# Capturing human behaviour through wearables by computational analysis of social dynamics

Gedik, Ekin

**DOI**
10.4233/uuid:8f61321c-24c8-4d00-aa9c-738a125e6c98

**Publication date**
2018

**Document Version**
Final published version

**Citation (APA)**
Gedik, E. (2018). *Capturing human behaviour through wearables by computational analysis of social dynamics*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:8f61321c-24c8-4d00-aa9c-738a125e6c98

# Capturing human behaviour through wearables by computational analysis of social dynamics

# Capturing human behaviour through wearables by computational analysis of social dynamics

## Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology,
by the authority of the Rector Magnificus prof. dr. ir. T.H.J.J. van der Hagen,
chair of the Board of Doctorates,
to be defended publicly on
Tuesday 4 December 2018 at 15:00 o'clock

by


**Ekin Gedik**

Master of Science in Computer Engineering,
Middle East Technical University, Turkey,
born in Ankara, Turkey.

This dissertation has been approved by the

promotor: Prof. dr. ir. M.J.T. Reinders and
copromotor: Dr. H. Hung

Composition of the doctoral committee:

| | |
|---|---|
| Rector Magnificus, | Chairperson |
| Prof. dr. ir. M.J.T. Reinders, | Delft University of Technology |
| Dr. H. Hung, | Delft University of Technology |
| *Independent members:* | |
| Prof. dr. C.M. Jonker | Delft University of Technology |
| Prof. dr. B.J.A. Krose | University of Amsterdam |
| Prof. dr. D.K.J. Heylen | University of Twente |
| Dr. J.M. Odobez | IDIAP Research Institute |
| Dr. ir. R.W. Poppe | Utrecht University |



This work was supported by the COMMIT/ community and the Delft Technology Fellowship.

Printed by: Proefschriftmaken.nl
Cover designed by: Argun Cencen and Ekin Gedik, inspired by [90]



ISBN 978-94-6380-143-0

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

# Contents

# 1

# Introduction

*Man is by nature a social animal; an individual who is unsocial naturally and not accidentally is either beneath our notice or more than human. Society is something that precedes the individual. Anyone who either cannot lead the common life or is so self-sufficient as not to need to, and therefore does not partake of society, is either a beast or a god.*

Aristotle, Politics

1

## 1.1. Humans as social beings

Understanding human behaviour, how individuals or groups respond to internal and external stimuli, has always been a fascinating topic for people from various professions including philosophers, theologians, artists, and scientists. It's hard to imagine a person who has not reflected on his/her own actions or questioned others' behaviour and the underlying motives. Many brilliant writers, from Shakespeare to Dostoevsky, from Russell to Camus, delved deep into the human mind and explored behaviour from a literary perspective, influencing many coming after them, even scientists. Jean-Luc Godard, the renowned French film director, says "Art attracts us only by what it reveals of our most secret self." [1]. In the 20th century, B.F. Skinner, a famous psychologist and the father of radical behaviourism, presented a scientific framework for studying, predicting, and controlling human behaviour [2]. These few examples and countless other attempts show how understanding human behaviour has always been an intriguing topic for mankind.

Behaviour is not completely individual though. The definition we used in the first paragraph includes the words 'groups and external stimuli', implying behaviour is also shaped by outside factors. Humans are inherently social beings. Even though there is an ongoing scientific debate about it, many scientists believe that we are wired to be social, as a product of evolution [3]. Aristotle argued that a person who prefers not to be a part of society is either a beast or a god. Karl Marx criticized the traditional conception of human nature as a species that incarnates itself in each individual, instead arguing that the conception of human nature is formed by the totality of social relations [4]. Peter Singer said we were social before we were even human [5]. Our behaviour is shaped by our interactions with the world and other people. Thus, in order to understand human behaviour completely and crack open the mysteries of being human, we need to look at social interactions. The traditional approach in psychology for analysing and understanding social behaviour has been by manual analysis of collected data by experts. Although social psychologists have presented insightful studies over the years, this process is extremely time-consuming [6] and not all social behaviour might be easily observable by annotators. With the increasing processing power of computers, it has become possible to facilitate computational and statistical methods that can automatically analyse and detect patterns in huge amounts of data relatively quickly. These methods have also shown themselves to be capable of detecting patterns that are not easily observable by humans. Even though automated methods might not replace manual analysis currently, they provide valuable and complementary information. Moreover, the increasing popularity of personal sensing, such as wearable devices and smartphones, has made it easier to collect data in every day, real life situations. Motivated by these advances in computing and sensing, this thesis aims to present novel computational approaches for social understanding and focuses on the use of wearable sensors in real life scenarios, specifically crowded mingling events and live performances.

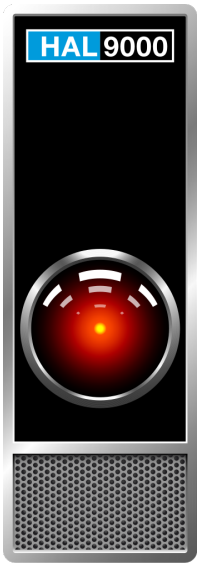


Figure 1.1: HAL 9000, a machine with 'Social Intelligence' from the science fiction book '2001: A Space Odyssey". The visual characteristics of HAL 9000 shown in this picture are taken from the movie adaptation of the book, directed by Stanley Kubrick in 1968.

## 1.2. Behavioural, affective and social computing through the years

Previously, we gave examples of philosophers, writers, artists, and psychologists who tried to investigate the social behaviour of humans. In the last quarter of the 20th century, a new profession became interested in social behaviour; computer scientists. Rosalind Picard argued that Spock, the first officer of USS Enterprise in the fictional TV series Star Trek, is seen as the patron saint of computer science, since he was highly rational, highly intelligent, and non-emotional [7]. This was certainly true for the early (and even medium) phases of computational studies, where researchers mainly focused on mathematical absolutes. One of the most prominent domains of computer science research is the 'Artificial Intelligence'. Humans have always been fascinated by humans; how we learn, represent, decide, and act. Artificial Intelligence (AI), coined by John McCarthy in 1956 for the now famous Dartmouth workshop [8], is the branch of computer science that aims to create 'intelligent' agents; machines that can mimic the cognitive functions of humans to solve problems, achieving 'General Intelligence'.

Even though AI has always been a divided field, with subfields regularly disagreeing with each other on the particular goals and the methods of research, we can say that most of the early studies focused on providing rigorous theory and methods of decision making, using either logical reasoning, statistical methods, or artificial neural networks [9]. In the following years, when research in theoretical AI became comparatively more mature, some scholars started to argue that

in order to achieve 'General Intelligence', a virtual agent should also have 'Social Intelligence' (Figure 1.1); it should be able to recognize the underlying factors of human behaviour, such as emotions and social factors. One of the first subfields of AI that focused on such aspects is 'Affective computing'.

### 1.2.1. Affective computing

Affective computing is defined as the computing that relates to, arises from or influences emotions by Rosalind Picard in her 1995 technical report [10]. As she stated in her invited introduction for the first issue of IEEE Transactions on Affective Computing; the idea of computing research related to emotions was found ludicrous by some and many people were sceptical in the beginning [11]. However, in the following years, Affective computing has become a respectable research area with its own journal and committed researchers.

From 1995 to 2018, the field of Affective computing has seen many developments, but the main questions that scholars tried to answer stayed generally the same. Most of the research in this field focused on the automatic detection of expressed affective states, either classifying them into discrete states of emotion or predicting continuous values on the dimensions of affect, such as valence and arousal. During the first years, the research generally focused on two of the most widely acknowledged forms of sentic modulation; facial expressions and voice which were generally investigated through video and audio modalities, respectively. We should note that research on the expression of emotions through different media was not new but Affective computing provided an automated way of detecting, measuring, and even transforming them [11].

The majority of affective computing research that employs video or audio, considered the estimation of basic emotions (anger, joy, sadness, disgust, fear, and surprise), or a subset of them as the target task. Facial expressions had been the primary cue for analysing and detecting affect through video. The Facial Action Coding System (FACS) proposed by Ekman et. al. was the backbone of such research [12]. Facial expression analysis focused on the automatic discovery, tracking, and representation of the Action Units(AU), as well as the detection of expressions based on these AUs. Throughout the years, many models for the representation [13–16] were presented and various classifiers [17–19] were used for the detection [13]. Audio analysis employed a variety of prosodic and acoustic features [13, 20, 21]. For both of the modalities, there has been a shift from controlled scenarios to spontaneous natural recordings in recent years. Alongside this, there has been a surge of representation learning approaches, rather than feature engineering as was done before, where Convolutional [22, 23] and Recurrent Neural Networks [24] have been employed to automatically obtain features. Also, there is an increasing preference for using facial expressions and voice together as cues to analyse the affective state, instead of relying on a single modality [13], which is compatible with how persons demonstrate affect in a multi-modal way.

Two other cues that have gained importance in recent years are body expressions and physiological signals. Modalities such as video, motion capture, and wearable sensing are being used to obtain spatio-temporal affective body features,

aiming to represent body postures, gestures, and movement [25]. Behavioural science studies showed that body expressions are much more important than previously thought and indeed are powerful affective communication channels [26, 27]. Although more affective studies started to investigate bodily expressions, either as the sole signal or in addition to more traditional modalities, most of the research focused on a limited set of acted body expressions in specific scenarios (dance, gait, posture, gesture) [25]. Physiological sensing is definitely an interesting direction for affect computing, since in comparison to other modalities, a computer has more access to motor output compared to humans [10]. For example, people tend to be relatively good at detecting affective signals encoded in facial expressions, voice, and bodily gestures, but they don't have direct access to physiological information, such as heart-rate, skin-conductance, respiration-rate, etc. With the increasing possibility of more pervasive physiological sensing systems and the known connection between some emotional states and physiological signals, more researchers have started to investigate in this direction [28].

From the aforementioned studies, we can see one strong commonality in the Affective computing research: A cognitivist approach is followed where the emotional states and expressions of people are modelled *individually*. The social context is generally not considered. As some argued before in the literature [29], we believe emotions are inherently social. Hence, to model human behaviour correctly, one must consider the social context too. This brings us to the emerging domain of social signal processing that aims to model, detect, and synthesise social signals and context.

### 1.2.2. Social signal processing

Social signal processing (SSP) aims to create socially aware computers. The vision is to provide computers with the ability to sense and evaluate social signals of people through various modalities of input, resulting in social intelligence. Social signals are defined as the expression of a person's attitude towards a social situation and manifest themselves through various non-verbal cues [30]. These non-verbal cues were presented under five main categories in the survey by Vinciarelli et. al. (based on existing social psychology literature): physical appearance, gestures and posture, face and eye behaviour, vocal behaviour, and space and environment [30]. We can then practically define the aim of SSP research to model and automatically detect social cues throughout various input modalities and facilitate these social cues to understand high level social behaviour.

#### Steps of SSP research

The main steps of a typical SSP study are data collection, detection of people in the scene, extraction of the informative cues from input data and their interpretation with respect to the social signals, and context sensitive classification of the social signals into the social-behaviour-interpretative target categories [30] (Figure 1.2). Of course, not all SSP studies need to have all of these steps. For example, a study with wearable sensors will already have the mapping of sensors to the individuals and will not require the step of person detection.

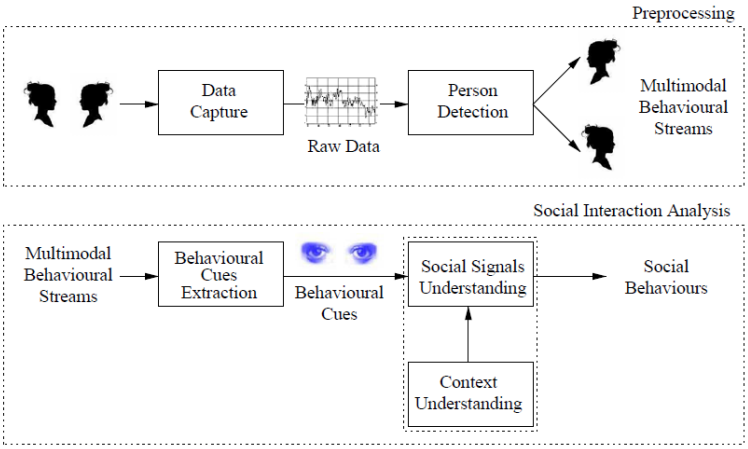


Figure 1.2: General workflow of a SSP study, as presented in [30].

For data collection, mostly cameras and microphones are being used; either as the single sensor capturing an event in a simple fashion [31], or arranged into synchronised multiple sensors to obtain multiple channels of recording [32]. Alternatively, data can be collected using mobile sensing, smartphones or wearable custom sensor packs that can include accelerometers, proximity sensors and microphones [33]. Also, physiological sensors might also be present in such customised sensor arrays, measuring heart rate, blood pressure, skin conductivity, etc. [34]. Less pervasive sensors, such as functional magnetic resonance imaging (fMRI) [35] and Electroencephalography (EEG) [36] have also been used to acquire social signal data.

The main challenges of the data collection relate to privacy aspects and the passiveness of the sensors in [30]. Any data collection experiment should ensure the privacy of the participants through means of informed consent and anonymisation. Also, it can be argued that some sensors intrinsically ensure privacy more than other and thus should be used if possible. For example, recording facial images and the voice of a person is perceived much more intrusive than recording body acceleration. Passiveness relates to the unintrusiveness of the sensors [37]. The intrusiveness of the sensors should be minimized when collecting data for SSP, to ensure that recordings are as natural as possible. With less intrusive and more privacy preserving sensors, participants will tend to forget that they are being recorded at all, ensuring that the data collected will resemble real life closely.

Person detection is generally done as a pre-processing step in the SSP domain when more than one person is being recorded. For the audio, this process is called speaker diarisation and first detects speech segments which are then clustered into personalised streams. Basic features such as energy and autocorrelation, as well as more specialised ones such as Mel Frequency Cepstral Coefficients (MFCCs) and Linear Predictive Coding (LPC) are being used for distinguishing between speech and non-speech segments. Recent approaches tended to combine the steps of

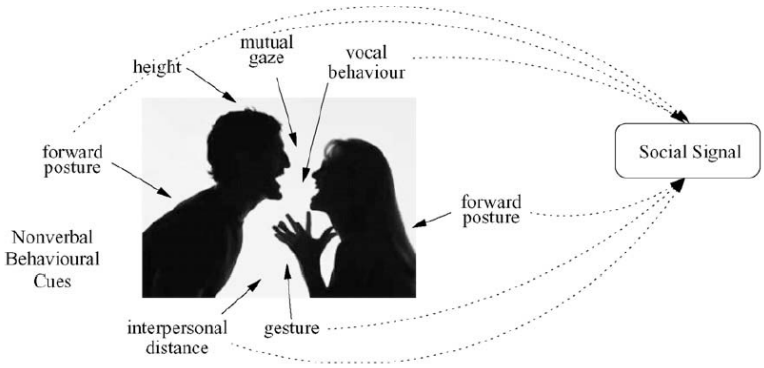


Figure 1.3: Social signals are produced by the combination of non-verbal behavioural cues. Image is taken from [30].

speech detection and clustering together by using graphical methods, such as Hidden Markov Models or Dynamic Bayesian Networks [38, 39]. Most approaches that employ video, either try to detect faces in the scene (using appearance features for representation and statistical learning for classification [40, 41]) or focus on the detection of the whole body (mostly using edge features and motion information [42]).

As mentioned, social signals can be grouped into five categories. There are only a few studies related to the first category of physical appearance, and they mainly focused on evaluating the beauty of faces [43, 44] (Figure 1.3). Even though some studies extract body related information (skin-hair color, body type, etc.) [45, 46], they lack the social aspect. This situation is similar for the second category of gesture and posture, where often the final goal is not understanding social behaviour. For example, gesture recognition is an active research topic in Computer Vision, but with the main application as Human-Computer Interaction [47, 48]. Although, there do exist a few studies that do investigate affect through gestures [25]. Alternatively, gait and posture are mainly investigated for biometric recognition [49, 50] and surveillance [51] purposes.

On the other hand, gaze and face, the third category of social signals, are being studied from an affective computing point of view; mainly by detecting facial expressions through AUs, which we already covered in the former subsection. For vocal behaviours (the fourth category), there exist few studies that aim to detect non-verbal vocalisations such as laughter [52, 53] and crying [54], whereas others generally focus on the analysis of linguistic behaviour [55, 56]. There are also quite a few studies that relate to the final category, social environment and space understanding. Generally, these focus on the detection of face-to-face interaction [33, 57].

#### Goals of SSP research

In recent years, SSP investigated several social phenomena with a focus on social relationships (roles in interaction) and social attitudes (dominance and personality)

[30, 58]. Social roles are generally dependent on the context and the nature of the interaction, and can be divided into two groups: formal and informal roles. Formal roles are predefined roles with specific functions such as the chair in a meeting (also known as functional roles). Informal roles are related to social structures.

The automatic detection of formal roles has received the most attention as these roles show more distinguishable behavioural cues with respect to others in an interaction. The two main detection approaches that are favoured are: the analysis of speaking behaviour and the choices of lexical content [58], mainly when analysing meetings [59] or news and radio broadcasts [60]. Cues from speaking behaviour, such as turn-taking, interventions, and overlaps, were found to be informative for distinguishing between roles [61, 62]. A few studies were multimodal (using video in addition to audio for detecting gestures as fidgeting [63]), but mainly the analysis is solely based on the audio.

Dominance is one of the social attitudes, defined as the impact one has on the group behaviour. Similar to role detection analysis, dominance detection is generally determined from speaking activity [58, 64], although movement and gaze (derived from the video input) seem to improve the results when analysed jointly with speaking [65].

The personality of an individual is another socially relevant concept that is expected to influence group interactions. Automatic detection of personality traits from non-verbal behaviour mainly uses prosodic features of audio (energy, pitch, etc.), movement and, spatial proximity [66, 67].

Social attitudes have also been studied to see how they influence negotiation [68], rapport through coordination [69] and agreement or disagreement [70]. In addition to the speaking behaviour, which was again essential, mimicry and correlation were found to be strong cues for the detection of social attitudes.

From this brief review we conclude that the goals of SSP research is varied but loosely connected. Not always is the goal to estimate the higher level social concept, but also essential steps towards this goal might be the topic of study. For example, when analysing speech behaviour ( which is an important source to detect various social phenomena) one often first needs to obtain speaking turns which makes the step of speech detection necessary. Similarly, if one wishes to analyse the dynamics of interaction in a group, first the groups and their members in the scene should be detected.

Besides detecting and analysing social signals, the synthesis of social actions, emotions and attitudes, is also receiving more and more attention but this is outside the scope of this thesis.

## 1.3. Mobile and wearable sensing for social computing

As we skimmed through the Affective Computing and Social Signal Processing literature, we have seen that various sensing media were employed for detecting socially relevant cues. We can say that the focus was mainly on video and audio input, obtained through stationary sensors. Even though the use of these modali-

ties was shown to be quite effective for the detection of affective and social signals, there are some conceptual limitations regarding privacy and passiveness. There are also some technical limitations regarding the places recordings take place in, for example, densely crowded gatherings.

### 1.3.1. Limitations of traditional audio and video sensing

Limitations related to the privacy and passiveness are connected to each other and related to how people perceive being recorded. People generally do not want to be recorded as they feel that it is too invasive to their privacy. These concerns are more extreme for media such as video and audio, since people are easily identifiable from them. Of course, there are methods of anonymisation, such as blurring the faces in a video, or recording descriptive features instead of raw audio. Still, these are pre- and post-processing steps and participants are expected to fully trust the researchers. Privacy invasive sensing media also affect the passiveness of the sensor. Even in cases where the sensor does not explicitly interfere with the people, such as a mounted camera, people tend to change their behaviour after realising they are being recorded in a non-privacy respective way, resulting in less naturalistic representations.

Video and audio sensors are often considered stationary, limiting the experiments to a particular spatial setup. This hampers a general applicability, since every application will require the same instrumentation (sensor placing). Although, we do understand the importance of these studies for the analysis of social phenomena in a controlled way, they severely limit the applicability in real-life applications. Another issue is how stationary sensors perform in specific real-life scenarios. For example, most of the work focusing on the analysis of facial expressions investigate constrained scenarios where accurate head, face, and facial tracking are obtainable. The application of such methods in a real-life event with spontaneously interacting people is quite hard and will require multiple carefully placed cameras.

The analysis of speaking behaviour is often applied in the setting of a meeting. In such a scenario, it is relatively easy to obtain accurate speaking turns of participants with correctly placed microphones. But in a larger event where multiple conversations can occur simultaneously, spontaneously and dynamically, stationary microphones are expected to result in low quality data collection, missing voices of people. The nature of the problem lies in the number of data channels per participant for both video and audio. If the ratio of participants to sensors is high, the person detection problem becomes harder.

### 1.3.2. Ubiquitous computing and mobile sensing

We believe a possible solution to the conceptual limitations of traditional sensing lies in the Ubiquitous computing paradigm. Mark Weiser, a chief scientist at Xerox Parc, founded the Ubiquitous computing domain with his 1992 essay 'The Computer for the 21st Century' [71]. He proposed that the most profound technologies are the ones that disappear and argued that the computers should be integrated into the daily life in a seamless way. He believed that the computing devices should fit the human environment instead of forcing humans to enter theirs. We support the

Figure 1.4: A custom made wearable sensor pack. This sensor pack is used in all studies presented in this thesis.

same for sensing devices. If a sensing device guarantees the users their privacy and disappears into the background, the users will forget that they are being recorded. Such an approach will result in a privacy-preserving and passive sensing experience, guaranteeing naturalistic recordings. Such technologies have been developed for years, both in terms of hardware and software, bringing Weiser's vision closer to reality [72, 73].

Mobile sensing has an organic connection with Ubiquitous computing. As Weiser argued, not everything mobile is ubiquitous and not everything ubiquitous is mobile but the intersection between these two domains is huge. We believe, as ubiquitous sensing was an answer to the conceptual limitations, mobile sensing provides a practical solution to the technical limitations of the stationary sensors. Social concepts in real life require time to arise and their analysis requires continuous sensing of multiple people in multiple places. Covering such scenarios with stationary devices is hard; requiring prior instrumentation of places where might people go. However, with a personal mobile sensing device, obtaining continuous sensing is much easier (if we disregard for now issues such as energy consumption). Also, the person detection step can be omitted; avoiding any noise that will be introduced by the detection process. These properties make mobile sensing a powerful candidate for computational social analysis studies.

### 1.3.3. Mobile sensing for SSP

With the increased popularity and accessibility of smartphones with various embedded sensors, mobile sensing has become a hot topic of research. In recent years, various applications of mobile sensing are proposed, with topics as diverse as transportation, environmental monitoring, health and well being, and social networking

[74]. The term SSP itself was first used by Pentland to group his and his group's pioneering studies related to the social understanding where mobile sensors were utilised [75]. They used the cell phones' proximity sensors and microphones to perform an automatic analysis of multiple persons' daily lives for months, which came to be known as 'Reality Mining' [33]. After this, the mobile sensors were employed for analysing various social concepts.

A specific version of mobile sensors are wearable badges (Figure 1.4). One of the first examples of a custom wearable badge, a Sociometer, is presented by Pentland's group in [76]. These badges are then used in various applications, ranging from the analysis of real life social networks [77] to the classification of people into personality traits [78]. Other wearables were also used in SSP for many different purposes; the detection of social cues as such speech [79], the detection of interacting partners [80], distinguishing between roles [81], detecting social attractors [82], etc.. In order to obtain less-intrusive and ecologically valid data, all studies of this thesis solely rely on the use of wearable sensors.

## 1.4. Current limitations: social actions, dynamics, coordination, and experimenting in the wild

This section focuses on identifying some conceptual and theoretical limitations of the current research approaches in SSP. We address these limitations with a similar categorization used for the steps of a SSP study, moving from data collection to the detection of social cues as proxies of signals, and then to the social behaviour understanding. We already briefly discussed the limitations of stationary sensors in the former section. In order to avoid those limitations, we have proposed the use of wearable sensing which provides a relatively passive, continuous, and personalised data collection procedure, eliminating the need for a person detection step.

### 1.4.1. Detection of social actions

The next step in a SSP study is the detection of behavioural cues that are expected to act as proxies for social signals. We will investigate this step in the scope of 'social action' detection. A social action is defined as an action of an Agent A performed in relation to an Agent B, who is perceived as a self regulated agent with its own goals by Agent A [83]. There are two main forms of social actions in real life interaction, those related to turn-taking and those related to backchannels [58]. Hence, we can argue that some of the most informative social actions are speaking (for the analysis of turn-taking behaviour and vocalisations related to backchannels) and gesturing (hand gestures accompanying speech and head movements related to the backchannels ).

In the literature, audio and video have been highly favoured for speaking turns and gesture detection [58]. This is understandable since the physical manifestation of speech shows itself in audio and gesturing should be distinguishable in videos. However, when we scrutinise the majority of work on speaking detection from audio, we see that the environment is fairly constrained; closer to lab conditions. In a real life scenario with many people, obtaining clean audio recordings becomes

harder. The background noise, possible music, etc. affect the quality of the audio recordings. One possible solution is to use personalised microphones but we already discussed privacy concerns related to audio recording. The majority of the efforts on gesture detection are directed towards alternative methods of human-computer interaction [47] and understanding sign language [48]. As expected, most of these works focus on hand gestures and uses unobstructed videos obtained in constrained settings, where the subject is clearly visible. Obtaining such videos using such a set up in real life scenarios is unlikely.

One overlooked source of information for the detection of social actions (in our case speaking and gesturing) is the body movement. This information is easily obtainable through mobile devices with embedded accelerometers. This sensing medium is cheap, privacy-preserving, energy efficient, and mobile. However, its capabilities are highly understudied in terms of the detection of social cues. In this thesis, we propose to facilitate this information source for the detection of social actions.

### 1.4.2. The role of dynamics in interaction

An intermediate step in SSP research that was not explicitly stated in the former categorisation is the detection of interacting partners. Social concepts generally arise in the existence of interaction. Hence, detecting who is interacting with whom in a scene should be a key step in understanding social concepts. However, works that study interaction in SSP generally assume the existence of interaction between people in the scene. For example, many investigated small group behaviour in the SSP domain [84] but such studies do not require the detection of interacting partners since all people in the meetings are assumed to be potential interaction partners.

Detecting groups in a scene was studied more for surveillance purposes rather than understanding social behaviour [30]. Traditionally, single images or videos were used as the main input and the approaches depended solely on the proxemics; the spatial position and orientation of participants in the scene. We already argued that obtaining such information might be hard for cases where social interactions are most likely to occur in crowded environments. SSP studies facilitating wearable sensors mainly relied on proximity sensing too, either through Infrared receivers or radio based sensors. Such studies generally do not provide a quantitative evaluation of the detection of groups and mainly assume any proximity detection shows a the existence of conversational interaction [33, 77]. Similar to the video based approaches, these studies mostly build their solutions on proxemics.

One really important aspect of interaction that has been generally overlooked is social dynamics. Proxemics is indeed important, since face-to-face interaction of two spatially distant persons is not possible. However, co-location and even mutually focused orientation might not exclusively point to the existence of the interaction. Over the years, a few researchers have focused on the dynamic aspects of interaction, such as synchrony, however the final aim was not the detection of groups but medical diagnosis [85, 86]. Another issue is how the temporal aspect is generally overlooked. Even while using video, the analysis tends to be performed on

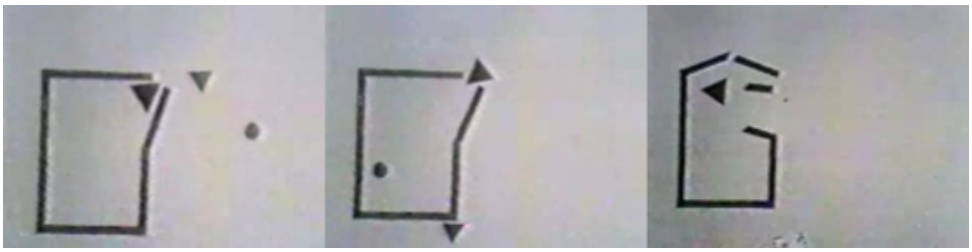

Figure 1.5: Three frames from the video shown to the participants in the experiment conducted by Heider and Simmel, taken from [90].

still images [87, 88] and studies considering the temporal aspect are few [82, 89]. However, viewing interaction as a static process does not align with how we perceive interaction as humans.

In a 1944 experiment carried out by psychologists Heider and Simmel [91], participants were shown a video of two triangles and one circle, moving around the space (Figure 4.1). Their movement were designed to be suggestive of humans moving around and interacting. Participants joining the experiment were inclined to treat these shapes as humans, primarily because of the dynamics of their movement. If the participants were just given a still image, they would not see humans in those shapes. We see interaction on a temporal basis and identify them through, in addition to proxemics, the coordination of the movements and actions of the participants. This also has a basis in social psychology, where coordinated movement and even mimicry between interacting partners were already identified by researchers [92]. Thus, we believe the dynamics of interaction and more specifically, the coordination of the partners' actions and movement, are precious information sources that are widely understudied. Therefore, in this thesis, we investigate methods of exploiting the coordination of peoples' movements and actions for detecting conversing groups.

### 1.4.3. Effects of coordination on appraisal

The dictionary definition of appraisal is defined as an act of assessing something or someone. Appraisal theory, a theory in psychology, essentially argues that one's evaluation of an external stimuli (appraisal of a situation) will cause relevant affective responses based on the appraisal [93]. Affective computing research drew from appraisal theory, where researchers investigated how a situation/event emotionally affected people and how these induced emotions connected to the people's evaluation of the situation/event. The analysis of a persons' reaction to an outside artistic stimulus such as movies, music and live shows are generally performed on individuals [94]. The main methodology is automatically detecting their affective states or measuring the changes in affective dimensions with respect to the stimuli presented to the subjects. Different sensing media are used; video and physiological sensing (generally non-pervasive signals such as EEG) being the most preferred ones. Studies are generally conducted in a constrained lab-like setting, where par-

ticipants are subjected to the stimuli individually and data is collected with carefully placed sensors. The main application areas were video summarisation [95], implicit tagging [96] and finding emotion-eliciting movies and music [94, 97].

This methodology, sensing an individual to analyse affective responses, is acceptable for cases where the participant is the only person who is subjected to the external stimuli; for example while they are watching a movie by themselves. However, when people go out and experience an event together with other people, such as going to a movie or a concert, an undeniable social aspect is introduced into the equation. There are immediate effects of being with people, reflected during the event itself. People tend to have coordinated responses (contagious laughter in comedy movies for example) during the event and their evaluation of the event can be affected by sharing the experience with other people. Another component is the delayed effects of sharing an experience with people. Just as being with other people can affect the immediate experience (affective responses), it can also affect one's social behaviour afterwards. These are all interesting questions that were not considered before. A part of this thesis investigates how the linkage of the body movements of people attending a live performance can be exploited to estimate their appraisal of a performance and how attending such events affect their subsequent social behaviour. The term linkage is originally presented for measuring the dependence of the physiological signals of two people [98]. In this thesis, we argue that this concept can be extended to the body acceleration.

### 1.4.4. Experimenting in the wild

Finally, we want to briefly discuss a dominant preference related to the data capture in the SSP domain. Even though there are many studies that collect data in real life settings, many researchers still conduct experiments on acted and controlled data. We believe controlled studies conducted in lab environments are important to validate hypotheses and measure the effectiveness of novel approaches. However, they should be treated as a step in scaling up to large real life scenarios where social interactions occur spontaneously in a natural way. Researchers should test their approaches in real life scenarios, in the wild, after perfecting them in acted and controlled settings. We believe that, only in this way can we understand how social interactions truly unfold in real life and thus provide actual solutions that will generalise to larger populations. To this end, every experiment presented in this thesis is conducted on data collected in real life events.

## 1.5. Challenges: addressing the limitations

In the former section, some limitations of the current methodologies in computational social understanding studies were identified and possible ways to overcome them were presented. We proposed to use accelerometers for detecting social actions, argued for focusing on social dynamics for identifying conversing groups, presented the linkage of peoples' actions in live performances as a possible solution for estimating their evaluations (appraisals) and pushed the need for experimenting in the wild. Each of these proposed solutions come with their own challenges. In

this subsection we discuss these challenges.

### 1.5.1. Detecting social actions with accelerometers

We advocated that accelerometers can be good candidates for detecting social actions such as speaking, gesturing, etc., since they are privacy preserving, energy efficient and provides continuous, personalised sensing. However, using accelerometers for this purpose is not optimal and comes with its own challenges. Accelerometers were generally used to detect every day activities, such as walking, running, transporting, etc. [99]. They were also used in health-care for automatic fall detection [100]. However, for such studies, the connection between the sensing medium and the physical manifestation of the action is direct. When a person starts to move, it will be directly visible in the acceleration. This is not the case for speaking and gesturing (if the sensor is not directly attached to the hand or head), where the connection is relatively indirect. However, in social psychology, the connection between speech and body movements is already studied, showing how speaker and listener behaviour differ. This opens a research direction where this (indirect) connection between the social action and sensing can be investigated for detection, which is previously understudied.

However, moving into this direction, sensing an action where the connection is indirect, requires a specialised approach. In addition to the indirectness of the connection between the action and the sensing medium, social actions tend to be highly person specific. The nature of the connection and the physical manifestation of the subsequent behaviour are could differ greatly for each person. Person specific training is a way of capturing such differences. For simple actions such as walking, person independent methods have been shown to perform well [99], but for more person specific actions such as speaking and gesturing, recognition results were relatively unsatisfactory [101]. Another challenging aspect is related to the placement and number of the accelerometers used. With more accelerometers that are placed at different parts of the body, it will become possible to detect more complicated social actions. However, such an instrumentation will not be realistic for a real world scenario. The experiments in this thesis are conducted with a single body-worn accelerometer per person which also increases the complexity of the problem. To devise a generalisable and satisfactory solution, a way of training personalised models without requiring data from new subjects should be investigated.

### 1.5.2. Interaction dynamics for detecting conversing groups

We have argued that the dynamics of interaction is a widely overlooked information source for the detection of interacting groups. As mentioned, most of the existing work relied on the proxemics; the detection of the relative positions and orientations of people in a scene. Obtaining this information has its own challenges that was briefly addressed in the former section. But after positions and orientations are estimated, the problem of detecting groups is relatively well defined. That is, athough the spatial orientations and positions of people in groups can vary, there are some conceptual and geometric constraints that makes it easier to formulate

a solution. For example, two people that are spatially far away cannot be in the same group. The possibility of two persons that are back to back being in the same group is quite low. On the other hand, the social dynamics that can arise in conversing groups are much more complex. Devising rules for interaction dynamics is quite hard since there are too many factors that can affect the characteristics of the interaction. These factors might be related to the characteristics of the people in the group, such as their personalities and mood, or can be linked to the cardinality of the group. Combining all these factors for a generalisable solution is challenging and is not attempted yet.

### 1.5.3. Estimating appraisal through linkage in real life events

We proposed to utilise the linkage of people's body movements to estimate their evaluations of real life performances. For the scenarios mentioned, the use of generally preferred sensing technologies (video, audio and EEG) become less effective. A cinema or a concert hall is generally dark, making it harder to capture the faces of people to analyse their spontaneous affective responses. Other than in a controlled experiment setting, no one would like to wear EEG sensors over their head when attending a concert. Affective computing has already started to investigate how emotions connect to the bodily expressions [25]. We believe this is an interesting direction for affective studies that aims to measure people's reactions to an outside artistic stimuli. As we stated before, wearable acceleration can be used to infer body movements which can be then used to capture spontaneous reactions of participants in live performances. This information can then be used to estimate higher level affective and social concepts, such as people's emotional states and their evaluations of the event.

Of course, not all parts of an event will be equally informative of a participant's evaluation. We aim to facilitate participants' coordinated spontaneous reactions throughout the event sensed through a body worn accelerometer. When the scenarios we are interested in are considered, where participants movements are expected to be somehow restricted, one might assume there won't be much movement, probably resulting in similar linkage values throughout the event. However, previous studies already showed that even in scenarios of limited movement, the movement behaviour is still indicative of high-level social information, such as the profession of a person [102]. Following from this result, we argue that it should be possible to automatically distinguish between parts of the event that are more informative of a participants evaluation and use this information to estimate the final appraisal.

### 1.5.4. Revising the concept of experimenting in the wild

Experimenting in the wild comes with its own difficulties. However it is a necessary evil that needs to be addressed if we truly want to understand how social behaviour is exhibited in real life. By moving to real life data analysis, the constraints and limitations of real life scenarios will be investigated more, hopefully resulting in satisfactory solutions that can be used for real life applications. The challenges start with the data collection. In order to obtain a realistic representation of a real

life scenario, intervention during the experiment should be kept to a minimum. This requires careful planning of the event since it will not be possible to stop the event, fix the problems and restart all over again. Even though the author of this thesis was a part of the data collection team for all the studies presented in this thesis, challenges related to the data collection and how they should be addressed are out of the scope of this thesis. For more information about issues related to the data collection in the wild, please refer to [103].

Even if the data collection process is considered to be successful, data obtained from realistic scenarios might be challenging for computational algorithms to analyse and use. Since the participants are free to act as they wish, it is highly probable that some classes in the data are going to be under-represented which might be problematic for pattern recognition approaches. The amount of noise in real life data is expected to be higher than controlled or acted scenarios in the lab. Many factors contribute to this, such as less restrictions affecting the sensing process (people playing with their sensors throughout the event for example). Since the event is not acted and flows freely, the actions of the participants cannot be known beforehand, thus manual annotation of the behaviours observed in the data might be required which can introduce a new layer of noise. So, any researcher that is working on data collected in real life events should be aware of these possible disruptions. They should first analyse the data to detect the existence of such disruptions, select/modify their methods accordingly and interpret their results with respect to them.

## 1.6. Contributions

The works presented in the following chapters of this thesis try to address the limitations mentioned in the former section in a novel way. All studies in this thesis explicitly use wearable sensors with embedded accelerometers, in order to cover the limitations of traditional video and audio recordings mentioned in the Section 1.3.1. The following four chapters (2, 3, 4, and 5) use data collected in a crowded mingle scenario where free standing conversational groups exist. The sixth chapter uses data collected in two live performances with different characteristics. All the data used in the experiments are collected in real life events with as few constraints as possible, aiming to provide recordings close to the real life as possible. The structure of the thesis and the contributions of each chapter are as follows:

Chapter 2 presents a novel transfer learning approach to obtain a generalisable and scalable solution to the detection of social actions through worn body acceleration and specifically focuses on the detection of speaking.

Chapter 3 builds on the results of the former section and focuses on the analysis of how the performance of the method presented in Chapter 2 is affected by the training set size. It also presents results for various social actions, discussing how the nature of the target action affects the performance.

Chapter 4 investigates how the dynamics of interaction, the coordinated actions and movements of people in a scene, can be used to estimate group membership. It presents an ensemble selection approach which provides group size awareness.

Chapter 5 acts as a proof of concept study, where we show the social signals

obtained by the former steps (social actions and interaction) can be used to estimate higher level social concepts, in this case personality. This chapter has also interesting insights into multimodal sensing, where it is shown that extracting a second behavioural modality (speaking status) from one physical modality (acceleration) results in better recognition performance.

Chapter 6 investigates how being affected by the same stimuli simultaneously in live performances results in coordinated responses of audience members and how it can be used to automatically infer the reappraisals of the participants. It also analyses how socialising might change the experience of such an event.

Works presented in Chapters 2 to 6 were already published or accepted for publication in conference proceedings and journals. For reasons of formality, we have kept the chapters identical to the publications. However, we acknowledge that it might be hard for the reader as some parts will be repeated. Below, I rest which parts are repeated and how the chapters are related in terms of the datasets used:

- The Transductive Parameter Transfer (TPT) method is first presented in Chapter 2.4 in detail. In chapters 3.4, 4.4.1, and 5.4.1 TPT is reintroduced in a less detailed manner. Reader can omit these sections after reading Chapter 2.4.

- Chapters 2, 3, 4 and 5 use different subsets of a large dataset, in terms of different annotations, time intervals and included participants. This large dataset is collected in three different days in real life speed dating events followed by mingling sessions. This thesis uses accelerometer and proximity data from the mingling sessions, exclusively. Chapters 2.5, 3.1, 4.3, and 5.3 include information about how the subsets used in each chapter are formed. We believe, the most general information regarding the large dataset is provided in Chapter 4.3. For a more detailed explanation of the dataset, we refer the readers to consult [103]. Chapter 6 uses entirely different datasets which are explained in Chapters 6.4 and 6.6.1.

# References

[1] B. Henderson, *Godard on godard: notes for a reading,* Film Quarterly **27**, 34 (1974).

[2] B. F. Skinner, *Science and human behavior* (Simon and Schuster, 1953).

[3] M. D. Lieberman, *Social: Why Our Brains are Wired to Connect* (Oxford University Press, USA, 2013).

[4] K. Marx, *The German ideology: including theses on Feuerbach and introduction to the critique of political economy* (Pyr Books, 1976).

[5] P. Singer, *The expanding circle: Ethics, evolution, and moral progress* (Princeton University Press, 2011).

[6] N. Lehmann-Willenbrock, H. Hung, and J. Keyton, *New frontiers in analyzing dynamic group interactions: Bridging social and computer science,* Small group research **48**, 519 (2017).

[7] R. W. Picard, *Affective computing: challenges,* International Journal of Human-Computer Studies **59**, 55 (2003).

[8] J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon, *A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955,* AI magazine **27**, 12 (2006).

[9] P. McCorduck, *Machines who think: A personal inquiry into the history and prospects of artificial intelligence* (AK Peters Natick, MA, 2004).

[10] R. W. Picard *et al.*, *Affective computing,* (1995).

[11] R. W. Picard, *Affective computing: from laughter to ieee,* IEEE Transactions on Affective Computing **1**, 11 (2010).

[12] P. Ekman and W. V. Friesen, *Manual for the facial action coding system* (Consulting Psychologists Press, 1978).

[13] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, *A review of affective computing: From unimodal analysis to multimodal fusion,* Information Fusion **37**, 98 (2017).

[14] A. Lanitis, C. J. Taylor, and T. F. Cootes, *Automatic face identification system using flexible appearance models,* Image and vision computing **13**, 393 (1995).

[15] Y. Yacoob and L. Davis, *Computing spatio-temporal representations of human faces*, Ph.D. thesis, research directed by Dept. of Computer Science.University of Maryland at College Park (1994).

