level of formality of the conversation was analysed by Matic et al. using both proximity and orientation sensors from a mobile phone [24]. Again, experiments were carried out in uncrowded situations.

Wearable sensors have been used as a means of supporting cooperative activities in working environments [29, 26], analysing the diffusion of political opinions or ideas [21], influence and centrality [4], or affiliation [12] amongst many others. In contrast, we focus on the detection of immediate aspects of interpersonal communication in crowded situations, at shorter timescales (and therefore relying on fewer measurements,) that require a reasonable estimation accuracy.

One of the most similar prior works that measured social behaviour using wearable sensors was proposed by Choudhury [4]. She carried out initial experiments to detect social interaction using infrared sensors attached to a "sociometer" device. This was worn on the shoulder blade of each participant for around 10 days. Experimental results showed that using infrared sensors as a proxy for interaction was quite inaccurate when trying to measure spontaneous interactions over short time scales, but performed better when accumulated over 10 days or more. Choudhury also obtained significantly better results at shorter time scales of one minute, but this was using the recorded audio signal from each device. In these experiments, the behaviour of 25 individuals was recorded, as a result of 3 different recorded time periods.

The primary distinguishing factor of our work is that the methods we propose are intended to address the specific challenges of analysing social behaviour in crowded and noisy environments, whereas most prior work, with the exception of Gips [11], have operated in less densely crowded environments. Gips [11] built on the work of Choudhury by using accelerometer readings to measure interactions using the "Uberbadge" [27]. Similar to the "sociometer", the "Uberbadge" contains an accelerometer and an IR transceiver but the badge was hung around the neck. Gips used accelerometer readings to extract the mutual information in the motion energy (MIME) to find interacting participants, rather than using an audio signal. Such a feature can find synchronous behaviour but could also find erroneous occurrences, for example if one person is following another down a corridor, or they are queuing. Unfortunately, while Gips provided an analysis of the extracted features with factors such as the affiliation of the participants at an event, no performance evaluation was made. So it is not clear how well the MIME feature performed on the various data sets that were tested on.

Olguin et al. [26] measured the behaviour of 22 colleagues in a company over one month using their "sociometric badge". The badge contained a number of different sensors and devices that are also used in conjunction with sensors within the wearer's mobile phone. For detecting and measuring social interactions, bluetooth was used to communicate between the wearers' mobile phone and the devices of others who were probably within an appropriate distance for conversation. In addition, a microphone for measuring non-verbal cues related to excitement and interest, and infrared (IR) sensors with a 30 degrees and 1m range of sensitivity for measuring interaction were also used. This badge was worn around the neck. The IR sensor readings were used as a proxy for interaction but was not evaluated explicitly during these experiments. Finally, an audio signal recorded from the badge was used to detect when someone was speaking. For a more detailed overview of existing methods that use wearable sensors for analysing social behaviour, we refer the reader to the recent book chapter by Olguin et al. [25].

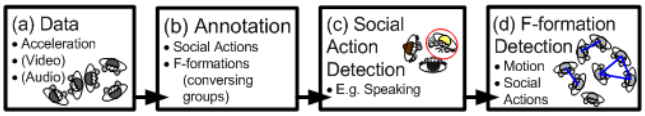

Figure 2: Flow diagram showing the experiments that were carried out in the rest of the paper.

In summary, all prior work listed above has tended to use IR sensors as a proxy for interaction, and used microphones for detecting when people are speaking. In contrast, we propose to do both using just one tri-axial accelerometer per person. Moreover, we explicitly address the problem of detecting F-formations [17] to ensure a clear labelling strategy in the presence of complex ambiguous data. So far, this has also not been addressed by existing research work.

3. OUR APPROACH

Figure 2 shows a flow diagram of our experiments. First the data is described in Section 4.1. Then in Section 4.2 the annotations for both social actions and F-formations will be described. Section 5 describes the methodology, experiments and resuts for estimating social actions. Section 6 describes how the social action labels are used to detect F-formations, the experimental procedure, and finally discusses the results. We conclude in Section 7.

4. DATA

4.1 The Scenario

For our experiments, we use the data presented in Hung et al. [13] where a social event was organised to obtain natural behaviour and in a dense crowded environment. 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. 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, prises (personal music players and book vouchers) were awarded to the top 3 winning teams.