[16] S. Kimura and M. Yachida, *Facial expression recognition and its degree estimation,* in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on* (IEEE, 1997) pp. 295–300.

[17] Y.-I. Tian, T. Kanade, and J. F. Cohn, *Recognizing action units for facial expression analysis,* IEEE Transactions on pattern analysis and machine intelligence **23**, 97 (2001).

[18] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan, *Dynamics of facial expression extracted automatically from video,* Image and Vision Computing **24**, 615 (2006).

[19] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, *Facial expression recognition from video sequences: temporal and static modeling,* Computer Vision and image understanding **91**, 160 (2003).

[20] T. Vogt and E. André, *Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition,* in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on* (IEEE, 2005) pp. 474–477.

[21] L. Devillers, L. Vidrascu, and L. Lamel, *Challenges in real-life emotion annotation and machine learning based detection,* Neural Networks **18**, 407 (2005).

[22] C. Xu, S. Cetintas, K.-C. Lee, and L.-J. Li, *Visual sentiment prediction with deep convolutional neural networks,* arXiv preprint arXiv:1411.5731 (2014).

[23] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, *Learning spatiotemporal features with 3d convolutional networks,* in *Computer Vision (ICCV), 2015 IEEE International Conference on* (IEEE, 2015) pp. 4489–4497.

[24] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, *Convolutional mkl based multimodal emotion recognition and sentiment analysis,* in *Data Mining (ICDM), 2016 IEEE 16th International Conference on* (IEEE, 2016) pp. 439–448.

[25] A. Kleinsmith and N. Bianchi-Berthouze, *Affective body expression perception and recognition: A survey,* IEEE Transactions on Affective Computing **4**, 15 (2013).

[26] A. Mehrabian and J. T. Friar, *Encoding of attitude by a seated communicator via posture and position cues.* Journal of Consulting and Clinical Psychology **33**, 330 (1969).

[27] M. Argyle, *Bodily communication* (Routledge, 2013).

[28] C. Peter, E. Ebert, and H. Beikirch, *Physiological sensing for affective computing,* in *Affective Information Processing* (Springer, 2009) pp. 293–310.

[29] B. Parkinson, A. H. Fischer, and A. S. Manstead, *Emotion in social relations: Cultural, group, and interpersonal processes* (Psychology Press, 2005).

[30] A. Vinciarelli, M. Pantic, and H. Bourlard, *Social signal processing: Survey of an emerging domain,* Image and vision computing **27**, 1743 (2009).

[31] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard, *Modeling human interaction in meetings,* in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, Vol. 4 (IEEE, 2003) pp. IV–748.

[32] A. Waibel, T. Schultz, M. Bett, M. Denecke, R. Malkin, I. Rogina, R. Stiefelhagen, and J. Yang, *Smart: The smart meeting room task at isl,* in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, Vol. 4 (IEEE, 2003) pp. IV–752.

[33] N. Eagle and A. S. Pentland, *Reality mining: sensing complex social systems,* Personal and ubiquitous computing **10**, 255 (2006).

[34] H. Gunes, M. Piccardi, and M. Pantic, *From the lab to the real world: Affect recognition using multiple cues and modalities,* in *Affective Computing* (InTech, 2008).

[35] P. R. Montague, G. S. Berns, J. D. Cohen, S. M. McClure, G. Pagnoni, M. Dhamala, M. C. Wiest, I. Karpov, R. D. King, N. Apple, *et al.*, *Hyperscanning: simultaneous fmri during linked social interactions,* (2002).

[36] L. Q. Uddin, M. Iacoboni, C. Lange, and J. P. Keenan, *The self and social cognition: the role of cortical midline structures and mirror neurons,* Trends in cognitive sciences **11**, 153 (2007).

[37] S. Mukhopadhyay and B. Smith, *Passive capture and structuring of lectures,* in *Proceedings of the seventh ACM international conference on Multimedia (Part 1)* (ACM, 1999) pp. 477–487.

[38] J. Ajmera, I. McCowan, and H. Bourlard, *Speech/music segmentation using entropy and dynamism features in a hmm classification framework,* Speech communication **40**, 351 (2003).

[39] J. Ajmera, I. A. McCowan, and H. Bourlard, *Robust audio segmentation*, Tech. Rep. (IDIAP, 2004).

[40] K. S. Huang and M. M. Trivedi, *Robust real-time detection, tracking, and pose estimation of faces in video streams,* in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, Vol. 3 (IEEE, 2004) pp. 965–968.

[41] P. Wang and Q. Ji, *Multi-view face detection under complex scene based on combined svms.* in *ICPR (4)* (2004) pp. 179–182.

[42] N. Dalal, B. Triggs, and C. Schmid, *Human detection using oriented histograms of flow and appearance,* in *European conference on computer vision* (Springer, 2006) pp. 428–441.

[43] P. Aarabi, D. Hughes, K. Mohajer, and M. Emami, *The automatic measurement of facial beauty,* in *Systems, Man, and Cybernetics, 2001 IEEE International Conference on*, Vol. 4 (IEEE, 2001) pp. 2644–2647.

[44] Y. Eisenthal, G. Dror, and E. Ruppin, *Facial attractiveness: Beauty and the machine,* Neural Computation **18**, 119 (2006).

[45] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, *Keep it smpl: Automatic estimation of 3d human pose and shape from a single image,* in *European Conference on Computer Vision* (Springer, 2016) pp. 561–578.

[46] C. BenAbdelkader and Y. Yacoob, *Statistical estimation of human anthropometry from a single uncalibrated image,* Computational Forensics , 200 (2008).

[47] S. S. Rautaray and A. Agrawal, *Vision based hand gesture recognition for human computer interaction: a survey,* Artificial Intelligence Review **43**, 1 (2015).

[48] S. Joudaki, D. b. Mohamad, T. Saba, A. Rehman, M. Al-Rodhaan, and A. Al-Dhelaan, *Vision-based sign language classification: a directional review,* IETE Technical Review **31**, 383 (2014).

[49] R. Poppe, *Vision-based human motion analysis: An overview,* Computer vision and image understanding **108**, 4 (2007).

[50] L. Lee and W. E. L. Grimson, *Gait analysis for recognition and classification,* in *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on* (IEEE, 2002) pp. 155–162.

[51] T. Gandhi and M. M. Trivedi, *Pedestrian protection systems: Issues, survey, and challenges,* IEEE Transactions on intelligent Transportation systems **8**, 413 (2007).

[52] L. Kennedy and D. Ellis, *Laughter detection in meetings,* in *NIST ICASSP 2004 Meeting Recognition Workshop* (2004) pp. 118–121.

[53] K. P. Truong and D. A. Van Leeuwen, *Automatic discrimination between laughter and speech,* Speech Communication **49**, 144 (2007).

[54] P. Pal, A. N. Iyer, and R. E. Yantorno, *Emotion detection from infant facial expressions and cries,* in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, Vol. 2 (IEEE, 2006) pp. II–II.

[55] E. E. Shriberg, *Phonetic consequences of speech disfluency*, Tech. Rep. (SRI INTERNATIONAL MENLO PARK CA, 1999).

[56] Y. Liu, E. Shriberg, A. Stolcke, and M. Harper, *Comparing hmm, maximum entropy, and conditional random fields for disfluency detection,* in *Ninth European Conference on Speech Communication and Technology* (2005).

[57] A. S. Pentland, *Automatic mapping and modeling of human networks,* Physica A: Statistical Mechanics and its Applications **378**, 59 (2007).

[58] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder, *Bridging the gap between social animal and unsocial machine: A survey of social signal processing,* IEEE Transactions on Affective Computing **3**, 69 (2012).

[59] N. P. Garg, S. Favre, H. Salamin, D. Hakkani Tür, and A. Vinciarelli, *Role recognition for meeting participants: an approach based on lexical information and social network analysis,* in *Proceedings of the 16th ACM international conference on Multimedia* (ACM, 2008) pp. 693–696.

[60] Y. Liu, *Initial study on automatic identification of speaker role in broadcast news speech,* in *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers* (Association for Computational Linguistics, 2006) pp. 81–84.

[61] H. Salamin, S. Favre, and A. Vinciarelli, *Automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction,* IEEE Transactions on Multimedia **11**, 1373 (2009).

[62] S. Favre, A. Dielmann, and A. Vinciarelli, *Automatic role recognition in multiparty recordings using social networks and probabilistic sequential models,* in *Proceedings of the 17th ACM international conference on Multimedia* (ACM, 2009) pp. 585–588.

[63] W. Dong, B. Lepri, A. Cappelletti, A. S. Pentland, F. Pianesi, and M. Zancanaro, *Using the influence model to recognize functional roles in meetings,* in *Proceedings of the 9th international conference on Multimodal interfaces* (ACM, 2007) pp. 271–278.

[64] R. Rienks, D. Zhang, D. Gatica-Perez, and W. Post, *Detection and application of influence rankings in small group meetings,* in *Proceedings of the 8th international conference on Multimodal interfaces* (ACM, 2006) pp. 257–264.

[65] D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, *Modeling dominance in group conversations using nonverbal activity cues,* IEEE Transactions on Audio, Speech, and Language Processing **17**, 501 (2009).

[66] D. O. Olguın, P. A. Gloor, and A. S. Pentland, *Capturing individual and group behavior with wearable sensors,* in *Proceedings of the 2009 aaai spring symposium on human behavior modeling, SSS*, Vol. 9 (2009).

[67] G. Mohammadi, A. Vinciarelli, and M. Mortillaro, *The voice of personality: Mapping nonverbal vocal behavior into trait attributions,* in *Proceedings of the 2nd international workshop on Social signal processing* (ACM, 2010) pp. 17–20.

[68] J. R. Curhan and A. Pentland, *Thin slices of negotiation: Predicting outcomes from conversational dynamics within the first 5 minutes.* Journal of Applied Psychology **92**, 802 (2007).

[69] L.-P. Morency, C. Sidner, C. Lee, and T. Darrell, *Head gestures for perceptual interfaces: The role of context in improving recognition,* Artificial Intelligence **171**, 568 (2007).

[70] K. Bousmalis, M. Mehu, and M. Pantic, *Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools,* in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on* (IEEE, 2009) pp. 1–9.

[71] M. Weiser, *The computer for the 21st century.* Mobile Computing and Communications Review **3**, 3 (1999).

[72] G. D. Abowd and E. D. Mynatt, *Charting past, present, and future research in ubiquitous computing,* ACM Transactions on Computer-Human Interaction (TOCHI) **7**, 29 (2000).

[73] G.-J. Hwang and C.-C. Tsai, *Research trends in mobile and ubiquitous learning: A review of publications in selected journals from 2001 to 2010,* British Journal of Educational Technology **42** (2011).

[74] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell, *A survey of mobile phone sensing,* IEEE Communications magazine **48** (2010).

[75] A. Pentland, *Social signal processing [exploratory dsp],* IEEE Signal Processing Magazine **24**, 108 (2007).

[76] T. Choudhury and A. Pentland, *The sociometer: A wearable device for understanding human networks,* in *CSCW'02 Workshop: Ad hoc Communications and Collaboration in Ubiquitous Computing Environments* (2002).

[77] T. Choudhury and A. Pentland, *Sensing and modeling human networks using the sociometer,* in *null* (IEEE, 2003) p. 216.

[78] 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) **39**, 43 (2009).

[79] E. Gedik and H. Hung, *Personalised models for speech detection from body movements using transductive parameter transfer,* Personal and Ubiquitous Computing **21**, 723 (2017).

[80] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, and W. Van den Broeck, *What's in a crowd? analysis of face-to-face behavioral networks,* Journal of theoretical biology **271**, 166 (2011).

[81] G. Englebienne and H. Hung, *Mining for motivation: using a single wearable accelerometer to detect people's interests,* in *Proceedings of the 2nd ACM international workshop on Interactive multimedia on mobile and portable devices* (ACM, 2012) pp. 23–26.

[82] X. Alameda-Pineda, Y. Yan, E. Ricci, O. Lanz, and N. Sebe, *Analyzing free-standing conversational groups: A multimodal approach,* in *Proceedings of the 23rd ACM international conference on Multimedia* (ACM, 2015) pp. 5–14.

[83] R. Conte, C. Castelfranchi, *et al.*, *Cognitive and social action* (Garland Science, 2016).

[84] D. Gatica-Perez, *Automatic nonverbal analysis of social interaction in small groups: A review,* Image and vision computing **27**, 1775 (2009).

[85] C. Saint-Georges, A. Mahdhaoui, M. Chetouani, R. S. Cassel, M.-C. Laznik, F. Apicella, P. Muratori, S. Maestro, F. Muratori, and D. Cohen, *Do parents recognize autistic deviant behavior long before diagnosis? taking into account interaction using computational methods,* PloS one **6**, e22393 (2011).

[86] E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen, *Interpersonal synchrony: A survey of evaluation methods across disciplines,* IEEE Transactions on Affective Computing **3**, 349 (2012).

[87] F. Setti, C. Russell, C. Bassetti, and M. Cristani, *F-formation detection: Individuating free-standing conversational groups in images,* PloS one **10**, e0123783 (2015).

[88] H. Hung and B. Kröse, *Detecting f-formations as dominant sets,* in *Proceedings of the 13th international conference on multimodal interfaces* (ACM, 2011) pp. 231–238.

[89] S. Vascon, E. Z. Mequanint, M. Cristani, H. Hung, M. Pelillo, and V. Murino, *Detecting conversational groups in images and sequences: A robust game-theoretic approach,* Computer Vision and Image Understanding **143**, 11 (2016).

[90] L. Demers, *Machine performers: Neither agentic nor automatic,* in *ACM/IEEE HRI Workshop on Collaborations with Arts* (2010).

[91] F. Heider and M. Simmel, *An experimental study of apparent behavior,* The American journal of psychology **57**, 243 (1944).

[92] T. L. Chartrand and J. A. Bargh, *The chameleon effect: the perception–behavior link and social interaction.* Journal of personality and social psychology **76**, 893 (1999).

[93] K. R. Scherer, A. Schorr, and T. Johnstone, *Appraisal processes in emotion: Theory, methods, research* (Oxford University Press, 2001).

[94] C. Chênes, G. Chanel, M. Soleymani, and T. Pun, *Highlight detection in movie scenes through inter-users, physiological linkage,* in *Social Media Retrieval* (Springer, 2013) pp. 217–237.

[95] M. Soleymani, G. Chanel, J. J. Kierkels, and T. Pun, *Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses,* in *Multimedia, 2008. ISM 2008. Tenth IEEE International Symposium on* (Ieee, 2008) pp. 228–235.

[96] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, *A multimodal database for affect recognition and implicit tagging,* IEEE Transactions on Affective Computing **3**, 42 (2012).

[97] M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang, *1000 songs for emotional analysis of music,* in *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia* (ACM, 2013) pp. 1–6.

[98] J. M. Gottman, *Detecting cyclicity in social interaction.* Psychological Bulletin **86**, 338 (1979).

[99] L. Bao and S. S. Intille, *Activity recognition from user-annotated acceleration data,* in *International Conference on Pervasive Computing* (Springer, 2004) pp. 1–17.

[100] C. Doukas, I. Maglogiannis, P. Tragas, D. Liapis, and G. Yovanof, *Patient fall detection using support vector machines,* in *IFIP International Conference on Artificial Intelligence Applications and Innovations* (Springer, 2007) pp. 147–156.

[101] H. Hung, G. Englebienne, and J. Kools, *Classifying social actions with a single accelerometer,* in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing* (ACM, 2013) pp. 207–210.

[102] G. Englebienne and H. Hung, *Mining for motivation: using a single wearable accelerometer to detect people's interests,* in *Proceedings of the 2nd ACM international workshop on Interactive multimedia on mobile and portable devices* (ACM, 2012) pp. 23–26.

[103] L. Cabrera-Quiros, A. Demetriou, E. Gedik, L. van der Meij, and H. Hung, *The matchnmingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates,* IEEE Transactions on Affective Computing (2018).

# 2

# Personalized models for speech detection

*You can get millions of dollars to drive a car through a desert, but you can not get money to try to do something that is more human.*

Marvin Minsky

This chapter is published as:
E. Gedik and H. Hung, **Personalised models for speech detection from body movements using transductive parameter transfer**, Personal and Ubiquitous Computing, 21(4), 2017.

Figure 2.1: A snapshot from the event

## 2.1. Introduction

This research addresses the analysis of social behaviour in crowded mingling events. Such events contain a large number of people interacting with each other closely. These scenarios are interesting since they are concentrated moments for people to interact, make new contacts, renew existing ones, or even influence each other.

In this paper, we focus on the detailed analysis of how to automatically detect whether someone is speaking in these dense crowded scenarios using just a single wearable triaxial accelerometer hung around the neck. Different challenges are introduced with the dense nature of such events, like the high non-stationary background noise from the audio and the heavy occlusion of people in the video. On the other hand, wearable sensors such as accelerometers are less affected by these challenges and their easy scalability makes their use appealing for such scenarios. Moreover, perceptions of privacy are often more sensitive to the recording of audio during conversations, even if the signal is immediately converted into privacy-sensitive features. In this paper, we focus on the use of accelerometers that could be embedded in a smart badge such as a conference badge and hung around the neck.

The use of accelerometers to detect speaking status is generally under-explored in the literature. However, limited amount of studies have shown that it is possible to detect whether someone is speaking based on just a single worn accelerometer [1, 2] by exploiting findings in behavioural psychology that speakers move (e.g. gesture) during speech [3]. One of the biggest challenges, which has not been addressed in the literature before, is accounting for the huge variation in ways in which people move while speaking. This person specific connection between movement and speech requires special approaches for detection, since relying on a single unified model to predict the speaking behaviour of everyone leads to large estimation errors as the size of the test population increases. We have chosen speech as the focus as our study since it is a vital unit of behaviour to analyse

social behaviour between people at the conversation level [4]. Some examples of further, higher level understanding that may follow from speech detection are the evaluation of an individual's social activeness, detection of conversing groups [1], dominance and group hierarchy [5, 6] and, cohesion [7]. In this paper, we propose to use transfer learning to enable the adaptation of a learnt ensemble model of speaking behaviour to a new unseen subject, based only on unlabelled data. The proposed method, Transductive Parameter Transfer [8], has never been used for this problem. With this method, we provide a solution that can generalise over large populations without requiring personal labelled data.



The key contributions of our work are: (i) we provide a study of speech detection through accelerometers, in a real world event (a snapshot is shown in Figure 2.1), with 18 participants; to our knowledge, no similar study at such scale exists. (ii) we delve deep into the connection between body movements and speech, showing how this problem differs from the traditional action recognition (e.g. walking) by providing results that compare the person dependent and independent models. (iii) we propose a transfer learning approach, which can generalise over large populations without requiring personal labelled data, overcoming the restrictions introduced by the person specific nature of speech. (iv) we present a detailed analysis of the parameter transfer that connects detection performance to personality which provides insight into the nature of both speech and transfer learning in this context.

## 2.2. Related work

### 2.2.1. Action recognition with accelerometers

Most research that has involved the detection of behaviour from worn accelerometers have tended to focus on the detection of daily activities. In 2004, Bao and Intille used five accelerometers worn on different body locations to detect 20 different actions which include activities like walking, sitting, running and vacuuming [9]. The data for the experiment was collected in a lab environment for 20 different participants. Statistical and spectral features extracted from acceleration data were used and different classifiers were compared for performance. Their results showed that, even without using person specific data, high recognition performance was possible for such actions.

The following year, Ravi et. al. presented their work that aims to detect eight similar daily activities with single worn accelerometer only [10]. The data collection was semi-controlled where the ordering of the activities was random. Their study showed that one accelerometer worn around the thigh area was sufficient for detecting many actions. With the rapid development of this domain, many different feature extraction techniques and classifiers are considered and compared with each other, providing a solid knowledge base for the detection of such activities [11].

Another research area that benefits from the utilisation of wearable sensors is health care, where people presented their work on automatic fall detection [12, 13]. As expected, also in these experiments, the data collection was carried out in a controlled environment where participants imitate falling. Both studies reported nearly perfect recognition scores. We show later in the "Comparing Controlled &

In-The-Wild Settings" section that there are significant differences in the nature of the data collected in controlled and acted settings compared to less controlled ecologically valid ones. Moreover, since such high accuracy was already obtained across a number of different participants, we can conclude that the nature of these tasks is much less sensitive to person-specific variations.



Unfortunately, none of these studies focuses on addressing the challenges of real life crowded environments or a social action like speech.

### 2.2.2. Transfer learning for behaviour recognition

Transfer learning is also used in some studies that focus on activity recognition for better performance but generally the setup of the transfer differs from our work. In their survey, Cook et. al. [14] grouped existing transfer learning studies with respect to the modalities used: video sequences [15], wearable [16, 17] and ambient sensors [18].

Some of these studies aim to transfer knowledge between different data acquisition setups, like van Kasteren et.al. [18]. This study is somewhat close to ours, since they used transfer learning to exploit existing labelled data sets to learn the parameters of a model applied in a new home. This was done to eliminate sensor placement and individual behaviour differences in each house. However, the sensors utilised (ambient sensors such as pressure mats, mercury contacts and passive infrared) and the detected actions (daily activities such as going to bed, brushing teeth, etc.) were entirely different than ours.

Another concept studied before is the transfer between actions. For example, Hu et. al. proposed a method, which focused on cross domain activity recognition [16]. They transferred the information from an available labelled data of a set of existing activities to a different yet still related set of activities. This was done by learning a similarity function between activities using Web search where web pages related to these activities are extracted and further processed to obtain a similarity measure (Maximum Mean Discrepancy). Similar to the former study, this study also presented its results on daily activities and used multimodal data streams as input.

Perhaps the closest study to ours was published by Zhao et. al. [17]. In this study, the authors presented a transfer learning based personalized activity recognition method. They used accelerometers embedded in mobile phones to gather data from different people while performing daily activities such as standing, walking, running and going upstairs or downstairs. In their method, they integrated decision trees (DT) and k-means clustering where decision trees were used to learn optimal parameters for labelled source data. Then, the DT model was transferred to a new user by classification and the initial parameters for k-means were set with respect to it. Finally, non terminal nodes of the DT were adapted to the new user, resulting in a personalized model. We discuss and experimentally show in our paper that the mentioned activities are less affected from interpersonal differences when compared to speech. Also, this method could only utilise a single source set for transfer while our approach can exploit multiple sources simultaneously. However, this study shows that transfer learning could be a good candidate for eliminating interpersonal differences.

### 2.2.3. Social computing with wearables

There are some studies in the literature that focus on analysing social phenomena using wearable sensors but most of them differ from ours in some aspects like the different modalities used as input, analysis of less crowded scenarios and lack of focus on fine time scale detection of social actions such as speech.



#### Large scale long-term studies

One of the first studies that utilises a wearable sensor for analysis of social phenomena was presented by Choudhury et. al. in 2003 [19]. Authors presented an automated method of analysing social network structures with the so-called sociometer, a wearable multimodal sensor that has a microphone, IR transceiver and two accelerometers. The data collection was done in two stages. In the first stage, 8 subjects from the same research group wore the sociometer during working hours for 10 days. The second stage included 23 participants from four different study groups wearing the badge for 11 days. In the study, audio data is used to detect speaking status, IR transceiver data was utilised for detecting interactions but acceleration information was not used. Using the frequency and duration of interactions detected, a social network of participants is formed. It is shown that by analysing this network, higher level information about the group structures, such as centrality of a participant, can be obtained.

Olguin et. al. obtained high level descriptions of human behaviour like physical and speech activity, face-to-face interaction, proximity and social network attributes using the sociometric badge mentioned earlier [20]. With this high level information, the authors classified the personality traits of participants, with respect to the "Big Five" model. The dataset included 67 participants and was collected for 27 days. Microphones and accelerometers were used to measure speech and physical activity, respectively. Although the study presented an excellent analysis of social phenomena throughout time, it did not focus on fine time grained detection of any action and aims to provide a higher level overview of social phenomena.

In a similar study conducted by Wyatt [21], social ties and collective behaviour of groups were investigated using a multimodal sensing device with 8 different modalities. Conversational characteristics of 24 people were analysed over 6 months. Similar to the former study, speech detection was applied to microphone data. Since social phenomena in a longer period of time is analysed in these studies, we expect the speech detection results to be quite rough. We believe participants current environment will greatly affect the actual detection performance. Such results are satisfactory for obtaining general statistics throughout time but if a fine grained analysis of speech and interaction is required, an approach that can provide more robust detection results of a fine scale is needed (e.g. over just a few seconds).

Apart from specialised sensor devices, some studies use mobile phones as social sensors like Madan et.al [22]. They used proximity, call data records and cellular-tower identifiers to investigate activities and interactions of individuals aiming to detect social behaviour changes with respect to illness. With the development of smart phones, this may eliminate the need for special sensing devices and makes scaling to bigger populations much easier.

**Studies of short-term dense crowded social events**

There are also studies that aim to analyse social behaviour in crowded mingling settings at a short-term level (i.e. minutes or hours rather than weeks or days). A recent study from Alameda-Pineda et.al. [23] showed that by combining sensor data from distributed cameras and wearable sensors, it was possible to obtain head and body pose estimation of people in a real life crowded event, with a fine time scale. The proposed method combined visual input from four cameras with noisy estimates of binary speaking status and proximity input obtained from wearable sensors and estimated the behaviour by learning from noisy incomplete observations using a matrix completion method. They went on to show that their automatically extracted head and body poses could be used to infer high level information such as detecting conversing groups or social attention attractors.

Cattuto et. al. [24] used conference badges equipped with RFID to analyse face-to-face interactions in crowded social gatherings. The exchange of radio packets between these badges were used to measure proximity and ultimately detect face-to-face interactions. The mentioned method was highly scalable and tested in three different events that include 25 to 575 people. Their analysis of the dynamics of interaction networks in these events showed a super-linear behaviour between the number of connections and their durations which can be used to define super connectors. However, this study automatically labelled interactions when two people came in close proximity but the accuracy of this was never evaluated.

Martella et. al. used accelerometers to predict implicit responses of an audience to a real life dance performance [25]. 32 spectators of the event were fitted with accelerometers hung around the neck. Aside from analysing their direct responses to the performance they also analysed the effects of the dance performance on the mingling behaviour of participants before and after the event using proximity sensing. Although the sensor pack was fitted with an accelerometer, no speech detection was carried out.

### 2.2.4. Speech detection with accelerometers

Although it is hard to find studies where wearable sensors were used for detecting speech and/or other social actions, there exists a few. Matic et. al. [26] used accelerometers for speech detection where accelerometers were tightly attached to the chest of participants in order to detect acoustic phenomena from speech. This methodology requires accelerometer to have a sample rate high enough to detect acoustic speech-based utterances and demands strict placement of the sensor which is impractical for many real life scenarios.

More similar to our work, Hung et. al.[2] presented their method for predicting social actions such as speaking, drinking, gesturing and laughter in a crowded environment with a single accelerometer hung around the neck. Spectral features were used to model these actions and HMMs were used for classification in a non adaptive learning approach. In a follow up to this study in 2014 [1], random forests were considered for classification and proved to perform better. In both studies, no detailed analysis to show variations of performance with respect to interpersonal differences were presented.

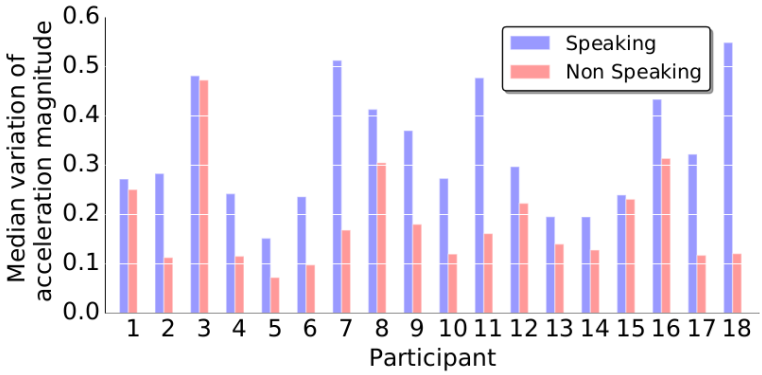


Figure 2.2: Median variance of acceleration magnitudes for speech & non-speech intervals for 18 people.

## 2.3. The nature of speech and body movements

In this section, we show how the person specific connection between speech and body movements shows itself in accelerometer readings by providing simple statistics computed from accelerometer readings of speech and non-speech intervals. These statistics, by proving the existence and personal nature of this connection, acts as a basis for our choice of an adaptive method that can eliminate interpersonal differences.

Similar to [1, 2], we aim to use movement information, obtained from accelerometers hung around the neck, as the proxy for speech. Fortunately, this assumption is partially backed by existing studies. Prior work has shown that it is possible to automatically classify conversing participants with an acceptable performance using acceleration information only [1, 2]. The connection between body movements and social behaviour is also extensively studied in social psychology [3, 27, 28]. For example, McNeill discussed that speakers tend to move noticeably more when compared to listeners [3]. It was discussed that gestures and speech are integrated parts of communication where gestures are used to complement the content of speech by providing visual stimuli acting as "symbols". Multiple studies also showed that there is a strong correlation and synchrony between speech and body movements in conversing groups [27, 28].

However, the connection between speech and body, specifically torso, movements is not theoretically well defined. Previous studies pointed to the existence of this connection but none made a precise description of the torso movement that can be exploited for automated detection that can generalise over large populations. We believe that this connection is highly personal and should be detectable from accelerometer readings. To test this assumption, we calculated the variation of accelerometer magnitudes over a sliding window (3 seconds with 1 second shift) of speech and non-speech intervals for 18 different people wearing accelerometers in a real life, crowded mingling event (see Section 2.5 for details).

Figure 2.2 shows the median values of the variation in accelerometer magnitudes

for speech and non-speech intervals. Each axis of raw acceleration is normalised using z-score standardization before computing the magnitude and extracting the variance values with sliding windows of the same length and shift size. We see huge differences between participants. One can easily see that one participant's median variation of accelerometer magnitude for speech intervals can be closer to another participant's non-speech feature. One-tailed t-tests applied to this feature during speech intervals for all pairwise combinations of participants showed that nearly 50% of these couples have significantly different distributions.

We also see that, for nearly all participants, the median of the variance in acceleration magnitude tends to significantly differ for speaking and non-speaking intervals. However, it can be also seen that the amount of this difference varies greatly per person. These two observations show that there is definitely a connection between speech and body movements but the nature of this connection is quite person specific.

This personal connection between speech and torso movement makes the problem entirely different and more challenging than traditional approaches to speech detection using audio. The connection between speech and audio is physically well defined via articulation of the vocal folds leading directly to resonances in the vocal tract. Of course, different speakers will have different spectral characteristics depending on their physiology[29] but satisfying speech detection results are already possible with person independent models [30].

With these findings, a traditional learning approach where the data of different subjects is amalgamated into a single training set will perform poorly since the decision surface obtained in this way will not be optimal. In our study, we propose to use Transductive Parameter Transfer [8, 31], an adaptive approach which uses transfer learning to overcome this issue by computing a personalised decision surface for each subject based on the similarity of a test subject's data distribution with those of multiple individuals in a training set.

## 2.4. The transductive parameter transfer method

With the findings of the last section, we propose to use an adaptive transfer learning approach, Transductive Parameter Transfer, presented in [8, 31]. The authors of [8, 31] used their method to compute personalized models for facial expression analysis from video input. To our knowledge, we present the first example of application of this method to action recognition and more specifically, speech detection from wearable sensors task. Although the main theory of the method stays the same, we have some different implementation choices than [8, 31] which we elaborate on below.

In this approach, with feature space $X$ and label space $Y$, $N$ source datasets with label information and the unlabelled target dataset are defined as $D_1^s, \ldots, D_N^s$, $D_i^s = \left\{x_j^s, y_j^s\right\}_{j=1}^{n_i^s}$ and $X^t = \{x_j^t\}_{j=1}^{n_t}$ , respectively. It is assumed that samples $X_i^s$ and $X^t$ are generated by marginal distributions $P_i^s$ and $P^t$, where $P^t \neq P_i^s$ and $P_i^s \neq P_j^s$. $P_t$ and $P_i^s$ are presumed to be drawn from $\rho$, the space of all possible distributions over $X$, with respect to meta distribution $\Pi$.

This approach aims to find the parameters of the classifier for the target dataset $X^t$, without using any label information of $X^t$, by learning a mapping between the marginal distributions of the source datasets and the parameter vectors of their classifiers. Main steps of the Transductive Parameter Transfer approach are shown in Algorithm 2 and each step is explained in detail below.

**ALGORITHM 1:** Transductive Parameter Transfer approach [31]

**Input:** Source sets $D_1^s, ..., D_N^s$ with labels and the target set $X^t$
**Output:** $\mathbf{w_t}, c_t$
Compute $\{\theta_i = (w_i, c_i)\}_{i=1}^N$ using (1).
Create training set $\tau = \{X_i^s, \theta_i\}_{i=1}^N$.
Compute the kernel matrix $K$ where $K_{ij} = \kappa(X_i^s, X_j^s)$ using (8).
Given $K$ and $\tau$, compute $\hat{f}(.)$ solving (6).
Compute $(w_t, c_t) = \hat{f}(X^t)$ with (7).

### 2.4.1. Obtaining personalized hyperplane parameters

First, person specific classifiers are trained on each source dataset individually to obtain the best performing parameter set $\theta$. Instead of a Linear SVM used in [8, 31], we have selected the well known binary class L2 penalized logistic regression classifier which minimizes Equation (2.1). Since both are linear classifiers and the format of the resulting parameters is similar, this selection does not require any extra steps.

$$\min_{(w,c)} \frac{1}{2} w^T w + C \sum_{i=1}^{n} log(exp(-y_i(X_i^T w + c)) + 1), \tag{2.1}$$

We have used Stochastic Average Gradient descent [32] to solve this optimization problem, obtaining the optimal parameter sets $\{\theta_i = (w_i, c_i)\}_{i=1}^N$ for each subject.[1] The optimal regularization parameter C is found through k-fold cross validation and the model is trained on the complete dataset of the participant with this C value.

### 2.4.2. Mapping from distributions to hyperplane parameters

The second step aims to learn the relation between the marginal distributions $P_i^s$ and the parameter vectors $\theta_i$. The assumption here is that for each participant, the hyperplane whose parameters are defined by $\theta_i$ are dependent on the underlying distribution $P_i$. By learning this relation, the optimal hyperplane parameters for the target dataset can be computed without any label information. The actual underlying distributions are not known, neither for the source datasets $\mathbf{P_i^s}$ nor the target $P^t$, however they can be approximated using the samples $X_i^s$ and $X^t$. Thus, the method aims to learn a mapping from samples to the parameters, $\hat{f} : 2^x \rightarrow \theta$, using the training set $\tau = \{X_i^s, \theta_i\}_{i=1}^N$, formed after the first step of the algorithm.

[1] $w$ and $c$ corresponds to regression coefficients and the intercept, respectively.

Since we assume that elements in $\theta$ are correlated, we employ Kernel Ridge Regression(KRR), instead of the multiple, independent regressors proposed in [8]. The primal problem for ridge regression is defined as follows [33]:

$$min((y - Xw)^T(y - Xw) + \|w\|^2) \tag{2.2}$$

where the optimal solution is given as:

$$w = (X^T X + \lambda I_D)^{-1} + X^T y = (\sum_i x_i x_i^T + \lambda I_D)^{-1} X^T y \tag{2.3}$$

The formulation for ridge regression can be kernelized with the following steps. First, Equation (3) is rewritten as

$$w = X^T (XX^T + \lambda I_N)^{-1} y \tag{2.4}$$

Term $XX^T$ in Equation (4) can be directly replaced with the Gram Matrix $K$, partially kernelizing the equation. In order to eliminate term $X^T$ and completely kernelize the formulation of ridge regression, following dual variables are introduced:

$$\alpha \equiv (K + \lambda I_N)^{-1} y \tag{2.5}$$

With the introduction of dual variables, Equation (4) becomes

$$w = X^T \alpha = \sum_i^N \alpha_i x_i \tag{2.6}$$

After solving for $w$, the solution for any variable x can be found as:

$$\hat{f}(x) = w^T x = \sum_i^N \alpha_i x_i^T x = \sum_i^N \alpha_i \kappa(x, x_i) \tag{2.7}$$

It can be seen from Equations (5) and (7), a kernel $\kappa$ that can define the similarities between two distributions is needed. Instead of the density estimate kernel defined in [8], we have selected an Earth Mover's Distance[34] based kernel which is discussed in [31]. In our implementation, each sample is treated to be a signature where all samples have uniform weights. The EMD kernel is defined as

$$\kappa_{EMD} = e^{-\gamma EMD(X_i, X_j)} \tag{2.8}$$

where $EMD(X_i, X_j)$ corresponds to the EMD between two datasets $X_i$ and $X_j$, the minimum cost needed to transform one into another. $\gamma$, a user defined parameter, is set to be the average distance between all possible pairs of datasets and experimentally shown to perform well.

### 2.4.3. Classification

By solving (6) for the source datasets, we learn the mapping $\hat{f} : 2^x \rightarrow \theta$. For any new target dataset, we can compute the parameter vector $\theta_t$ by plugging $X^t$ into the mapping function $\hat{f}$. Classification of the samples in the target dataset is then obtained by $y = sign(w_t x + c_t)$.



## 2.5. Dataset & feature extraction

### 2.5.1. Dataset

We recorded data in a real pub with 16 male and 16 female volunteers during a speed dating social event. The first phase involved having three-minute dates with each member of the opposite sex. After this, participants could get to know each other better in a mingling session. This phase has the characteristics of a crowded mingling scenario which we needed for our experiments. All throughout the event, participants wore a specialised sensor pack around their necks which collects acceleration and proximity information. The accelerometer in the sensor pack provides 20 samples per second. In our experiments, we only used accelerometer data. The area was fitted with multiple video cameras facing down on the scene, covering all the area participants were present. The video footage was used for labelling the ground truth.

### 2.5.2. Annotations & features

#### Annotation procedure

In this study, we will be focusing on the mingling phase. The mingling session lasted for approximately an hour. Due to hardware malfunctions, only 28 of the sensor packs recorded data in this session. Although we would have preferred to use all the data we have for the classification experiments, the annotation of social actions (in our case, speech) is extremely time consuming and costly. Also, some of the participants were at the blind spots of our cameras for the majority of the event, making robust annotation of their data extremely challenging. These factors forced us to use a subset of 18 participants for our experiments. This is in keeping with the numbers of test subjects typically used for studies in activity recognition, where datasets of varying sizes from 1 to 24 participants are reported [2, 9].

Thus, speaking status for these 18 participants were carefully labelled using the video for 10 minutes of the mingling phase with a time resolution of one twentieth of a second. A qualitative inspection revealed a rich dataset including participants with differing levels of expressiveness, interacting in dyads, larger groups or hardly interacting with someone at all, covering different types of personal characteristics and interactions possible in such an event. Detailed inspection of the annotations also showed that the speaking turn lengths per person vary greatly, from few seconds to more than half a minute, further showing the variety captured in the dataset.

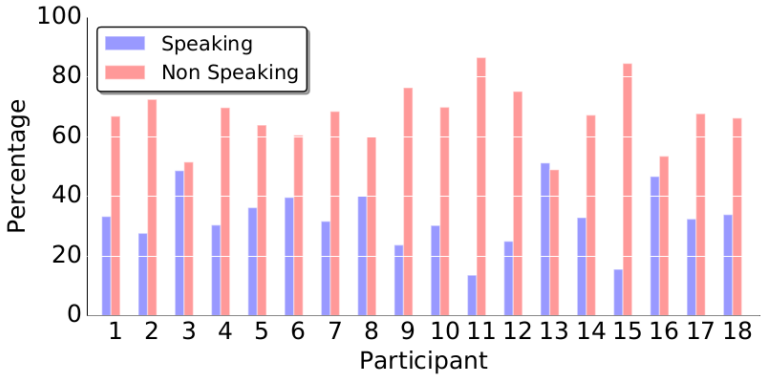


Figure 2.3: Percentages of speaking-non speaking samples

### Feature extraction

Before feature extraction, each axis of the acceleration input is standardised to have zero mean and unit variance. We selected our features from the literature and ensuring that were as simple as possible so as to avoid overfitting the data of the participants. The selected features can be grouped into two categories; statistical and spectral. As our statistical features, we calculated mean and variance values. As the spectral features, the power spectral density (PSD) was computed in the same way as [2], using 8 bins with logarithmic spacing from 0-8 Hz. These were extracted from 3s windows with one third overlap for each axis of the raw acceleration, absolute value of the acceleration, and magnitude of the acceleration. The length of the window was selected to be big enough to capture the speaking action while preserving a fine temporal resolution. All features were concatenated to obtain a 70-dimensional feature vector per window.

### Dataset analysis

Using the annotations and acceleration from this 10 minute interval, we have extracted features for each participant. This resulted in the total of 18 feature vectors, each having 299 samples with 70 dimensions, with varying class distributions. The class distributions for each participant are shown in Figure 2.3. The mean percentage of the positive samples (speech) across all participants is found to be 33, with a standard deviation of 10%. Participant 11 had the least number of positive samples (14%) whereas, person 13 had the highest percentage (51%). This imbalance in class distribution, which is also person specific, introduces a new difficulty for robust detection of speech.

In order to see how person specific nature of speech affects the distribution of samples in feature space, we have applied dimensionality reduction to samples of four participants and plotted them for the first two principal components. To standardize the plots, samples from the four participants were collectively normalized with z-score standardization. We can see from Figure 2.4 even after preprocessing, distributions are close to each other in the feature space, while the distribution

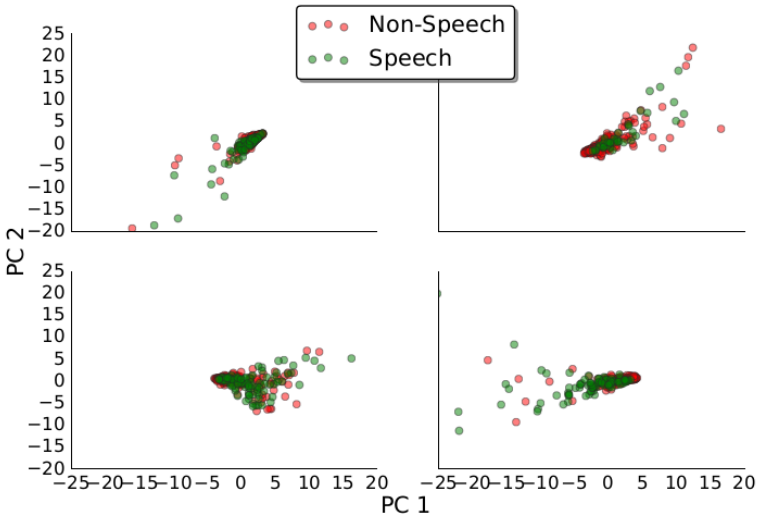


Figure 2.4: First two principal components of four participants (18,17,10,11)

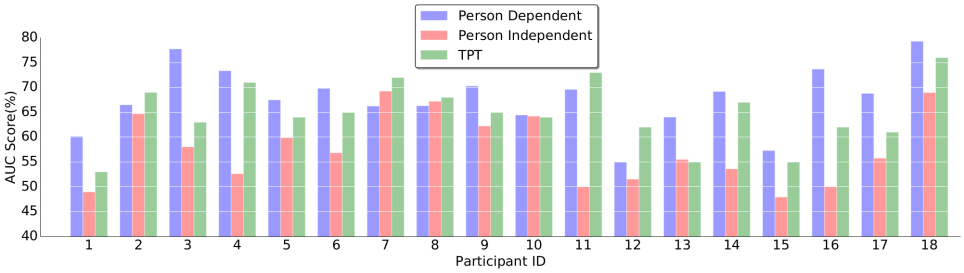


Figure 2.5: Performance in terms of AUC for speech detection. Person dependent setup uses data from the same participant for training and testing and expected to act as an upper bound for the performance. Person independent and TPT setups use data from other participants in a leave-one-subject-out manner.

of samples and the characteristics of the data still vary greatly between different participants.

## 2.6. Experimental results

In this section, we will discuss and compare the performance obtained with different classification setups and approaches. When presenting classification performance, we have specifically selected Area Under Curve (AUC) since it provides a more valid performance estimate in the presence of our imbalanced binary classification problem. Also, while training any classifier, class weights are set to be inversely proportional to the number of samples in the class so as to remove any bias caused by imbalanced class sizes. Of the all setups discussed in this section, only person dependent one uses the data from a single participant, for training and testing, in a leave-one-sample out manner. Other setups, person independent and TPT, use

data from other participants. Thus, person dependent setup is expected to act as an upper bound on the performance since it is a personalised setting by nature.



### 2.6.1. Person dependent performance

In the person dependent setup, each participant is trained and tested on their own data. Since we don't have enough data to come up with distinct training and test sets, we applied Leave-One-Sample-Out cross validation scheme for performance evaluation.[2] Based on the findings reported in [35], we made sure that training set is not contaminated. This means for each fold, any adjacent samples to the test sample are eliminated from the training set. With this elimination, we aim to provide an unbiased performance estimate. We have used a logistic regressor as classifier where the optimal regularization parameter C in Equation (1) is found by nested k-fold cross validation.

The procedure is applied to each participants' data separately, obtaining performance evaluations for each. This resulted in varying performance scores, ranging from an AUC score of 55% to 79%. The mean performance across all participants is $68\% \pm 6$. Individual scores for each participant are shown in Figure 2.5.

The variation in performance scores can be linked to two different factors we have already discussed. The first is the personal connection between speech and body movements read through the accelerometer. As expected, the problem becomes harder for people with more subtle movements, resulting in lower performance. Still, each participants' performance score is higher than random (50% AUC), proving that our features are still discriminative.[3]

Second factor is related to the class distributions. As shown in Figure 5, some participants' class distributions are highly skewed towards the negative class. We can not say that such imbalance always guarantees low performance, since it may still be possible to train robust models from small numbers of highly informative samples. However, we already see negative effects of this imbalance in our results. The participants with the lowest performance scores have small number of positive samples. There are only two participants with AUC scores lower than 60% (P12: 55% and P15: 57%) and they have the second and third lowest percentages of positive samples (25% and 16%, respectively) in the whole dataset. So, for these two participants, we can not be sure if the low performance is caused by subtle movement while speaking or the small number of positive samples.

We expect these results to act as an upper-unbiased limit for speech detection performance.

### 2.6.2. Person independent performance

In the person independent setup, we have used Leave-One-Subject-Out cross validation for performance evaluation, where each participants' samples are classified with the model obtained from other participants' data. So, the training set is formed by concatenating and standardising all other participants data. Similar to the

[2] In LOSO-CV, classifier is trained on n-3 samples and tested on one in each fold.

[3] A random classifier gives 50% AUC in expectation regardless of the class balance [36]

person dependent setup, logistic regressor is used as the classifier and optimal regularization is then found on the training set with cross validation.

With this setup, we obtained an average AUC score of 58%, with a standard deviation of 7%. The individual scores for participants varied from 45 to 60%. The individual scores obtained with the person independent setup are also shown in Figure 2.5, together with the results of other setups. Apart from two participants (7 and 8), where the person independent setup yielded slightly better AUC scores than the dependent one, the person dependent setup always outperforms the independent setup. We compared the performances of person dependent and independent setups per person using a paired one-tailed t-test. As expected, the result of the t-test showed that the person dependent setup yields significantly better results than the independent one ($p < 0.01$).

In the ideal learning paradigm, training with more samples should yield a better, more robust model, contradicting what we see. However, it is also assumed that the samples in the dataset are coming from the same independent and identically distributed (i.i.d.) probability distribution. From what we see from Figures 2.3 and 2.4, it is more likely that every participant has their own probability distribution that their samples are drawn from. Thus, concatenating the data of all participants and training a model on this dataset results in an unreasonable and impractical decision boundary. These person independent results strengthen our claim of the personal nature of connection between speech and body movements and motivate the requirement of an adaptive model.

### 2.6.3. Transductive parameter transfer performance

Our TPT experiments also employed a Leave-One-Subject-Out setup, where each participant is treated to be the target dataset while all other participants acted as source sets. This setup is similar to the person independent one, since the labels of only other participants are used for classification. With TPT, an average AUC of $65\% \pm 6$ is obtained. Individual performance values are included in the Figure 2.5, in addition to those of the person dependent and independent setups.

It is clearly seen that TPT outperformed the person independent setup for majority of the participants (16 out of 18), providing an AUC score close to the person dependent setup. One-tailed t-test between the TPT and the person independent scores showed that TPT is significantly better than the other ($p < 0.01$). For few cases, TPT even outperforms the person dependent setup (participants 2, 7, 8, 11), however, the person dependent results are still significantly better than TPT ($p < 0.02$). This result is quite interesting and might be caused by different factors. When the performance for participants 7 and 8 are inspected, it can be seen that even the person independent setup outperforms that of the person dependent one. This suggests that for these participants, using more data (even belonging to other participants) provides a better estimation of the decision boundary. In such a case, we may expect TPT to outperform all other setups. Although the same pattern is not present for participants 2 and 11, we might still argue that these participants benefited from the use of the data of other participants, most probably the ones having a similar distribution.

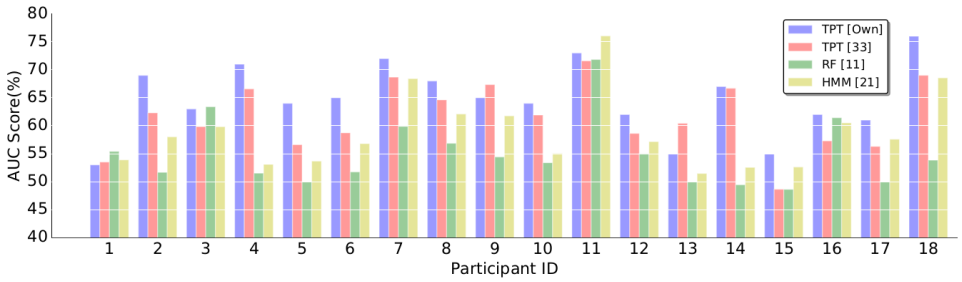


Figure 2.6: Comparison with the state-of-the-art as presented in [11] (RF and HMM) & [33] (TPT)

These results prove that it is still possible to generalise over unseen data, with an acceptable performance, if an adaptive method like TPT is employed. In 10 minutes one might argue that there is relatively little variation in an individuals' behaviour. However, assuming that between-person variation remains fairly high over this interval, as it can be seen from Figures 2 and 3, it is particularly interesting that we get good results, showing the robust generalisation ability of our method even with a limited amount of data. With the proposed transfer learning approach, performance results that are always better than the random baseline are obtained and statistical significance tests showed that our proposed method guarantees to perform better than traditional non-adaptive person independent learning.

### 2.6.4. Comparison with the state-of-the-art

This section compares the performance of our Transductive Parameter implementation with the state-of-the-art approaches. Firstly, we present the person independent results obtained with Random Forests (RF) and Hidden Markov Model (HMM) based approaches proposed in [1]. Secondly, we present the results obtained with the TPT implementation given in [8] and discuss in detail how our different choices affected the final performance. Individual performance scores obtained with all four methods, including ours, can be seen in Figure 2.6.

#### Non-adaptive person independent methods

We have implemented the methods presented in [1]. We have used the exact same setup they defined which includes the features they used (PSD 0-8Hz), window sizes for feature extraction (5s for RF, 3.5s for HMM), number of trees in Random Forest classifier (500) and number of states in HMM (2). We compare with the Leave-One-Subject-Out cross validation setup reported in [1].

With the RF, we obtained an average AUC score of $55\% \pm 6$. The HMM performed slightly better, providing an average AUC of $59\% \pm 6$. When compared to our person independent results obtained with logistic regression, neither RF nor HMM provided a significantly better result. This is an interesting finding since it shows that a linear model is as powerful as a nonlinear model for the speech detection problem, in a Leave-One-Subject-Out setup. Our proposed TPT method, on the other hand, significantly outperforms both of these methods. There are only 3 participants that have better performance scores than our proposed implementation of TPT;

Table 2.1: Performance and significance of the four modified TPT implementations compared to ours, which had an average AUC score of $65\% \pm 6$ (**(p<0.01),*(p<0.05)).

| | Modification (Our implementation) | | | |
|---|---|---|---|---|
| AUC $\pm$ Std | SVM (LR) | SVRs (KRR) | DK (EMD) | SV (WD) |
| | $60 \pm 4$ ** | $63 \pm 7$ * | $65 \pm 7$ | $61 \pm 5$** |



participants 1 and 3 for RF and participants 1 and 11 for HMM. One tailed t-tests between our TPT results and both RF and HMM showed TPT performs significantly better ($p < 0.01$ for both RF and HMM). The authors of [1] applied their non-adaptive method on a limited dataset that includes only 9 people. We believe, with the increasing number of participants, the person specific nature of speech is magnified and the requirement for adaptive methods increases.

**Detailed comparison with state-of-the-art TPT implementation**

Our proposed TPT implementation improves upon that presented in [8]. Although the basic framework of the method remains, our implementation choices made the method more suitable to the nature of our problem, as demonstrated by the performance results. We have used the implementation provided by [8] and obtained performance results with that setup, resulting in an AUC of $62\% \pm 6$. Our implementation outperforms it for 15 out of 18 participants.The paired one-tailed t-test between performance scores shows that our implementation is significantly better than [8] ($p < 0.01$).

There are four main differences between our implementation and the one in [8]. TPT implementation in [8] uses: (i) a SVM instead of logistic regression (LR), (ii) independent Support Vector Regressors (SVRs) instead of KRR, (iii) a density kernel (DK) instead of EMD kernel, (iv) support vectors (SV) instead of the whole data (WD) to estimate distributions of source sets. To investigate which modification affected the performance most, we carried out four follow-up experiments. In these experiments, we replaced one of our choices with the original one in [8]. Table 2.1 shows the average AUC and standard deviation over all participants obtained with each of these modifications. One tailed t-tests were used to quantify differences between our full implementation and one of the modified approaches.

Table 2.1 shows that the most effective change uses a logistic regressor instead of a Linear SVM. The two setups where our logistic regressor is replaced by a SVM (SVM and SV in Table 2.1) have the lowest performances. It is an unexpected result since the two classifiers are quite similar. However, the logistic regressor was more successful than the Linear SVM when person specific classifiers were being trained which we believe resulted in this performance difference.

Since our features are often correlated with each other, we preferred to use a KRR instead of the SVRs which is also supported by [31]. The performances shown in Table 2.1 backs our decision since our method with KRR performed significantly better than the SVRs method. The average performance difference between two methods could be low but our method provides significantly better results.

Finally, we can see that replacing EMD with a density kernel (DK) does not affect

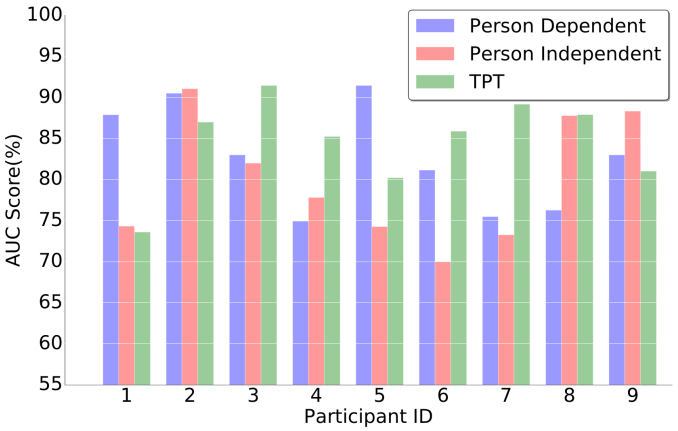


Figure 2.7: Performance in terms of AUC for walking

the performance at all. For our data, a density kernel was as successful as the EMD kernel in estimating the similarities of distributions. This is quite different than the findings in [31] but we believe it is related to the distribution characteristics of our data.

## 2.7. Comparing speech detection with walking

We investigated how different the nature of the speech detection problem was compared to other more traditional actions in the action detection literature to see if speech detection from body motion really requires a different approach. To address this question, we have conducted a follow-up experiment where we compared the speech detection results to an action which is widely studied in the action detection literature, walking.

Here, we used the same setup from our speech detection experiments. Similar to the former section, we obtain two performance scores for each participant; one for each of the person dependent and independent setups. We used a subset of the participants from 9 people who had enough walking samples. In order to obtain an acceptable number of samples, we only included participants that continuously walked more than 3 seconds with at least 15 seconds total walking time. To make the problem similar to our speech detection experiments, we added a random number of non-walking samples to each participant, creating possibly imbalanced distributions. The performances for this experiment are shown in Figure 2.7.

The person dependent setup yielded an average AUC of $83\% \pm 6$. With the person independent setup we have obtained an average AUC of $80\% \pm 7$. We have also applied TPT to the walking data with the same leave-one-participant-out setup of the former experiments where data from other participants acted as sources for the transfer. TPT obtained an average AUC of $84\% \pm 7$. The pairwise t-tests between setups showed that no single setup is significantly better than the others and all might provide better performance for an unseen participant. From Figure 2.7,

we can see that the pattern here is entirely different than the speech detection one. First, both person independent and person dependent setups yielded relatively high performances, when compared to performances of the speech detection experiments reported in Section 2.6 (average AUC of $68\% \pm 6$ versus $83\% \pm 6$ for the person dependent setup and $58\% \pm 7$ versus $80\% \pm 7$ for the person independent one). This is an expected result, since the connection between speech and body movements are not as universally characterizable as the connection between walking and body movements. Secondly, in many cases, better performances than the person dependent setup are actually obtained with the independent one.

Interestingly, the best overall performance score is obtained with the TPT, resulting in an average score slightly higher than the person dependent one. This is definitely different than the speech detection problem where the person dependent setup and TPT performed significantly better than the person independent one. We can still argue that the relatively smaller sample sizes compared to the speech detection experiments might have caused the person dependent setup to perform sub-optimally, explaining the cases where person independent and TPT setups outperformed the dependent one. Yet, these experimental results show that the detection of walking is less challenging, is not influenced by personal differences as much as speech-related body movements and it is still possible to achieve high performance with a non-adaptive model, unlike our speech detection task. In addition, high performances obtained with the TPT, even for a problem that seemed to be less person specific, show that the proposed method is quite robust and still preferable to the traditional person-independent setup in such cases.

## 2.8. Comparing controlled & in-the-wild settings

To experimentally demonstrate the restrictions introduced by a real event, we organised a small controlled experiment where one participant imitated speaking, walking and standing in a structured way while wearing an accelerometer. The participant alternated between actions where each action is performed for at least 15 seconds, resulting in a dataset that has 125, 139 and 110 seconds of standing, speaking and walking, respectively. The participant did not exaggerate any action to make them distinguishable from others. It should be noted that, the standing parts also include the imitation of listening, where head-hand gestures and body shifts natural to listening were randomly acted by the participant.

We have used the same experiment setup of the person dependent experiments discussed in Section 2.6. Thus, the logistic regressor is used as the classifier, the same set of features and Leave-One-Sample-Out evaluation scheme is utilised. Even though we had three different classes, we treated the problem as a binary classification task, where the samples corresponding to walking and standing formed the negative class. This results in roughly one third of the samples being positive.

Using this controlled data, we achieved an AUC score of 84%. More detailed analysis shows that 4% of walking samples and one third of standing samples

are misclassified as speech. This is consistent with former experiments, showing that distinguishing between speech and walking is relatively easy. On the other hand, listening-standing is often confused, probably because similar gestures occur in both. Still, the majority of standing samples are classified correctly. Also, the trained model is quite robust in detecting speech, only misclassifying 8% of speech samples as non-speech.

The performance score obtained in a controlled environment outperforms all our previous experiments with real in-the-wild data. We believe this is related to the two main differences between the setups. First, in the controlled environment we have precise annotations for each action. The noise introduced in the annotation procedure tends to affect the learning procedure. Even not guaranteed, since we don't have a robust way of measuring the quality of the annotations we have for the real life event; better annotations may increase performance. However, the annotation quality may also not be related to the essence of the difference between the real life and controlled events.

Secondly, the actions performed by the participant in the controlled experiment is highly structured and limited. However, the actions of participants in a real life event is completely unstructured. There is no limit to the type of actions they may perform and the transitions between them. Participants can even perform multiple actions at the same time. It is nearly impossible to cover all the possibilities that may happen in a real life event in a controlled environment. So, we believe that the results obtained from controlled experiments will be always positively biased and would not reflect the true phenomena as it occurs in the wild.

## 2.9. Analysis of transfer source quality

While using TPT, we employed a Leave-One-Subject-Out learning scheme where data of all other participants acted as sources. Some source sets might be more informative than others. Conversely, some source sets may negatively affect the mapping function learned, dropping the final performance. Thus, we hypothesised that there might be optimal source subsets for each participant. To check this hypothesis, we classified each participant with every possible triad of source sets. Then, we selected the top 10 best performing triads for each participant. We should note that, all of these setups were somewhat optimal, performing better than the setup where all sources were used.

Figure 2.8 visualises links between the best performing subset where the size of each node indicates the number of times it was in one of the best performing source sets. A directed edge from node A to B (where the end of the edge is slightly wider) means that participant B was at least in one of participant A's best performing source sets. The width of the edges are proportional to the number of times B was in A's source sets.

From the Figure 2.8, we can see that participants 3, 4, 8 and 13 are the optimal sources for the majority of others. Still, the directed edges show that there is no single perfect source for everyone, meaning multiple sources are needed to cover a larger population. When we inspected the person dependent performances and class distributions for these participants, we did not see any distinguishing features

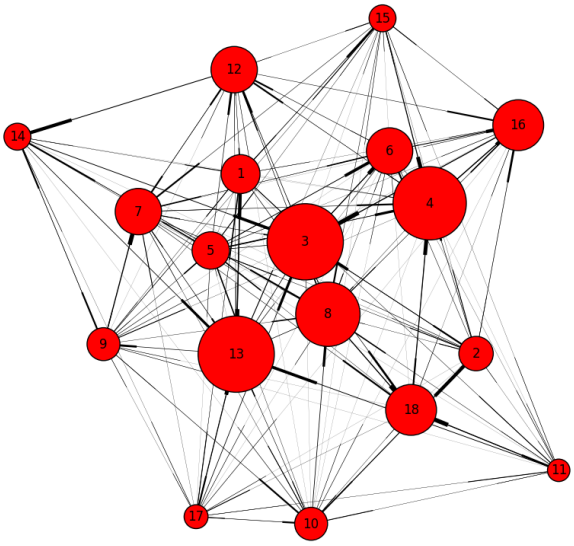


Figure 2.8: Visualisation of optimal source sets for each person.

to indicate their quality as sources. Closer inspection of the video of the event confirmed no spatial connection or presence of interaction was necessary for one participant to act as a good source for another. We believe these findings show that the success of these participants as sources comes from something more inherent, most probably related to connection between speech and torso movements.

We analysed whether being a good source might be related to personality. Each participant filled in the HEXACO personality inventory [37] before the event. The HEXACO scale measures personality in 6 dimensions and can broadly be considered similar to the more well-known Big Five personality traits except with an additional sixth dimension measuring humility or honesty. The dimensions are mapped onto a 5-point likert scale. We observed that all these four participants have relatively high extraversion (3.8, 3.6, 3.9, 3.6) and openness (4.1, 3.6, 4.2, 3.4) scores which may contribute to them being good sources. Further analysis of the connection between personality and transfer is left for future work.

## 2.10. Analysis of gender differences in transfer

One interesting aspect we haven't investigated in the former sections is how gender specific attributes affect the proposed method. In all of the former TPT experiments, we either used all remaining participants as sources or fetched all possible triads without considering the gender of the participants. In a traditional speech detection setup, where audio recordings are used as input, gender is expected to be a distinctive feature because of the frequency differences in male and female voices. In this section, we present a detailed analysis to see if such a difference exist for our method which relies on accelerometer readings instead of sound. Luckily, we have

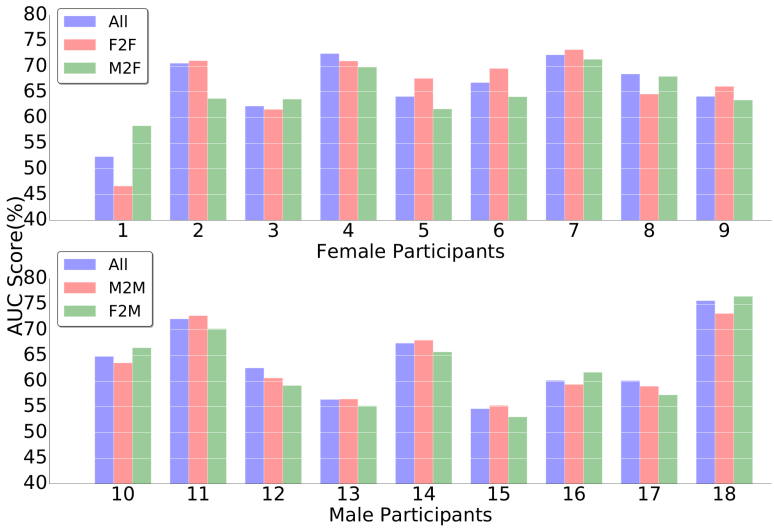


Figure 2.9: Performance scores of TPT for gender based transfer (Participant IDs are same with the ones shown in previous figures)

a balanced dataset in terms of gender, 9 females (Participant IDs 1-9 in Figure 5 and 6) and 9 males (Participant IDs 10-18 in Figure 5 and 6).

In order to check if there are any gender specific characteristics affecting our method, we devised three different experiment setups. The first setup is entirely the same as the one presented in Section 2.6.3, where we use all other participants as sources. For the second setup, we only use participants as sources who are the same sex with the target participant. The last setup is the reverse of the second one where all sources are the opposite sex of the target participant. Figure 2.9 shows the performances for all these three setups applied on male and female participants. The items in the legend correspond to all three setups where All corresponds to setup 1, F2F (female sources, female targets) and M2M (male sources, male targets) are setup 2 (same gender transfer) and M2F (male sources, female targets) and F2M (female sources, male targets) are setup 3 (transfer from the opposite gender).

As it can be seen from the Figure 2.9, there seems to be no significant difference between any of these setups. When we use the all participants as sources, the average AUC scores for female and males are $66\% \pm 6$ and $64\% \pm 6$, respectively. For the same gender transfer, F2F and M2M setups, the average AUC scores are $66\% \pm 8$ and $63\% \pm 6$. Finally, for the opposite sex transfer, M2F and F2M, we have obtained average AUC scores of $65\% \pm 4$ and $63\% \pm 7$, respectively. The individual performances of participants seem to be slightly changing with respect to the setups, however, there is no apparent pattern suggesting a convincing effect of gender on the transfer quality. This is further proved by the t-tests between all different pairs of setups that showed no significant difference.

These results are somehow expected since we are not trying to infer speech from vibrations in the chest which might be strongly affected by the frequency differences

of sound between genders. Our method is based on the connection between body (mostly torso) movements and speech which is expected to be affected less by any gender specific differences. These results are also on par with the analysis of the last section where we identified optimal sources for transfer. Three out of four optimal sources were found to be females in this analysis, however, they were good sources for participants from all genders. Thus, we can conclude that even though there might be some gender specific gesturing, we haven't seen any strong effects of it on the success of transfer in our data.



## 2.11. Conclusion and future work

In this study, we presented a transfer learning approach for detecting speech in real world crowded environments, using accelerometers. By comparing speech detection task to a traditional action recognition problem (e.g walking), we have shown the requirement for a specialised approach that can address the person specific nature of the speech and body movements. As a novel contribution, for the first time, Transductive Parameter Transfer [8] was used to address the person specific patterns of estimating speech from body acceleration. We also analysed the parameter transfer in detail by considering different source sets, providing insights into the nature of transfer and the task of speech detection.

Results obtained with the proposed method outperformed the state-of-the-art, providing performance scores close to person dependent setups. We discussed the challenges that are introduced by a more ecologically valid setting when compared to controlled experiments and experimentally showed how they affected the detection performance. Analysis of transfer quality demonstrated that an optimal subset of sources could be identified for each target set. Moreover, we found that some participants generally acted as good sources for subsets of the population in our data. We observed that this connection was not related to the spatial distance or to their corresponding interaction partners but something more inherent in the individuals.

As future work, we plan to explore automated methods of selecting source sets for each target. Another direction we would like to pursue is testing our method in a different environment, for example, a seated scenario where different variety of actions can be examined.

## References

[1] H. Hung, G. Englebienne, and L. Cabrera Quiros, *Detecting conversing groups with a single worn accelerometer,* in *Proceedings of the 16th international conference on multimodal interaction* (ACM, 2014) pp. 84–91.

[2] H. Hung, G. Englebienne, and J. Kools, *Classifying social actions with a single accelerometer,* in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing* (ACM, 2013) pp. 207–210.

[3] D. McNeill, *Language and gesture*, Vol. 2 (Cambridge University Press, 2000).

[4] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder, *Bridging the gap between social animal and unsocial machine: A survey of social signal processing,* IEEE Transactions on Affective Computing **3**, 69 (2012).

[5] D. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, *Modeling dominance in group conversations using nonverbal activity cues,* Audio, Speech, and Language Processing, IEEE Transactions on **17**, 501 (2009).

[6] M. S. Mast, *Dominance as expressed and inferred through speaking time,* Human Communication Research **28**, 420 (2002).

[7] H. Hung and D. Gatica-Perez, *Estimating cohesion in small groups using audio-visual nonverbal behavior,* Multimedia, IEEE Transactions on **12**, 563 (2010).

[8] G. Zen, E. Sangineto, E. Ricci, and N. Sebe, *Unsupervised domain adaptation for personalized facial emotion recognition,* in *Proceedings of the 16th International Conference on Multimodal Interaction* (ACM, 2014) pp. 128–135.

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

[10] N. Ravi, N. Dandekar, P. Mysore, and M. Littman, *Activity recognition from accelerometer data,* AAAI , 1541 (2005).

[11] S. J. Preece, J. Y. Goulermas, L. P. Kenney, and D. Howard, *A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data,* IEEE Transactions on Biomedical Engineering **56**, 871 (2009).

[12] 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**, 277 (2006).

[13] C. Doukas, I. Maglogiannis, P. Tragas, D. Liapis, and G. Yovanof, *Patient fall detection using support vector machines,* in *IFIP International Conference on Artificial Intelligence Applications and Innovations* (Springer, 2007) pp. 147–156.

[14] D. Cook, K. D. Feuz, and N. C. Krishnan, *Transfer learning for activity recognition: A survey,* Knowledge and information systems **36**, 537 (2013).

[15] J. Yang, R. Yan, and A. G. Hauptmann, *Cross-domain video concept detection using adaptive svms,* in *Proceedings of the 15th international conference on Multimedia* (ACM, 2007) pp. 188–197.

[16] D. H. Hu, V. W. Zheng, and Q. Yang, *Cross-domain activity recognition via transfer learning,* Pervasive and Mobile Computing **7**, 344 (2011).

[17] Z. Zhao, Y. Chen, J. Liu, Z. Shen, and M. Liu, *Cross-people mobile-phone based activity recognition,* in *IJCAI*, Vol. 11 (Citeseer, 2011) pp. 2545–250.

[18] T. van Kasteren, G. Englebienne, and B. J. Kröse, *Transferring knowledge of activity recognition across sensor networks,* in *Pervasive computing* (Springer, 2010) pp. 283–300.



[19] T. Choudhury and A. Pentland, *Sensing and modeling human networks using the sociometer,* in *null* (IEEE, 2003) p. 216.

[20] 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) **39**, 43 (2009).

[21] D. Wyatt, *Collective Modeling of Human Social Behavior.* AAAI Spring Symposium: Human Behavior Modeling (2009).

[22] A. Madan, M. Cebrian, D. Lazer, and A. Pentland, *Social sensing for epidemiological behavior change,* Proceedings of the 12th … (2010).

[23] X. Alameda-Pineda, Y. Yan, E. Ricci, O. Lanz, and N. Sebe, *Analyzing free-standing conversational groups: A multimodal approach,* in *Proceedings of the 23rd ACM international conference on Multimedia* (ACM, 2015) pp. 5–14.

[24] C. Cattuto, W. Van den Broeck, A. Barrat, V. Colizza, J.-F. Pinton, and A. Vespignani, *Dynamics of person-to-person interactions from distributed rfid sensor networks,* PloS one **5**, e11596 (2010).

[25] C. Martella, E. Gedik, L. Cabrera-Quiros, G. Englebienne, and H. Hung, *How was it?: Exploiting smartphone sensing to measure implicit audience responses to live performances,* in *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference* (ACM, 2015) pp. 201–210.

[26] A. Matic, V. Osmani, and O. Mayora, *Speech activity detection using accelerometer,* in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE* (IEEE, 2012) pp. 2112–2115.

[27] A. Kendon, *Conducting interaction: Patterns of behavior in focused encounters*, Vol. 7 (CUP Archive, 1990).

[28] T. L. Chartrand and J. A. Bargh, *The chameleon effect: the perception–behavior link and social interaction.* Journal of personality and social psychology **76**, 893 (1999).

[29] D. Reynolds, *An overview of automatic speaker recognition,* in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, S 4072 (2002) p. 4075.

[30] J. Dines, J. Vepa, and T. Hain, *The segmentation of multi-channel meeting recordings for automatic speech recognition,* in *Int. Conf. on Spoken Language Processing (Interspeech ICSLP)*, LIDIAP-CONF-2006-007 (2006).

[31] E. Sangineto, G. Zen, E. Ricci, and N. Sebe, *We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer,* in *Proceedings of the ACM international conference on multimedia* (ACM, 2014) pp. 357–366.

[32] M. Schmidt, N. L. Roux, and F. Bach, *Minimizing finite sums with the stochastic average gradient,* arXiv preprint arXiv:1309.2388 (2013).

[33] K. P. Murphy, *Machine learning: a probabilistic perspective* (MIT press, 2012).

[34] Y. Rubner, C. Tomasi, and L. J. Guibas, *The earth mover's distance as a metric for image retrieval,* International journal of computer vision **40**, 99 (2000).

[35] N. Y. Hammerla and T. Plötz, *Let's (not) stick together: pairwise similarity biases cross-validation in activity recognition,* in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (ACM, 2015) pp. 1041–1051.

[36] T. Fawcett, *An introduction to roc analysis,* Pattern Recogn. Lett. **27**, 861 (2006).

[37] K. Lee and M. C. Ashton, *Psychometric properties of the hexaco personality inventory,* Multivariate Behavioral Research **39**, 329 (2004).

# 3

# Analysing social actions with a single body worn accelerometer

*Computing is not about computers any more. It is about living.*

Nicholas Negroponte

This chapter is to be published as: Section 7 of Chapter 11 (H. Hung, E. Gedik, and L. Cabrera-Quiros, Complex Conversational Scene Analysis Using Wearable Sensing), in the book **Multi-modal Behavior Analysis in the Wild: Advances and Challenges**, *Elsevier*, X. Alameda-Pineda, E. Ricci, N. Sebe, 2018. All contents of this chapter is authored by Ekin Gedik.

## 3.1. Introduction

In this chapter, we will present a case study on social behaviour analysis, focusing on automatic social action detection in complex conversational scenes. Various (social) actions will be discussed with respect to their physical manifestation and their connection to the worn sensing device, the required approaches, and available data size. In our case the sensing device we focus on is a single tri-axial accelerometer which is embedded in an ID badge hung around the neck.



When analysing human behaviour, past analysis has tended to assume that this is more or less person-independent. Throughout the text, 'person-independent' will be used for settings where data for training a model comes from different sources (people in our case) compared to the test data. Much work has been done on estimating daily activities such as walking or running from accelerometer data, showing promising results with a person independent setup [1, 2]. There is a direct connection between the sensing medium and the physical manifestation so that behaviours such as walking and stepping results in acceleration readings that are easy to discriminate directly from the magnitude of the signal. This makes making a person independent setup for discriminating such behaviour quite easy to implement.

However, some of the actions observed in crowded social settings tend to be much more person specific and the connection between the existence of these actions and the accelerometer readings is more ambiguous. In our case, the physical manifestation of speaking comes from vibration of the vocal chords, so unless the subject has a very sensitive accelerometer attached tightly to the body (e.g. the chest [3, 4]), there won't be a direct connection between the action and the sensing. However, speaking also has a physical non-verbal aspect, and it has been shown in previous works that the connection between body movements and speech can still be exploited for detecting if someone is speaking or not [5, 6]. Actions like speaking, which are loosely connected with the sensing medium, are expected to be harder to detect and may require specialised approaches.

To examine this, we conducted a number of experiments on a dataset that is collected from a real life, 'in the wild' event. The dataset is comprised of mingling events from 3 separate evenings where each evening includes data from approximately 32 people. Each participant wore a sensor hung around the neck that records individual triaxial acceleration at 20Hz. Note that the sample rate is not high enough to detect vocal chord vibration. However, it is high enough to capture body movements such as gestures. Different social actions are manually labelled by trained annotators for 30 minutes of the mingling sessions. For more information about the dataset, please refer to [7]. We have focused on the mingling session from the first day for the experiments presented in this chapter.

## 3.2. Feature extraction and classification

We have extracted features for each of the 26 subjects with valid accelerometer data. Statistical and spectral features are extracted from each axis of raw and absolute values of the acceleration and the magnitude of the acceleration, using 3s

Table 3.1: AUC scores for various actions

| | AUC(%) | Std(±) | Annotator Agreement |
|---|---|---|---|
| Stepping | 76.0 | 10.5 | 0.51 |
| Speaking | 69.5 | 8.3 | 0.55 |
| Hand Gestures | 70.4 | 9.1 | 0.61 |
| Head Gestures | 64.4 | 7.4 | 0.25 |
| Laughter | 67.8 | 12.5 | 0.39 |



windows with 1.5s overlap. As the statistical features, mean, and variance values are calculated. The spectral features consist of the power spectral density binned into 8 components with logarithmic spacing between 0-8 Hz.

We have used L2 penalized Logistic Regressor as the classifier. Performance evaluation is done with leave-one-subject-out cross-validation. Hyperparameter optimisation for regularisation is carried out with nested cross-validation. Stepping, speaking, hand and head gestures, and laughter are selected as the target actions. Since the class distributions for each participant are different, we have chosen the AUC (area under the ROC curve) as the performance metric. Performances obtained with the aforementioned setup is presented in Table 3.1. We also present the mean annotator agreement for each action using Fleiss'-Kappa for 3 annotators. Values higher than $0.4$ are considered to be of moderate agreement.

We can see that the results presented in Table 3.1 support the claim that actions that are loosely connected to the physical manifestation of the behaviour are harder to detect. Stepping, as expected, has the highest performance of all. We also see that performance tends to drop with the reducing connection between the physical manifestation of the action itself and the acceleration. For example, head gestures labels in the dataset, social action with the lowest detection rate, include many subtle nods which are harder to capture via acceleration, compared to a step or hand gesture.

It should be noted that there might be a second factor at play here. In real life events, it is generally harder to obtain annotations. Thus, the annotations must be made later manually.[1] This of course introduces some differences in annotator agreement which differs with respect to the type of the action. Table 3.1 shows the annotator agreements as reported on a subset of the data taken from [7]. It can be seen that the lowest annotator agreement values are for the head gestures, followed by laughter. Variation in agreement (due to behavioural ambiguity or visual occlusion of the person being annotated) in the labels might have also contributed to the low performance of these actions, in addition to the nature of the connection between the action and the sensing medium. Thus, noisy labels, at least for some actions, are a reality of data collection in wild which needs to be taken into account when evaluating the perception performance. A further discussion of the trade-offs between using crowd sourced annotations compared to on-site annotators are also discussed in [7].

[1] Our annotators labelled the social actions by watching a top-down video of the event.

## 3.3. Performance vs. sample size

In the former experiment, thirty minutes of data from each participant was used. The results obtained showed that thirty minutes was enough to capture a variety of actions with various different situational contexts (i.e. differing conversing partners with different levels of conversational involvement), obtaining acceptable performance even for more subtle actions. But what is the minimum required amount of data for acceptable performance? Will the patterns be similar if we had less data? Since it is not guaranteed to have a continuous stream of 30 minutes of data, we conducted another experiment, where we used the earlier setup but with gradually increasing amounts of data for each participant, starting from 5 samples to a total of 1198 that covers the whole 30 minutes. As mentioned in the former section, each sample is extracted with a sliding window of 3 seconds with 1.5 seconds shift. Thus, we can say that 5 samples corresponds to 9 seconds of data, 40 samples roughly correspond to one minute, and so on. We still used a leave-one-subject-out setup where for each fold, all the data from one participant corresponded to the test set. However, the training set is formed randomly by selecting $n$ samples from each of the other participant's data. Since the selection is random, the process is repeated $m$ times which was also dependent on the number of samples selected. For computational reasons, we gradually reduced the number of repetitions from 150 to 15 and from 5 to 1000 samples.We have selected two relatively well performing actions, stepping and speaking. These actions have different characteristics as described earlier with stepping being more closely connected to the physical manifestation of the behaviour compared to speaking, which relies on detecting bodily gestures that are related to speech. In addition, this selection is based on former studies that showed the connection between speech and acceleration is highly person specific compared to stepping-walking[6]. The mean of the AUC scores of all repetitions, with increasing data size, are shown in Figure 3.1 with standard deviation.

First, from Figure 3.1 we observe the higher standard deviation for smaller sample sizes. This is related to the decreasing number of repetitions but we argue that is not the only factor. We believe there are parts of the event that are less informative than others and if the selected samples are coming from such intervals the performance tends to be low, and therefore fails to generalise over the whole event. This issue will be discussed further later in this chapter where we will present results of an experiment where the samples are not randomly sampled but selected chronologically. We also also observe that the standard deviation for both actions converge to small values with the increasing sample size.

We can see from Figure 3.1 that the pattern for both actions are quite similar. Performances for the actions increase with a steep curve in the beginning and after 120 samples the increase gets smaller. This suggest that 3 minutes of data from each person is enough to cover the variations in each type of action in such an event. The question then becomes if it is possible to provide a specialised solution which can guarantee better results even if the number of samples is relatively low.

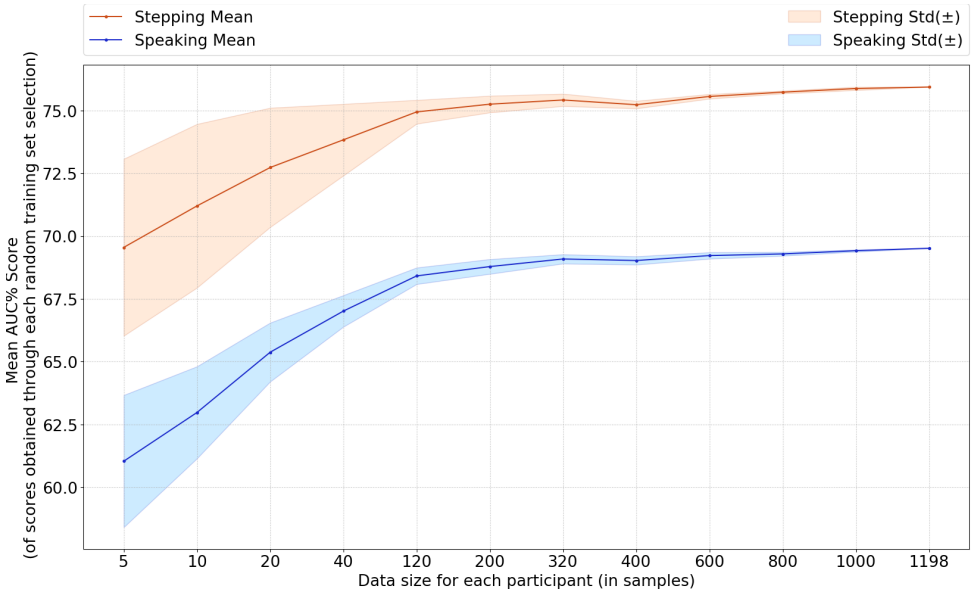


Figure 3.1: AUC scores of stepping and speaking with respect to data size

## 3.4. Transductive parameter transfer (TPT) for personalised models

Following on from the results of [6], where it was shown that a transfer learning approach that guarantees personalised models in a person independent setup tends to perform better for person specific actions, we repeated the former experiment with a personalised model. The method is named Transductive Parameter Transfer (TPT) and was first proposed for personalised facial expression recognition [8] and then modified for social action detection from a body worn accelerometer in [6].