Each participant wore a sensor pack hung around the neck, which contained a triaxial accelerometer, proximity sensor, and an indoor positioning device. The proximity and position information was not used in these experiments though investigating the combination of such modalities as well as the trade-offs are left for feature work. 12 wireless microphones were randomly distributed amongst the participants and three overhead fish-eye cameras recorded the experimental area (5m $\times$ 6m). All participants were requested to stay within this marked area during the recording. Both audio and video data was used only to collect ground truth labels for training the classifiers. An example snapshot of the scene is shown in Figure 3.

Unfortunately, of the 12 subjects wearing microphones, only 10 had accelerometer data due to a firmware bug. The behaviour of these subjects was already annotated every 2s for the actions: *speaking*, *laughing*, *gesturing* (either hand or head), *stepping* (or walking) and *drinking*. An additional 17 subjects wore working accelerom-

Figure 3: Overhead snapshot of the scenario used in our data.

eters but wore no microphone and so no ground truth annotations were recorded for these people. We carried out experiment with both fully annotated social action data (**10 person data**), and the partially annotated social action data **26 person data**.

4.2 Data Annotation

We used the 10 minute mingling part of the event as annotated by Hung et al. [13] for which annotations of social actions were already available. We extended the work of Hung et al. [13] by manually annotating for F-formations. The F-formations were annotated as follows. This involved manually associating each person in the video with their corresponding sensor readings and then labelling the social actions appropriately. The same 10 minute segment for all 32 subjects was annotated every 2s, depending on whether people were in the same F-formation or not, leading to 300 time frames of annotated data. Unfortunately only 26 subjects had working accelerometer data. Annotating those with missing data was important for understanding the affect that these people might have on the behaviour of those with working accelerometers.

5. ESTIMATING SOCIAL ACTIONS

5.1 Method

Similar to Hung et al. [13], we extract spectral features from the accelerometer readings and evaluate how well a Hidden Markov Model (HMM, [30]) and random forests (RF,[2]) models these. Unlike the work of Hung et al. [13], who attempted to label pre-segmented data sequences as corresponding to an action or not, in this work, we perform segmentation and classification simultaneously. These tasks are different, but they are clearly related: we can use an HMM to solve both cases, the difference is in the meaning of the states: in [13], the states of the HMM model different aspects of an action, such as its onset or final phase. The optimal number of states is, therefore, related to the action to be recognised, and not intrinsically limited. In this work, there are only two states: whether the action is being performed in the current time slot or not. This task is clearly more challenging, as the model needs to recognise both on the segment boundaries and the corresponding actions.

We have, therefore, modified the earlier approach by modelling the emission probabilities of the "action" and "non-action" states with mixtures of Gaussians instead of single Gaussians. Whereas the complex distribution of the observations was handled by many states in "action" and "non-action" HMMs, in this work the complexity is handled by the mixture elements within the "action" and "non-action" states of a single HMM. The resulting model is equally powerful in terms of its capacity to model the complex distributions of the observations, but is slightly less powerful in its capacity to model the transitions between parts of an action. For example, if we imagine that the action of speaking consists of an onset, varying behaviours during speech, and termination, the current model is oblivious to the order of these phases even though it correctly models each.

As the expressiveness of the transition model decreases, the question arises whether a standard classifier, without transition model, is not more suitable. To test this, we have compared the HMM's performance to a powerful ensemble method, random forests [2], trained on the individual time slices.

Table 1: Summary of social action estimation performance using continuous data. (see text for details)

HMM-based

Action	Class Prop.	length (s)	elem.	Prec.	Rec.	F1	Acc.
Gesturing	0.67	5	1	0.70	0.62	0.66	0.67
Stepping	0.09	5	3	0.32	0.28	0.30	0.85
Drinking	0.05	2.5	1	0.20	0.19	0.19	0.88
Laughing	0.03	3.5	2	0	0	0	0.97
Speaking	0.42	3.5	2	0.85	0.84	.84	0.50

Random-forest-based

Action	Class Prop.	length (s)	Trees	Prec.	Rec.	F1	Acc.
Gesturing	0.67	5	500	0.58	0.56	0.57	0.83
Stepping	0.09	5	500	0.54	0.53	0.53	0.91
Drinking	0.05	5	1000	0.43	0.43	0.43	0.95
Laughing	0.03	5	1000	0.23	0.24	0.23	0.96
Speaking	0.42	5	500	0.69	0.68	0.69	0.76

5.2 Results

We performed our experiments on the accelerometer readings of 26 participants from [13]. We replicated the feature extraction of Hung et al., by processing the signal from the three-dimensional acceleration readings by computing a Discrete Fourier Transform of the acceleration along each dimension individually, and used log-spaced bins of varying size, so as to obtain a higher resolution for the low frequencies.