TPT aims to find the parameters of the classifier for the target dataset $X^t$, without using any label information of $X^t$, by learning a mapping between the marginal distributions of the source datasets and the parameter vectors of their classifiers. $N$ source datasets with label information and the unlabelled target dataset are defined as $D_1^s, ..., D_N^s$, $D_i^s = \left\{x_j^s, y_j^s\right\}_{j=1}^{n_i^s}$ and $X^t = \{x_j^t\}_{j=1}^{n_t}$, respectively. The main steps of the TPT are shown below (for a detailed explanation, please refer to [6]):

1. Compute $\{\theta_i = (w_i, c_i)\}_{i=1}^N$ using L2 penalized Logistic Regression.
2. Create training set $\tau = \{X_i^s, \theta_i\}_{i=1}^N$.
3. Compute the kernel matrix $K$ that defines the distances between distributions where $K_{ij} = \kappa(X_i^s, X_j^s)$.
4. Given $K$ and $\tau$, compute $\hat{f}(.)$ with Kernel Ridge Regression.

5. Compute $(w_t, c_t) = \hat{f}(X^t)$ using the mapping obtained in former step.

We conducted the performance vs. sample size experiment explained in the former section, with the addition of TPT. TPT is also used in a person independent setup, where data from other participants are treated as source datasets with label information whereas the data to be classified is the target dataset. Although [6] suggests the use of an Earth Mover's Distance (EMD) kernel for computing the distance between distributions, we employed a Density Estimate kernel [8] since it is computationally less complex and more suitable for many random repetitions. The resulting AUC scores are plotted in Figure 3.2.

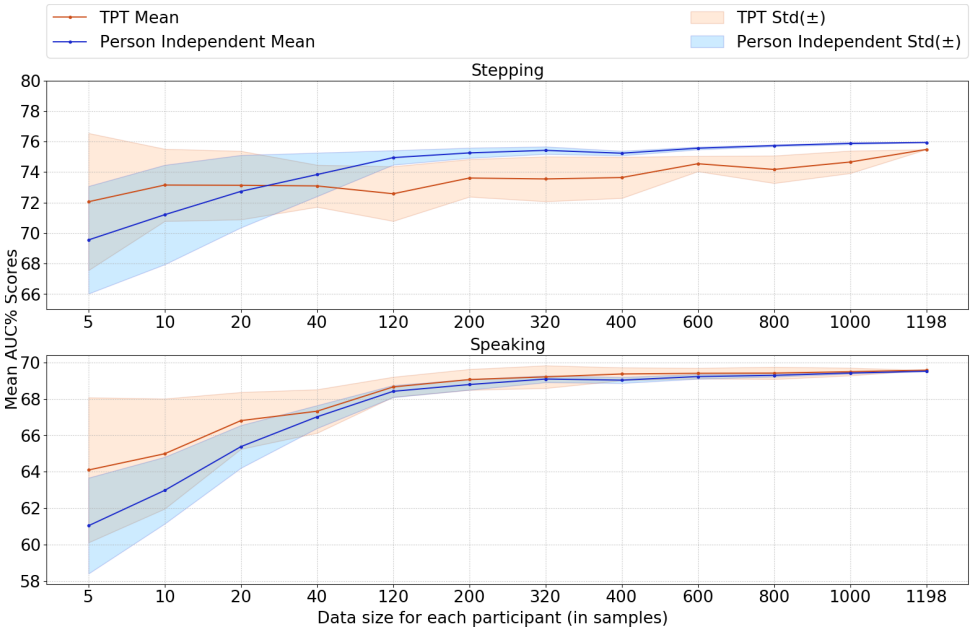


Figure 3.2: AUC scores of stepping and speaking with respect to data size

According to Figure 3.2, TPT outperforms a traditional person independent setup when using small sample sizes for both actions. It seems to generalise better even with a small amount of data. For speaking, with the increasing data size, the gap between the two methodologies starts to close, showing that the single logistic regressor in the person independent setup has seen enough diverse cases to generalise better. A one tailed paired t-test between AUC scores showed that up until 320 samples, TPT provides significantly better performance ($p < 0.05$ for 40 samples and $p < 0.01$ for the rest). After that point, the mean scores provided by TPT seemed to be still higher than the person independent setup but the significance is not guaranteed (some results such as those at 400 and 600 samples are still significant though). We can say that with the increasing data size, both methods converge to similar performances. However, especially for smaller sample sizes, we

can still conclude that for estimating an action in a person specific manner, TPT is more robust.

For stepping, the trend shown is different. For extremely small amounts of data of 5, 10 and 20 samples, TPT outperforms the traditional person independent setup (significantly for 5 and 10 samples). With increasing data sizes, the person independent setup clearly outperforms TPT. It can be argued that this is related to the nature of the action. Stepping is less person specific than speaking and the connection between the sensor and the physical manifestation of the action is more direct. Thus, it can be expected that the representations of such an action should not vary too much between participants. With the increasing number of samples, the person independent classifier will see more samples and since samples from different participants can be expected to be equally informative for all, a more optimal and general decision boundary can be obtained, unlike for speaking. So although we can advocate the use of TPT for really small sample sizes, a traditional person independent setup seems to be a more robust selection for less person specific actions.



Now, we want to go back to our claim that some parts of the event are more informative than others. The first parts of the dataset correspond to the beginning of the event, when groups are just starting to be formed. We might expect people to be less involved in the conversation as the discussions are not yet in full flow. This might result in samples that are not representative of all variations of actions that can occur in a real life event, throughout time. So, we did a follow up experiment where we compared the performances of TPT and the traditional person independent setup for speaking detection. However, this time for each participant in the training set, we increased the number of samples in chronological order. Thus, $n$ samples for a participant correspond to the first $n$ samples in time. Since there are no repetitions, the means and the standard deviations are computed on the individual performances of all participants. The results of this experiment are shown in Figure 3.3.

The first thing we observe from Figure 3.3 is how the performances of the person independent method is lower compared to those from Figure 3.2. Using random selection of the samples throughout the event, the person independent method was providing an AUC of roughly 61% for 5 samples. However, in the temporally increasing setup, the performance for the same number of samples is roughly 56%. The pattern is similar for the following sample sizes and the performance of temporally increasing selection is only able to reach the level of random selection if at least 320 samples are used for training. TPT on the other hand still provides similar results to the random selection method and provides relatively satisfactory results even with samples that were less informative for a traditional person independent approach.

One other interesting observation is the relatively high standard deviations for both methods, even with increasing number of samples. This shows that, for some participants, classifying the action is harder compared to others regardless of the sample size, further showing the person specific characteristics of speaking. These results further strengthen the claim that TPT should be considered for person spe-

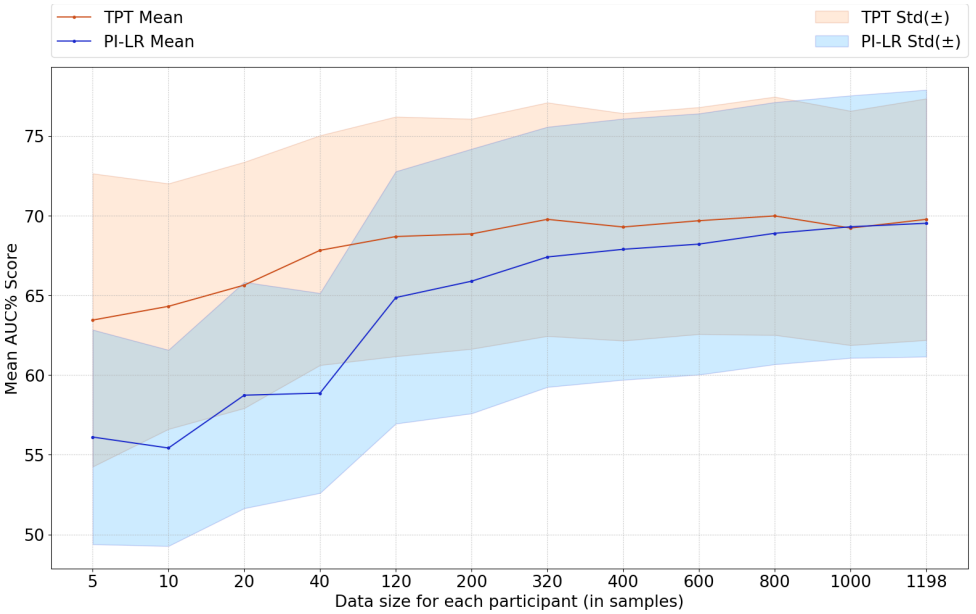


Figure 3.3: AUC scores of speaking with temporally increasing data size

cific and indirect actions such as speaking.

## 3.5. Discussion

With the presented perception analysis results, a few issues emerge that are all related to the 'in the wild' nature of the experiment. When collecting data from real life events, many challenges arise. Some of these restrictions and difficulties come from the unrestricted nature of the event: the variety and frequency of actions might cause some cases to be under or over-represented making detection harder. The difficulty of the annotation process (either due to the ambiguity of the behaviour or occlusion) can also result in label noise. Thus, when designing and conducting experiments on real life data, a researcher should always first consider how these issues will affect the machine perception problem to be solved.

Specifically, for the case study presented in this chapter, when focusing on the detection of actions through wearables, there are some important points to consider. First, one should understand the connection between the physical manifestation of the action, and the sensing medium they are using. This is required for the valid selection of features and models that will be used for classification. In real life scenarios, it is not guaranteed to have each action perfectly represented in all its possible variations for each participant. This is particularly true because natural 'in the wild' behaviour samples only come into being as the result of the dynamics of a conversation as it unfolds over time. That is, a monologue in a group would yield more positive examples of speaking for the speaker of the group but

no speaking samples for the members of the group who are just listening. So, the experimental setup and methodology chosen should encapsulate this together with the physical nature of the action. The experiments presented in this chapter are a good examples of this, where two approaches for the detection of two actions tend to perform differently, because of the physical nature of the actions in relation to the sample sizes.

## References



[1] L. Bao and S. S. Intille, *Activity recognition from user-annotated acceleration data,* in *International Conference on Pervasive Computing* (Springer, 2004) pp. 1–17.

[2] N. Ravi, N. Dandekar, P. Mysore, and M. Littman, *Activity recognition from accelerometer data,* AAAI , 1541 (2005).

[3] A. Matic, V. Osmani, and O. Mayora-Ibarra, *Mobile Monitoring of Formal and Informal Social Interactions at Workplace,* in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, UbiComp '14 Adjunct (ACM, 2014) p. 1035–1044.

[4] A. Matic, V. Osmani, A. Maxhuni, and O. Mayora, *Multi-modal mobile sensing of social interactions.* in *PervasiveHealth* (IEEE, 2012) p. 105–114.

[5] H. Hung, G. Englebienne, and J. Kools, *Classifying social actions with a single accelerometer,* in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing* (ACM, 2013) pp. 207–210.

[6] E. Gedik and H. Hung, *Personalised models for speech detection from body movements using transductive parameter transfer,* Personal and Ubiquitous Computing **21**, 723 (2017).

[7] L. Cabrera-Quiros, A. Demetriou, E. Gedik, L. van der Meij, and H. Hung, *The matchnmingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates,* IEEE Transactions on Affective Computing (2018).

[8] E. Sangineto, G. Zen, E. Ricci, and N. Sebe, *We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer,* in *Proceedings of the ACM international conference on multimedia* (ACM, 2014) pp. 357–366.

# 4

# Detecting conversing groups through social dynamics

*New knowledge is the most valuable commodity on earth. The more truth we have to work with, the richer we become.*

Kurt Vonnegut

This chapter will be published as:
E. Gedik and H. Hung, **Detecting Conversing Groups Using Social Dynamics from Wearable Acceleration: Group Size Awareness**, *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4), 2018.

## 4.1. Introduction

In most social scenarios, be it a small private gathering or a crowded music festival, people tend to interact with each other, forming groups of varying size. Automatic detection of such conversing groups has a wide variety of possible applications, ranging from surveillance to detailed analysis of socially relevant behaviour. For example, a deeper understanding of how people interact throughout an event can provide valuable information regarding the success of the event, such as the mood of the participants, their evaluations, whether they will return and recommend the event to others[1]. Such information is potentially important for organizers. Also, examples in the literature show that when an interacting group is identified, it is possible to estimate social attributes such as dominance [2], leadership[3] and cohesion[4] through behaviour of participants, which are especially valuable for organizational scenarios. In order to obtain such information, a deeper understanding of social interactions between people is definitely required, which we address in this paper.



This paper focuses on the automatic detection of conversing groups using a single accelerometer in real life crowded social scenarios, more specifically mingling events. Differing from the majority of the existing work that relies solely on the proxemics (spatial distance and relative orientation of the participants), our proposed method focuses on a widely overlooked rich information source: interaction dynamics; the coordinated behaviour of people during a conversation. The approach presented in this paper aims to represent interaction dynamics through people's actions and movement patterns which are inferred with a single body worn triaxial accelerometer. Importantly, our proposed approach considers how interaction patterns vary in different sized groups. This is achieved by training multiple classifiers with respect to the group cardinality and classifying a new sample by a meta-classifier which is trained with the probability outputs of the group based classifiers. This ensemble fusion technique is known as stacked generalization (stacking in short) in the literature [5]. Unlike the traditional stacking approaches, we use only the data from the local neighbourhood of the test sample while training the meta-classifier. Influenced by the findings in social psychology related to the speaker and listener behaviours in groups, we aim to provide a scalable and ubiquitous solution to the detection of conversing groups while preserving the privacy of the participants. Our method works solely on accelerometer data obtained by a custom sensor pack worn around the neck by participants like an ID badge. This makes the approach ubiquitous and viable for dense crowded scenarios, where other modalities such as video and sound may fail due to the crowded characteristics. Results are presented on a real life, in-the-wild mingling dataset, collected during three unique instances, showing the generalization of the proposed method. No recording of raw speech is done; participants' privacy is not invaded. Since only a single sensor is being used, our method is highly power-preserving, making it deployable for large and long scenarios.

We can list the novel contributions of this paper as follows: We (i) propose and utilise a new feature set (overlap statistics) together with some others that were previously used in other domains to capture interaction dynamics through social

behaviour, (ii) propose an approach that considers interaction dynamics of groups related to cardinality and show how it beats the state-of-the-art,(iii) perform various analyses to investigate the nature of this problem further: performance in relation to the group sizes, feature effectiveness, comparison of feature types in terms of performance and the effects of meta-level classifier on the local neighbourhood instead of the whole dataset. Following subsections aim to provide more insight into the problem and the proposed method by presenting a formal problem formulation and discussions related to the use of social dynamics instead of proxemics and the necessity for group size awareness.

### 4.1.1. Proxemics vs. dynamics

Explicitly, we are trying to detect pairwise F-formation membership; if two participants are in the same interacting group or not. We built our problem formulation on the definition of F-formations from the social psychologist Adam Kendon [6]. F-formations are specific types of focused encounter, where participants spatially and orientationally organize themselves into a group which makes it possible to facilitate conversation.



There are many examples of existing work in the literature that aim to F-formations. However, many of these works solely focus on the proxemics, either by working on the proximity information obtained with infrared (IR) sensors [7, 8] or by employing still images and videos for obtaining head and body orientations in addition to the spatial location [9–11]. This focus on proxemics is understandable and coherent with the definition of F-formation itself. If a person was asked to decide from a still image if a participant is part of a group, most probably they will first check the physical proximity and the orientation of the people. However, we argue that even though spatial distance and orientation are extremely strong cues, they overlook one important aspect of the interaction; the dynamic behaviour and actions of the participants.

Social scientists have already shown that interacting people tend to coordinate their movements [12] and even start to mimic each other in terms of posture, mannerism and other behaviours [13]. We believe such patterns can be used as informative cues for distinguishing between interacting and non-interacting partners. Thus, in this study we show how solely the dynamics related to the social behaviour of participants can be exploited to infer the conversational group membership of participants in real life scenarios.

Another reason to use the proposed method in this study is the practical limitations of inferring the spatial location and orientation of the people. In a real life scenario, images or video of the scene might not be always available. A method that relies on video input for conversing group detection will require a similar instrumentation in every event, making the solution less generalizable and non-ubiquitous. Also, characteristics of the event, for example the density of the crowd or location of the cameras can affect the performance of the method. Another reason is the privacy, not every participant will be comfortable with their videos being taken, since it is more intrusive than recording body acceleration. Another option is to use indoor localization solutions but they have been shown to perform poorly in a high

density scenario, where a commercial indoor GPS system was tested in [14].

A more pervasive solution than video based methods can be obtained with wearable sensors that provide proximity information but they also have weaknesses. Two technologies are generally used for inferring proximity with wearable sensors; infrared (IR) and bluetooth. A recent study that evaluated the use of wearable sensors in organizational settings investigated a custom sensor pack, sociometric badges [15], which uses these sensors for proximity detection [16]. Their experiments show that bluetooth had greater accuracy in detecting the co-location of participants. However, detections were generally overly optimistic, indicating proximity even when there were obstacles between the sensors such as a separating wall. IR, on the other hand, was more pessimistic in its detections, requiring clean line of sight and strict face-to-face orientation. In the types of crowded scenarios that we are interested in, a bluetooth based approach will most probably put many participants in the same group whereas an IR based method will tend to miss participants in larger groups since it requires strict face-to-face orientation, making those less suitable for precise group detection.



### 4.1.2. Group cardinality

We hypothesise that the dynamics of the interaction differ greatly with respect to the cardinality of the group. Previous studies in computer vision have already shown how a cardinality sensitive approach can improve performance [17]. We believe, the effect of group cardinality on the interpersonal dynamics is even greater. The assumption of one conversational flow per F-formation may not be always true. For large groups, sustaining a single informal conversation is not possible [18]. Even though multiple participants can be in the same F-formation of a larger cardinality, they might still form sub-groups inside, exhibiting behaviour of smaller cardinalities. For example, a four person F-formation can have many different interaction characteristics. It could be an egalitarian group where everyone contributes to the interaction equally. However, there can be also two sub-groups exhibiting dyadic interaction characteristics. We hypothesise that with the increasing cardinality, possibilities for different interaction characteristics in the group increases significantly.

We empirically demonstrate that interactions in different sized groups should be considered separately for meaningful results. Thus, we propose to use a multi-stage approach where multiple classifiers are trained with respect to the group size. Prediction for a newly observed sample is then obtained by a linear combination of these classifiers, which is dynamically learned using only the training samples from the local neighbourhood of this newly observed sample. We will refer to our method as GAMUT, short for 'Group bAsed Meta-classifier learning using local neighbourhood Training'. Our method can therefore learn appropriate pairwise interaction dynamics tuned to the particular group cardinality directly from the data while having no prior knowledge about the interaction status or the group size of a given pair of people.

## 4.2. Related work on the detection of conversing groups

Automatic detection and analysis of interacting groups through computation has been a hot topic for computer science researchers under various names such as F-formation detection, modelling of human networks, detection of free standing conversing groups, etc. In this section we aim to provide a brief overview of such existing studies to show the foundations of our approach. Of course, we should also note that the analysis of human behaviour in groups and interaction dynamics are extensively studied in other disciplines, especially social psychology [6, 13]. Findings and insights from those studies greatly affected and influenced computer science researchers.

Categorization of the existing work on group detection can be made with respect to various criteria, such as the temporal length of the studies, employed sensor modalities, and the focus on proxemics or dynamics related to social behaviour. Most of the existing studies have multiple of these aforementioned characteristics. Thus, the categorization we present in this section is by no means strict.



### 4.2.1. Long-term studies with pervasive devices

Earlier studies generally analysed large scale long-term (in the order of months) social phenomena. Such studies did not explicitly aim to detect conversing groups but they focused on obtaining a rough estimation of face-to-face interaction to analyse long term social concepts. Choudhury et. al. presented one of the first studies on the topic in 2003 [7]. They used custom built wearable sensor packs called sociometers, which have accelerometers, IR transreceivers and a microphone. Data collection was done in two different stages, first one including 8 subjects and covering a time period of 10 days and the second one with 23 subjects for 11 days. IR transreceiver data was mostly used for detecting face-to-face interactions and shown to be quite noisy. As a slight shift to the dynamics, authors used audio to fetch speaking status, which was then used to refine the results of interaction obtained with the IR. As the final step of understanding social concepts, they showed that by analysing this interaction network, it was possible to obtain information related to group structures, such as centrality of a user.

Using a similar device, a sociometric badge, Olguin et. al. focused on analysing and measuring organizational behaviour in their 2009 work [19]. They employed IR to detect face-to-face interactions and bluetooth for measuring physical proximity. They also made use of accelerometers for detecting physical activity levels and microphones for speech detection. Data collection took 27 days and included 67 participants. Then, interaction characteristics sensed through these multiple modalities were used to classify personality traits of participants into the "Big Five" model.

There are also examples of work focusing on long-term characteristics that do not employ specific wearable sensors. Eagle and Pentland came up with their study "Reality Mining" in 2006, that focused on the utilization of mobile phones for sensing complex social structures [20]. They collected data from 100 mobile phones

over the period of 9 months. They aimed to infer various social concepts, such as recognizing social patterns in daily life, identification of significant relations, modelling organizational rhythms and, most interesting to us, recognition of social interactions. They relied on bluetooth communication of mobile phones to detect people in close proximity. No quantitative evaluation of the proximity networks was provided, but they showed that various social groups, such as friends and daily occurrences can be distinguished and this information can be used for further social understanding.

Similarly, Madan et. al. used mobile phones as social sensors in their 2010 work that aimed to detect behaviour changes with respect to illness [21]. For detecting social interactions of participants, they relied mainly on spatial distance, inferred from proximity and cellular-tower identifiers.



Wang et. al. presented their continuous sensing application, StudentLife, in 2014 [22]. A class of 48 students used the application for a 10 week term on their phones. The main aim of the study was to connect the automatic sensor data to the mental health and educational outcomes of the participants. Activity data and indoor and outdoor mobility were inferred from the accelerometer recordings. Audio from microphones are used to extract conversation data. A mix of cues related to light, activity, phone usage and sound are used to detect the sleeping patterns of the participants. Finally, location data is gathered from the GPS and co-location with other students are inferred through bluetooth. Number of significant correlations with various mental well-being surveys were found. Assuming the co-location and conversation related measures are proxies for interaction, the results indicate that students with frequent social interactions tend to be less depressed and more flourishing.

### 4.2.2. Short-term studies with pervasive devices

There are studies in the literature that uses pervasive devices, custom sensor packs or mobile phones, for the analysis of short term social interactions. By short term, we mean studies that focus on a single event that generally spans multiple hours. Gips and Pentland employed a custom sensor pack, UbER-Badge, embedded in a badge worn by conference attendees, to analyse interest and affiliation [23]. Specifically for affiliation, Gips argued for the use of cues related to wearer activity, inferred from accelerometers, in addition to proximity information obtained from IR encounters. In agreement with this study, he proposed to use pairwise measures, more specifically mutual information of the motion energy between pairs (MIME), to detect interacting partners in his thesis [8]. However, no qualitative performance evaluation was presented for the detection of interacting partners.

Similarly, Cattuto et. al. analysed face-to-face interactions in various crowded social settings, including 25 to 575 people, by using custom conference badges equipped with RFID [24]. Exchange of radio packets between badges were treated as a proxy for inferring spatial distance between participants and ultimately used to detect face-to-face interactions. They focused on analysing the dynamics of interaction networks, showing a super-linear behaviour between the number of connections and their durations. However, this study assumed an interaction bet-

ween two people in close proximity and the actual performance of this assumption was not quantitatively evaluated.

A relatively recent study from Matic et. al. used mobile phones to detect two parameters, interpersonal distance and relative body orientation, which were then used as proxies for inferring social interaction between participants [25]. Authors used a time frame of 10s, aiming to capture dynamic changes in social interactions. Their experiments showed that, using the standard deviation of body orientation throughout 10s windows as a feature in addition to the distance and body orientation, improves the correct detections of social interactions.

### 4.2.3. Static image based methods

More recently, with the increasing success of computer vision methods, researchers started to use images and video as main input modalities. Cristani et. al. focused on unconstrained scenarios and employed solely visual cues to detect social interactions in their 2011 work [26]. Their proposed method took the positions and head orientations of people in the scene as input and employed a voting strategy built on the Hough transform. They presented their results on synthetic data and videos of real life indoor and outdoor scenarios, discussing how automatic detection of positions and head orientations in real life affects the performance. Promising performance on both real life datasets (outdoor and indoor) were presented.



In the same year, Hung and Krose presented their work on detecting F-formations with a graph clustering algorithm that was formulated as the identification of dominant sets [27]. In addition to the proximity between people, body orientation information was used as a cue for detection. They proposed to use socially motivated estimate of focus orientation (SMEFO), which was calculated from location information only, as the body orientation feature and experimentally showed that the addition of this feature to location increased the performance.

In 2013, Setti et. al. compared these two main approaches of detecting F-formations from images and presented their advantages and disadvantages over different scenarios [9]. They concluded that the Hough-based method [26] performs better when using position and orientation together, showing good robustness to noise; whereas dominant set based method [27] is better for scenarios when only position information is available.

In the same year, Setti et. al. published another work that advocates a multi-scale approach for F-formation discovery [17]. It is one of the first papers that takes the cardinality of the interacting groups into consideration in the detection process, as we also advocate. The proposed approach was built on the Hough voting policy of [26] and based on a competition of different voting sessions, specialized for a specific group cardinality, which are then evaluated with an information theoretic criteria to obtain final set of groups. They showed promising results on various datasets, synthetic and real life.

Setti's work in 2015 [28] presented a detailed review of current group detection algorithms for single images, including the ones mentioned here, and proposed a graph-cut based approach that outperformed others. They reported their results in five datasets with various characteristics and presented a deep analysis of methods

robustness to noise. This is also one of the few studies (that we are aware of) that includes a performance analysis related to the cardinality of the target groups.

### 4.2.4. Video based analysis

All the computer vision studies mentioned in the former subsection lack the temporal information. Even though they took video as input, the detection of groups as performed on single frames. Vascon et. al. proposed a game-theoretic approach for detection of F-formations and presented their results on single and multiple frames, integrating temporal information with the multi-payoff evolutionary game theory [10]. They showed that the integration of multiple frames augments the overall group accuracy, especially in cases of strong noise in the positions and orientations.

Another study that uses the temporal information was Alameda-Pineda et.al.' s work[11]. They presented a multimodal approach that combined data from cameras and wearable sensors for estimating head and body pose of participants in the scene. Estimated head and body poses were then used for detection of F-formations and social attractors. Wearable sensors were used to obtain noisy estimates of speaking status and proximity input, which was then used in combination with the visual features in a matrix completion formulation for obtaining head and body poses. Their optimization included a coupling of body and head pose estimates and a temporal constraint, making it certain that detected head and body estimates are jointly estimated and temporally viable.

Depth sensors were also utilized in the literature for the estimation of spatial distance and orientation of the participants in a scene. An example study was presented by Gan et. al. in 2013, that used multiple Kinects for obtaining spatial location and orientation of each participant in the scene [29]. This information was then employed by the heat-map based feature representation proposed in the paper. Qualitative evaluation of the proposed feature representation was done on a synthetic dataset only, where the authors found their temporal encoded IS performed slightly worse than the one without temporal information. Authors then argued that this result was mainly caused by the characteristics of their ground truth.

### 4.2.5. Moving beyond just group detection

There are also works in the literature that aim more than the detection of interacting groups. Tran et. al. employed a dominant sets based approach for F-formation detection, which was followed by group activity representation and recognition [30]. For group discovery, authors presented and compared two social cues, personal distance and visual focus of attention, which are basically spatial distance and head orientation. For representing group activity, they used a bag-of-words approach that represented videos as a histogram of codewords and employed Support Vector Machine for activity classification.

Zhang and Hung presented their method to detect differing levels of social involvement, more specifically discovering associates of F-formations, people who are attached to an F-formation but do not have the status of full members [31]. They introduced novel multi-annotator annotations of the associates and compared two methods for detecting them. They also proposed a spatial-context-aware F-

formation detector, that focuses on modelling people's frustum of attention. They showed that detecting and cleaning in-group associates improved the performance of F-formation detections.

### 4.2.6. Dynamics related to social behaviour

Up until now, all the works we mentioned solely focused on the proxemics of the interaction (aside from [7] and [23] that used speaking status and movement energy, respectively, as cues for detection). Although some studies included temporal information, they mainly focused on modelling the temporal changes in the spatial location only. Perhaps the closest study to our work was published by Hung et. al. in 2014, where the authors used a single accelerometer to classify social actions of participants in a crowded gathering [32]. These classified actions were used to extract pairwise mutual information that aims to capture interaction characteristics between dyads. Interacting partners were then detected by thresholding these values. The authors found that the mutual information computed from the pairwise speaking turns performed the best when using 40 second windows, compared to other social actions, raw acceleration and window sizes. The results were presented on a relatively limited dataset that included 10 minutes data from 26 subjects.



## 4.3. Dataset

To test our method and assess its generalization capabilities, we used a dataset collected during a speed dating event, in a real pub, for 3 days [33]. The first phase of each day involved members of the opposite sex having three minute seated dates. This phase was then followed by a mingling session that for approximately an hour. These second phases of the events had free-standing conversational groups in a crowded environment; the scenario we are interested in. The mingling area was limited to ensure a high spatial density of people. However, people were not instructed in any specific way, so they could freely move and interact with their peers. For more detailed information about this dataset, please refer to [33].

Throughout the event, participants wore a custom made sensor pack around their necks, that records tri-axial acceleration at 20Hz. Top-view videos of each event are also captured, which is only used for the labelling of the ground truth, both participants' social actions and F-formations. Example screenshots are shown in Figure 4.1.

Figure 4.1: Example Snapshots from the mingling phase of Day 3. Taken from [33]

### 4.3.1. Dataset statistics

The dataset includes working sensor readings of the mingling session for 26, 22 and 22 participants, respectively for days one, two and three. For each day, a ten minute segment was manually annotated for the F-formations, resulting in three different segments from each day. A variety of social actions of participants were also manually annotated, including the speaking status, hand gestures and head gestures that are used in our experiments. We should note that our proposed method is fully automatic, so the ground truth related to these social actions is only used for the training phase of social action classification (see Section 4.4.1 for further details).

Table 4.1 shows the number of unique interactions and mean and standard deviation of interaction lengths in seconds per group cardinality, for each day.



Table 4.1: Number of unique interactions, average length and standard deviation of interactions with respect to the group cardinality, for each day. They are extracted from the ground truth labels of F-formations by counting each unique occurrence of a group. All statistics related to length of interactions are in seconds.

| | Day1 | | | Day 2 | | | Day3 | | |
|---|---|---|---|---|---|---|---|---|---|
| Cardinality | # Int | Avg length | Std length | # Int | Avg length | Std length | # Int | Avg length | Std length |
| 2 Persons | 24 | 160 | 163 | 10 | 126 | 108 | 21 | 152 | 169 |
| 3 Persons | 17 | 53 | 51 | 11 | 157 | 119 | 5 | 89 | 56 |
| 4 Persons | 9 | 79 | 77 | 4 | 134 | 104 | 14 | 110 | 161 |
| 5 Persons | 3 | 29 | 18 | 3 | 127 | 97 | 0 | - | - |
| 6 Persons | 1 | 26 | - | 5 | 55 | 48 | 0 | - | - |
| 7 Persons | 0 | - | - | 5 | 81 | 123 | 0 | - | - |

The number of groups with a specific cardinality vary greatly with respect to the day of the event. For example, we can see for Day 1, the number of unique interactions reduces with increasing cardinality. Similarly, for Day 3, most of the interactions are in two, three and four person groups. The number of dyadic interactions and the length of these interactions for Day 1 and 3 are much higher than Day 2.

When we inspect the statistics related to the speed dates [33], we can see that the number of matches (pairs where both participants stated they would like to see each other in the future) for Day 1 and 2 (70 and 79 respectively) are higher than Day 3 (61). This might be the reason why we see more dyadic interactions in Day 1 and 3, where participants tended to stay in mostly dyads. Day 2 has much more variation with respect to the group cardinality, having the only examples of seven person groups. These statistics show that even with the same setup of events, various configurations of the group cardinalities can arise and a preferred solution should be able to generalise over the dynamics of all cases.

Another interesting statistic is related to the mean and standard deviation of the interaction lengths. For nearly all cardinalities and days, standard deviations of the interaction lengths are quite high with respect to the mean. This suggests that the dataset contains a wide distribution of conversation lengths. Manual inspection of the lengths of various unique interactions supported this observation; there are groups staying together for matter of seconds (splitting, one person leaving, etc.) and groups staying together for the entire 10 minute interval.

# 4.4. Methodology

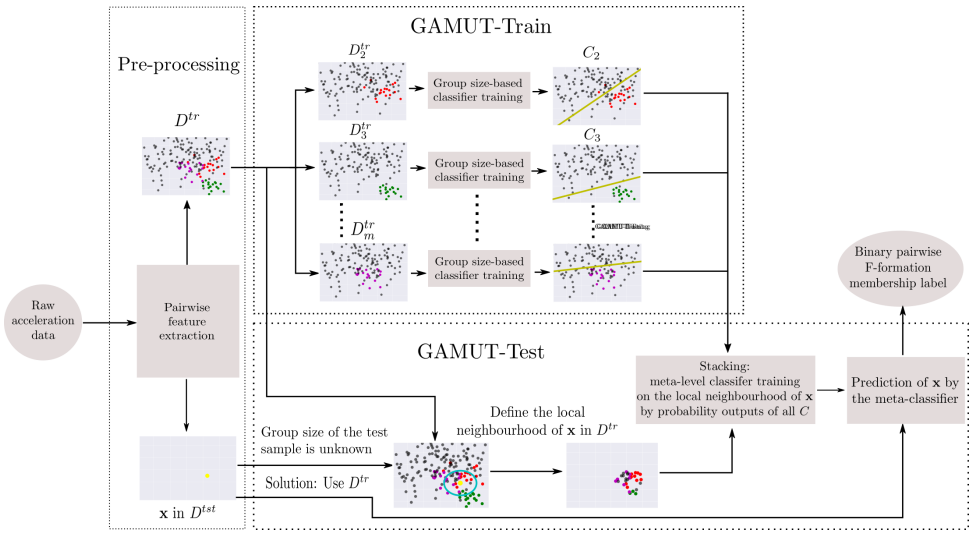


Figure 4.2: Flow diagram of the proposed method. $D^{tr}$ and $D^{tst}$ correspond to entire training and test sets, respectively. $D_n^{tr}$ is a subset of $D^{tr}$ which is formed by the positive samples coming from groups of cardinality $n$ and all the negative samples. $C_n$ is the group-based classifier that is trained with $D_n^{tr}$.

Our method aims to estimate pairwise F-formation membership for each pair in the scene. We define the problem as a binary classification task, where the final aim is to classify whether a pairwise feature representation $P_{ij}$ indicates that person $i$ and person $j$ belong to the same conversing group or not. Thus, data from all participant pairs are used to obtain a joint representation which corresponds to the samples in the classification process. A flow diagram of the proposed method is shown in Figure 4.2. Here are the basic steps of the proposed method:

1. Preprocessing
   (a) Social action classification
   (b) Pairwise feature extraction
2. Group-based meta-classifier learning using local neighbourhood training (GAMUT)
   (a) Training multiple classifiers with respect to the group cardinality.
   (b) Prediction of a new test sample with meta-level classifier training using the local neighbourhood of the test sample [1]

Each of these steps will be explained in detail in the following subsections.

[1] We define local neighbourhood as the K-nearest samples of the training set, in the feature space.

### 4.4.1. Preprocessing

The preprocessing steps convert the raw triaxial acceleration signal from participants into pairwise feature representations. Some of the pairwise features are computed on the social action streams of the participants, so the first step of preprocessing is the classification of social actions; speaking, hand gestures and head gestures. This step is then followed by the actual feature extraction, where the raw acceleration and social action streams are used to obtain pairwise representations. These pairwise features are our samples in GAMUT.

#### Social action classification

To provide a generalized solution, our action classification method should be person independent, where data from the test subjects are not used in the training. There are examples of person independent methods for action classification in the literature [34, 35] but they mainly focus on daily activities such as walking and running. On the other hand, manifestations of actions such as speaking and gesturing are highly person specific, making their detections harder tasks for generalisation. We employed a transfer learning method, Transductive Parameter Transfer(TPT), which is experimentally shown to outperform traditional person independent approaches for person specific actions [36].

Transductive Parameter Transfer(TPT) is an adaptive transfer learning approach that aims to learn a mapping between the distribution of a dataset and the parameters of the optimal classifier for it. Sangineto et. al. proposed the method in 2014, for personalized facial expression detection from portrait images [37]. A specialized version for social action detection was then proposed by Gedik and Hung [36], which we employ in this study. TPT finds the parameters of the optimal classifier for a target dataset $X^t$ (test set in a traditional setting) by learning a mapping between the marginal distributions of the source datasets (the training set in a traditional setup) and the parameter vectors of their optimal classifiers. A formal definition of the $N$ source datasets (with label information) and the target dataset (without label information) can be made as $D_1^s, ..., D_N^s$, $D_i^s = \left\{x_j^s, y_j^s\right\}_{j=1}^{n_i^s}$ and $X^t = \{x_j^t\}_{j=1}^{n_t}$, respectively . Algorithm 2 presents the main steps of TPT (A more detailed explanation can be found in [36]).

We used the same feature extraction and classification setup of [36] for obtaining the social action labels for all the participants used in the experiment. Statistical (mean and variance) and spectral (power spectral density with 8 logarithmically spaced bins between 0-8 Hz) features were extracted from each axis of raw and absolute values of acceleration and the magnitude of the acceleration. 3s windows with 1.5s overlap, experimentally shown to perform well in [36], were used.

Classification was done in a Leave-one-subject-out fashion. So, in each fold, we treated one participant as the target set and all others (including participants from other days) as the source sets. This procedure was replicated for each corresponding social behaviour type; speaking, hand gestures and head gestures. Since the labels were imbalanced for many participants, we chose to evaluate using Area Under Curve(AUC).The performances obtained with TPT are shown in Table 4.2. For each participant, we obtained classified labels corresponding to a 3s window for

**ALGORITHM 2:** Transductive Parameter Transfer approach (Taken from [36])

**Input:** Source sets $D_1^s, ..., D_N^s$ with labels and the target set $X^t$

**Output:** $\mathbf{w_t}, c_t$

Compute $\{\theta_i = (w_i, c_i)\}_{i=1}^N$ using L2 penalized Logistic Regression.

Create training set $\tau = \{X_i^s, \theta_i\}_{i=1}^N$.

Compute the Earth Mover's Distance (EMD) kernel matrix $K$ that defines distances between distributions where $K_{ij} = \kappa(X_i^s, X_j^s)$.

Given $K$ and $\tau$, compute $\hat{f}(.)$ by Kernel Ridge Regression.

Compute $(w_t, c_t) = \hat{f}(X^t)$ using the mapping obtained by step 4.

| Social Action | Mean AUC(%) | Std +- |
|---|---|---|
| Speaking | 66 | 6 |
| Hand Gestures | 67 | 10 |
| Head Gestures | 60 | 9 |

Table 4.2: Performances of social action detection with TPT



each action, with 1.5s overlap. In terms of the 1.5s overlap of labels, we favoured positive ones. Specifically, if a positive label was followed by a negative one, or vice versa, the overlapping 1.5s was considered to be positive.

### Pairwise feature extraction

We mentioned in the beginning of this section that each possible pair of participants in a scene is treated as a single entity in the classification process. Each of these features are extracted from the pairs of data (either social actions or raw acceleration) coming from the two participants in the pair and generally aims to define a measure of behavioural coordination.

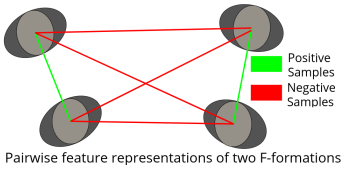


Figure 4.3: Synthetic visualisation of two F-formations with cardinality of two and their pairwise representations (lines). All possible pairwise representations form a sample.

| Feature | Dim. | Computed on | ID |
|---|---|---|---|
| Correlation | 1 | Acceleration[mag1-mag2,Y1-Y2,Z1-Z2] | 0-2 |
| (N)Mutual information | 1 | Acceleration[mag1-mag2, Y1-Y2, Z1-Z2, Y1-Z2 ], S1-S2 S1-Ha2, S1-He2, S2-Ha1, S2-He1 | 3-21 |
| Boolean turn activity[38] | 6 | S1-S2, S1-Ha2, S1-He2, S2-Ha1, S2-He1 | 22-52 |
| Overlap statistics | 4 | S1-S2, S1-Ha2, S1-He2, S2-Ha1, S2-He1 | 53-73 |
| Event synchrony[39] | 2 | S1-S2, S1-Ha2, S1-He2, S2-Ha1, S2-He1 | 74-84 |

Table 4.3: Pairwise features for joint representation. Acronyms used: Mag: Magnitude, Y: Y axis, Z: Z axis, S: Binary speaking status, Ha: Binary hand gesture status, He: Binary head gesture status, (N): Shows both the mutual information and the normalized mutual information are calculated, Numbers: Shows which participant in the pair that the stream comes from.

Figure 4.3 is a synthetic visualization of a simple possible scene, where four people are interacting in two groups. We assume that every participant is connected to all other participants in the scene with hypothetical connections and these hypothetical connections represent samples in the classification process. The aim of the classification is to decide if these hypothetical connections connect two participants that are in the same conversing group or not. Here, green lines indicate positive samples or true connections and red lines are negative samples or false connections.

These hypothetical connections correspond to our pairwise features that provides a joint representation of the data of the participants connected by the line. We have used various features that are already employed in the literature and proposed a new one, which we named overlap statistics. Table 4.3 shows all the features that are used in our experiments. It also presents the dimensionality of the aforementioned feature, from which pairwise data streams they are extracted and their assigned IDs. The table is followed by the detailed explanation of each feature.
***Correlation***: Pearson correlation coefficient of two input streams. Calculated as:

$$\rho = \frac{cov(X,Y)}{\sigma X, \sigma Y} \tag{4.1}$$



***(Normalized) Mutual information***: Mutual information of two input streams. Calculated as:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{4.2}$$

where $H(X)$ and $H(Y)$ are the marginal entropies and $H(X|Y)$ is the joint entropy. For the calculation of entropies, we used binned individual and joint counts. From there, normalized mutual information per sample is also calculated as:

$$NI(X;Y) = \frac{I(X;Y)}{\sqrt{H(X)H(Y)}} \tag{4.3}$$

Both normalized and non-normalized mutual information values are used in the experiments.
***Boolean turns activity(BTA)***: Selection of measures presented in [38], that aims to provide statistics related to turn taking in dyadic interaction. This is applied to two binary social action streams where a 1 indicates the presence of the action. For the sake of easier representation, we will call it being 'active'. Here are the measures defined for this feature set:

1. Ratio of participant 1 being active (over the whole period).
2. Ratio of participant 2 being active (over the whole period).
3. Ratio of total active time for both participants over the whole period.
4. Ratio of total inactive time for both participants over the whole period.
5. Synchrony ratio of participant 1 with respect to participant 2, which is computed as the ratio of times participant 1 became active after a predetermined time (3s) that participant 2 was active to the total number of times participant 1 was active.
6. Synchrony ratio of participant 2 with respect to participant 1, computed as 5.