Different actions have varying degrees of complexity and different durations. It makes sense, therefore, to use different numbers of mixture elements and different lengths for the windows used for feature extraction (for the HMM), and different numbers of trees for the random forests. Table 1 lists the number of mixture elements and the window lengths in seconds used for the different gestures (fourth and third columns respectively). These numbers were chosen based on cross-validated selection using pre-segmented data. The table further lists the precision, recall, F1 measure and accuracy (columns five to eight). The listed parameters in columns three and four were trained on cross-validated pre-segmented data.

From these results, we note that the HMM performs better at recognising *gesturing* and *speaking*, which are longer-duration actions in our dataset, while the random forest is better at recognising *stepping*, *drinking* and *laughing*. in this work we focus on the analysis of F-formations from the obtained actions, but in practice, of course, there is no problem with choosing the best tool for the task and applying different models for the different actions.

Notice how *stepping*, *drinking* and *laughing* are not recognised well at all by the HMM, although the random forest performs quite a bit better despite the class imbalance. In particular, *stepping* is remarkably badly recognised. This is due to the definition of *stepping* used in the data, which can consist both of walking or *stepping*, but also of switching one's weight from one foot to another during a conversation. Although relevant, this form of body language is very challenging to detect with our current model. *Drinking* and *laughing* suffer from the small dataset and few positive examples. Table 1 shows the performance, allowing up to a 2 seconds shift of the an-

Figure 4: Matrices showing the average mutual information for each pair of social actions for people who were annotated to be in the same F-formation (left) and not in the same F-formation (right) over a 20s window.

notations either way. The resulting figures show that *gesturing* and *speaking* are segmented and recognised much better than random but that, as expected from the few positive examples of these classes, *stepping*, *drinking* and *laughing* perform badly.

6. IDENTIFYING PAIRWISE F-FORMATION MEMBERSHIP

6.1 Method

To understand more about the problem of automatically identifying F-formations, we hypothesised that people have strongly co-ordinated behaviour during conversations that would highlight the moments when they were conversing compared to times when they are not. Furthermore, motivated from findings in social psychology[17], we hypothesise that in addition to the co-ordination of speaking turns, other socially relevant behaviours could also be indicative of being in an F-formation. Moreover, when the accuracy in estimating speaking status is reduced, the use of other socially relevant conversational behaviours could enhance the performance.

To verify this hypothesis, we adjusted the method of Wyatt et al. [32], who applied this method on audio data. Mutual information is calculated over a window for both social action streams. The pair are labelled as being in an F-formation using a threshold on the mutual information value (trained on separate data). So pairs of people whose binary streams yield a higher mutual information are considered to be more likely to be talking together.

To illustrate how each social action can indicate who is in an F-formation with whom, we plotted the mean mutual information per ground-truth-labelled action pair, differentiated by whether they were in the same F-formation or not. Figure 4 shows a colour-coded matrix of the within-group and without-group mean mutual information per pair of social actions. Light colours indicate where pairs of actions have higher mean mutual information. The left half of the figure shows the mean mutual information of actions for people in the same F-formation and the right half shows the same information for people in different F-formations. We see that overall, the mean mutual information between actions of pairs of people in the same group tend to be higher than those of people in different groups. As expected, speaking from both people yields the highest mean mutual information. However, it is worth highlighting that other action pairs also have higher mutual information within the same F-formation. Note that these results highlight patterns for the dyads that exist in this labelled data.

Interestingly, we also see that pairwise speaking activity for both participants also yields the highest mean mutual information for people who are not in the same F-formation. In fact, the mutual information for two speakers is almost twice as high as the next ranked action pairs (speaking and gesturing) in the in-group case. This suggests that although speaking is a key characteristic of people being in a conversation together, other social actions also contribute to the dynamics of activities that are indicative of F-formations.