***Overlap statistics***: We propose a new feature set, mainly inspired by BTA. It aims to statistically represent co-occurring events:

1. Number of unique times that participant 2 was active while participant 1 was active.
2. Mean length of the active intervals of participant 2 where participant 1 is active.
3. Median length of the active intervals of participant 2 where participant 1 is active.
4. Standard deviation of the length of the active intervals of participant 2 where participant 1 is active.

***Event synchrony***: A method to measure synchronicity and time delays between two univariate signals was presented in [39]. The synchronicity and time delay patterns of two signals are represented by symmetrical ($Q_\tau$) and anti-symmetrical combinations ($q_\tau$) of events happening in the signals. Events correspond to unique continuous active regions of the signals. The formulation for two univariate signals $x$ and $y$ is as follows:

$$c^\tau(x|y) = \sum_{i=1}^{m_x} \sum_{j=1}^{m_y} J_{ij}^\tau \tag{4.4}$$

where $c^\tau(x|y)$ is the number of times that an event happens in x shortly after y and vice versa, $\tau$ is a predefined lag between the signals and

$$J_{ij}^\tau = \begin{cases} 1 & \text{if } 0 < t_i^x - t_j^y \leq \tau \\ 1/2 & \text{if } t_i^x = t_j^y \\ 0 & \text{else} \end{cases} \tag{4.5}$$

where $t_i^x$ and $t_j^y$ $(i = 1, .., m_x; j = 1, .., m_j)$ correspond to event times. With this formulation, $Q_\tau$ and $(q_\tau)$ are then computed as:

$$Q_\tau = \frac{c^\tau(y|x) + c^\tau(x|y)}{\sqrt{m_x m_y}}, q_\tau = \frac{c^\tau(y|x) - c^\tau(x|y)}{\sqrt{m_x m_y}} \tag{4.6}$$

With this feature extraction setup, we end up with samples with the dimension of 85. As it can be seen, each feature tries to represent some type of pairwise measure between data streams of participants, may it be correlation, synchrony, lag, commonly occurring events, etc. The IDs correspond to the order of the streams given in Table 4.3 and their explanations. For example, IDs 53 to 57 maps to overlap statistics, as in given order in the explanation, of the speaking streams.

### 4.4.2. GAMUT: Group-based meta-classifier learning using local neighbourhood training

As mentioned earlier, the scenarios we are interested in can include a variety of groups. Differing interaction dynamics are expected to arise in groups of different cardinalities. For example, we will expect two groups in Figure 4.3 to have relatively

similar characteristics with identifiable turn-taking patterns. But, the pairwise interactions in a group of three with participants A, B and C might differ compared to a group of two. On the other hand, participants A and B can be in a dyadic interaction where both are active in the conversation, still resembling the dynamics of the two person groups mentioned. However, in such a case, the pairwise interaction dynamics of participant C with the others is expected to be different, since C will be in a listener role. Variations in interaction dynamics increase with the increasing cardinality.

Moving on from this assumption, we form the hypothesis that a classifier trained specifically for capturing the dynamics of a single cardinality should perform better on test samples of the same cardinality compared to a classifier trained on the data from other cardinalities. Moreover, this cardinality specific classifier might also perform better in capturing subgroups in larger group sizes. In order to address varying interaction dynamics of different sized groups, we propose to train different classifiers for different group sizes. It is not possible to directly choose the optimal classifier (or classifiers) for a new sample since the group size of a sample is unknown. We propose to overcome this difficulty by employing a transductive approach where the local neighbourhood of the the test sample in the training set is used as an additional information source in the test phase. We expect the local neighbourhood of a test sample in the feature space to be more informative than the entire training set. Thus, a meta-level classifier is trained only with the probability outputs of the group based classifiers based on the training samples from this local neighbourhood. This meta-level classifier is then used to classify the test sample.

Formally, the whole training dataset is defined as $D^{tr} = \{\mathbf{x}_i, y_i, g_i\}_{i=1}^{n^{tr}}$ where $\mathbf{x}_i \in \mathbb{R}^d$ , $y \in \{0,1\}$, $g_i \in \{0,2,...,m\}$ and $m$ is the largest possible group size. Here, $\mathbf{x}_i$ is the pairwise feature vector, $y_i$ is the binary labels of pairwise F-formation membership, $g$ is the cardinality of the group that the sample is coming from and $n^{tr}$ is the total number of samples in the training set. Then we define the set of negative samples in training as $D_0^{tr} = \{\mathbf{x}_i \mid y = 0\}_{i=1}^{n^{tr}}$ and positive samples in training coming from a specific group size $k$ as $D_k^{tr} = \{\mathbf{x}_i \mid g = k\}_{i=1}^{n^{tr}}$ where $g > 0$. With this setup, the steps for training and testing is shown in Algorithm 3. The following two subsections explain the training and testing procedures in detail.

### Training multiple classifiers with respect to group cardinality

Figure 4.4, which includes two conversing groups of size two and one of size three, is provided to visualize the training of group level classifiers. It uses a similar representation to Figure 4.3, where lines correspond to pairwise features; a sample in the classification process. As it can be seen, while training a classifier for detection of groups of size two, pairwise samples from the two person groups are treated as positives and all samples that are extracted from a pair that is not in the same group as negative. Positive samples from the three person group are not included in the training. Similarly, the right side of the image visualizes the case for the classifier of size three, where positive samples come from the three person group and positive samples from two person groups are not used.

**ALGORITHM 3:** Training and testing phases for the GAMUT

**Input:** Training and test sets, $D^{tr}$ and $D^{tst}$
**Output:** Classified labels (and/or probabilities) for $D^{tst}$
**Training:**
Train a set of classifiers $C = \{C_2, C_3, .., C_m\}$ where $C_k$ is trained on the dataset $D_k^{tr} \cup D_0^{tr}$.
**Test:**
**for** *each sample* $\mathbf{x_j}$ *in* $D^{tst} = \{\mathbf{x}_j\}_{j=1}^{n^{tst}}$ **do**
Find $\Psi$, the K-nearest neighbours of $\mathbf{x_j}$ in $D^{tr}$.
Train a meta-level classifier, $C_{meta}$, on the probability outputs of $C$, all pre-trained group size based classifiers, on $\Psi$.
Use $C$ and $C_{meta}$ to classify (or obtain probabilities of) $\mathbf{x_j}$.
**end**



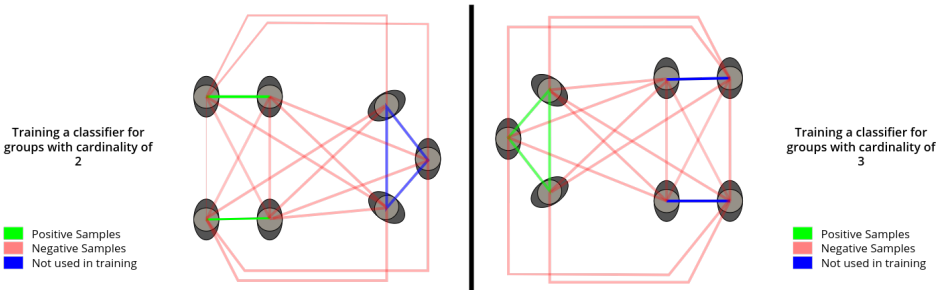


Figure 4.4: Visual explanation of the training of group cardinality based classifiers, namely $C_2$ (left) and $C_3$ (right). The same convention as Figure 4 is used where all possible pairwise representations are shown with connecting lines between participants. The scene includes three conversing groups, two of cardinality two and one of cardinality three. As the lines suggest, while training $C_2$, positives samples from the three person group are excluded and vice versa for $C_3$.

We have selected the L2 Regularized Logistic Regression for training, which minimizes the unconstrained optimization problem shown as follows:

$$\min_{(w,c)} \frac{1}{2} w^T w + C \sum_{i=1}^{n} log(exp(-y_i(\mathbf{x}_i^T w + c)) + 1), \tag{4.7}$$

where $x_i$ is the pairwise feature vector, $y_i$ is the binary pairwise F-formation membership labels, $c$ and $C$ are the bias and regularization terms, respectively. We used stochastic average gradient descent as the optimizer in our experiments[40]. The pairwise formulation causes the classes to be extremely imbalanced. Participants are not in the same conversing group with the majority of the others at the events, resulting in many negative samples. This phenomena can be easily seen from Figures 4.3 and 4.4, where the number of negative samples is much higher than the positive ones, even in these simple scenarios. To account for the imbalance, the weights are adjusted to be inversely proportional to the class frequencies.

**Class prediction of a new sample by meta-classifier training (stacking) using its local neighbourhood**

For a test sample, we first define its local neighbourhood in the training set, similar to a transductive setting. We first obtain probability outputs for training samples in this local neighbourhood with each of the (already trained) group size based classifiers. Then, we train a meta-level classifier using these probability outputs and the labels of these samples in the local neighbourhood. This process, stacking, is known to reduce the generalisation error rate of multiple classifiers by reducing their biases [5]. Using only the samples from the local neighbourhood makes it possible to consider samples with similar characteristics to the current test sample, for further tuning the weights learned for the meta-level classifier. This process is repeated for every test sample. In other words, different meta-level classifiers are trained for each test sample. Similar to group size based classifiers, we chose a L2 Regularized Logistic Regressor as the meta-level classifier.



## 4.5. Results

### 4.5.1. Experiments setup

We tested GAMUT on the dataset of Section 4.3. We randomly kept 10% of the dataset as the test set while the remaining samples were used in the training. In order to test the generalization capabilities of the proposed method, this random selection process and the following evaluation is repeated 500 times, each producing a performance score of it's own. While forming the training and test sets, we made sure that there is no data from the same pair of people in both training and test sets to avoid contamination.

Our proposed approach is compared to various baselines. Firstly, we implemented the method proposed in [32], which is closest to our setting in terms of approach and modality; the state-of-the-art in our problem. This approach finds the optimal threshold value using the mean mutual information values calculated over speaking and gesturing streams of pairs in the training set. This threshold value is then used to classify the samples in the test set. Secondly, we considered an approach where group cardinalities were not considered in the training. In this approach, all the features we propose are used, but training is performed with Logistic Regression on all the dataset without any distinction related to the group sizes.

For selecting the size of the local neighbourhood (K value), we used an empirical approach. Since we have samples coming from six different group sizes, we expected $n^{tr}/6$ samples should be representative as a local neighbourhood. In an optimal case where the number of samples are equally divided between cardinalities and the samples from same group sizes have similar representations in the feature space, this neighbourhood will be formed by the training samples of the same corresponding group cardinality as the test sample. However, this is not always the case and there might be regions where distinguishing between the characteristics of different sized groups are harder, so we experimented with various neighbourhood sizes, ranging from $n^{tr}$ (no local information, all samples are used in the stacking

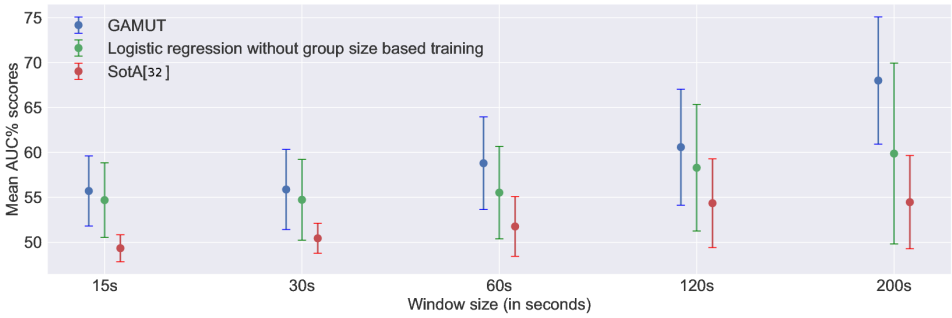


Figure 4.5: Performance scores (mean AUC±STD (%)) of various approaches on pairwise F-formation membership detection. Mean and standard deviation of the AUC scores of 500 runs are visualised with the points and error bars, respectively.



process) to $n^{tr}/6$.

We have imbalanced data, thus we chose Area Under Curve (AUC) as the performance metric. Since we expect various dynamics to arise in different temporal resolutions, we present our results for different pairwise feature extraction window lengths, ranging from 15 to 200 seconds. If the two participants of a sample are in the same group for at least two thirds of this interval, the sample is treated as a positive. Empirically the best performing local neighbourhood sizes for window sizes of 15, 30, 60, 120 and 200 seconds were found to be $n^{tr}/4$, $n^{tr}/3$, $n^{tr}/3$, $n^{tr}/3$ and $n^{tr}/5$, respectively. The mean AUC and the standard deviation of 500 runs of the aforementioned methods are presented in Figure 4.5.

### 4.5.2. Performance scores

The method presented in [32] performed worst, even providing AUC scores lower than the random baseline (AUC of 50%). The reason of this becomes clear when the statistics of the dataset presented in [32] are investigated. The authors reported a performance value better than random on a subset of their dataset that includes nine participants with only dyadic interactions. The low performance obtained with this approach is extremely important since up until this point, existing work solely relied on pairwise mutual information of various streams for investigating speaking turns and conversing groups detection through dynamics [32, 41]. Our empirical results show that when a realistic scenario with groups of various sizes is considered, such approaches fail to provide satisfying results.

The contribution of our newly proposed features is already demonstrated by the performance of the logistic regressor without the group size based training. Even with this setup, AUC scores that are better than random and outperforming state-of-the-art are always obtained. A more detailed analysis of the effectiveness of features will be presented in Section 4.6.

Our proposed approach, GAMUT, performs significantly better than all other approaches regardless of the window size ($p < 0.01$, with a paired t-test computed on the performance values of 500 runs). The contribution of our method is more

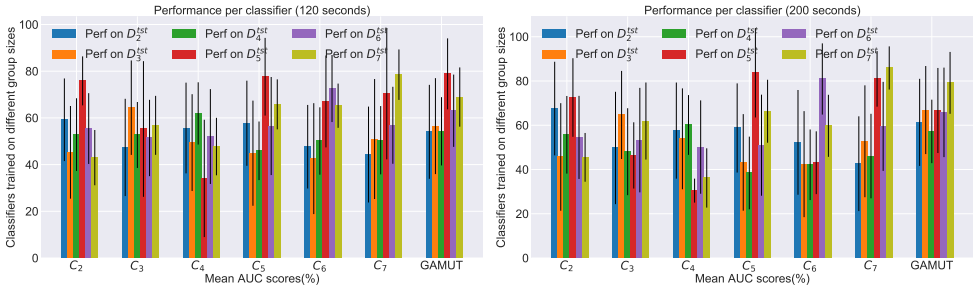


Figure 4.6: Mean and standard deviation of AUC% scores of $C_k$ and GAMUT on $D_k^{tst}$ for all $k$ with window sizes of 120 (left) and 200 (right) seconds. Each group of bars visualises one classifiers performance on various group cardinality based subsets of the data.



clearly visible for larger window sizes. Also, we generally see a pattern for all methods: Increasing window size results in better performance. This is expected, since various interaction dynamics can arise in longer intervals and it becomes easier to capture such dynamics in longer window sizes. Even though the performance was increasing, we stopped our experiments at 200 seconds, since the number of usable samples from some group sizes would reduced drastically, making training and testing impossible. These results clearly show that in order to capture variations in real life, a method that is group size aware is definitely required.

## 4.6. Further analysis

In this section, we further investigate the nature of the problem by presenting various analyses and ablation studies.

### 4.6.1. Performances of group size based classifiers and GAMUT on datasets of different group cardinalities

In this subsection, we provide an analysis of how group size based classifiers and GAMUT perform on datasets that include positive samples only from a specific group cardinality. This way, we empirically show that our hypothesis in Section 4.4.2 (that a classifier trained specifically for capturing the dynamics of a single cardinality should perform optimally for the samples of the same group size) holds.

Formally, we present the performances of classifiers $C_k$ and GAMUT on $D_k^{tst} \cup D_0^{tst}$ , where $k \in \{2, 3, 4, 5, 6, 7\}$, all possible group cardinalities in our dataset. Similar to the training phase, we create subsets of our test dataset, where positive samples come from one specific cardinality. For simplicity, we will refer to the whole test subsets, that also includes the negative samples, as $D_k^{tst}$ in this section.

Figure 4.6 presents the results for two window sizes, 120 and 200 seconds, for the sake of space. In the plots, each collection of bars corresponds to the mean AUC score of one classifier ($C_k$ or GAMUT) on six different group cardinality based subsets of the data, calculated over 500 runs of leaving 10% of the data out for testing. The error bars correspond to the standard deviation of the AUC scores.

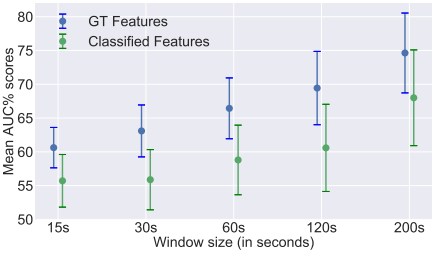


Figure 4.7: Performances of GAMUT with the ground truth or classified social action labels.

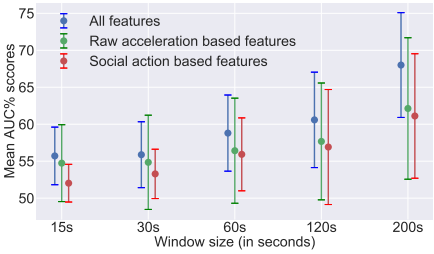


Figure 4.8: Performances of GAMUT with raw acceleration or social action based features



Figure 4.6 supports our hypothesis and shows the power of our proposed approach. For both window sizes, the best performing group size based classifier for a subset is the one with the matching cardinality. We also see that some group based classifiers performed relatively well on subsets with different cardinalities. This supports our second hypothesis that for some cases, a classifier might capture dynamics of the groups of other sizes. More importantly, nearly for each subset regardless of the window size, GAMUT guarantees to be the second best performing classifier after the matching group size based classifier. Note that in practice, GAMUT is therefore the best performing classifier as the group size from which a test sample is drawn is not known.

### 4.6.2. Effects of social action classification performance on pairwise F-formation membership detection

As mentioned in the former section, the first step of our proposed approach is the classification of social actions. The performance of the social action classification is by no means perfect and faulty labellings in this step are expected to have an effect on the final pairwise F-formation membership detection.

In order to see the effects of the performance of social action detection on the final goal, a follow up experiment was performed where we used the human annotated labels (ground truth) for speaking, hand gesturing, and head gesturing to extract pairwise features, instead of the automatically generated labels with TPT. Figure 4.7 shows the pairwise F-formation membership detection scores with ground truth and classified social action labels.

As expected, GAMUT with ground truth social action labels always performs better. This result shows that our features tend to perform better in capturing interaction dynamics if the social action labels are entirely correct. Fortunately, the difference in performance is not a lot, showing that our social action detection approach still provides valuable information that can be used to infer pairwise F-formation membership relatively satisfactorily.

### 4.6.3. Contributions of raw acceleration and social action based features

Pairwise features used in GAMUT can be grouped into two categories with respect to which type of streams are used in their extraction: Raw acceleration (IDs 0-11) or social action labels (IDs 12-84). In this section, we present an ablation study where we use these feature groups separately in the training, to further understand their contribution to the final performance. Figure 4.8 shows the performances obtained when these features are used separately and together.

Using both sets results in higher performance compared to using either separately. This is an interesting outcome showing how additional higher level information extracted from the same source can act in a complementary manner. We also see that the raw acceleration based features tend to perform better than the social action based ones. This is especially true for small window sizes where raw acceleration based features clearly outperform the social action based ones. However, with the increasing window size, the gap between the performances of two feature sets seems to close, showing that the social action based features require more time to be informative. This is expected since many social concepts require time to unfold and it is harder to capture them in shorter time resolutions. Also, the social action labels used for extracting the features are results of a classification process (Section 4.4.1) and by no means perfect. As discussed in Section 4.6.2, if the ground truth labels are used for the extraction of social action based features, the overall performance increases significantly. This is another factor that can explain the gap between the performances of the two feature sets.

### 4.6.4. Correlation analysis of features and pairwise F-formation labels

To have a deeper understanding about the contributions of each feature, we conducted a correlation analysis between the vectors of single features and the ground truth for pairwise F-formation membership. In order to investigate how the correlation of features change with respect to the group cardinality, subsets of the whole dataset that had positive samples coming from a single group cardinality were used. We used the whole dataset for computing the Pearson correlation coefficients, without any distinction between training and test sets. $D_k \cup D_0$, a subset where positive samples are coming from the groups of cardinality $k$, will be denoted as $D_k$ for simplicity. The correlation coefficients are calculated for all window sizes per subset but to preserve space, only the results for two window sizes are presented in Figure 4.9.

The highest correlation coefficients tend to be around 0.1 or -0.1 which are considered to be weak correlations. Still, even weak correlations have information and we expect our classifier to combine such weakly informative features to tackle the problem. Thus, we will be considering features with at least weak correlation coefficients as marginally informative and analyse their occurrences. Only the features with statistical significance ($p<0.05$) are investigated below.

The most striking observation from Figure 4.9 is how the features with highest correlation coefficients vary with respect to the group cardinalities. This supports

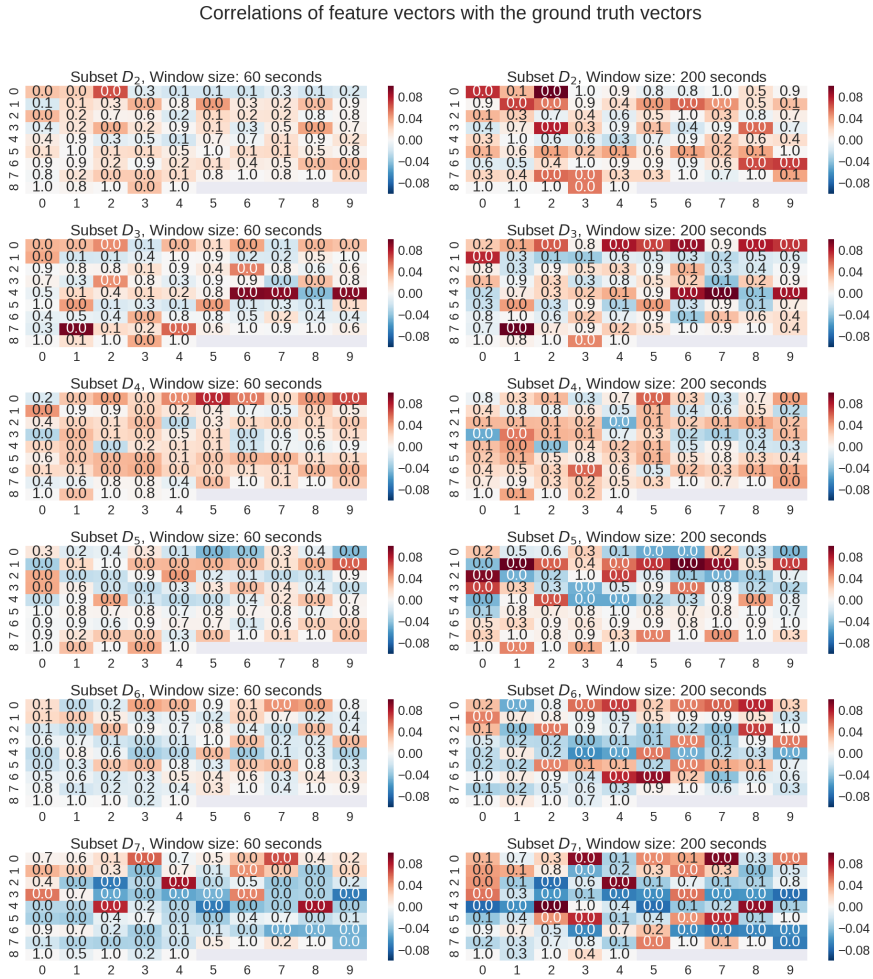


Figure 4.9: Pearson correlation coefficients (r) and significance (p) values calculated between feature vectors and the F-formation membership ground truth on group cardinality based subsets ($D_k$) of the data. Correlation values (r) are presented as the color of the cells where as the value inside the cells correspond to the significance. While presenting the significance values, we have used one decimal places, thus a value of 0.0 corresponds to $p < 0.05$. Each row in one matrix has correlation and significance values for ten features. For example, the first row corresponds to feature IDs 0 to 9 and so on. Matrices in the left and right columns correspond to the correlation coefficients computed on the feature vectors extracted from 60 and 120 second windows, respectively. The correspondence of IDs to features are presented in Table 2.

our claim that groups with differing sizes have different interaction characteristics and might be more discriminative with different features. Another interesting aspect is how correlations of some features increase (or decrease) with the increasing window size, supporting that some dynamics of interactions are only captured in specific temporal resolutions. In the following paragraphs, we will analyse informative features (according to the correlation values) per group size in detail.

- **$D_2$** : Over the range of all window sizes, the correlation of the Z-axis of the accelerometer readings (ID 2) seems to have the highest correlation coefficients. Z-axis captures the forward-backward acceleration of the body. The high correlation value is not surprising, since in two person groups, people are expected to move synchronously. Couple of features that are easily noticeable in 200 second windows are the synchrony ratio of speaking and hand gestures (ID 32) and the median length and standard deviation of the hand gestures co-occurring between the participants (IDs 68 and 69). With the increasing window size, correlation coefficients of the features related to the co-occurrence of speaking and hand gesturing increase, pointing to a more involved interaction.

- **$D_3$** : It can be directly seen that there are four features with high coefficients that are consistent over different windows, IDs 46, 47, 49 and 71. The first three correspond to the Boolean Turn Activity features between streams of speaking and hand and head gesturing, more specifically the synchrony ratio and co-occurrence of these actions. Feature 71 is the mean length of head gestures occurring while the other person is speaking. Features with high correlations seem to be more representative of the listening behaviour, such as head nods occurring while the other participant is speaking. Features based on mutual information of the raw acceleration improve in correlation with the increasing window size, such as IDs 4, 5, 6, 9 and 10. This might be connected to the increasing variance in interaction characteristics, pointing to mimicry and synchrony emerging between the pairs within longer intervals.

- **$D_4$** : Features with relatively high correlations are sparse for $D_4$. The correlations of features 5 and 9, non-normalized and normalized mutual information between the Z-axes of the acceleration, seem to be comparatively higher than the rest for the window size of 60 seconds. Correlations for features 31 (co-occurrence of not speaking and not hand gesturing) and 63 (standard deviation of the length of the head gestures during speaking) marginally improve with the window size of 200 seconds. This collection of comparatively highly correlated features covers concepts both from $D_2$ (measures related to the Z axis of raw acceleration) and $D_3$ (measures between social actions of speaking and gesturing), and represents the dyadic interactions where both participants are active in conversation and the listener behaviour in a three person group, that might both occur in a group of four.

- **$D_5$** : Similar to $D_4$, there are not many features with high correlations for the window size of 60 seconds. However, correlation coefficients of various

features improve with the increasing window size. The most notable features are 11, 16, 17 and 20, which are all (normalized) mutual information measures between two speaking status streams, speaking-hand gesturing streams and speaking-head gesturing streams. As the group size increases, we see that many social action pairs are more strongly correlated.

- $\mathbf{D_6}$ : Comparatively high correlations are only present in the window size of 200 seconds, most notably features with IDs of 4, 8, 64 and 65. The first two are (normalized) mutual information values between the Y-axis, the right to left acceleration of the participant. The second two are the overlap statistics calculated from the streams of speaking and gesturing axes. This is the only group cardinality where the movement in the Y-axis has a higher correlation value than any other acceleration based measure. This can be related to the more frequent occurrences of listener-listener pairs in such large groups, where synchronized posture shifts are captured through side to side movements of the body.

- $\mathbf{D_7}$ : Unlike other large sized groups, we a see a few features with comparatively higher correlation coefficients at the window size of 60 seconds, such as 3, 7, 24, 42 and 48. This collection of features already covers different concepts, such as the mutual information between raw acceleration streams and boolean turn activity features between two speaking statuses, speaking status and head and head gestures. These features also either retain or increase their correlation values with increasing window size. There are also features with relatively higher negative correlation coefficients in both window sizes. Two examples are features 39 and 45, synchrony ratios of speaking with hand and head gestures. This result suggests that in larger groups, there might be multiple parallel interactions happening at the same time, resulting in pairs with non-synchronized social actions.

In summary, our proposed feature sets, perhaps apart from event synchrony, intrinsically carry information about the different aspects of the problem. In particular, for dyadic interactions measures between the raw acceleration readings are seen as desirable cues. Compatible with the observations of Section 4.6.3, even with the second level of information encoded in the social actions, raw acceleration still holds much information. Especially for groups with high cardinalities, measures that include gestures, especially head gestures, seem to gain importance, probably representing active listening behaviour.

### 4.6.5. Comparison of ensemble learning methods

GAMUT combines the prediction probabilities of the group based classifiers by training a meta-classifier (stacking). There are other options for combining predictions of multiple classifiers in the literature [42]. In this section, we compare the performance of GAMUT to two other ensemble learning techniques; maximum and mean fusion. In these methods, final prediction for a test sample is obtained by computing the maximum or mean of the probabilities provided by the multiple classifiers.

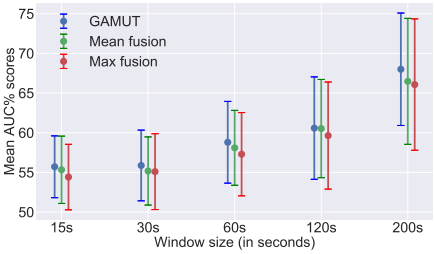


Figure 4.10: Performances of GAMUT, mean and max fusion

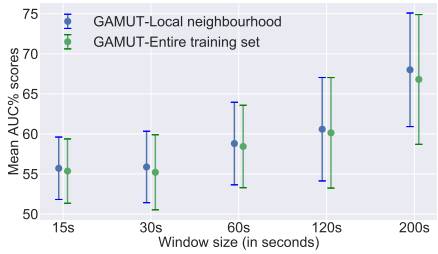


Figure 4.11: Performances of GAMUT with stacking on local neighbourhood and the entire training data



Figure 4.10 shows the performances of maximum and mean fusion in addition to GAMUT.

GAMUT outperforms both ensemble learning methods regardless of the window size. Since the performance difference between the methods are relatively low, we applied a paired t-test (computed on the results of 500 repetitions for each) to see if these differences are statistically significant. Apart from the mean fusion for 120s, all results are found to be statistically significant with $p < 0.01$. In other words, GAMUT guarantees better performance for almost all cases in comparison to other ensemble learning methods. When compared to Figure 4.5, we can see that even the mean and maximum fusion methods outperform logistic regression without group size based training and the method of [15]. Thus, we can conclude that, regardless of the combination technique, employing group based classifiers will always provide better performance than methods that ignore this.

### 4.6.6. Effects of using the local neighbourhood in meta-classifier training

GAMUT uses data only from the local neighbourhood of a test sample while training the meta-classifier, expecting the local neighbourhood to be more informative than the entire training set. This subsection investigates the effects of this setup by comparing the performances of GAMUT where the meta-classifier training is either performed on the local neighbourhood or the entire training set. Results are shown in Figure 4.11.

GAMUT with the local neighbourhood meta-classifier training always outperforms the cases where the entire training set is used. Paired t-tests for each window size showed that the differences are significant with $p < 0.01$. Still, the differences are not too high in terms of percentage. We believe this is caused by the dataset statistics and the random selection process used while forming the training and testing splits. Investigation of the distributions of group sizes in different runs have shown that in some runs, the number of samples in the training set coming from some group sizes are too low to be representative. In such cases, the probability of having representative samples in the local neighbourhood reduces and using the entire training set might prove to be superior. However, with more data, it will be possible to have a training set that includes enough samples from all group si-

zes. Then, the use of the local neighbourhood should be more optimal, providing a higher increase in the performance.

## 4.7. Conclusion and future work

### 4.7.1. Conclusion

We presented our study that focuses on the detection of pairwise F-formation membership in real life crowded scenarios. Our solution exploits a widely overlooked information source for this problem; the interaction dynamics. Instead of relying on spatial distance and orientation, our method is based on the interaction patterns between pairs of participants, inferred by a single tri-axial body worn accelerometer. The main idea was that two people in the same group should exhibit distinguishing interaction dynamics embedded in their movement and social actions, which will not be present in unrelated pairs.



We argued that the dynamics of interaction is expected to vary with respect to the cardinality of the group in which interaction is taking place. We hypothesised that a classifier that is trained on the data coming from a group of a specific cardinality should perform better for samples obtained from groups with the same cardinality. Our solution, GAMUT, was based on training multiple group size based classifiers. Final prediction of a new sample was then performed by combining the predictions of these classifiers by training a meta-classifier with the samples from the local neighbourhood.

Our proposed method was fully automatic; taking the acceleration readings as the input and providing the binary pairwise F-formation membership labels as output. We defined a new feature set (overlap statistics) and utilised some others that previously used in other domains for the joint representation of the participants interaction. They are extracted using raw acceleration and automatically classified social action labels.

We tested our approach on a real world mingling dataset, that includes groups of different sizes and various types of interactions between people. Our proposed method outperformed the state-of-the-art, methods without group size based training, and other ensemble fusion methods and guaranteed the best performance regardless of the window size.

We presented various analyses for further understanding. Performances obtained when ground truth labels are used showed there is room for improvement. We then focused on how different group size based classifiers $(C_k)$ perform on subsets of the data containing positive samples from a single cardinality, $D_k$. We saw experimental proof of our hypotheses, where all group sized based classifiers performed best on subsets with the same group cardinality.

Our experiments showed that feature sets extracted from raw acceleration and social action streams to be complementary. To further understand the contribution of the features, we analysed how individual feature vectors correlate with the ground truth labels. We have seen that the majority of the features, apart from event synchrony, had some correlation, suggesting that they are indeed informative. Weakly correlated features tended to differ for different group cardinalities,

further showing varying interaction dynamics of different sized groups.

GAMUT was shown to be superior to other ensemble learning techniques in terms of performance. The higher performance of these ensemble learning techniques compared to the approaches not considering group sizes further proved the importance of group size based training. Finally, when compared to using the entire training set, only using samples from local neighbourhood always provided better results. This suggests that samples with similar interaction characteristics and group cardinalities tend to be also closer in the feature space.

### 4.7.2. Future work

We believe there are still many possibilities for the improvement of the method. The analysis of the features showed that each group size based classifier has different optimal feature sets. This information can be exploited in the method, where the classifiers are trained with a subset of the features, automatically selected in the training phase. This way, redundant and weak features can be eliminated, providing group size based classifiers truly specific to one cardinality.

The main aim of the method was to provide pairwise F-formation membership. This information can be used as a starting point for creating a connectivity graph, that includes all the participants in the scene. While doing so, incorrect estimations of our method can be refined by introducing constraints related to the group membership and temporal consistency. A possible option is to use the posterior probability estimates that our method provides as edge strengths in the connectivity graph. There are already successful methods in the literature mainly used for optimizing connectivity graphs, that can be modified to be suitable for our problem formulation [27].

The size of the local neighbourhood used in GAMUT is set empirically per window size and for all the test samples the same neighbourhood size is used. A dynamic local neighbourhood selection step might be beneficial and considered as a future addition to GAMUT. In such an approach, the size of the local neighbourhood will be automatically inferred for each test sample, possibly with an informativeness criteria. This way, for each test sample, an optimal local neighbourhood reflecting the characteristics of the said sample can be used while training the meta-classifier.

## References

[1] C. Martella, E. Gedik, L. Cabrera-Quiros, G. Englebienne, and H. Hung, *How was it?: Exploiting smartphone sensing to measure implicit audience responses to live performances,* in *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference* (ACM, 2015) pp. 201–210.

[2] D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, *Modeling dominance in group conversations using nonverbal activity cues,* IEEE Transactions on Audio, Speech, and Language Processing **17**, 501 (2009).

[3] D. B. Jayagopi and D. Gatica-Perez, *Mining group nonverbal conversational*

*patterns using probabilistic topic models,* IEEE Transactions on Multimedia **12**, 790 (2010).

[4] H. Hung and D. Gatica-Perez, *Estimating cohesion in small groups using audio-visual nonverbal behavior,* IEEE Transactions on Multimedia **12**, 563 (2010).

[5] D. H. Wolpert, *Stacked generalization,* Neural networks **5**, 241 (1992).

[6] A. Kendon, *Conducting interaction: Patterns of behavior in focused encounters*, Vol. 7 (CUP Archive, 1990).

[7] T. Choudhury and A. Pentland, *Sensing and modeling human networks using the sociometer,* in *null* (IEEE, 2003) p. 216.

[8] J. P. Gips, *Social motion: Mobile networking through sensing human behavior*, Ph.D. thesis, Massachusetts Institute of Technology (2006).

[9] F. Setti, H. Hung, and M. Cristani, *Group detection in still images by f-formation modeling: A comparative study,* in *Image Analysis for Multimedia Interactive Services (WIAMIS), 2013 14th International Workshop on* (IEEE, 2013) pp. 1–4.

[10] S. Vascon, E. Z. Mequanint, M. Cristani, H. Hung, M. Pelillo, and V. Murino, *Detecting conversational groups in images and sequences: A robust game-theoretic approach,* Computer Vision and Image Understanding **143**, 11 (2016).

[11] X. Alameda-Pineda, Y. Yan, E. Ricci, O. Lanz, and N. Sebe, *Analyzing free-standing conversational groups: A multimodal approach,* in *Proceedings of the 23rd ACM international conference on Multimedia* (ACM, 2015) pp. 5–14.

[12] A. Kendon, *Movement coordination in social interaction: Some examples described,* Acta psychologica **32**, 101 (1970).

[13] T. L. Chartrand and J. A. Bargh, *The chameleon effect: the perception–behavior link and social interaction.* Journal of personality and social psychology **76**, 893 (1999).

[14] H. Hung, G. Englebienne, and J. Kools, *Classifying social actions with a single accelerometer,* in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing* (ACM, 2013) pp. 207–210.

[15] T. Kim, E. McFee, D. O. Olguin, B. Waber, A. Pentland, *et al.*, *Sociometric badges: Using sensor technology to capture new forms of collaboration,* Journal of Organizational Behavior **33**, 412 (2012).

[16] D. Chaffin, R. Heidl, J. R. Hollenbeck, M. Howe, A. Yu, C. Voorhees, and R. Calantone, *The promise and perils of wearable sensors in organizational research,* Organizational Research Methods **20**, 3 (2017).

[17] F. Setti, O. Lanz, R. Ferrario, V. Murino, and M. Cristani, *Multi-scale f-formation discovery for group detection,* in *Image Processing (ICIP), 2013 20th IEEE International Conference on* (IEEE, 2013) pp. 3547–3551.

[18] R. I. Dunbar, N. Duncan, and D. Nettle, *Size and structure of freely forming conversational groups,* Human nature **6**, 67 (1995).

[19] 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) **39**, 43 (2009).

[20] N. Eagle and A. S. Pentland, *Reality mining: sensing complex social systems,* Personal and ubiquitous computing **10**, 255 (2006).

[21] A. Madan, M. Cebrian, D. Lazer, and A. Pentland, *Social sensing for epidemiological behavior change,* Proceedings of the 12th … (2010).

[22] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell, *Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones,* in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (ACM, 2014) pp. 3–14.

[23] J. Gips and A. Pentland, *Mapping human networks,* in *Pervasive Computing and Communications, 2006. PerCom 2006. Fourth Annual IEEE International Conference on* (IEEE, 2006) pp. 10–pp.

[24] C. Cattuto, W. Van den Broeck, A. Barrat, V. Colizza, J.-F. Pinton, and A. Vespignani, *Dynamics of person-to-person interactions from distributed rfid sensor networks,* PloS one **5**, e11596 (2010).

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

[26] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, D. Tosato, A. Del Bue, G. Menegaz, and V. Murino, *Social interaction discovery by statistical analysis of f-formations.* in *BMVC*, Vol. 2 (2011) p. 4.

[27] H. Hung and B. Kröse, *Detecting f-formations as dominant sets,* in *Proceedings of the 13th international conference on multimodal interfaces* (ACM, 2011) pp. 231–238.

[28] F. Setti, C. Russell, C. Bassetti, and M. Cristani, *F-formation detection: Individuating free-standing conversational groups in images,* PloS one **10**, e0123783 (2015).

[29] T. Gan, Y. Wong, D. Zhang, and M. S. Kankanhalli, *Temporal encoded f-formation system for social interaction detection,* in *Proceedings of the 21st ACM international conference on Multimedia* (ACM, 2013) pp. 937–946.

[30] K. N. Tran, A. Gala, I. A. Kakadiaris, and S. K. Shah, *Activity analysis in crowded environments using social cues for group discovery and human interaction modeling,* Pattern Recognition Letters **44**, 49 (2014).

[31] L. Zhang and H. Hung, *Beyond f-formations: Determining social involvement in free standing conversing groups from static images,* in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016) pp. 1086–1095.

[32] H. Hung, G. Englebienne, and L. Cabrera Quiros, *Detecting conversing groups with a single worn accelerometer,* in *Proceedings of the 16th international conference on multimodal interaction* (ACM, 2014) pp. 84–91.

[33] L. Cabrera-Quiros, A. Demetriou, E. Gedik, L. van der Meij, and H. Hung, *The matchnmingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates,* IEEE Transactions on Affective Computing (2018).

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

[35] N. Ravi, N. Dandekar, P. Mysore, and M. Littman, *Activity recognition from accelerometer data,* AAAI , 1541 (2005).

[36] E. Gedik and H. Hung, *Personalised models for speech detection from body movements using transductive parameter transfer,* Personal and Ubiquitous Computing **21**, 723 (2017).

[37] E. Sangineto, G. Zen, E. Ricci, and N. Sebe, *We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer,* in *Proceedings of the ACM international conference on multimedia* (ACM, 2014) pp. 357–366.

[38] E. Delaherche, M. Chetouani, F. Bigouret, J. Xavier, M. Plaza, and D. Cohen, *Assessment of the communicative and coordination skills of children with autism spectrum disorders and typically developing children using social signal processing,* Research in Autism Spectrum Disorders **7**, 741 (2013).

[39] R. Q. Quiroga, T. Kreuz, and P. Grassberger, *Event synchronization: a simple and fast method to measure synchronicity and time delay patterns,* Physical review E **66**, 041904 (2002).

[40] M. Schmidt, N. L. Roux, and F. Bach, *Minimizing finite sums with the stochastic average gradient,* arXiv preprint arXiv:1309.2388 (2013).

[41] D. Wyatt, T. Choudhury, and J. Bilmes, *Conversation detection and speaker segmentation in privacy-sensitive situated speech data,* in *Eighth Annual Conference of the International Speech Communication Association* (2007).

[42] D. M. Tax, M. Van Breukelen, R. P. Duin, and J. Kittler, *Combining multiple classifiers by averaging or by multiplying?* Pattern recognition **33**, 1475 (2000).

# 5

# Estimating self-assessed personality with wearable sensing

*Nobody phrases it this way, but I think that artificial intelligence is almost a humanities discipline. It's really an attempt to understand human intelligence and human cognition.*

Sebastian Thrun

This chapter is published as:
L. Cabrera-Quiros*, E. Gedik*, and H. Hung. **Estimating self-assessed personality from body movements and proximity in crowded mingling scenarios**, *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016. (*:Equal contribution)

(a) (b)

Figure 5.1: Example snapshots of mingling events (a) a less crowded event taken from [2], (b) the more crowded mingle event from our scenario.

## 5.1. Introduction

In the past 15 years, the automatic recognition of displayed personality has received increasing interest due to the pursuit of intelligent systems that can adapt to every individual [1]. In this paper, we focus on crowded mingling events, such as cocktail parties (see Figure 5.1), which are intriguing scenarios for investigation due to their dynamic nature, the large number of simultaneous interactions, and the varied goals of each individual.

Specifically in the domain of self-assessed personality recognition during dynamic face-to-face social interactions, many previous applications focused on scenarios with a pre-defined task (eg. meetings). Also, the majority of prior work studied scenarios with a limited number of simultaneously occurring social interactions (generally just one such as meetings [3]), and certainly lower than 5 [2]. In contrast, in this paper, we investigate a scenario with on average over 15 simultaneous interactions occurring freely and dynamically.

Furthermore, audio-visual approaches are predominant in the field for predefined task scenarios due to the low number of people involved [1]. The characteristics of such scenarios enables them to set up several cameras , typically directed frontally or near frontally on participants, and microphones that capture relatively clean audio data.

However, for crowded mingle events, the visual boundaries between persons become harder to discriminate in the video and the noise of the event itself makes the extraction of speech features robustly from audio more challenging. Moreover, one could imagine that recording each individual's voice could have higher perceptions of privacy invasion. Thus, wearable devices are an appealing alternative, as they inherently encapsulate the sensor data of a single individual and are pervasive enough to avoid disturbing normal behaviour and easily scalable to larger populations.

In this paper, we present an approach to automatically recognize the self-assessed personality traits from the HEXACO inventory using an accelerometer and proximity sensors embedded in a wearable device hung around the neck. HEXACO is a personality inventory which includes items analogous to the well known Big-Five [4]. In addition, HEXACO also includes the trait for Honesty-Humility, which measures sincerity, fairness, greed avoidance, and modesty.

Our main contributions in this study are: (i) we address the problem of classifying self-assessed personality recognition in more complex and crowded mingle scenarios than previous work, where several social interactions are occurring dynamically; (ii) our approach is solely based on sensors that can be embedded in a wearable device which makes it easily scalable, and (iii) we propose a reliable approximation of speaking status from acceleration using a transfer learning approach, resulting in improved recognition performance even when fusing cues from two *behavioural* modalities originating from a single *digital* modality.

## 5.2. Related work

Here, we focus our discussion on works estimating *self-assessed* personality, although many efforts have been made in automated third-party attribution-based personality recognition [5]. There has also been much work focused on personality estimation in social media, which is also beyond the scope of this paper. A comprehensive review of the related personality computing literature can be found in the review by Vinciarelli and Mohammadi [1]. Within the domain of automated self-assessed personality estimation, works can be grouped mainly into those considering meetings and mingle scenarios.



As an example of the meeting setting, Pianesi et al. [3] proposed a method to recognize Extraversion and the Locus of Control during multi-party meetings of 4 people. The setting in this study has a pre-defined task and a controlled environment, where cameras and microphones were recording every participant individually. Batrinca et al. [6] presented a method to analyse self-presentations performed by participants in front of a camera during a Skype call, which simulated an interview, to recognize all traits in the Big-Five. Although they collected data for 89 people, they only interact with the interviewer for a part of the call while the main segment for non-verbal cue extraction was a monologue.

To the best of our knowledge, we are the first to address the complexity of crowded mingle scenarios using solely wearable devices. The closest work to our own was presented by Zen et al. [2]. In a considerably less crowded mingle event than ours, the authors proposed a classification method to recognize Extraversion and Neuroticism (from the Big-Five) using proximity related features extracted from multiple cameras. These features were motivated by findings from social psychology about the relationship between proxemics and the 2 personality traits in question. Compared to this work, with a total of 7 participants, we present a significant increase with experiments evaluated on 71 people. Finally, unlike their distance-based proximity features, ours rely on binary neighbour detection from simple a radio-based sensing mechanism in each wearable device (see Section 5.3).

## 5.3. Our data

We collected data during three separate 2-hour social evenings in a public bar-restaurant involving a real speed dating event followed by a mingle session.

During each event, between 30 and 32 different participants, with a total of 94 participants for the 3 events, were asked to use a wearable device hung around

their neck , which recorded triaxial acceleration at 20Hz. Each device communicated with other devices using a radio-based beacon communication by emitting its own ID to all other devices around it in a 2-3 meter radius, allowing them to synchronize with each other every second . The detection of a device is considered as a binary proximity detection.

A 30 minutes segment from the mingle was selected to maximize the number of people interacting. We used this part for the experimental validation in our paper. Due to hardware malfunction, only 71 of the devices recorded data during this segment. Finally, 5 GoPro Hero +5 cameras recorded the event from above. Note that the video data was only used to label the speaking status (ground truth) of 18 participants for 10 minutes to train our speaking status detector. This 10 minute segment was extracted in a non-overlapping part of the mingle from the 30 minute segment we used for testing. A snapshot of the event can be seen in Figure 5.1, where we contrast the density of our event with that used by Zen et al. [2].

Prior to the event, each participant filled in the HEXACO personality inventory [4], for which six dimensions are extracted: Honesty(H), Emotionality (E), Extraversion (X), Agreeableness (A), Conscientiousness (C), and Openness to Experience (O), by means of the HEXACO-PI-R survey [7].



## 5.4. Non-verbal cues

We can group our cues, originating from 2 digital modalities (wearable acceleration and proximity), into 3 behavioural modality categories: speaking turns, body movement energy, and proximity. A detailed description of each set of cues is presented below. Table 5.2 summarizes our derived features per cue type with a reference number.

### 5.4.1. Speaking turns

Building on prior findings that people's speaking status is representative of their personality [1, 3, 6], we extracted them from each individual's accelerometer signal. The use of this non-traditional modality to detect speech is motivated by the well-studied relationship between bodily gestures and speaking [8]. We have used a novel transfer learning method, Transductive Parameter Transfer (TPT) [9], which is experimentally shown to perform significantly better than a traditional machine learning approach. We hypothesize that TPT is much better in capturing the person specific nature of the connection between body movements and speech. Speaking turns are then used to extract high-level features representing the interaction characteristics of a participant.

#### Transductive Parameter Transfer (TPT)

For a feature space $X$ and label space $Y$, $N$ source datasets with label information $D_i^s = \left\{x_j^s, y_j^s\right\}_{j=1}^{n_i^s}$ and an unlabelled target dataset $X^t = \{x_j^t\}_{j=1}^{n_t}$ are defined. It is assumed that samples $X_i^s = \{x_j^s\}_{j=1}^{n_s}$ and $X^t$ are generated by marginal distributions $P_i^s$ and $P^t$, where $P^t \neq P_i^s$ and $P_i^s \neq P_j^s$. This approach aims to find the parameters

of the classifier for the target dataset $X^t$ by learning a mapping between the marginal distribution of the datasets and the parameter vectors of the classifier in the three following steps:

1. **Train source specific classifiers on each source set $\mathbf{D_i^s}$:** Instead of using the Linear SVM presented in [9], we have selected a L2 penalized logistic regressor as our classifier which is experimentally shown to perform better with our data. Chosen classifier minimizes Equation (1).

$$\min_{(w,c)} \frac{1}{2} w^T w + C \sum_{i=1}^{n} log(exp(-y_i(X_i^T w + c)) + 1) \tag{5.1}$$

   Thus, for every source dataset $D_i^s$, parameters $\theta_i = (w, c)_i$ are computed.

2. **Learn the relation between the marginal distributions $\mathbf{P_i^s}$ and the parameter vectors $\theta_i$ using a regression algorithm:** Training set $T = \{X_i^s, \theta_i\}_{i=1}^N$ is formed by samples $X_i^s$ and parameters $\theta_i$ obtained from each source dataset. A mapping $\hat{f} : 2^x \rightarrow \theta$, which takes a set of samples and returns the parameter vector $\theta$ needs to be learned. Assuming that elements in $\theta$ may be correlated, we have employed Kernel Ridge Regression [10], instead of the independent Support Vector Regressors used in [9]. Since we need to define the similarities between distributions $X_i^s$ instead of independent samples, we employ an Earth Mover's Distance [11] based kernel. EMD kernel is computed as:

$$\kappa_{EMD} = e^{-\gamma EMD(X_i, X_j)} \tag{5.2}$$

   In Equation (2), $EMD(X_i, X_j)$ corresponds to the minimum cost needed to transform $X_i$ into $X_j$. The user defined parameter $\gamma$ is set to be the average distance between all pairs of datasets.

3. **Use $\hat{f}$ to obtain the classifier parameters on the target distribution:** After computing $\hat{f}(.)$, we directly apply this mapping to target data $X^t$ to obtain $\theta^t$. With $\theta^t$ known, we can infer the labels for the target dataset.

#### TPT for extracting speaking turns

For detecting speaking turns with TPT, we selected simple statistical (mean and variance) and spectral features (power spectral density, using 8 bins with logarithmic spacing from 0-8 Hz as presented in [12]) that are expected to be representative of speech related body movement. These features were extracted from each axis of the raw acceleration, the absolute values from each axis of the acceleration, and magnitude of the acceleration using 3s windows with a 2s shift. Using the labelled data of 18 participants as sources, we obtained speaking turns for all 71 participants during 30 minutes. Finally, derived features were extracted from the speaking turns as shown in Table 5.2.

### 5.4.2. Body movement energy

For each wearable device, a single acceleration magnitude from the 3 axes is computed. Next, we apply a sliding window calculating the variance over the magnitude of the acceleration, using a 3s windows with a 2s shift (similar to Section 5.4.1). This gave us a better representation of *movement energy* over time than the acceleration magnitude. To obtain a single value for the 30 minute segment, we calculate 2 features to represent the movement energy; the mean and variance of the energy values in all windows. Finally, we create 2 multi-modal behavioural features from the mean and variance of the energy values in all windows during the detected speaking turns.

### 5.4.3. Proximity

As stated before, each wearable device has a binary proximity detector based on beacon communication with other devices. So, each device emits its own ID to all other devices and a detection of a particular ID is treated as a neighbour. From these binary detections, a dynamic (in time) binary proximity graph can be generated for each participant. To eliminate false neighbour detections, the method proposed by Martella et al. [13] was applied.

Then, 2 features were calculated for each participant from the proximity graphs: the largest size of group participated in and the total number of people interacted with during the event. Since we do not have actual distances, these features allow us to represent statistics related to the number of people's interactions during the event. To consider stable interactions in our proximity features, 2 nodes are only accounted as neighbours if they detect each other for more the one minute in the proximity graphs.

## 5.5. Experimental results

Table 5.1: Mean accuracy (%) $\pm$ std. error. M:Movement; S:Speaking turns; MS: Movement+Speaking turns; P:Proximity. Statistical significance against a random baseline is indicated:- $^{**}$($p<0.01$), $^{*}$($p<0.05$).

| | Concatenated Features Combinations | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **M** | **S** | **MS** | **P** | **M+S** | **M+MS** | **M+P** | **S+MS** | **S+P** | **MS+P** | **M+S+MS** | **M+S+P** | **M+MS+P** | **S+MS+P** | **M+S+MS+P** |
| **H** | $59 \pm 22$ | $66 \pm 17^{**}$ | $68 \pm 17^{**}$ | $44 \pm 12$ | $62 \pm 20^{*}$ | $\mathbf{69 \pm 15}^{**}$ | $47 \pm 20$ | $58 \pm 16$ | $57 \pm 14$ | $62 \pm 14^{*}$ | $58 \pm 18$ | $61 \pm 22$ | $63 \pm 13^{**}$ | $56 \pm 17$ | $62 \pm 18^{*}$ |
| **E** | $47 \pm 7$ | $43 \pm 13$ | $52 \pm 3$ | $52 \pm 3$ | $48 \pm 12$ | $48 \pm 7$ | $52 \pm 3$ | $45 \pm 13$ | $46 \pm 13$ | $52 \pm 3$ | $48 \pm 10$ | $46 \pm 13$ | $52 \pm 3$ | $49 \pm 11$ | $52 \pm 3$ |
| **X** | $52 \pm 12$ | $46 \pm 9$ | $48 \pm 12$ | $53 \pm 15$ | $51 \pm 4$ | $48 \pm 10$ | $59 \pm 17$ | $46 \pm 13$ | $50 \pm 12$ | $\mathbf{60 \pm 12}^{*}$ | $49 \pm 7$ | $51 \pm 7$ | $\mathbf{61 \pm 14}^{*}$ | $50 \pm 12$ | $54 \pm 9$ |
| **A** | $54 \pm 9$ | $52 \pm 10$ | $54 \pm 8$ | $55 \pm 14$ | $53 \pm 15$ | $55 \pm 6$ | $56 \pm 15$ | $53 \pm 17$ | $58 \pm 18$ | $59 \pm 15$ | $62 \pm 10^{*}$ | $53 \pm 12$ | $54 \pm 20$ | $60 \pm 15$ | $\mathbf{65 \pm 14}^{*}$ |
| **C** | $46 \pm 19$ | $49 \pm 19$ | $57 \pm 13$ | $46 \pm 8$ | $52 \pm 16$ | $55 \pm 13$ | $42 \pm 19$ | $56 \pm 12$ | $53 \pm 13$ | $50 \pm 13$ | $66 \pm 15^{**}$ | $55 \pm 14$ | $49 \pm 16$ | $55 \pm 20$ | $\mathbf{69 \pm 15}^{**}$ |
| **O** | $58 \pm 1$ | $56 \pm 5$ | $58 \pm 1$ | $\mathbf{69 \pm 17}^{*}$ | $55 \pm 9$ | $53 \pm 9$ | $63 \pm 17$ | $58 \pm 1$ | $66 \pm 14$ | $60 \pm 19$ | $53 \pm 13$ | $48 \pm 17$ | $65 \pm 18$ | $51 \pm 12$ | $56 \pm 19$ |

### 5.5.1. Performance of TPT on detecting speaking turns

First, we tested the performance of the TPT method against a traditional person independent machine learning approaches on the subset of 18 participants with labels for speaking turns. In this test, we used Leave-one-out cross validation. With the TPT method, each participant acted as target and all others acted as sources, once. For the traditional approaches, the other participants' data was concatenated to form the training set for each participant. Different linear (logistic regression)

Table 5.2: Summary of our features. S.T.= Speaking turns, E.T.=entire event

| | Feature | Modality |
|---|---|---|
| 1 | mean of accel. magnitude var. per window during E.T | Movement **(M)** |
| 2 | var. of accel. magnitude var. per window during E.T | |
| 3 | maximum length of S.T. | Speaking turns **(S)** |
| 4 | mean length S.T. | |
| 5 | variance of length for S.T. | |
| 6 | maximum length of non-S.T. | |
| 7 | mean length non-S.T. | |
| 8 | variance of length for non-S.T. | |
| 9 | total length of S.T. | |
| 10 | mean of accel. magnitude var. per window for S.T. | Movement + Speaking turns **(MS)** |
| 11 | var. of accel. magnitude var. per window for S.T. | |
| 12 | largest size of group interacted with | Proximity **(P)** |
| 13 | total number of people interacted with | |

Table 5.3: Correlations between selected features and traits ($p < 0.05$ for all correlations)

| Feature | 7 | 8 | 9 | 12 | 13 |
|---|---|---|---|---|---|
| H | -0.419 | -0.235 | 0.261 | x | x |
| X | x | x | x | 0.254 | 0.307 |
| O | x | x | x | -0.291 | x |

and non-linear classifiers (Hidden Markov Models and random forests) were used in the comparison. Paired one-tailed t-tests between performances (Area under the curve (AUC)) of these methods (Mean AUC for LR:58%, HMM:59%, RF:56%) and TPT (%65) showed TPT significantly $(p < 0.01)$ outperforms all of them. Compared to the implementation in [9], which yielded an average AUC of %60, our implementation provided significantly better results $(p < 0.05)$. These tests show that using TPT to extract speaking turns provides more robust and reliable results, which will allow us to have a proxy for speaking status without needing audio.

### 5.5.2. Feature-trait correlation

Table 5.3 shows the correlations of the features. In this Table, only those comparisons between features and traits with a significant value are summarized. For the trait of Honesty (H), those cues related with speaking turns tend to have an inverse correlation with the trait, suggesting that honest people may tend to be more vocal. Interestingly, all proximity features are directly correlated with the Extraversion (X) trait. This supports the impact of proxemics (management of spatial relationship and personal space) on this trait, as found by Zen et al. [2].

### 5.5.3. Classification of HEXACO traits

We treated the personality detection as a binary classification problem, where each item of the HEXACO inventory yielded one label for each participant, as positive or negative. This labelling is obtained by finding the median value for each item and placing participants with higher (and equal) values than the median in the positive class and the rest in the negative one. This labelling procedure resulted in fairly balanced class distributions. The distributions of the values for each item is shown

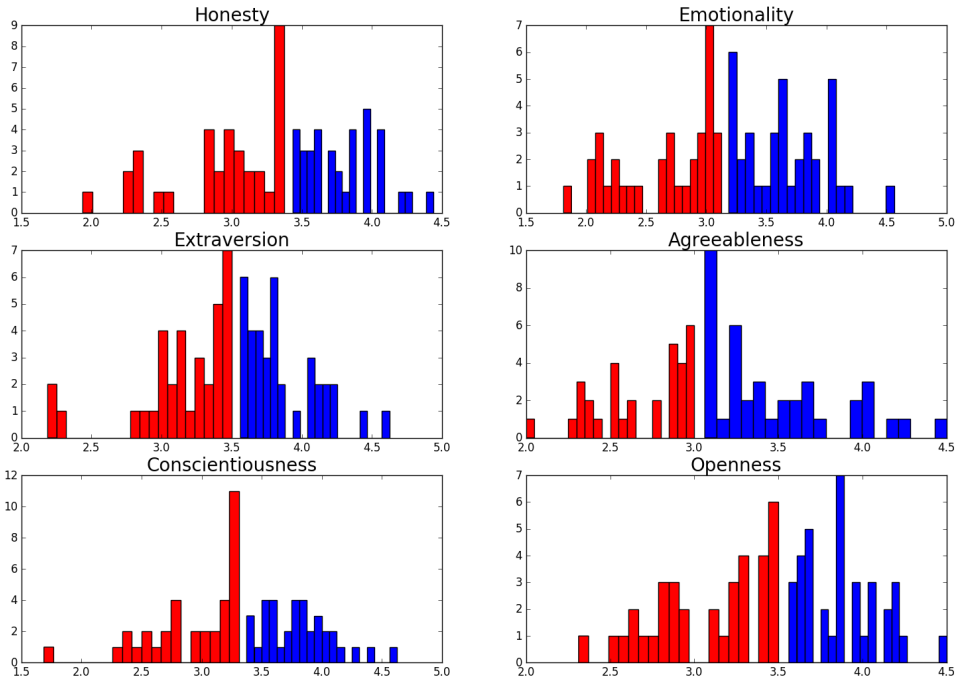


Figure 5.2: Distributions of personality scores per trait (Red: Negative class Blue: Positive class)

in Figure 5.2, where the red and blue parts correspond to negative and positive classes, respectively.

By extracting the features in Table 5.2, we obtained 71 samples with 13 dimensions each (when all the features were used). Since we have a low number of samples and feature dimensions, we selected the logistic regressor as our classifier. For performance evaluation, we have used 10-fold cross validation. The optimal regularization parameter C for the logistic regressor was set using nested cross validation. The accuracies obtained with this setup, for each item and with different feature combinations are provided in Table 5.1.

Table 5.1 shows that apart from Emotionality, we were able to classify items in the HEXACO inventory significantly better than a random baseline classifier. This random baseline classifier assigns all samples the most frequent label in the training set. To test significance for a given trait detection task, we applied a paired one-tailed t-test to the performance values of our method and the random baseline classifier which are computed from each stratified test fold.

From Table 5.1, it can be seen that using the multimodal feature set that includes all features (M+S+MS+P) provides the best general result where significant performances are obtained for three items: Honesty (H), Agreeableness (A) and Conscientiousness (C). For Honesty, significant results are obtained when speaking turn based features are in the feature set. This is quite interesting, when compared to the non-significant result obtained with just the movement energy features

since it shows that we were able to extract distinguishing information (that imitates another modality) from acceleration only. On closer inspection, we have seen that Feature 10, the mean of the acceleration magnitude variance in speaking turns, has the largest weight of all the features in the feature set M+MS.

Compatible with the correlation analysis of Section 5.5.2, we see that significant results for Openness (O) and Extraversion (X) are obtained with feature sets that include proximity based features. Significant results for Extraversion (X) are obtained when movement and proximity features are used together. This is most probably caused by the fact that extroverts tend to (i) interact with more people (which is captured by the proximity data), and (ii) to display more body movement energy. For Openness (O), using proximity based features only were enough to obtain significant results. The contribution of multimodality is more apparent for Agreeableness (A) and Conscientiousness (C), where satisfying results are only obtained by using all features (corresponding to different behavioural modalities but extracted from the same digital modality; acceleration) in combination and adding features from another digital modality (proximity) to this combination resulted in noticeable increased performance.



## 5.6. Conclusion

We presented a novel approach to recognize self-assessed personality during crowded mingling events using accelerometers and proximity sensors embedded in wearable devices. To the best of our knowledge, we are the first to address this complex problem using wearable devices alone and with such a high number of subjects in such a scenario. We also applied a novel transfer learning method, TPT [9], to our problem to extract reliable speech information from acceleration. This allowed us to have a proxy for speech in a noisy environment like a crowded mingle event and improve our performance by fusing cues from two behavioural modalities originating from the same digital modality. Our best performing traits were Honesty (H) with a 69% accuracy when using movement (M) in combination with speech-based movement (MS) and Conscientiousness (C) with 69% accuracy when using all modalities. When estimating all other traits, except for Emotionality, our method performed significantly above a random baseline. Finally, we show that adding the information from proximity and therefore exploiting multiple digital modalities increases the accuracy of almost all traits. A more detailed analysis of the contribution of the behavioural cues to the different personality traits is left for future work.

## References

[1] A. Vinciarelli and G. Mahammadi, *A survey of personality computing,* IEEE Trans. on Affective Computing (2014).

[2] G. Zen, B. Lepri, E. Ricci, and O. Lanz, *Space Speaks-Towards Socially and Personality Aware Visual Surveillance ,* MPVA (2010).

[3] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro, *Multimodal Recognition of Personality Traits in Social Interactions,* ICMI (2008).

[4] M. Ashton, K. Lee, M. Perugini, P. Szarota, R. De Vries, L. Di Blas, K. Boies, and B. De Raad, *A six-factor structure of personality-descriptive adjectives: Solutions from psycholexical studies in seven languages,* Journal of personality and social psychology (2004).

[5] O. Celiktutan, F. Eyben, E. Sariyanidi, H. Gunes, and B. Schuller, *MAPTRAITS 2014: The First Audio/Visual Mapping Personality Traits Challenge ,* ICMI (2014).

[6] L. Batrinca, N. Mana, B. Lepri, F. Pianesi, and N. Sebe, *Please, tell me about yourself: automatic personality assessment using short self-presentations,* .

[7] K. Lee and M. Ashton, *Psychometric properties of the HEXACO personality inventory,* Multivariate Behavioral Researc (2004).

[8] D. McNeill, *Language and gesture*, Vol. 2 (Cambridge University Press, 2000).

[9] G. Zen, E. Sangineto, E. Ricci, and N. Sebe, *Unsupervised domain adaptation for personalized facial emotion recognition,* in *Proceedings of the 16th International Conference on Multimodal Interaction* (ACM, 2014) pp. 128–135.

[10] K. P. Murphy, *Machine learning: a probabilistic perspective* (MIT press, 2012).

[11] Y. Rubner, C. Tomasi, and L. J. Guibas, *The earth mover's distance as a metric for image retrieval,* International journal of computer vision **40**, 99 (2000).

[12] H. Hung, G. Englebienne, and J. Kools, *Classifying social actions with a single accelerometer,* in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing* (ACM, 2013) pp. 207–210.

[13] C. Martella, M. Dobson, A. van Halteren, and M. van Steen, *From Proximity Sensing to Spatio-Temporal Social Graphs,* PerCom (2014).

# 6

# Predicting how live performances are experienced from crowd movement with wearable sensing

*The history of human thought recalls the swinging of a pendulum which takes centuries to swing. After a long period of slumber comes a moment of awakening.*

Peter Kropotkin

Parts of this chapter are published as:
C. Martella*, E. Gedik*, L. Cabrera-Quiros*, G. Englebienne, and H. Hung, **How was it?: Exploiting smartphone sensing to measure implicit audience responses to live performances**, *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*, 2015 (*:Equal contribution)
E. Gedik, L. Cabrera-Quiros, C. Martella, G. Englebienne, and H. Hung, **Towards Analyzing and Predicting the Experience of Live Performances with Wearable Sensing**, *Transactions in Affective Computing*, 2018

## 6.1. Introduction

Institutions that organise live performances (be it artistic, cultural or generic), increasingly require being able to quantify the response to the service they provide. Such quantification enables institutions to design more targeted events, make better monetary decisions and provide members of the public more polished and enhanced experiences. Quantitative data about audience response should eventually allow us to demonstrate the contribution of live performances to the lives of individuals and their well-being. While art and cultural events may appear to be a luxury to have in society, numerous studies have shown the benefits of such events for stimulating the social life of public spaces [1], health and mental well-being [2–5] and perceived quality of life [6]. In this paper, we investigate ways to automatically measure the response to live performances as a means of enhancement for both consumers and practitioners.

According to the appraisal theory, one's evaluation of a situation (in our case, the performance a person attends) causes related affective responses [7]. In other words, a person's appraisal of an event will be reflected, to an extent, in the emotional responses the person exhibits throughout the event itself. In this study, we present a method that uses this connection to detect an audience's appraisal of a live performance, based on the assumption that an audience's individual and jointly characterised body movements capture some form of affective response. We will be using a similar language used in implicit tagging literature [8] to distinguish between self reported evaluations of the event and immediate responses obtained through sensing. Questionnaire answers correspond to explicit responses provided by the participants, indicative of their reappraisal of the event. We use the term reappraisal since questionnaires are not immediate and filled after the event finishes. Sensors on the other hand, capture immediate responses and act as implicit cues for the appraisal of the event. We use the term implicit for evaluations obtained through sensing since it exploits the non-verbal reactions of the participant instead of direct responses. Thus, we aim to automatically predict explicit evaluations of participants through sensor recordings that captures their non-verbal reactions. We do not explicitly detect any affective tags or emotional states but we try to connect immediate body movements to explicit evaluations of the event.

The automatic detection of peoples' affective states is a widely studied topic in affective computing, with a majority of works in literature focusing on facial expressions [9] and/or speech [10]. However, most of these studies are generally conducted in lab environments and have restricted or controlled characteristics, both in terms of data acquisition (high-quality video and audio collection) and generation (posed facial expressions, carefully designed stimuli) when compared to the real-life performances we are interested in. The practical characteristics of real-world performances are different from the pre-designed lab experiments and introduce important restrictions on the use of aforementioned modalities. For example, robustly detecting audience members' facial expressions in a dark concert hall from video input is a challenging task. Luckily, recent studies have shown that body movements also convey affective expressions which might be exploited for the detection of emotional states [11, 12]. Even though most of the existing studies

investigating affective body expressions use either video [13, 14], motion capture [15] or pressure sensors [16], we hypothesised that it is still possible to capture enough of these body movements through the commonly available wearable sensors that are suitable for audiences in real-world settings.

We further realised that, in live performances, multiple people are simultaneously exposed to the same stimuli. This makes it possible to analyse and exploit the collective spontaneous response to the stimuli, as it has been shown that the link between multiple people's responses can be exploited to detect salient moments of movies using physiological sensing [17]. Building on these findings, we propose a novel method to measure the audience's collective response to live performances. In contrast to prior work that exploits fairly reliable but less pervasive biosignals or physiological sensing [18, 19], we hypothesise that individual and collective body movement patterns of audience members, as measured through the accelerometers, could also be used to measure affective responses to a performance. The proposed method exploits the linkage between audience members' body movement to detect distinctive time intervals in the performance. Individual movement patterns of participants in these distinctive parts are then used to classify the general evaluation of the performance.

By working closely for the last 2 years with Holland Dance (HD), an organization whose role is to promote dance in The Netherlands, we have identified some key challenges to measuring audience response in live performances:



**The limits of survey responses:** Organisers of live performances are always interested to gauge audience opinions about the performance they organise — if they enjoyed a performance or enjoy similar types of performances in general, they are more likely to recommend it to others, thus sustaining the popularity of the art form. Survey responses must be obtained after the performance, at a time when audience members are not necessarily eager to fill in questionnaires, and do not capture the audience's spontaneous response to specific moments of the performance. Note that the sentiment about a performance can also be assessed via social media, but this again requires audience members to actively participate in putting forward an opinion publicly [20].

**Obtaining implicit measurements on a large scale:** To our knowledge, most related work that tries to use implicit responses to visual stimuli such as movies [18, 19] or live performances [21] have tended to rely on physiological or brain activity measurements. While such signals are considered fairly reliable, the equipment to sense this data is still not particularly pervasive. As social norms would dictate, one tries to stay as quiet and still as possible when sitting and watching a live performance, making the measurements from pervasive sensors such as accelerometers less noisy and more meaningful.

**Obtaining detailed audience responses on a large scale:** Even when survey responses are available for a performance, typical Likert scale questions cannot provide detailed insights into what parts or aspects of a performance could have triggered someone to dislike or like it. One way to circumvent this problem involves using free text answers, which can provide richer — yet still incomplete — information about someone's experience, but these need to be manually proces-

sed. Interviews are another possibility and provide a very rich medium for those few who are willing to spend more time reflecting on their experience. They are, therefore, at best limited to an even smaller subset of an entire audience.

**Quantifying the impact of a performance on our social lives:** To our knowledge, most work focuses exclusively on measurements obtained during the performance, measuring direct responses to it. However, the value of a performance can stretch beyond this period and affect people long after they witnessed it. In particular, it could affect one's mood immediately after the experience or serve as a topic of stimulating discussion over drinks, thus leading to positive feelings about the entire performance and socialising experience. In an ideal case, its effects should last far beyond the performance itself, perhaps even providing lasting memories that are recalled collectively by friends. Identifying this is perhaps the greatest challenge but, if answered even in part, would provide a broader metric to quantify the value of arts and cultural events.

We address these challenges by making the following novel contributions in this study: we show, using multiple real-life events (two modern dance performances and a collection of music-art-technology presentations which has similar characteristics to a TED-X event), that (i) when people are watching a live performance, their spontaneous reactions result in body movements that can be captured with a standard acceleration sensor, (ii) some moments of spontaneous collective reaction correspond to memorable events of high affective output in the performance as can be verified by survey responses relating to the performance, (iii) audience members' reactions can be used to predict their enjoyment of the performance, whether they felt immersed in the experience, would recommend it to others, or thought dance performance changed their mood positively, (iv) the physical distance and joining the event with acquaintances might have an effect on the evaluation of the event, (v) the side neighbours of audience members can be approximated with an acceptable performance with proximity sensing, (vi) and finally by considering the social context that surrounds the activity of going to a live dance performance, we also provide initial results, using acceleration and proximity sensors, that suggest that a change in the mood of a person as a result of watching a live dance performance is reflected in their general body behaviour while mingling.



## 6.2. Related work

When the measurement of responses to a performance is approached from an appraisal theory angle, where affective responses are considered to be linked to the final evaluation [7], it becomes important to first investigate affect recognition itself. The automated detection of human affective behaviour has been gaining increasing interest, mostly because of its implications on better human-computer interaction and affect-related research including behavioural science, psychology, etc. A large number of studies have been published on this topic in the last decades [10]. Most of the early work on this area was focused on video (for detecting facial expressions) and/or audio inputs [9], and the datasets used in these early studies tended to use a single input modality [22], include a limited set of deliberate affective displays (six prototypical emotions) [23] and be recorded under highly constrained conditions,

generally covering exaggerated-artificial affective expressions [24]. More recent studies, on the other hand, generally aim to detect spontaneous affective displays [25], prefer to use multimodal information [26] and focus on detection of non-basic affective states [14].

With these new approaches in affective computing domain, cues other than facial expression and speech have started to gain importance, bodily expressions being one. The idea to use bodily expressions for affect detection is supported by existing work in social psychology literature that shows the strong connection between body movements and affective expressions [27, 28]. The decreasing cost and increasing availability of whole-body sensing technologies made it feasible to investigate the recognition of bodily expressions for affect perception and detection. This is reflected in the increasing number of studies that are discussed in recent surveys [11, 12] which rely on various approaches for capturing bodily expression such as computer vision [13, 14], motion capture [15] and pressure sensors [16], and generally aim to automatically map bodily expressions into well-known affective states. These affective states might be categorical (such as anger, happiness and neutral [29]) or continuous (valence and arousal [30]). Most of the datasets used in such studies include acted bodily expressions [29, 30], however, recent studies tend to focus more on non-acted, real life data [31]. The used methodology tends to be similar for most of the studies, where features (representative of bodily expressions) are extracted from sensor data, followed by the training of statistical models for automatic affect detection. Importantly, one key distinction between these and our approach is that while such studies explicitly focus on detecting affective states, our approach does not. Instead, we build on the results of these studies and show that spontaneous bodily movements captured by simple accelerometers (which may implicitly reflect affective states) are sufficient to infer one's experience of a live event. We do not try to discriminate between complex bodily movements or map them to affective states in our work.



Existing literature on the evaluation of events traditionally investigates the response of an audience to a live performance using self-reports, such as surveys and interviews [32, 33]. Digital technologies can overcome some limitations of surveys and interviews and give more direct and fine-grained insights into the response of an audience. For example, the explosion in popularity of social media such as Twitter, and mobile computing have broadened the borders of a live performance, as fans comment and post information and opinions live to the online community [34]. Practitioners are interested in measuring the activity of their audience in social media, both to understand their response and to leverage their activities as marketing tools for their performances [35, 36]. For example, some theatres, including Broadway, have experimented with so-called "tweet-seats" reserved for customers who promised to tweet about the performance live [37].

Other sensor technologies, albeit rather less pervasive, have also been used to overcome the granularity issues of surveys. For example, work in neuroaesthetics use fMRI scanning to relate viewer responses to the aesthetics of the performance [38–40]. Similarly, the tracking of eye gaze from video has been used when trying to distinguish novice from expert observers of dance [41]. Finally, physiolo-

gical sensing such as galvanic skin response (GSR) sensors have been investigated to measure the arousal of individuals watching a video of a dance performance, and its relationship with the individuals' self-reports [42]. GSRs have also been used to measure the response to other types of live performance, such as comedy [21] and movies in a cinema [19]. One specific example we would like to point is the work of Chenes et. al., where GSRs had been used to detect highlights in movie scenes [17]. The focus of this study was to exploit the inter-user physiological linkage which is calculated with simple sliding window correlation over pairs of participants' GSR readings. This study shows that when people are exposed to the same stimuli (even at different times), they tend to give synchronous physiological responses which can be used to detect salient parts of those stimuli. We build our study on a similar base where we hypothesize such linkage might be also computed with body movements, yielding a similar result.

These attempts show an increasing interest in quantifying the experience of live performances. Unlike these approaches, we advocate the use of pervasive sensors which are readily available in smartphones. As such, they enable less obtrusive measurements, on a massive scale, compared to those obtained via physiological sensing. This makes them much more readily deployable, and vastly increases their practical use.



In this work, we rely on acceleration and proximity sensors to measure people's reactions to live performance. These sensors have thus far been limited to other contexts, and have been used to measure very different phenomena. Specifically, most work that considers accelerometers and people addressed the problems of recognising daily activities such as walking, running, sitting, climbing the stairs [43], recognising daily household activities such as eating, drinking, vacuuming, scrubbing or lying down [44], and identifying modes of transportation taken [45]. There is a trend moving towards the detection of medically relevant events, such as fall detection [46, 47], but all of these approaches focus resolutely on physical activities where the behaviour can be represented directly by quite specific movements of the body.

It is possible to classify these types of activities with excellent performance, yet these activities are very different to analysing the response to a live performance. Little work exists where less specific body movements have been classified. For example, 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 [48], Hung et al. [49] used body movements to predict socially relevant actions with a device hung loosely round the neck or for detecting conversations[50]. Gedik et al. proposed a personalised solution for detecting speaking turns from acceleration in 2017[51]. Such works highlight the potential of measuring spontaneous bodily responses to external stimuli using more pervasive sensing.

Apart from focusing on different activities and tasks, the above mentioned works measure behaviour in environments that are far less challenging than a theatre, where the audience sits in silence and where the link between activity and behaviour is not as direct. The most similar work to our own was presented by Engle-

bienne and Hung [52] who found that they were able to identify audience members as professors and non-professors from their behaviour while attending an inaugural lecture. Although they were sitting, the small movements made in reaction to parts of the lecture demonstrated implicit responses of interest to particular moments and content delivered during the lecture. However, they did not analyse whether reactions from the audience to the lecture correlated with enjoyment of the lecture, for example. Another closely related work where the audience response was measured was presented by Bao et al. [53] who investigated how users watching movies on a tablet could have their implicit responses sensed by a wide variety of modalities from the tablet itself including the video, audio, tablet interactions, and accelerometer. In this case, movements from the tablet that the user was holding were used to gauge responses. Using a multimodal approach, they were able to predict the user's ratings of the movies they watched. However, in this case, the user sat alone to watch the movies and was not inhibited by the social norms usually adhered to in a public space.

Proximity sensors have been used to study the interactions between individuals with approaches more similar to complex network analysis. Cattuto et al. [54] used wearable sensors to analyse social interactions in crowded social settings, by means of proximity data collected through RFIDs. Martella et al. [55] used data collected by a series of wearable proximity sensors to identify the different communities attending a multi-disciplinary ICT conference. Roggen et al. [56] and Wirz et al. [57] proposed the usage of wearable sensors to discover spatio-temporal relationships between a number of individuals in the context of crowd dynamics. While these studies show that social relationship between individuals can be captured by means of spatio-temporal information, none of these works focus on the measurement of spatio-temporal relationship information in the context of live performances as we aim to do.



## 6.3. Data collection

### 6.3.1. Dataset 1: Dance performance

**The sensor set-up:** This study took place during a live dance performance that lasted almost an hour and a half without intermission. It consisted of mainly dancing, interspersed with monologues, in Italian, by the performers. The music was mainly based on live cello arrangements but also included pre-recorded songs. We recorded 41 participants watching the performance with triaxial accelerometers and IR cameras (for additional data verification). The accelerometers were located in a custom-made device hung around each participant's neck. These devices recorded acceleration at 20Hz and were kept synchronised to a global time through wireless network communication. Due to various hardware malfunctions, however, only 32 devices recorded acceleration data. In addition, the performance was recorded using a GoPro Hero +3 to analyse salient moments (i.e. favourite moments that were reported by the participants). We used ~79 minutes of sensor data in our experiments, starting just before the first piece, when all participants are seated and ending when the final piece of the performance finishes.

The custom-made device used in the study is also equipped with a wireless radio, which broadcasts the device's unique identifier (ID) every second up to a distance of some 2-3 meters. The reception of such broadcast by the devices nearby is considered a proximity detection. The device logs each detection on the on-board storage along with their timestamps. We used an energy-efficient MAC protocol [58] to allow the devices to communicate their IDs and detect each other's proximity.

**Survey responses:** All 41 participants filled in a questionnaire after the performance. These questionnaires consisted of 12 questions on four topics (three questions per topic), measuring "enjoyment", "recommendation (to a friend)", "immersion" and "mood changes". All questions used a ten-point Likert scale, where one means "I completely disagree" and ten means "I completely agree". For measuring "enjoyment", we adapted and selected questions presented in [59]; for "immersion", we selected involvement questions from the Igroup Presence Questionnaire [60]; for "recommendation" we used items from O'Brien's questionnaire [61]. Each of these questions was carefully chosen to measure each task and slightly adapted to match our scenario. We formed the questions regarding mood by ourselves. Given that the majority of the audience members were Dutch, we used a back-translation procedure to ensure that each questionnaire item was accurately matching the original English wording. This involved finding three different Dutch speakers to translate the questions from English to Dutch, then from Dutch to English and then from English to Dutch again ensuring that the finally chosen words best matched the original English. The complete set of questions asked in this questionnaire in both English and Dutch are listed in the Appendix. From the total number of participants, 32 responded with the Dutch questionnaire and 9 to the English one.



Of the 32 participants with working accelerometers, 25 reported a favourite moment of the performance. Two moments were particularly memorable: the *motorcycle sequence* was declared as favourite by 32% of the participants, and the *bolero finale*, favourite of 52% of the subjects. Note that in some cases participants declared more than one favourite moment.