From the analysis shown in Figure 4, we chose a simple late fusion approach by taking the mean mutual information for every possible action pair in a given window.

Table 2: F-formation estimation accuracy using the mean MI. GT: ground truth, RF: Random forest.

Win. Size (s)	5s	10s	20s	40s	80s
10 person GT	0.51	0.59	0.62	0.62	0.57
10 person RF	0.51	0.56	0.63	**0.64**	0.60
26 person RF	0.49	0.50	0.50	0.50	0.50

6.2 Experimental Results

To experimentally validate the F-formation detection, the performance was cross-validated by a leave-one-pair out approach. For a given pair of people and window size, the mutual information threshold was trained using the labelled binary social action streams of all other pairs of the remaining people. The threshold was selected by finding the value that maximised the classification accuracy from the training data. For the 10 annotated people, at any one time, there were far fewer instances of pairs of people in the same F-formations than in different F-formations. Therefore, to handle the significant imbalance in the classes, the smaller class was reweighted during training and testing. All the presented results are therefore presented based on upsampling the smaller class to be identical in size.

6.2.1 Using Ground Truth Social Actions

To understand what the upper bound of the performance of the F-formation detector would be, we first present results using the ground truth labels of the social actions. Figure 5 summarises the classification accuracy for detecting people being in the same F-formation for different window sizes. Lighter colours indicate better performance. In practice, the optimal window size can also be trained for, but we wanted to analyse how differing window sizes might affect performance when using ground truth compared to estimated social actions.

The first thing to note is that the best accuracy of 0.65 is obtained when using the mutual information of speaking activity of both participants using a 20s window. There appears to be a peak in performance of the *speaking- speaking* action pair at this window length, after which the performance decreases. This suggests that within 20s, turn exchanges were sufficiently captured to obtain discriminative mutual information between the two binary streams. However, the performance of the *speaking-gesturing* action pair increases with longer window lengths, which on further investigation of the social action label estimates can be explained by *gesturing* occurring more sparsely than *speaking*. Importantly, all the remaining social actions, with the exception of *laughing* all pair with other social actions to produce F-formation estimates which are better than the random baseline (50% classification accuracy). This may be due to there being very few examples of *laughing* in the data. Moreover, peaks in performance of different social action pairs occur at varying window lengths, which is in keeping with the nature of some social actions being longer than others. Overall, in line with prior work [32], speaking status is very important for identifying conversing groups. However, our experiments show that other social actions are also informative of being in an F-formation but their infrequence in the labelled data suggests that more data is needed to further investigate this result. To understand how the performance might improve by incorporating the mutual information between all social action

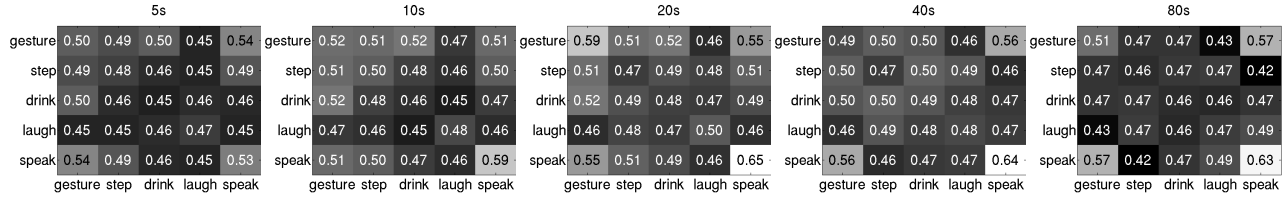

Figure 5: Ground-truth social action labels: F-formation detection performance in terms of classification accuracy per action pair using the 10 people from the fully annotated data.

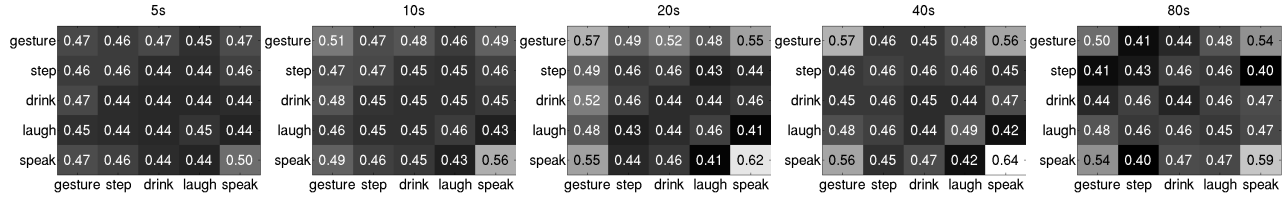

Figure 6: Estimated social action labels using the random forests method: F-formation detection performance in terms of classification accuracy per social action pair for the 10 people from the fully annotated data.

pairs. The same cross-validated experiments were carried out using the mean mutual information of all action pairs over a given window. The balanced classification accuracy is shown in Table 2 for the same window sizes. In the first row, a peak classification accuracy of 0.62 is achieved for the 20s and 40s window. This is a reduction in performance compared to just the *speaking-speaking* action pair. Given that other social action pairs can be indicative of F-formation membership, this suggests that more complex classifiers and more data for these other social actions are required to understand the problem more deeply.

6.2.2 Using Estimated Social Actions

Results on the Fully Annotated 10-person Data.

Having seen what the upper bound on performance is when using the ground truth labels, we now present the performance results when using the social actions as estimated using the Random Forest method as described in Section 5. We present just results with this method since its social action estimation performance was better than the HMMs. Again, we initially carried out tests by observing the performance differences for different action pairs. The results are summarised in Figure 6. If we compare the performance when using the ground truth social action labels with the estimates from the random forest, the performance is comparable but achieves the best accuracy (0.64) when using a 40s window and the *speaking-speaking* social action pair. Comparing this to the results using the ground truth labels of the social actions, where the best performance of 0.65 classification accuracy was obtained using a 20s window, we see that significantly longer window sizes are necessary, probably to account for the noisier estimates of the social actions. Note that the colour coding is normalised across Figures 6, 7 and 5 for easier comparison.

In terms of the performance when using the mean mutual information of all action pairs, the results are summarised in Table 2 again in terms of the class-balanced classification accuracy. In the second row, the best performance (0.64) is again obtained with a 40s window. However, now a more comparable performance of 0.63 classification accuracy is also achieved with the 20s window, which is also quite competitive compared to the results when using the ground truth labels. This suggests that when the social action estimates are more noisy, then exploiting other social action streams becomes more useful. Using the mean MI values across all social action pairs leads to an equal or slight increase in performance over any single action pair for all window sizes. This suggests again that considering other social action pairs is beneficial.

Table 3: Frequency of occurrence (in seconds) of different group sizes for both data sets.

	1	2	3	4	5
all 32 persons	1528	2624	2210	1300	126
26-person data	1694	3150	1736	874	10
10-person data	2848	1576	0	0	0

Results on the 26 person data.

We performed F-formation estimation on the full set of people (with working sensor readings) during the same 10 minute interval. Since not everyone was wearing a microphone, this data was generated from the 10-person training data only. The results for each action pair and using the mean of the MI are shown in Figure 7 and Table 2 respectively. In both cases, the performance of the F-formation detection drops to below the random baseline. There are a number of possible explanations for this.

First, the social action performance could be worse given that the size of the training data is significantly less than the test set. Second, a deeper analysis of the data reveals very different distributions of each group size instance in the 10person and 26 person data (see Table 3). Note that the table also shows statistics for those who were not labelled to be in any F-formation and are therefore singletons.

As shown in Table 3, the distribution of group sizes was substantially different. It is also likely that in the 10-person data, many of the singletons and some of the dyads were people talking others who were not wearing microphones and therefore not annotated for social actions. Likewise in the 26-person data, only the data from those who had accelerometer readings were used to generate the distribution. By comparing with the statistics for all 32 subjects, we see that there were deviations between the true and estimated F-formation sizes. Note that these differences do not affect the F-formation estimation performance because it was trained on pairwise mutual information values. However, there are clearly more groups of size 3-5 and therefore estimating differing F-formation sizes is a much more challenging task. It is likely that the mutual information between dyads is much higher because both have to be constantly actively involved in the conversation. On the other hand, for smaller group sizes, the behaviour is perhaps less strongly co-ordinated.