### 6.3.2. Dataset 2: A day of Wonder

**The sensor set-up:** As a follow up to the first dance performance event, we organised a second study in the 'A day of Wonder' festival that took place at the TU Delft. This one-day festival is advertised as a combination of technology, music, food and art where different events take place in parallel, on different stages. Our study focused on one specific event that comprises two adjacent sets; namely 'Tales for the Curious Mind' and 'Enhancing Classical Music'. The first set included three presentations from various researchers and designers, who shared stories about their latest findings and inventions. The first presenter talked about a minimally-invasive surgical instrument, the second one explained his smart wedding dress with controllable LED lights and the final speaker introduced a micro air vehicle, a drone that weighs a mere 20 grams. The second set was an innovative classical concert experience, a lecture-performance focusing on enhancing the experience

of both performers and the audience using technology. This set started with a solo piano performance, followed by the talk of the performer and concluded with the classical music piece Zigeunerreisen, performed by a duo of violin and piano. The whole festival was free to attend and was open to the public, attracting a vast demographic of participants. Participation in the data collection was voluntary and participants were allowed to leave whenever they wanted. Limited seats were available, thus some of the participants were seated while others were standing.

Like the 'Dance Performance' dataset, participants wore our custom-made sensor pack hung around their necks, recording tri-axial acceleration and proximity information with the same setup (20 Hz and synchronised globally). A GoPro Hero +3 camera recorded the stage for further verification. We have treated the two sets, 'Tales for the Curious Mind' and 'Enhancing Classical Music', as two separate events. In total, 56 accelerometers are used in the experiments. After filtering non-valid data, either due to a technical problem, a participant leaving too early or lack of a questionnaire for the participant, we end up with valid data for 23 people in the first part and 21 in the second part. For the 'Tales for the Curious Mind', we used a ∼42 minutes interval in our experiments that started with the introduction of the presenters and ended with the final presentation. 'Enhancing Classical Music' lasted a bit shorter, totalling at ∼22 minutes.

**Survey responses:** After a participant left the event, we asked them to fill a questionnaire which had six questions, identical to the ones used in 'Dance Performance' for 'enjoyment' and 'immersion'. The same ten-point Likert scales were used. Questionnaires were taken separately for the two sets, 'Tales for the Curious Mind' and 'Enhancing Classical Music'. Thus, a participant joining only one of these events filled in the relevant questionnaire only. For the first set, 48% of the participants stated they really enjoyed the drone presentation (delfly) while 62% of them chose the 'real' presentation of the surgical device as the top moment.[1] For the second set, only 6 participants noted a favourite moment. This was same for everyone, which was the musical performance at the end of the presentation, the piece Zigeunerreisen.



## 6.4. Data analysis

In this section, we analyse the datasets in terms of shared experience and shared movement. Our assumption was that both the participants' subtle and more expansive movements are related to the experience of the event. We used the variance of the magnitude of the accelerometer readings to act as a proxy for the physical activity level of the participants. We calculate the variance in a sliding window of 2 seconds (40 samples) with 1 second shift (20 samples) to capture the subtle variations in motion while preserving a fine time scale. Before calculating the variance, z-score of the magnitude is computed to remove interpersonal differences. Then, for each dataset, we computed the Mutual Information (MI) of the variance of magnitude signals from every possible pair of participants, creating a pairwise co-occurrence measurement of the physical activity over time. These signals were

[1] Participants were allowed to choose multiple favourite moments.

computed over a sliding window with a size of 60 samples and shifted by one sample, resulting in a vector reflecting co-occurrence of motion, over time, between two participants.

For the first dataset, we also have the information of where people are seated. Thus, for this dataset we also provide an analysis of how spatial distance of people affects their movement patterns and evaluations and investigate ways of automatically obtaining neighbourhood information through proximity sensing. Before moving on to the analysis, we first explain how we obtain binary labels from the questionnaire answers, since binary labels for enjoyment are used for distinguishing between people who enjoyed the event or not in Section 6.4.2.

### 6.4.1. Binary labels for evaluation

As described in Section 6.3, we set up the questions to be redundant with three questions per aspect, which we averaged to obtain a single numerical value. This way, for each participant, we obtain four and two different labels each, for Datasets 1 and 2. We divided the participants into two classes for each task, corresponding to a "positive" and "negative" report on their experience of the performance. Participants whose averaged answer was below 5 for a task's three questions was placed in the negative class for that task, meaning the participant, respectively, did not enjoy the event, did not feel immersed throughout the performance (for both of our events), would not recommend the performance or did not think the performance uplifted their mood (only for the Dataset 1 ). The class distributions of all tasks for each event obtained with this setup are given below.



**Dataset 1:** For "enjoyment" and "recommendation", the majority of participants (26 out of 32) gave positive answers. 22 participants thought "the performance affected their mood positively". The distribution for the "immersion" task is relatively more balanced with 17 participants in the positive class.

**Dataset 2:** 21 out of 23 participants and 18 out of 20 participants gave positive responses to the "enjoyment" questions for the first and second parts. The imbalance for the "immersion" task was not as bad as the "enjoyment" task, where 16 out of 23 and 9 out of 20 participants responded positively for the first and second sets, respectively.

### 6.4.2. Dataset 1

There are three things we wanted to investigate for this performance: 1)Do moments when people move in synchrony correspond to salient moments of the performance? 2)Is the proximity of people in the audience a factor that also triggers synchronous motion and does it affect the evaluations? 3)Will it be possible to automatically identify sitting neighbours through proximity sensors?

#### Synchrony and salient moments

We hypothesised that salient moments should correspond to a high MI among all participants. We used an Otsu threshold [62] on the the mean pairwise MI of all possible pairs (computed as explained in Section 6.4) to select parts where co-occurrence of the physical activity is relatively high. Traditionally, Otsu thresholding

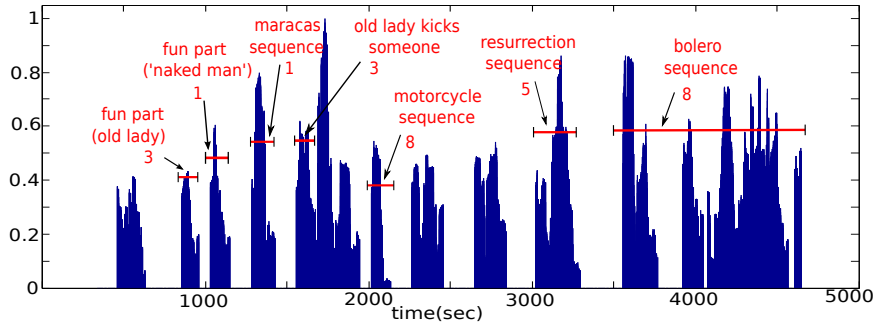


Figure 6.1: Mean co-occurrence measurement distance over time for all participants using Mutual Information (MI) for Dataset 1. Moments that were reported as salient are highlighted in red, together with number of times they were reported.

is used for converting grayscale images (continuous pixel values from 0 to 1) to black and white (binary). Since our MI values also lay between 0 and 1, we employed this method to detect moments of high physical activity co-occurrence. Figure 6.1 shows parts where the average MI for all pairs is more than the threshold, in blue, as well as the favourite moments reported, in red, together with their reporting frequency. Notice that all of the reported favourite moments show up in the MI, including the two moments declared as favourites for the majority of participants (*motorcycle* and *bolero finale*), and that most moments of high MI correspond to reported moments. This shows that memorable moments for people during these events can be captured by their coordinated movements, as they share the experience.



### Role of music

Second, the role of music during the performance is also interesting and we want to understand its effect, if any, on the collective behaviour of the participants. To do so, we looked at the sound intensity of the performance as obtained from the video and annotated song changes. Figure 6.2 shows the sound intensity of the performance (green) compared to the normalised co-occurrence measurement for MI (blue). The performance's songs are also highlighted in this figure in red. Here, it can be seen that although the music had a correlation with the response of the public in a performance in certain sequences, other moments of high mean MI are also correlated with acts with no music. This suggests that the music may not have been the only factor stimulating coordination between our participants. We decided, therefore, that the song changes would not be useful for the classification experiments (reported in Section 6.5) and focused solely on acceleration data instead.

### Impact of proximity

In this section, we analyse the impact of proximity in the enjoyment of the event. For this event, the participants were seated throughout the performance, making

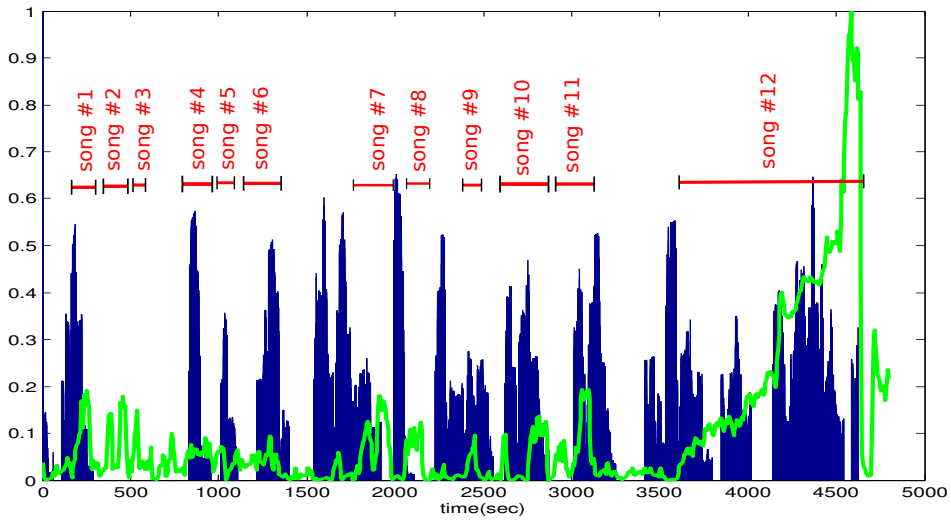


Figure 6.2: Sound intensity of the performance (green) compared against the normalised co-occurrence measurement calculated by MI (blue) for Dataset 1.



people's relative location static. During the setup phase, we identified where each participant was sitting during the performance and used this ground truth information for the analysis in this subsection. Figure 6.3 shows the mean MI (calculated over the whole event) between neighbouring participants (side, front and back neighbours). In addition, red subjects represents those who did not enjoy the event while the green ones did.

Figure 6.3 has 41 connections between neighbouring participants. Similar to the former analysis, MI between two people is considered low if the value is less than the Otsu threshold computed on all connected pairs. When all four neighbours are considered, there are 15 and 12 connections of high and low average MI between people who enjoyed the event, respectively. The values are 7 and 6 if only side neighbours are considered. Higher number of connections with high MI shows that proximity might have an effect on the evaluation but the low difference between numbers of high and low MI connections makes it harder to come up with hard conclusions.

We must also account for the people that came together to the event. The groups of participants that are known to come together to the event are shown in Figure 6.3 as dashed black lines. Although the pairwise MI and enjoyment of the event is comparatively high for some of the participants that came together, this does not generalise for all groups of acquaintances. Also, there are five cases where two participants shared a high MI but their enjoyment of the event differed. We hypothesise that such high co-occurrences in their movements are due to shared comments or other shared actions that had no relation with the performance. We cannot directly prove this hypothesis since we do not have video recordings of the audience.

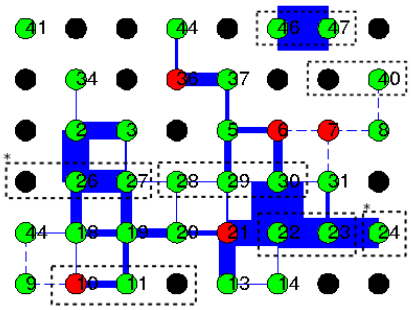


Figure 6.3: Mean MI between participants sitting together during the Dataset 1. Green dots indicate subjects who did enjoy the performance, red dots indicate subjects who did not, and black dots indicate empty seats (or people for which no data is available). The width of the blue bars indicate the average MI value throughout the performance, while dashed lines are non relevant MI relations.

### Identifying sitting neighbours

In this section, we investigate whether we can leverage the proximity data to identify who is sitting close to whom. Basically, we are trying to see if it is possible to construct a connectivity graph similar to Figure 6.3 automatically, using the proximity detections of our sensors. The proximity sensing is omnidirectional, however how the shielding effect of the body influences the detection of individuals sitting sideways, front or behind is unclear. Furthermore, even assuming neighbours can be detected, it is unclear how far they can be sensed and how this relationship can be characterised since no signal-strength is recorded by the sensors.

To start, one would assume that the closer two individuals sit together, within the detection range of the sensor of 2-3 meters, the more frequently their nodes will detect each other. With this assumption, we investigate which neighbours are frequently detected through sensing by the following methodology:

1. For every node $u_{i,j}$ (participant sitting at row $i$ and column$j$), count how often each ID was detected over the duration of the event,
2. Keep top $K$ IDs as the candidate neighbours,
3. Check if these $K$ candidate neighbours correspond to:
   (a) 1-Hop side neighbours($u_{i,j-1}$,$u_{i,j+1}$ )
   (b) Front and back neighbours ($u_{i-1,j}$,$u_{i+1,j}$)
   (c) 1 and 2-Hop side neighbours ($u_{i,j-1}$,$u_{i,j+1}$,$u_{i,j-2}$, $u_{i,j+2}$)
   (d) Diagonal neighbours ($u_{i-1,j-1}$, $u_{i-1,j+1}$, $u_{i+1,j+1}$ ,$u_{i+1,j-1}$)

For evaluating a and b, $K$ is selected to be 2 and for c and d, 4. Frontal and diagonal neighbours yield low recalls of $0.37$ and $0.24$ respectively, while 1-hop neighbours yield precision of $0.62$ and recall of $0.86$. When we add also 2-hop neighbours, we obtain a precision of $0.59$ and a recall of $0.84$. These suggest that

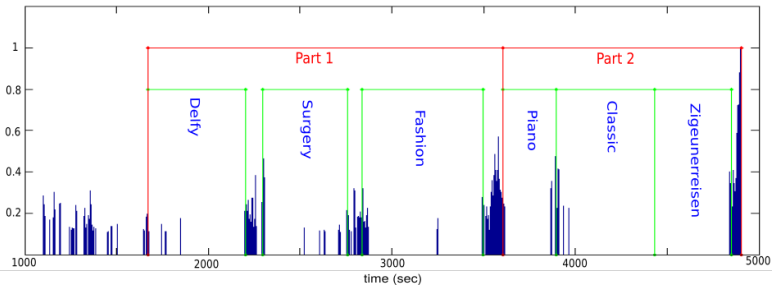


Figure 6.4: Mean co-occurrence measurement distance over time for all participants using Mutual Information (MI) for Dataset 2. The 2 main sets of the event are highlighted in red and the talks in green.

some of the neighbours detected for the 1-hop neighbours (with $K = 2$) are 2-hop neighbours (lowering the precision), but 2-hop neighbours are not consistently detected such that precision and recall are still similar with $K = 4$. The other source of error in precision in both cases are the rare detections of frontal and diagonal neighbours, which are not detected consistently but sometimes appear in the top-K list for some individuals.

To conclude, it is not possible to satisfactorily detect diagonal, front and back neighbours through proximity sensing. However, the precision and recall values obtained when classifying 1-hop and 2-hop neighbours show that it is possible to detect who is sitting at the sides of an individual with some sampling of frontal and diagonal neighbours. This information is valuable in analysing events where people are seated but the seating arrangement is unknown.

### 6.4.3. Dataset 2

Same analysis of synchrony and salient moments in Section 6.4.2 was carried out for this dataset. Figure 6.4 shows the mean MI among all participants along with the separations between the sections of the event (parts and talks). One key difference between the two datasets is their structure. Dataset 1 is collected in a continuously flowing event with smooth transitions between scenes, whereas the Day of Wonder has clearly delimited talks on different topics. This structure of the event can be clearly seen in Figure 6.4, where after each talk a high MI value is observed. These correspond to the rounds of applause by the audience and possible relocations of people between talks. This behaviour was not present in the Dance Performance as that event only had a single round of applause at the end of the performance. We also see the highest peaks between the two talks and after the second talk ends. People were allowed to leave at these points, corresponding to global high co-occurrences of physical activity.

However, different than the same analysis of Dataset 1, we don't see that many peaks during the talks. There are possible factors related to the nature of the event that can cause this. First of all, the crowd in this event was a mix of seated and standing people. This might cause an overall drop at the global pairwise MI,

since the volume of the reactions of seated and standing people are expected to be different. Secondly and more importantly, there are many parts where everyone in the audience reacts, such as the ending of the talks. Such parts are shown to be have high global MI, and they might suppress co-occurring subtle responses to the event (representative of salient moments in the talks) by increasing the threshold. So, if the aim is to find salient moments in an event like this, moments like applause or people leaving should be excluded from the analysis.

## 6.5. Immediate effects: Analysing the performance

We investigate on both datasets, whether it is possible to predict questionnaire responses about the performance from data collected with wearable sensors. The participants of Dataset 1 were seated and watched an actual dance performance, while the participants of Dataset 2 were not forced to sit and could come in and leave at any time. In the following sections, we perform classification experiments, where we present our methodology for automatically predicting a participant's evaluation of the events.

### 6.5.1. Classifying experience

#### Methodology

To emphasise the connection between the information contained in the motion data and the participants' experience of the event, in our classification experiments we focus on a simple set of features and a well-understood classifier. Our features are the variance of acceleration along each axis and of acceleration magnitude and our classifier is a Linear Support Vector Machine (SVM [63]). Since the number of samples is limited, we opted for a model with few parameters. We evaluated the performance of our method with leave-one-participant-out cross validation. The hyperparameters of the SVM are selected using nested cross validation on the training set. Variance values of each window are treated as a features, resulting in high dimensional feature vectors. Since we do not expect all intervals to be equally informative, we used a feature selection filtering approach which selects the features corresponding to informative intervals. The steps of feature extraction, feature (interval) selection and classification are presented below:

**Feature Extraction**

1. For each participant, compute variances of the X,Y,Z axes and the magnitude of the acceleration using a sliding window of 2s with 1s shift, resulting in 4 features representing a specific time window of 2s.

2. Concatenate computed variance values for each window to obtain a single representation (feature vector) of the whole event, per participant.

**Feature (Interval) Selection**

1. Compute Dynamic Time Warping(DTW) [64] values over the variances of the magnitude of accelerations (computed with a sliding window of 2s with 1s

shift) for each possible pair with a new sliding window. Each sample included in this new window corresponds to a specific interval of 2 seconds. So, if 5 sample windows are used in the DTW computation, each computed DTW value corresponds to 20 features in the feature vectors (4 features per each 2s window.)

2. Obtain a threshold with OTSU using all computed DTW values.

3. Select windows where the computed DTW scores are higher than the OTSU threshold.

4. For each participant, keep the features that correspond to the selected windows. For example, if 3 windows are selected in this phase, with DTW computed over 5 samples, resulting feature vector for each participant will have a dimension of 60.

**Classification**

1. For further dimensionality reduction, apply principal component analysis (PCA) on the feature vectors. Keep the principal components which preserve the 99 percent variance of features.

2. For each participant $p$:

   (a) Train a Linear SVM using the feature vectors of $n-1$ participants where $n$ is the total number of participants, excluding $p$.

   (b) Classify the feature vector of $p$.

Our assumption for feature selection is that the intervals with significantly high average DTW distances between each pair are more discriminative than the rest. In an ideal scenario, intra-class distances should stay relatively stable throughout the event, so the parts where the average DTW distance between pairs is high correspond to intervals where the inter-class distances are maximised. We expect this metric to provide better discrimination between classes than mutual information, where windows with high mutual information would correspond to moments where the classes would be almost indistinguishable, but requiring better synchronicity in people's movements. Empirical results using MI supported this claim, with performance scores significantly lower than the proposed method for the majority of the tasks.

Features (variance values) are computed over 4705, 2503 and 1293 windows for the Dataset 1 and Dataset 2 Parts 1 and 2, respectively. Each window corresponds to an interval of 2 seconds, where 4 features are extracted. The number of remaining intervals after feature selection depends on the window size for the computation of the DTW values, where we experimented with window sizes ranging from 1 sample to 80 samples, each with a 1 sample shift. For Dataset 1, the number of selected intervals ranged from 44 to 1065. For the first and second parts of Dataset 2, number of selected intervals ranged from 166 to 802 and 55 to 935.

After the PCA, dimensions of the feature vectors used in the classification experiments of Dataset 1 ranged between 18 and 28, whereas the range for Dataset 2 was 15 to 22.

**Results and discussion**

In this subsection, the performances obtained for both datasets will be presented and discussed in detail. Table 6.1 reports the performance results for both datasets, for different window size selections and using the thresholded DTW distance for pre-filtering salient intervals. This table also includes the performance scores obtained without interval selection. The results that are statistically significantly better than using the whole event are indicated with an asterisk. Significance was computed using an asymptotic McNemar's test with misclassification costs that are inversely proportional to the class distributions. As mentioned earlier, the class distributions are highly imbalanced in our dataset. In order to avoid artificially inflating our results by favouring the most common class, we used weighting in the training of our Linear SVM where the samples are weighted inversely proportional to the class frequencies.

**Dataset 1:** The first thing we see that, without interval selection, the results (final row of Table 6.1) are generally unsatisfactory. Any task other than predicting recommendation has a balanced accuracy score at, or below, chance level. We should note that we did apply PCA to the feature vectors for the non-filtered method. These scores showed that, without interval selection, PCA requires many more components to keep the same amount of variance in order to model the many non-informative intervals, supporting our claim of interval selection is necessary.

We were able to get perfect classification results for the "Enjoyment" task when performing interval selection, with window sizes ranging from 1 to 20 samples. In addition, all other window sizes still yielded significantly better performance ($p<0.05$) compared to using the whole event or to a random classifier. It can be seen that the performance tends to drop with increasing window size, suggesting a small window size might be more suitable for detecting enjoyment. Further supporting this claim, using data from the whole event fails to give results better than random, strengthening the decision of pre-filtering with DTW. Even though computing DTW over single-sample windows might sound counter-intuitive, the filtering approach is still able to find informative intervals. This works, probably because even a single sample has temporal information, since its value is extracted from a 2 second window.

Results for the Recommendation task show similar characteristics to the task of Enjoyment. Perfect classification, significantly better than the whole event approach ($p<0.05$), is achieved with window sizes of 5 and 10 and the performance tends to drop with the increasing window size. Interestingly, using features from the whole event still provides performance better than random with a balanced accuracy of 65%. This might simply mean that recommendation can be inferred from the whole event with an acceptable performance but some parts of the event might be still more indicative, providing finer results.

The performance for the tasks of Immersion and Mood is relatively poor compared to others. These tasks are less immediately related to the performance itself,

| Method \ BAcc (%) | Enjoyment | Recomm. | Immersion | Mood |
|---|---|---|---|---|
| DTW IS(1 Sample) | 100** \| **48** \| *50* | 92* | 58 \| **58** \| *100*** | 46 |
| DTW IS(5 Sample) | 100** \| **50** \| *50* | 100** | 65* \| **65** \| *100*** | 47 |
| DTW IS(10 Sample) | 100** \| **48** \| *50* | 100** | 59 \| **68** \| *90** | 53 |
| DTW IS(20 Sample) | 100** \| **63** \| *50* | 92* | 65* \| **58** \| *90** | 56 |
| DTW IS(40 Sample) | 92** \| **53** \| *47* | 90* | 52 \| **71** \| *94*** | 47 |
| DTW IS(80 Sample) | 81** \| **48** \| *44* | 73 | 52 \| **68** \| *84** | 49 |
| Whole Event | 48 \| **48** \| *44* | 65 | 46 \| **68** \| *52* | 51 |

*(* → p<0.1) (** → p<0.05)*

Table 6.1: Prediction performances for both datasets. Scores for Dataset 2 parts 1 and 2 are shown in **bold** and *italic*, respectively as second and third values at cells of Enjoyment and Immersion.

and may be harder to report objectively. For Immersion, highest performance is 65 percent, obtained with 5 and 20 sample windows which is still significantly better than using the whole event ($p<0.1$). The performance for this task does not seem to be changing too much between 1 to 20 samples, and fluctuates between 58 and 65. However, similar to the former tasks, using larger windows result in poor performance, ultimately reaching 46% with the data from whole event. For the Mood task, highest obtained performance is 56% with a window size selection of 20 samples. Most of the other window sizes resulted in performances worse than random and the performance of the whole event (51%).



The changes in performance obtained with different window sizes for different tasks have some interesting implications. Optimal window size for each task tends to differ. This could suggest that some tasks are reflected in shorter time scales than others. However, it can be also seen that most tasks performed best when small to medium sized windows, indicating that large window sizes fail to capture the connection between participants' movements.

We experimented with computing DTW distances on the raw accelerometer magnitude signal instead of the variance over a window. This experiment resulted in performance scores that were worse than random for "immersion" and "mood". Highest balanced accuracy scores for tasks of "enjoyment" and "recommendation" were 58 and 68 percent, respectively. For all tasks, using the variance rather than raw signal in DTW distance computation resulted in relatively better performance. We can conclude that variance in acceleration is a useful feature, both as a feature for prediction and for the interval selection using the thresholded DTW distance. This is probably because the variance of acceleration reflects the *amount* of movement rather than the precise movement and its direction, leading to more robust recognition.

**Dataset 2:** As shown in Table 6.1, for the first part of the event, we were able to obtain better-than-random performance for both tasks, but the very limited number of negative examples make it impossible to make hard conclusions. The highest performance for the Enjoyment task was 63%, obtained with a window size of 20 samples. Compared to the balanced accuracy of 48% obtained with the whole event setup, this result supports the pre-filtering with DTW. However, all other window sizes failed to capture any meaningful information, providing either slightly higher or lower performances than a random baseline. This result is quite different than

the first dataset one where high performances are obtained with many different choices of window size, suggesting that optimal window size for a task might also change with the characteristics of the event. For Immersion task, window size choice seems to be quite arbitrary. Highest performance, 71%, is obtained with 40 samples. However, using features from the whole event also results in a balanced accuracy of 68% which is not significantly different than the best score. Thus,for the first part of this event, Immersion can be detected with an acceptable performance without requiring filtering.

Results are quite different for the second part of the event. For the Enjoyment task, most of the window sizes resulted in a balanced accuracy of 50%, showing the classifier fails to learn anything from the data and always favours the majority class. Multiple factors related to the characteristics of the dataset might have caused this. First of all, we only had 2 negative samples in the whole dataset. Even though it was also the case for the first part, having only one negative sample in the training makes classification extremely hard. We believe the negative samples for the first part were more informative than the second one, making it possible to obtain better performance. Secondly, the length of the second part, is the shortest of our all datasets. In order to capture a complex concept as enjoyment, temporally large data might be required. Finally, this part was the closing act. Even though majority of the people reported this part as one of their favourites, 1) there may be a memory effect in play, where people report the event that's most fresh in their mind as the favourite, and 2) movement patterns of people might tend to change when nearing the end of events, explaining the poor performance.



We were able to get perfect classification for the Immersion task with windows of 1 and 5 samples. Contrary to the first part, using the features from the whole event results in a balanced accuracy of 52% and the results with filtering are significantly better. This supports our claim that the optimal window size depends not only on the task, but also on external factors to the task.

These follow-up experiments with an event of differing characteristics show that whether people are standing or sitting does not really affect our capacity to analyse people's response to the event. Our proposed methodology still provides competitive results, even in the quite unruly, noisy, real-world situation of these festival-style event.

### 6.5.2. Further analysis of salient moments with respect to enjoyment

It is interesting to revisit the salient moments of the performance in the light of the classes identified in Section 6.4.1. For space reasons, we focus on the enjoyment task. The pairwise similarity measurements from the previous qualitative analysis were separated into two groups for each task: the ones who completely agreed with the statement and everyone else. For each group, we computed the same unified similarity measurement and obtained the salient moments as in Section 6.4. Since the goal is to assess the similarity of people within the same class, we focused on pairs within the class and left out pairs from different classes.

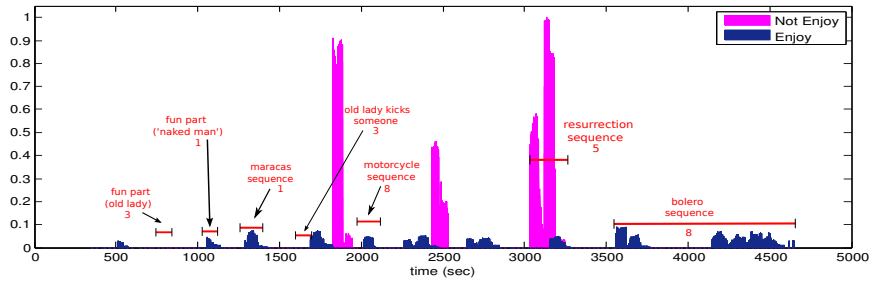


Figure 6.5: Salient moments for the Dataset 1 from mean MI discriminating people in class 'Enjoy' and 'Not Enjoy'

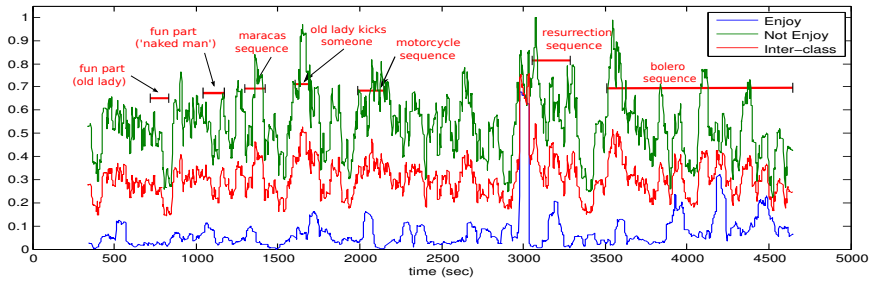


Figure 6.6: DTW distance using a sliding window for each class in the enjoyment task in the Dataset 2.

### Dataset 1

Figure 6.5 shows the measurements of mean MI over time (still using a sliding window, same as Figure 6.1) for both classes in the enjoyment task. Notice how the two moments considered as favourites for the majority of participants reappears for the mean MI in the group that enjoyed the performance but not for those who disliked it. Actually, there is almost no overlap between the salient moments for the classes. This reaffirms that specific acts or sequences in a performance can have a significant impact in the final assessment of enjoyment.

Furthermore, Figure 6.6 shows the mean DTW distance calculated over a sliding window (similar to Figures 6.1 and 6.5) for members within the 'Enjoy' class (blue), the 'Not Enjoy' class (green) and all pairs in opposing classes (red), separately. The 'Not Enjoy' class resulted in a higher overall DTW distance over the complete performance, compared to the 'Enjoy' class. This might indicate a lack of direct synchrony among people who dislike the performance, which echoes findings by Wang and Cesar with Galvanic Skin Response measures to an audience's reaction to a live performance [65].

Notice that there is a subtle difference between the measurements in Figures 6.5 and 6.6. The DTW distance measures the direct similarity between the two signals, while the MI measures the co-occurrence in movement between signals which also includes movements that are concurrently consistently different. This might explain why not in all the moments where the DTW distance is high, the MI

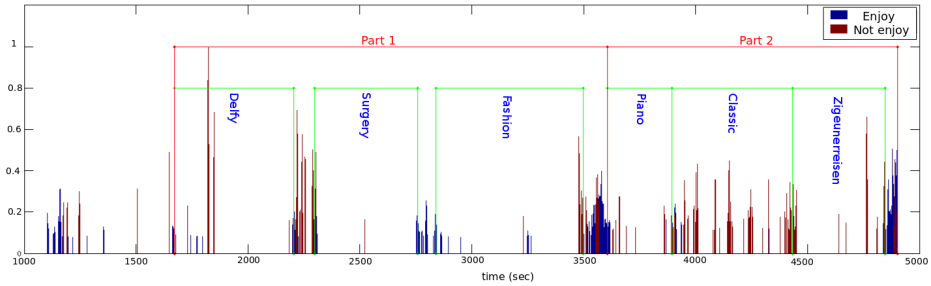


Figure 6.7: Salient moments for the Dataset 2 from mean MI discriminating people in class 'Enjoy' and 'Not Enjoy'

is low.

**Dataset 2**

Figure 6.7 shows a similar analysis of pairwise MI, separating the responses between people that did and did not enjoy the event. Due to the separation of this event into 2 parts, the mean MI over time of each group was calculated separately (for each part) and then concatenated for visualisation. The questionnaires have scores for each part but not separations between talks. Thus, a talk by talk analysis was not possible. Similarly to Figure 6.5, Figure 6.7 shows a difference between the MI from the people who enjoyed the event and those who did not. However, for this event, the peaks in pairwise MI do overlap for the classes implying that some parts of the event caused a mixed reaction. In addition, a difference between parts 1 and 2 can be seen for the movement of the people that did not enjoy the event. Overall, the MI peaks are more sparse in part 2 compared to part 1. As mentioned, this first part corresponds to 3 TED-X style talks, where the most of the dislike peaks are localised in the surgery talk, which might imply that the topic was unpleasant for these participants. On the contrary, part 2 corresponds to classical music presentations and the peaks are sparse within the performance. This suggests that the participants who dislike the performance might have an overall dislike of the classical music than those who did enjoy the presentation.

## 6.6. Delayed effects: Analysing social behaviour

There is clearly a context that surrounds the event itself — typically, people will attend a performance with friends and/or family, may come for a drink beforehand and stay for more afterwards. We hypothesised that people's social behaviour could also be affected by watching a performance. As a business model, HD was already co-organising networking events around dance performances together with two local networking organisations. The idea was that the dance performance could be an occasion to enhance the networking event, and the co-located networking event would encourage more people to watch dance.

Figure 6.8: Snapshots of the instrumented mingling room.

To investigate this hypothesis, we decided to investigate whether we could measure differences in how people socialized during the event. Hence, we measured mingling behaviour during two networking sessions, one right before a dance performance and another right after it. With HD and regional networking groups, we co-organised a networking event with 48 (35 of whom had valid data) volunteers, which were instrumented with the same sensing devices described in Section 6.3. An example snapshot of the mingling data is shown in Figure 6.8. Although a networking event differs in some ways from the more casual meetings that people might attend socially after live performances, we believe this initial investigation provides a feasibility study for larger-scale, less controlled studies in the future.

Similar to previous work using proximity sensors to analyse social behaviour in conferences [55], musea [54], and work-places [66], we used proximity sensors as proxies for face-to-face social interactions, together with accelerometer data.

### 6.6.1. Setup

Volunteers participated, in this order, to an initial networking session, a dance performance, and a second networking session. After the second session, people were asked to fill in a questionnaire. We conjectured that attending the dance performance may cause differences in 1) duration of the interactions, 2) the number of people that a person might interact with at a certain time, i.e., the size of interacting groups, and 3) how many people a person would interact with over the duration of the networking session. In addition, we evaluated whether we could predict self-reported changes in their mood from our participant's measured behaviour.

We measured whether people were interacting or not by processing the proximity detections (which is explained Section 6.3.1) collected by the devices as follows. We used a density-based filtering technique to increase the sensitivity of the signal for detecting face-to-face proximity [67]. For each pair of individuals, we computed the intervals where pairs were continuously facing each other, formally $[t_i, t_j]$ where $t_i$ is the timestamp of the first detection and $t_j$ is the timestamp of the last detection of the interval. Because pairs can be close for multiple time intervals

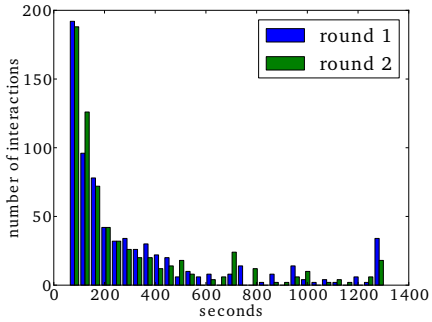


Figure 6.9: Distribution of the lengths of the interactions during the two rounds.

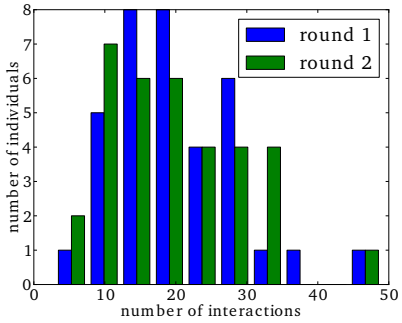


Figure 6.10: Distribution of the number of interactions for round 1 and round 2 across all the individuals.

during the same measurement, we computed multiple intervals for the same pair. Here, we refer to an interval of proximity between any pair as an *interaction*. For our experiments we considered only intervals of proximity longer than 60 seconds to indicate interactions.

### 6.6.2. Proximity-based results

Our first conjecture is that there might be a difference in the length or the number of interactions between the two sessions. In Figure 6.9, we present the distribution of the length of the interactions for the two networking sessions (from here on referred to as round 1 and round 2) across all the individuals. In both rounds shorter interactions are predominant. During both rounds, individuals often left a conversation to fill their glass and went back right afterwards to the same conversation, which would be measured as two distinct interactions. No significant mean difference was seen between the distribution in interaction length for the two rounds. In Figure 6.10 we present the distribution of the number of distinct interactions for round 1 and round 2 across all the individuals.

Our second conjecture is that the size of conversational groups might change. For example, people could be engaged in conversations involving more people, or conversely more one-to-one conversations, perhaps to discuss the content of the performance. We define a *neighbourhood* as the set of nodes a sensor $a$ detects at a given moment in time, i.e. the individuals in physical proximity of the individual wearing sensor $a$. In Figure 6.11 we present the distribution of neighbourhood size with respect to the amount of time they were observed together, expressed as a ratio over the round duration. In other words, it represents the amount of time individuals have spent in proximity to another $n$ individuals. The results show a peak around four individuals, a reasonable group size for a conversation. Similar to the interaction lengths, the two distributions look very similar.

The third hypothesis regarded changes in conversational partners. For example, people could be interacting with the same individuals as before the performance, or be stimulated to engage with others. To this end, for each individual we computed

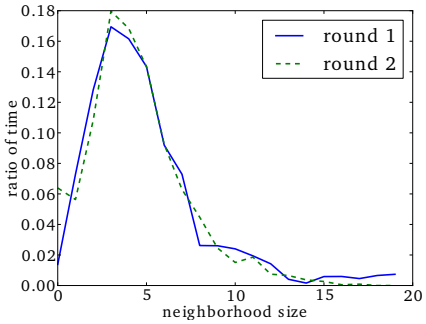


Figure 6.11: Relative amount of time sensors detected a certain number of other sensors (at a specific moment in time).

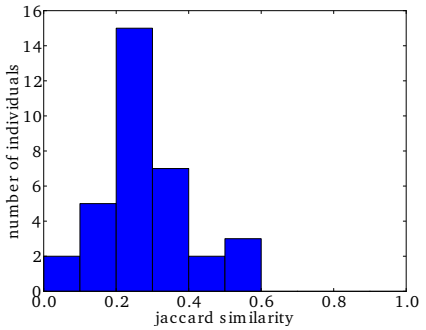


Figure 6.12: Distribution of the jaccard similarity across the individuals.

the *Jaccard similarity* between the set of participants an individual has interacted with during the two rounds. Given two sets of IDs $R_1$ and $R_2$, the Jaccard similarity function is defined as $J(R_1, R_2) = \frac{|R_1 \cap R_2|}{|R_1 \cup R_2|}$ and computes a value in the interval $[0, 1]$. Figure 6.12 presents the distribution of the Jaccard similarity across all individuals between round 1 and round 2. The results show that although the mingling pattern of the individuals did not change between the two rounds, they did interact with different individuals. In particular, they changed at least $50\%$ of their interaction partners between round 1 and round 2 (mean $0.278$ and standard deviation $0.121$).

### 6.6.3. Acceleration-based results

The image emerging from the above proximity measurements is that of an ordinary mingling event. Overall, these results indicate that the volunteers, as a group, applied a consistent pattern in their mingling behaviour during the two rounds, a pattern that they used, however, to target different conversational partners between the two rounds. The measurements picture a socialising context, but it is difficult to reach conclusions about the impact of the performance. For this reason, we focused on the acceleration data as well.

Similar to the direct approach, we used variance in the acceleration magnitude as the main feature. We then correlated the participant's self-reported behaviour with our findings. Correlation between the answers to question "Do you think the performance had an effect on your mood? Yes|No" and the difference between the acceleration magnitude variance in round 1 and 2 is computed. Since not all participants filled in the post event-survey and some accelerometers failed due to a firmware bug, we were only able to use the accelerometer data from 14 participants.

The variance values are extracted using the whole intervals for round 1 and 2. A statistically significant ($p = 0.02$) positive correlation value of 0.60 was obtained between the variance in acceleration and a self-reported effect on the mood. This correlation supports our hypothesis that the mood change can be linked to implicit behaviour as measured by acceleration. In conclusion, the results suggest that

while individuals acted similarly as a group in terms of networking behaviour captured by the proximity sensors, the quality of those interactions seemed different between the two sessions, as captured by the accelerometers.

## 6.7. Conclusions

In our study, we have investigated how an audience's perception of a performance can be perceived and measured from their body movements using an accelerometer that exists typically in smart phones. We have presented our results on two datasets which were collected during two live performances with different characteristics, both in terms of the performance itself and the audience demographics. Using findings from appraisal theory and affective studies that show how a stimulus creates an affective response and that response can be connected to experience, we analysed whether subtle and complex concepts — such as "enjoyment", "immersion", an improvement in mood as a result of the performance, and whether participants would "recommend" dance in general — would be reflected in the body motion measured by a simple accelerometer hung around the neck. Using the variance of the acceleration, we were able to predict the audience's self-reported experience in both events, in terms of aforementioned complex concepts.

Importantly, joint coordination in the variance in acceleration, the linkage in body movements of participants, helps to distinguish salient from non-salient moments of the performance. Knowing these leads to significant improvements over using each person's body movements from the entire performance period. We analysed how the spatial layout of a seated audience might affect its members' experience of the performance and presented a proximity-based method that can automatically detect neighbouring participants with satisfying performance.

As well as the obvious usefulness to the entertainment industry of such direct measurements of an audience's reaction, we have also made an attempt to measure the role that a live performance can have on the social behaviour that precedes and follows it. Our experiments shows huge promise in enabling us to measure the implicit responses of people while watching a live performance without the need for more traditional sensing approaches using physiological or brain signals. However, and perhaps more importantly, our experiments demonstrate the potential of quantifying the experience of 'a cultural night out', highlighting the relevance of the social context in moderating an individual's enjoyment of an event.