We investigated further by computing the accuracy per group size, as shown in Figure 8 for 40s and 80s window lengths. We can see that above random performance is mostly concentrated in the groups of size 2. This further suggests that different group sizes should be treated differently during training and testing and indeed could

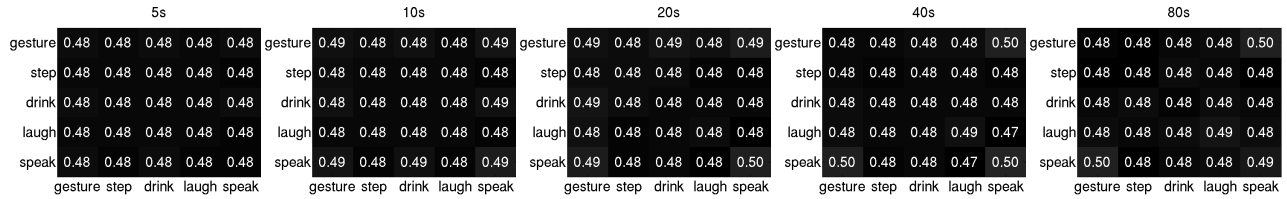

Figure 7: F-formation detection performance in terms of classification accuracy per social action pair, using action pairs estimated with random forests for all 26 people from the 10 minute interval annotated for F-Formations.

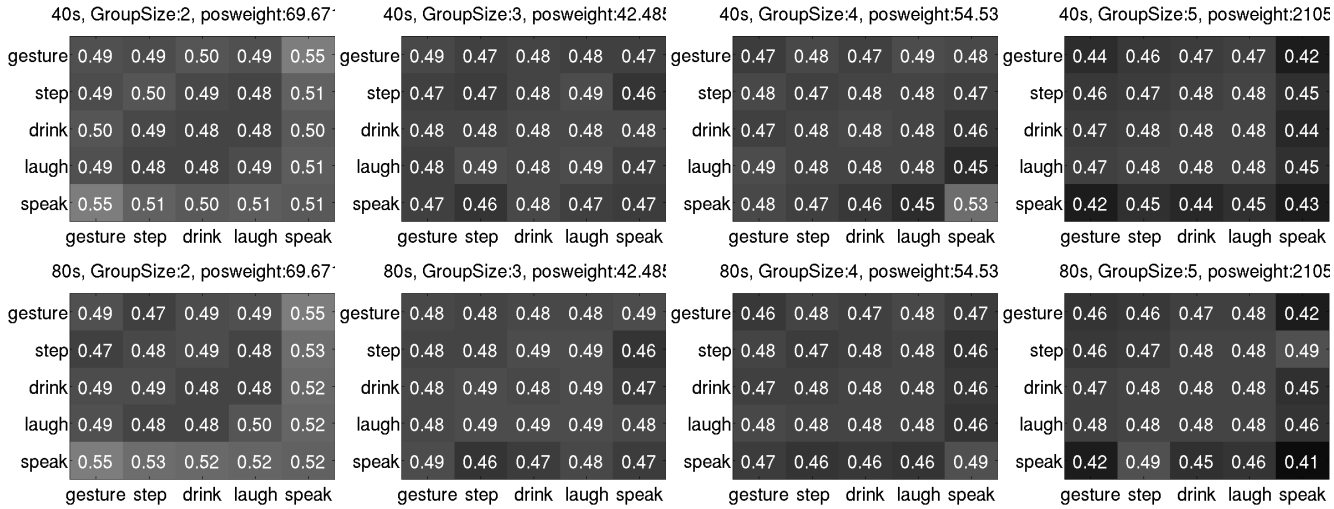

Figure 8: F-formation detection performance in terms of class balanced classification accuracy per action pair in different group sizes. Posweight shows the ratio of the number of negative to positive data points for each corresponding group size.

depend on certain action pairs more than others.

6.2.3 Ground Truth vs. Estimated Social Actions

On comparing the performance of the F-formation detection using ground truth compared to estimated labels, in Figures 5 and 6, a general decrease in performance is observed except for the *gesturing-gesturing* social action pair at the 40s window length. This shows that each of the defined social actions have a salient and measurable meaning in the co-ordination of behaviour in F-formations. The exception is due to errors in the social action estimate: closer inspection of the estimated social actions revealed that *gesturing* is most often mistaken for speaking, which is the single most informative action in the data.

When using the true labelled social actions, taking the mean mutual information did not improve the performance over considering the mutual information of any individual action pair. In comparison, taking the mean mutual information tended to improve the performance of the F-formation detection using estimated social actions. Also when using both the estimated labels, taking the mean of the mutual information of all social action pairs led a reduced latency in obtaining above close to the best performance. This suggests that in the presence of noisy labels, relying on multiple action pairs rather than a single one to improve accuracy and latency is a sensible choice. It also suggests that combining multiple action pairs at shorter time scales can provide stronger evidence for F-formations.

6.3 Social Actions vs. Motion

It is not clear whether it is purely the co-ordination of motion rather than specific social actions that indicate whether people are in an F-formation or not. We carried out additional experiments where we calculated the mutual information on the raw tri-axial motion signal per person. Since the mutual information is sensitive to the binning choice when approximating a probability distribution, we wanted to chose a more stable method of estimating the mutual information from continuous data. To this end, we used the implementation of the mutual information proposed by Peng et al. [28] for both these experiments, and those with binary streams. For the motion data, this method approximates the distribution of the data using kernel density estimation.

Table 4: F-formation accuracy from raw acceleration

Window Size (s.)	10-person data	26-person data
5	0.55	0.52
10	**0.56**	0.51
20	**0.56**	0.51
40	0.55	0.50
80	0.52	0.48

We carried out the same experiments calculating the mutual information on the magnitude of the raw acceleration streams. The results are shown in Table 4. Here we see that for the 10-person data, the performance does not surpass the social action approach, which suggests that there is something salient about the social actions that provides clear distinctions between behaviour which is relevant to conversations, and other motion which does not. For the 26-person data, the motion features alone do provide some discrimination, which suggests that the reason that our social action approach did not work is possible related to not having enough data for training the social actions.

There are, of course, other ways of pre-processing the motion features before calculating the mutual information for every pair. However, more complex representations of the motion features could have resulted indirectly in more semantically meaningful representations of the motion that would go beyond treating it as the raw signal, thus defeating the purpose of this particular comparative study.

7. CONCLUSIONS

In this paper, we have demonstrated that complex social behaviour related to conversations can be estimated using just a single body-worn accelerometer. The goal of the paper was to understand how the co-ordinated body movements can be indicative of being in a conversation. Thus far, it has been used to augment audio-based sensing of conversations in less crowded and acoustically clean situations. First, we have shown improved social action estimation performance on non pre-segmented data over prior work [13]. Second, we have also shown that it is possible to detect F-formations by measuring the co-ordination of automatically extracted conversationally relevant behaviours. Importantly, we have highlighted that using noisy estimates of speaking activity alone to detect F-formations is informative but that taking into account other conversationally related behaviours leads to an improved performance. A best classification accuracy of 64% was achieved for the *speaking-speaking* social action pair where mutual information between these streams was accumulated over a 40s window for the fully annotated 10-person data. Combining multiple social actions to estimate F-formation membership showed slightly improved results and warrants further investigation. However, when considering the 26-person data, overall F-formation estimation performance dropped below the random baseline. Further investigation showed that this was because more different group sizes were taken into account than in the 10-person data where only dyads were present. Future work will investigate how to develop methods that take into account group size.

8. ACKNOWLEDGEMENTS

Thanks to Matthew Dobson, Claudio Martella, and Maarten van Steen (VU University of Amsterdam) for the use of their wearable sensors and assistance during the data collection. We also thank Jeroen Kools and Ben Kröse (University of Amsterdam) for their help during the data collection and the annotation process. This work was partially supported by the Delft Technology Fellowship, European Commission under contract number FP7-ICT-600877 (SPENCER) and grant agreement number 601033 - MOnarCH, the Dutch National Program COMMIT, and the Costa Rican Institute of Technology and is affiliated with the Delft Data Science consortium.

9. REFERENCES

[1] L. Bao and S. Intille. Activity recognition from user-annotated acceleration data. *Pervasive Computing*, pages 1–17, 2004.

[2] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[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] T. Choudhury. *Sensing and Modeling Human Networks*. PhD thesis, 2004.

[5] T. Choudhury and A. Pentland. Characterizing Social Networks using the Sociometer. In *PROCEEDINGS OF NAACOS 2004*, 2004.

[6] T. M. Ciolek and A. Kendon. Environment and the Spatial Arrangement of Conversational Encounters. *Sociological Inquiry*, 50(3-4):237–271, 1980.

[7] 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.

[8] R. Dunbar, N. Duncan, and D. Nettle. Size and structure of freely forming conversational groups. *Human Nature*, 6(1):67–78, 1995.

[9] N. Eagle and A. S. Pentland. Reality mining: sensing complex social systems. *Personal Ubiquitous Computing*, 10(4):255–268, 2006.

[10] D. Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: A review. *Image and Vision Computing*, 27(12):1775–1787, 2009.

[11] J. Gips. Social Motion: Mobile Networking Through Sensing Human Behavior. Master's thesis, 2006.

[12] J. Gips and A. Pentland. Mapping Human Networks. In *PerCom*, pages 159–168. IEEE Computer Society, 2006.

[13] H. Hung, G. Englebienne, and J. Kools. Classifying Social Actions with a Single Accelerometer. In *Ubicomp*, pages 207–210. ACM, 2013.

[14] 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.

[15] 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.

[16] D. B. Jayagopi and D. Gatica-Perez. Mining Group Nonverbal Conversational Patterns Using Probabilistic Topic Models. *IEEE Transactions on Multimedia*, 12(8):790–802, 2010.

[17] A. Kendon. *Conducting Interaction: Patterns of Behavior in Focused Encounters*. Cambridge University Press, 1990.

[18] T. Kim, A. Chang, L. Holland, and A. Pentland. Meeting mediator: enhancing group collaborationusing sociometric feedback. In *CSCW*, pages 457–466, 2008.

[19] 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.

[20] C.-C. Lian and J. Y.-j. Hsu. Probabilistic Models for Concurrent Chatting Activity Recognition. In *IJCAI*, pages 1138–1143, 2009.

[21] A. Madan, K. Farrahi, D. Gatica-Perez, and A. Pentland. Pervasive sensing to model political opinions in face-to-face networks. *Pervasive Computing*, pages 214–231, 2011.

[22] P. Marshall, Y. Rogers, and N. Pantidi. Using F-formations to analyse spatial patterns of interaction in physical environments. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, CSCW '11, pages 445–454, New York, NY, USA, 2011. ACM.

[23] A. Matic, V. Osmani, A. Maxhuni, and O. Mayora. Multi-modal mobile sensing of social interactions. In *Pervasive computing technologies for healthcare (PervasiveHealth), 2012 6th international conference on*, pages 105–114. IEEE, 2012.

[24] A. Matic, V. Osmani, and O. Mayora-Ibarra. Analysis of social interactions through mobile phones. *Mobile Networks and Applications*, 17(6):808–819, 2012.

[25] D. O. Olguin, A. Madan, M. Cebrian, and A. Pentland. Mobile Sensing Technologies and Computational Methods for Collective Intelligence. In N. Bessis and F. Xhafa, editors, *Next Generation Data Technologies for Collective Computational Intelligence*, Studies in Computational Intelligence, pages 575–597.

[26] D. O. Olguín, 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: Cybernetics*, 2009.

[27] J. A. Paradiso, J. Gips, M. Laibowitz, S. Sadi, D. Merrill, R. Aylward, P. Maes, and A. Pentland. Identifying and facilitating social interaction with a wearable wireless sensor network. *Personal and Ubiquitous Computing*, 14(2):137–152, 2010.

[28] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1226–1238, 2005.

[29] A. S. Pentland, T. J. Kim, et al. *Enhancing distributed collaboration using sociometric feedback*. PhD thesis, Massachusetts Institute of Technology, 2011.

[30] L. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *PROCEEDINGS OF THE IEEE*, 77(2):257, 1989.

[31] 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.

[32] D. Wyatt, T. Choudhury, and J. Bilmes. Conversation detection and speaker segmentation in privacy-sensitive situated speech data. In *In Proc. of Interspeech*, 2007.

[33] T. Zhang, J. Wang, P. Liu, and J. Hou. Fall detection by embedding an accelerometer in cellphone and using KFD algorithm. *Int. Journal of Computer Science and Network Security*, 6(10):277–284, 2006.