## 6.8. Appendix

**Post-event Questionnaire (English)**

1. The dance performance was interesting.(*)
2. The dance performance was exciting.(*)
3. The dance performance was enjoyable.(*)
4. I lost track of the world while I was watching the dance performance.(**)

5. I still paid attention to my surroundings while I was watching the dance performance.(**)
6. I was completely captivated by the dance performance.(**)
7. I will definitely want to come to another dance performance again.(***)
8. I will recommend dance performances to my friends.(***)
9. Dance performance was worthwhile.(***)
10. This dance performance uplifted my mood.(****)
11. This dance performance energized me.(****)
12. This dance performance made me feel more cheerful.(****)
13. It was natural for me to wear the sensors during the performance
14. Did you came with friends or family?
15. Did you have a favourite moment? If yes, please describe it.

**Post-event Questionnaire (Back-translated Dutch)**

1. De voorstelling was interessant.(*)
2. De voorstelling was opwindend.(*)
3. De voorstelling was aangenaam.(*)
4. Ik vergat de wereld om me heen gedurende de voorstelling.(**)
5. Ik had gedurende de voorstelling aandacht voor mijn omgeving.(**)
6. Ik was volledig in de ban van de voorstelling.(**)
7. Ik kom zeker terug voor een andere dansvoorstelling.(***)
8. Ik zal dansvoorstelling aan mijn vrienden aanraden.(***)
9. Dansvoorstellingen zijn de moeite waard.(***)
10. Deze dansvoorstelling heeft me opgebeurd.(****)
11. Deze dansvoorstelling heeft me een energetisch gemaakt.(****)
12. Deze dansvoorstelling eeft me blij gemaakt.(****)
13. De sensoren voelden gedurende de voorstelling niet onnatturlijk ann
14. Bent u met wrienden of familie gekomen?
15. Had u een favoriet moment? Zo ja, gelieve dit te omschrijven:

(*)Enjoyment [59], (**)Immersion [60], (***)Recommendation [61], (****)Mood.

# References


[1] W. H. Whyte, *The Social Life Of Small Urban Spaces* (Project for Public Spaces Inc, 1980).

[2] C. L. Nightingale, C. Rodriguez, and G. Carnaby, *The impact of music interventions on anxiety for adult cancer patients: a meta-analysis and systematic review,* Integrative cancer therapies **12**, 393 (2013).

[3] M. Ritter and K. G. Low, *Effects of dance/movement therapy: A meta-analysis,* The Arts in Psychotherapy **23**, 249 (1996).

[4] D. Fujiwara, L. Kudrna, and P. Dolan, *Quantifying and Valuing the Wellbeing Impacts of Culture and Sport*, Tech. Rep. (UK Department of Culture, Media and Sport, 2014).

[5] T. Nenonen, R. Kaikkonen, J. Murto, and M.-L. Luoma, *Cultural services and activities: The association with self-rated health and quality of life,* Arts & Health **6**, 235 (2014).

[6] A. C. Michalos and P. M. Kahlke, *Arts and the perceived quality of life in british columbia,* Social indicators research **96**, 1 (2010).

[7] K. R. Scherer, A. Schorr, and T. Johnstone, *Appraisal processes in emotion: Theory, methods, research* (Oxford University Press, 2001).

[8] M. Soleymani and M. Pantic, *Human-centered implicit tagging: Overview and perspectives,* in *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on* (IEEE, 2012) pp. 3304–3309.

[9] M. Pantic and L. J. M. Rothkrantz, *Automatic analysis of facial expressions: The state of the art,* IEEE Transactions on pattern analysis and machine intelligence **22**, 1424 (2000).

[10] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, *A survey of affect recognition methods: Audio, visual, and spontaneous expressions,* IEEE transactions on pattern analysis and machine intelligence **31**, 39 (2009).

[11] M. Karg, A.-A. Samadani, R. Gorbet, K. Kühnlenz, J. Hoey, and D. Kulić, *Body movements for affective expression: A survey of automatic recognition and generation,* IEEE Transactions on Affective Computing **4**, 341 (2013).

[12] A. Kleinsmith and N. Bianchi-Berthouze, *Affective body expression perception and recognition: A survey,* IEEE Transactions on Affective Computing **4**, 15 (2013).

[13] G. Castellano, S. D. Villalba, and A. Camurri, *Recognising human emotions from body movement and gesture dynamics,* in *International Conference on Affective Computing and Intelligent Interaction* (Springer, 2007) pp. 71–82.

[14] H. Gunes and M. Piccardi, *Automatic temporal segment detection and affect recognition from face and body display,* IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **39**, 64 (2009).

[15] M. Karg, K. Kuhnlenz, and M. Buss, *Recognition of affect based on gait patterns,* IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **40**, 1050 (2010).

[16] S. D'Mello and A. Graesser, *Automatic detection of learner's affect from gross body language,* Applied Artificial Intelligence **23**, 123 (2009).

[17] C. Chênes, G. Chanel, M. Soleymani, and T. Pun, *Highlight detection in movie scenes through inter-users, physiological linkage,* in *Social Media Retrieval* (Springer, 2013) pp. 217–237.

[18] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, *A multimodal database for affect recognition and implicit tagging,* IEEE Transactions on Affective Computing **3**, 42 (2012).

[19] J. Fleureau, P. Guillotel, and I. Orlac, *Affective benchmarking of movies based on the physiological responses of a real audience,* in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on* (IEEE, 2013) pp. 73–78.

[20] M. Thelwall, K. Buckley, and G. Paltoglou, *Sentiment in twitter events,* Journal of the American Society for Information Science and Technology **62**, 406 (2011).

[21] C. Wang, E. N. Geelhoed, P. P. Stenton, and P. Cesar, *Sensing a live audience,* in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems* (ACM, 2014) pp. 1909–1912.

[22] I. Cohen, N. Sebe, A. Garg, M. S. Lew, and T. S. Huang, *Facial expression recognition from video sequences,* in *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*, Vol. 2 (IEEE, 2002) pp. 121–124.

[23] T. Kanade, J. F. Cohn, and Y. Tian, *Comprehensive database for facial expression analysis,* in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on* (IEEE, 2000) pp. 46–53.

[24] Y.-L. Tian, T. Kanade, and J. F. Cohn, *Facial expression analysis,* in *Handbook of face recognition* (Springer, 2005) pp. 247–275.

[25] N. Sebe, M. S. Lew, Y. Sun, I. Cohen, T. Gevers, and T. S. Huang, *Authentic facial expression analysis,* Image and Vision Computing **25**, 1856 (2007).

[26] A. Kapoor and R. W. Picard, *Multimodal affect recognition in learning environments,* in *Proceedings of the 13th annual ACM international conference on Multimedia* (ACM, 2005) pp. 677–682.

[27] H. G. Wallbott, *Bodily expression of emotion,* European journal of social psychology **28**, 879 (1998).

[28] J. K. Burgoon, L. K. Guerrero, and K. Floyd, *Nonverbal communication* (Routledge, 2016).

[29] D. Bernhardt and P. Robinson, *Detecting affect from non-stylised body motions,* in *International Conference on Affective Computing and Intelligent Interaction* (Springer, 2007) pp. 59–70.

[30] A. Kleinsmith and N. Bianchi-Berthouze, *Recognizing affective dimensions from body posture,* in *International Conference on Affective Computing and Intelligent Interaction* (Springer, 2007) pp. 48–58.

[31] A. Kleinsmith, N. Bianchi-Berthouze, and A. Steed, *Automatic recognition of non-acted affective postures,* IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **41**, 1027 (2011).

[32] A. S. Brown and J. L. Novak, *Assessing the intrinsic impacts of a live performance* (WolfBrown San Francisco, CA, 2007).

[33] M. Reason and D. Reynolds, *Kinesthesia, empathy, and related pleasures: An inquiry into audience experiences of watching dance,* Dance research journal **42**, 49 (2010).

[34] L. Bennett, *Patterns of listening through social media: online fan engagement with the live music experience,* Social Semiotics **22**, 545 (2012).

[35] M. R. Lockstone, L. Olga Junek, S. Hudson, and R. Hudson, *Engaging with consumers using social media: a case study of music festivals,* International Journal of Event and Festival Management **4**, 206 (2013).

[36] A. Leask, A. Hassanien, and P. C. Rothschild, *Social media use in sports and entertainment venues,* International Journal of Event and Festival Management **2**, 139 (2011).

[37] *USA Today. Providence theater experiments with 'tweet seats',* (2013), http://www.usatoday.com/story/tech/2013/01/27/theater-tweet-seats/1868693/.

[38] B. Bläsing, B. Calvo-Merino, E. S. Cross, C. Jola, J. Honisch, and C. J. Stevens, *Neurocognitive control in dance perception and performance,* Acta psychologica **139**, 300 (2012).

[39] E. S. Cross, L. Kirsch, L. F. Ticini, and S. Schütz-Bosbach, *The impact of aesthetic evaluation and physical ability on dance perception,* Frontiers in human neuroscience **5**, 102 (2011).

[40] B. Calvo-Merino, C. Jola, D. E. Glaser, and P. Haggard, *Towards a sensorimotor aesthetics of performing art,* Consciousness and cognition **17**, 911 (2008).

[41] C. J. Stevens, H. Winskel, C. Howell, L.-M. Vidal, J. Milne-Home, and C. Latimer, *Direct and indirect methods for measuring audience reactions to contemporary dance,* Dance Dialogues: Conversations Across Cultures, Artforms and Practices: Proceedings of World Dance Alliance Global Summit 2008: Brisbane, 13–18 July, 2008 (2008).

[42] C. Latulipe, E. A. Carroll, and D. Lottridge, *Love, hate, arousal and engagement: exploring audience responses to performing arts,* in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (ACM, 2011) pp. 1845–1854.

[43] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, *Activity recognition using cell phone accelerometers,* ACM SigKDD Explorations Newsletter **12**, 74 (2011).

[44] L. Bao and S. S. Intille, *Activity recognition from user-annotated acceleration data,* in *International Conference on Pervasive Computing* (Springer, 2004) pp. 1–17.

[45] 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**, 13 (2010).

[46] C. Doukas, I. Maglogiannis, P. Tragas, D. Liapis, and G. Yovanof, *Patient fall detection using support vector machines,* in *IFIP International Conference on Artificial Intelligence Applications and Innovations* (Springer, 2007) pp. 147–156.

[47] 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**, 277 (2006).

[48] 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* (IEEE, 2012) p. 105?1?7114.

[49] H. Hung, G. Englebienne, and J. Kools, *Classifying social actions with a single accelerometer,* in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing* (ACM, 2013) pp. 207–210.

[50] H. Hung, G. Englebienne, and L. Cabrera Quiros, *Detecting conversing groups with a single worn accelerometer,* in *Proceedings of the 16th international conference on multimodal interaction* (ACM, 2014) pp. 84–91.

[51] E. Gedik and H. Hung, *Personalised models for speech detection from body movements using transductive parameter transfer,* Personal and Ubiquitous Computing **21**, 723 (2017).

[52] G. Englebienne and H. Hung, *Mining for motivation: using a single wearable accelerometer to detect people's interests,* in *Proceedings of the 2nd ACM international workshop on Interactive multimedia on mobile and portable devices* (ACM, 2012) pp. 23–26.

[53] X. Bao, S. Fan, A. Varshavsky, K. Li, and R. Roy Choudhury, *Your reactions suggest you liked the movie: Automatic content rating via reaction sensing,* in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing* (ACM, 2013) pp. 197–206.

[54] 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**, e11596 (2010).

[55] C. Martella, A. van Halteren, M. van Steen, C. Conrado, and J. Li, *Crowd Textures as Proximity Graphs,* in *IEEE Communications Magazine* (2014).

[56] D. Roggen, M. Wirz, D. Helbing, and G. Tröster, *Recognition of Crowd Behavior from Mobile Sensors with Pattern Analysis and Graph Clustering Methods,* Networks and Heterogeneous Media **6** (2011).

[57] M. Wirz, D. Roggen, and G. Tr?ster, *A Methodology towards the Detection of Collective Behavior Patterns by Means of Body-Worn Sensors,* in *Workshop at the 8th International Conference on Pervasive Computing (Pervasive 2010)* (2010).

[58] M. Dobson, S. Voulgaris, and M. van Steen, *Merging ultra-low duty cycle networks,* in *Proceedings of the 41st International Conference on Dependable Systems & Networks (DSN 2011)* (2011).

[59] E. W. See-To, S. Papagiannidis, and V. Cho, *User experience on mobile video appreciation: How to engross users and to enhance their enjoyment in watching mobile video clips,* Technological Forecasting and Social Change **79**, 1484 (2012).

[60] T. Schubert, F. Friedmann, and H. Regenbrecht, *The experience of presence: Factor analytic insights,* Presence: Teleoperators and virtual environments **10**, 266 (2001).

[61] H. L. O'Brien and E. G. Toms, *The development and evaluation of a survey to measure user engagement,* Journal of the American Society for Information Science and Technology **61**, 50 (2010).

[62] R. Gonzalez and R. Woods, *Digital Image Processing* (Prentice Hall, 2008).

[63] C.-C. Chang and C.-J. Lin, *LIBSVM: A library for support vector machines,* ACM Transactions on Intelligent Systems and Technology **2**, 27:1 (2011), software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[64] D. J. Berndt and J. Clifford, *Using dynamic time warping to find patterns in time series.* in *KDD workshop*, Vol. 10 (Seattle, WA, 1994) pp. 359–370.

[65] C. Wang and P. Cesar, *Do we react in the same manner?: comparing GSR patterns across scenarios,* in *Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational* (2014).

[66] A. Pentland, T. Choudhury, N. Eagle, and P. Singh, *Human dynamics: computation for organizations,* Pattern Recognition Letters **26**, 503 (2005).

[67] C. Martella, M. Dobson, A. van Halteren, and M. van Steen, *From Proximity Sensing to Spatio-Temporal Social Graphs,* in *Pervasive Computing and Communications (PerCom), 2014 IEEE International Conference on* (2014).

# 7

# Discussion

*The ultimate end of all revolutionary social change is to establish the sanctity of human life, the dignity of man, the right of every human being to liberty and well-being.*

Emma Goldman

In the former chapters of this thesis, we identified some limitations of computational social behaviour studies and addressed these limitations by providing novel solutions. We focused on the use of wearable sensing (mostly accelerometers) and presented our results on data captured in-the-wild from real life experiments.

In Chapter 2, as an answer to the limitations of traditional audio and video sensing in crowded scenarios, we proposed to use accelerometers for the detection of social actions and focused specifically on speaking. We showed empirically how peoples' movements vary greatly while speaking, demonstrating that the manifestation of speaking through body movements is highly person specific. In order to address this issue, we used a transfer learning approach, Transductive Parameter Transfer (TPT), which significantly improved upon traditional person independent approaches. We compared the detection of speaking to the less person specific activity of walking and found that performance differences between a traditional method and TPT were indeed smaller. In order to experimentally demonstrate the challenges of experimenting in-the-wild, we organised a small experiment where a participant imitated speaking, walking and standing in a structured way. The high performances obtained for this setup showed how controlled lab experiments fail to capture all the possible variations of actions in a real life scenario. We also analysed the transfer source quality and identified that there is no single perfect source for everyone. However, some participants were found to be optimal sources for the majority of others. Interestingly, no connection between the quality of a source and it's person dependent performance, actual physical distance to the target, gender, or similarities of the data distributions were found, showing it is most probably related to something more inherent such as personality.



Following on from the study in Chapter 2, we compared the classification of different actions such as speaking, stepping and gesturing in Chapter 3. Our results further demonstrated that there is a direct connection between the physical manifestation of the action, annotation quality and the classification performance. We tried to identify the minimum amount of data required for covering all variations in actions by experimenting with gradually increasing training set sizes. Performances for both TPT and a traditional setup seemed to converge when at least 3 minutes of data from each participant were used for training, identifying an empirical minimum limit. The higher performance of TPT with a low number of samples in the training, for both speaking and stepping, made it preferable for scenarios where obtaining large amounts of training data is not possible. TPT outperformed the traditional person independent setup for classifying speech regardless of the data size but failed to do so for the less person specific action of stepping, supporting the results of the former chapter. Performance scores obtained with the setup where the number of samples were increased in a chronological order showed that there are indeed parts of the event that are more informative than the others. To conclude, our results showed that identifying such intervals and using a personalised approach such as TPT, acceptable performance can be still obtained for small training sizes.

In Chapter 4, we focused on the detection of conversing groups using social dynamics, coordination of partners' actions and movements, rather than the proxemics widely preferred in the literature. In order to account for various interaction

dynamics that can arise in non-controlled real life scenarios, we presented a novel approach that has 'group size awareness'. Our proposed method out-performed the state-of-the-art by dynamically selecting classifiers trained on data from different sized groups and estimating final membership of a new sample by fusing the selected classifiers' probability outputs. Further analysis of how group size based classifiers perform on data from groups of specific cardinalities where classifiers and test sets matched in terms of group size performed the best, strengthened the need for such an approach. We also found that the importance of features changes with respect to the group cardinality suggesting that when modelling the characteristics of different sized groups, different representations might be needed. Finally, we presented an analysis of our results based on the roles in the interaction, suggesting new directions for the research on the detection of groups.

Chapter 5 acted as a proof of concept study where we used speaking statuses and group memberships for classifying a higher level social concept; personality. We focused on a real life mingle event, which is not traditionally considered in the literature for personality detection, and showed it is still possible to infer this information in such a complex scenario with a non-traditional sensing medium. Performance differences in the estimation of various traits showed that some sensing media might not be optimal for recording all behavioural aspects of the personality. Most importantly, our results showed that it is possible to imitate two behavioural cues (speaking and movement) from a single digital modality (acceleration) and their joint use results in better estimation performance.

Chapter 6 investigated how the social context moderates an individuals' evaluations of an event. During various live performances, we captured reactions of the audience with mobile sensors and used the linkage between these spontaneous reactions to evaluate different aspects of participants' appraisals of the event, such as enjoyment and immersion. We showed that the linkage of participants' body movements are representative of an events' informative parts. Intervals with high linkage were found to be highly correlated with the salient moments reported by the participants. We then used the linkage levels to detect intervals that are highly informative and used samples from only these intervals to automatically estimate participants' evaluations. This procedure was shown to be successful since the method without interval selection performed poorly compared to the method using it. Finally, we investigated how attending a live performance might affect an individuals' social behaviour. We did so by organising a two session mingling event with a dance performance in the middle and quantifying the differences between participants behaviour in the two sessions. The high correlation between an individuals' physical activity level and the self reported effects of the performance on their mood suggested that the experience of an event might change the subsequent social behaviour.



While conducting the research presented in this thesis, we were able to identify current weaknesses that are yet to be addressed, new challenges, and possible directions to follow. In the following subsections, we will mention and briefly discuss some limitations of our work, present possible directions for future research, and identify general issues in the computational social behaviour understanding domain

that need to be addressed.

## 7.1. Who to transfer from? Finding good sources when estimating socially relevant behaviour

In Chapter 2, we argued that personalised concepts, such as the connection between speaking and body movements, require approaches that can provide personalised models. Although we studied the specific context of speaking, we believe this to be valid for any person specific concept. TPT, the method we proposed as the solution, computes personalised models with transfer learning by regressing over the parameters of source datasets of other participants. Theoretically, TPT should inherently suppress bad sources since it computes the similarities between data distributions of participants and the transfer occurs with respect to this similarity. However, our further analysis of source quality showed that the distance between data distributions is not enough for finding the optimal sources for a specific target. By exhaustive search, we were able to identify optimal source sets for each target and found some participants to be better sources than the others. Thus, as future work, the following questions should be answered: Which properties make a good source? If the similarity of data distributions is not enough for identifying the best subset for training, what can?

We analysed the importance of gender and the actual physical distance on source quality but could not find any meaningful connections. We were able to find some cues related to the personalities of good sources for them to be extroverts and open to experience but these cues were not enough for solid conclusions to be made. However, these cues point to a highly possible direction: The qualities that make a good source are inherently personal and might be more complex then expected. The question to answer then becomes: What factors can make two people act similarly when they are speaking (or behave similarly in general)?

There are various possibilities covering personal and social characteristics of people, both short and long term. We already briefly investigated how personalities of individuals can make them good sources. If we were to speculate, we can say that it might be more than just one persons' personality that make them good sources but the similarity between personalities of people might cause them to act similarly. Another possibility is the mood of the participants, how they are experiencing the interactions that they are in. This might change how people behave and people with similar moods and experiences might behave in a similar manner. At a finer granularity, even the roles in the current interaction might temporarily change behaviour characteristics. We can then argue the detection approach for behaviour should be more dynamic and temporally aware. Perhaps the differences are related to longer term constructs, such as the cultures of different societies. Some cultures are known to be more expressive than others in interaction. The crowd of the event of Chapter 2 was formed by many international students and this property might have an effect on the source quality. None of the possibilities mentioned here are investigated thoroughly yet and stand as open questions. Answering them will give us a better understanding of the underlying factors of social behaviour and allow

us to provide satisfactory solutions.

The problem of finding people that express their behaviours similarly, can be formulated more generally, outside of the scope of transfer learning. If we were to identify people that are expected to act similarly, we can specify a subset to use for training our models. Recommender systems already do something similar by suggesting content that are known to be liked by friends and people with similar tastes, neglecting data from millions of others. However, the problem we have is much more complex. Since there are no direct categorisations of behaviour, correctly identifying similarities or concepts contributing to the similarity of behaviour is quite challenging. Even assuming we can do this satisfactorily, there is also another dimension to consider; ethical constraints. Take personality for example. Even it is common practice in today's work application procedures, categorising one's personality traits without their permission can be considered as profiling and obtaining strictly personal information. Even identifying one's nationality is quite sensitive and can be interpreted as intrusive. One possible direction is directing the research on context sensing. Without going into areas that might be private to some, if we manage to model the context that the people are in, we can analyse how it affects their behaviour and use this knowledge for the task at hand. However, the problem of sensing context in real life scenarios is quite difficult to solve. We believe, analysing behaviour in the longer terms and in relation to others can be a good point to start. The context information will also allow us to detect similarities between events with different characteristics, aims, and atmospheres and make knowledge transfer between them possible.

## 7.2. Social dynamics in group detection: Challenges for a new frontier



In Chapter 4, we showed that social dynamics, coordination of people's actions, and movement, are indeed rich information sources for the detection of conversing groups. However, the results we obtained were still flawed in some aspects. Even though our proposed approach is specialised to capture various interaction dynamics arising in different sized groups, detailed analysis of the results showed that we are still far from satisfactorily capturing some types of interaction, such as speaker-listener behaviour. In order to use the full potential of the rich information embedded in dynamics and provide solutions that will significantly outperform traditional proxemics based approaches, the current weaknesses need to be addressed.

One aspect that is not investigated in Chapter 4 is the role of a participant in the interaction. This information can be used to analyse the performance of pairs or can be employed in the training in a similar manner to group cardinality used in Chapter 4. We believe that utilising the role information is an interesting future direction that might provide insights into the nature of the problem. Most basic approach in this direction is defining two roles (speaker and listener) which results in three possible interaction types: Speaker-Speaker, Speaker-Listener and Listener-Listener. In order to define the roles, simple heuristics on the speaking statuses of the participants can be used. However, we believe for capturing all possibilities

of interaction, a more detailed representation might be needed. We can argue that there are various types of speakers and listeners, some speakers are more dominant, some listeners are more responsive and active than others, etc. The behaviour of specific roles might also change with respect to the group size the participants are in. Not considering these varieties in behaviour will result in poor representation. However, it should be possible to model these varieties at least to some extent and obtain a more detailed role representation. One can expect active listeners to be more physically active which can be measured through acceleration. They should use backchannels such as head nods or short vocalisations more which can be extracted through their social action streams. For the speakers, a similar approach that takes their physical activity levels and statistics of their social actions, individually and with respect to other group members, into consideration will make it possible to obtain a more detailed categorisation (dominant-nondominant speakers, for example). With a more detailed representation of the roles, it might be possible to move from role based analysis to actually using this information in the training phase of the model itself. A mixed representation that considers both complex roles and group sizes should be able to cover more variations of interaction dynamics.

As mentioned in Chapter 4, the majority of the work that aims to model social interaction generally focused on strict dyadic interaction and proposed ways of representing this behaviour through concepts like synchrony, coordination of actions, and mimicry [1]. Even most of the features we used aim to represent one to one interactions where a specific pattern of action is followed (first participant speaks, signals the other, stops speaking and the second one starts speaking, etc.). By the addition of gesturing into the equation and providing some statistical measures related to how they coincide with speaking, we tried to model backchannels that are typical of listener behaviour. It was successful to some extent since our method provided better performance than approaches that only focuses on speaking related joint measures.



However, we believe that our attempts still do not fully represent Speaker-Listener and Listener-Listener pairs. To obtain a better representation, as a first step, the importance of our features with respect to the roles can be analysed, in a similar manner to the group size analysis done in Chapter 4. In this way, we can identify subsets of features that are more informative for each role pair. While doing so, having a better categorisation of roles will help to identify more precise and specialised subsets. Finally, we can define new features that are more powerful representations of these concepts. There are two possible paths to take here. First, we can rely on representational learning methods where features will be automatically extracted by an artificial neural network. This will of course require huge amounts of data and the interpretability of extracted representations will be low. The second possibility is to consult social science findings on group behaviour. By identifying distinguishing aspects of speaker and listener behaviour with the help of social science and devising ways to computationally represent them, we may finally be able to model all possible interactional role pairs.

At the end of Chapter 4 as future work, we mentioned the use of a post-processing step that moves from pairwise representations to a connectivity graph

including all participants which is a more general representation of the scene. In our current methodology, all the representations are pairwise, training and estimation is performed on pairwise data samples. We already discussed some ways of achieving this final connectivity graph using existing methods in the literature [2] and providing a possible formulation where posterior probabilities obtained by classification are used as edge weights in the graph.

A more interesting question is having a complete representation of the scene from the start. Instead of performing classification on pairwise representations, we could try to train and test our computational models on the whole data of a scene represented directly as groups. Such an approach will need to consider all possible groupings for all participants for the feature extraction, training and testing phases and will be extremely computationally complex. However, there could be ways to reduce this complexity by considering only plausible groupings that are detected through additional information. One additional information source is proximity. Instead of completely disregarding proxemics for conversing group detection, we can utilise this information for better estimation. Only considering possible groupings of participants that are spatially close will already reduce the complexity a lot. We could then fuse the two main components of interaction, proxemics and dynamics, and obtain a complete representation of a scene directly.

## 7.3. Joint estimation of actions and interactions

In our experiments of Chapters 2, 3, 4, and 5 we treated the connection between social actions and interactions to be one-way. That is, we assumed that the social actions of an individual were estimated solely from that person's sensor data. So no information from the behaviour of others who they are interacting with was used to estimate their behaviour. The first thing to investigate is how the information from interacting partners can be utilised for the better estimation of social actions. If we know the actions of other people in the group, we can use this information to refine the social action estimations of the current participant. For example, it is quite unlikely that two people in the same group are speaking at the same time. We implicitly used this while detecting conversing partners but did not try to use it for the detection (or the refinement) of the social actions. However, while trying to detect social actions of a person, requiring the knowledge of the social actions of their interacting partners is a chicken and egg problem. A more realistic approach will be using the raw data from interacting partners as additional cues in the detection process. Of course, such an approach will still require the identification of interacting partners. A noisy estimation of this can be obtained through proximity sensing.



However, the most interesting and challenging problem is the actual joint estimation of social actions and interactions, where no information regarding the both is present beforehand. Since we argue that they influence one another, estimating them simultaneously in the same classification procedure might be the optimal way for detection. This joint estimation is of course more complex than separately estimating actions or interactions but might be achievable by a well defined joint cost function and an optimisation procedure. Currently, we are not proposing any final

formulation but an optimisation procedure that alternates between minimising two cost functions corresponding to the existence of an action and the interaction status seems like a plausible option. As a starting point, this joint estimation problem can be realised in a pairwise manner where two individuals' social actions and their interaction status are simultaneously estimated. A harder challenge will be incorporating this into the direct group sensing we mentioned in the former subsection. We believe, this is one of the directions that needs to be investigated for obtaining more refined estimations of both actions and interactions.

## 7.4. Socially relevant appraisal analysis: Interpretations to facts

The study presented in Chapter 6 acts as a first step in utilising the social aspect of being together with others while attending a cultural event. It mainly focused on the detection of the informative parts of the event through the linkage of audience responses and utilised such moments to distinguish between audience members with differing evaluations. Various interpretations of the results are presented and possible hypotheses related to the factors affecting these results are proposed. One example is why there were some people coming to the event together that had high mutual information throughout the event but their evaluations were different (the hypothesis was that their high linkages were unrelated to the what was happening during the event). There are various examples of such hypotheses throughout the study and most of them are not experimentally confirmed. This study identified a possible direction that is worth investigating rather than drawing solid conclusions.



One possibility for reaching solid conclusions is capturing more information with more modalities of sensing. We actually installed IR cameras for one of the experiments, aiming to visually capture the movement patterns and reactions of the audience. Sadly, the video quality was not good enough to reach any robust conclusions. Better quality video data of the audience should make it possible to visually confirm some of the hypotheses. Especially for unconstrained scenarios like the second case study presented in Chapter 6, such information will be quite valuable since it will be easier to identify audience members that are sitting, standing or leaving, etc. . In this way, specialised approaches for these possible variants can be developed. Another complementary modality that will be helpful in verifying our approach is physiological sensing. Physiological sensing is traditionally used in the literature for detecting emotional responses of people to external stimuli [3, 4]. The connection between some physiological data (heart-rate, skin-conductance, etc.) and emotions are well studied in Affective Computing. Thus, using this modality together with acceleration will help us identify more inherent reasons for reactions we detect in acceleration, allowing us to have a better understanding of our results. However, we should be clear that we are not proposing to use video or physiological sensing as parts of the final solution. We propose instead to use these complementary modalities to verify the information we obtained with accelerometers, experimentally confirm the hypotheses, and finally to develop our acceleration based algorithms with respect to these findings.

Apart from insufficient verification of the results, there were also some experimental shortcomings in Chapter 6 which are mainly related to the generalisation capabilities of the method. First of all, in all the experiments, the number of negative evaluations were generally quite few. From a pattern recognition perspective, this might result in poor representation of the negative class. We might not be capturing all possible variations of audience members with negative evaluations. Also the method is only tested on a limited number of datasets, two to be exact. Even though we had organised a follow up experiment to verify our methods, the setup of this second experiment had different characteristics. So, this follow up experiment showed how the method generalises to events with different characteristics but provides no information on how the method will work on a different iteration of the same event. Conducting more experiments, with similar and different characteristics, will allow us to capture a more varied set of audience evaluations and will provide a better assessment of the robustness of the approach. To conclude, this study paves the way for further study but in order to generate more conclusive findings, more detailed analysis and further experimentation is needed.

## 7.5. Computational social behaviour research: General advice and concerns

As the closing section of this thesis, we first want to briefly mention what the contributions of this thesis mean for the domain of behavioural computing more generally. This thesis mainly focused on the use of wearable sensors, more specifically accelerometers, for social behaviour understanding, showing how informative body movements are. So the main aim is to direct future research into this specific direction. More specifically, we (i) advocated focusing on generalisable personalised models for inherently personal concepts, (ii) suggested a new direction for group detection studies based on the dynamics of social interaction by providing a group size awareness based solution to the complexity introduced by the many possible variations of interaction dynamics, (iii) showed how personality traits can be classified in a complex real life scenario using the social actions and interaction information, and (iv) showed the importance of the social aspect of attending an event and how it affects the evaluations of participants. We believe that all the chapters in thesis provide novel ideas to follow for researchers in social behaviour computing.



As the final remarks, we would like to discuss more general ideas, recommendations, and concerns regarding the future of computational social behaviour research and even the AI domain in general. First of all, we would like to mention again the importance of experimenting in-the-wild. As the small experiment of Chapter 2 already showed, capturing all the variations of social phenomena is rather difficult in controlled lab experiments. However, capturing and working on real life data is quite difficult for many reasons [5] and because of it, a concern is that researchers are reluctant to do so.

This brings us to the second point of our discussion; data sharing and reproducibility of the experiments. Reproducibility is already a concern for nearly all

domains of research. We might argue that socially related computing is one of the domains in AI where data collection is the hardest. It is not possible to obtain real life datasets for the analysis of social concepts by crawling through thousands of web pages, extracting images with respect to the tags and retrieving annotations by assigning HITs on Amazon Mechanical Turk. We are not trying to undermine the efforts of such studies but capturing a real to life event requires more precise work and can not be automated easily. The experiments need to be precisely designed by considering all possible scenarios, people need to be recruited to attend the event and annotators need to be trained for labelling the data. If some part of the experiment design fails, the whole experiment might fail, resulting in no valid data at all. In such a case, repeating an experiment is not easy. With these properties in mind, we might argue that data sharing is even more important for social understanding studies. With more and more people collecting and sharing data, it will be possible to test the generalisation capabilities and the reproducibility of the proposed approaches. Thus, the data collection procedure should not be undermined and it should be seen as an essential part of the research.

Finally, we would like to say a few words about the ethics of the studies in social behaviour computing and AI in general. We already mentioned how the sensing procedure should be private but we think it is just a part of the whole picture. The ethical implications of AI research are already debated heavily where the main focus is on subjects like what AI should do in an ethical dilemma, automation of jobs, and if robots will kill us one day or not. Even though we believe each one of these are valid discussion points, they partially miss what is already happening today. AI is already being used for many different purposes in various domains such as business, politics and security, having vast societal impacts. Hopefully, this aspect is seeing more interest nowadays, even having its own specialised conference being organised (e.g. the ACM Conference on AI, Ethics and Society).



If the input of the computational models are human data and the output is the interpretations of their behaviour, just like in social behaviour computing, ethics become more important. Apart from the requirements for data collection, we believe one of the most important aspects is how the results of computational methods are interpreted. Computational methods or AI in general are not perfect. They 'learn' through data and if the data is already biased, its results will be also biased. Interpreting the outcomes of computational models directly as facts is then extremely dangerous, if the data is not properly analysed beforehand. Especially with the rise of representation learning where all the features are automatically learned and hard to interpret for humans, more and more studies with poor analysis are being published. So, we believe before making any final conclusions, researchers should know the data well with all its aspects, identify its shortcomings and biases, and construct their narrative with respect to this knowledge. Thus, as the concluding few words, we can say that we support more shared, real to life, ethical and privacy preserving sensing and analysis, and propose these properties as necessities for further research.

# References

[1] H. Hung, G. Englebienne, and L. Cabrera Quiros, *Detecting conversing groups with a single worn accelerometer,* in *Proceedings of the 16th international conference on multimodal interaction* (ACM, 2014) pp. 84–91.

[2] H. Hung and B. Kröse, *Detecting f-formations as dominant sets,* in *Proceedings of the 13th international conference on multimodal interfaces* (ACM, 2011) pp. 231–238.

[3] M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang, *1000 songs for emotional analysis of music,* in *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia* (ACM, 2013) pp. 1–6.

[4] C. Chênes, G. Chanel, M. Soleymani, and T. Pun, *Highlight detection in movie scenes through inter-users, physiological linkage,* in *Social Media Retrieval* (Springer, 2013) pp. 217–237.

[5] L. Cabrera-Quiros, A. Demetriou, E. Gedik, L. van der Meij, and H. Hung, *The matchnmingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates,* IEEE Transactions on Affective Computing (2018).

# Summary

Understanding human behaviour has sparked the minds of many throughout centuries. One intriguing aspect of human behaviour is the social part; how humans react to each other and their environment. Scientifically studying such behaviour is hampered because of the need for manual annotations, so that social scientists limited themselves to observing only short time intervals in limited settings. With the growing processing power of computers and increasing possibilities of robust, continuous, and mobile sensing, collecting and analysing large amounts of real-life behaviour data has become possible. Moreover, computational methods make it possible to go beyond traditional approaches for social understanding, since they detect patterns that are not easily distinguishable for humans.

However, even with powerful computational models, investigating human behaviour is quite challenging as behaviour is personal and contextual, resulting in huge variations. This thesis proposes novel computational solutions for analysing human social behaviour. It focusses on data collected from people with wearable accelerometers in crowded events where people freely mingle with each other. It provides solutions to robustly detect actions and interactions, as well as how to use the detected information to derive higher level social understanding.

The thesis starts by introducing novel ways of detecting social actions and interactions. To deal with intra personal variations, we show how general action predictors can be adapted to become personalized models using the transfer learning methodology. Further, we show that the detection of conversing groups can be deduced from interaction dynamics, instead of the mainly preferred modality of proximity. Large variations of interaction patterns that might arise in unrestricted scenarios are addressed by a novel method that considers the sizes of the groups; both in training and detection phases.

The thesis continues with a proof-of-concept study that shows how detected action and interaction patterns of people can be used to infer an individuals' psychological construct. We show that it is possible to detect the construct of personality in a real life event by imitating two behavioural cues (speaking and movement) from one digital modality (acceleration). Additionally, we describe a detailed investigation of how social context moderates an individuals' evaluation of a live performance. Through a novel approach, we infer audience members' evaluations from informative parts of the event, identified by the linkage of body accelerations.

Taken together, with this thesis we show that with the increased sensing and computing power, the understanding of human social behaviour in more dynamic social situations is within reach.

# Samenvatting

Het begrijpen van menselijk gedrag heeft velen aan het denken gezet in de geschiedenis. Een intrigerend aspect van menselijk gedrag is het sociale deel; hoe mensen op elkaar en op de omgeving reageren. Onderzoek naar sociaal gedrag is lastig doordat observaties gedaan moeten worden door wetenschappers 'met de hand' (bijv. door het kijken naar een video en aantekeningen te maken). Om deze rede moeten sociale wetenschappers zichzelf vaak beperken tot onderzoeken van korte duur, of moet onderzoek gebeuren in gecontroleerde omstandigheden (zoals in een lab). Met de opkomst van betere mobiele sensoren is het mogelijk geworden om grote hoeveelheden data van menselijk gedrag in een natuurlijke omgeving te verzamelen. Met de opkomst van snellere rekenkracht van computers is het mogelijk geworden om deze grote hoeveelheden data automatisch te analyseren. Met moderne computermodellen is het zelfs mogelijk om sociaal gedrag nog beter te begrijpen dan met traditionele methoden, omdat deze modellen patronen kunnen detecteren in data die lastig zijn te herkennen voor mensen.

Zelfs met zulke krachtige modellen is het onderzoeken van menselijk gedrag nog steeds zeer uitdagend. Dit komt doordat gedrag verschilt van persoon tot persoon en afhangt van context. Hierdoor is er grote variatie in menselijk gedrag. Dit proefschrift introduceert nieuwe computationele methoden voor het analyseren van sociaal gedrag. Dit werk richt zich op een experiment waarbij mensen met draagbare sensoren gezellig met elkaar omgaan op een druk evenement. In dit werk worden technische oplossingen gegeven om op een robuuste manier acties en interacties te herkennen aan de hand van sensordata. Deze informatie wordt vervolgens gebruikt om tot abstractere sociale inzichten te komen.

Dit proefschrift introduceert nieuwe manieren om sociale acties en interacties te herkennen. Om slim om te gaan met intra-persoonlijke variatie, laten we zien hoe algemene modellen om acties te voorspellen kunnen worden gepersonaliseerd door middel van transfer learning. Vervolgens laten we zien dat conversatie groepen afgeleid kunnen worden van interactie patronen, in plaats van het gebruik van afstand tussen personen zoals gebruikelijk. Grote variatie in gedragspatronen die voorkomen in een natuurlijke omgevingen vormen een uitdaging. Een nieuwe computationele methode die rekening houd met groepsgrootte (gedurende de train en detectie fase) pakt dit probleem aan.

Vervolgens gebruiken we een proof-of-concept studie om te laten zien dat interactie patronen van mensen gebruikt kan worden om het psychologisch profiel van een persoon te achterhalen. We laten zien dat de persoonlijkheid in te schatten is door middel van twee gedragspatronen (spreken en beweging) en van één sensor (acceleratie). Vervolgens beschrijven we een gedetailleerd onderzoek dat laat zien hoe sociale context invloed heeft op hoe men een life-optreden beoordeelt. Met een nieuwe methode kunnen we voorspellen hoe mensen in het publiek het

life-optreden gaan beoordelen door middel van bewegingssensoren.

Samenvattend illustreert dit proefschrift dat met de toegenomen kracht van sensoren en computerkracht het begrijpen van menselijk gedrag in dynamische sociale situaties binnen handbereik is gekomen.

# Acknowledgements

If someone told me four years ago that writing the acknowledgements part of my thesis is going to be quite challenging, I would not have believed them. "It should be fairly easy, I'll just write couple sentences to thank a couple of people and that's it", I would have said. Currently, I am looking at a blank page and I am quite terrified that I might not do it justice and miss something. Many people throughout these years contributed to this thesis by guiding and helping me, both academically and mentally. Now, I'm here to thank all these people and I'm not sure if my words will suffice. If I somehow end up forgetting you, you are free to bug me continuously.

I would like to start by thanking my supervisor, Hayley. You were much more than a 'boss' for me. Your understanding and sympathetic behaviour always motivated me to do better. Even though I'm not much of a fan of some endless meetings that we had, I should admit that they helped me to obtain a deeper understanding of science and the problems we were trying to tackle. Your limitless enthusiasm about research and ability to come up with interesting questions always fascinated me. All in all, working with you was (and luckily still is) a privilege that I was (and is) lucky to have. I hope we will continue to tackle new and exciting problems in the future with the same enthusiasm.

Marcel, my promotor, your no-nonsense and sometimes practical attitude about science and management is motivating. With your strict deadlines, I was able to focus and finish my thesis. I must admit, I was quite surprised when you went over my introduction on a single night and provided suggestions to make it better, sentence by sentence. It showed me how a professor can have control over micro issues while still managing on a macro level.

Marco, you are an inspiration for me. Few weeks after I start my new job in the Netherlands, I somehow found myself drinking beers with you. How could I know that this will turn into our routine and will extend over the working hours? Your perspective on science and life influenced me and your extremely witty humour never ceases to amaze me. I'm glad to know you and call you a friend.

I would like to extend my gratitudes to the all of the PRB group, without any specific order: Alexey and Amin, my first office mates. Sjoerd, the golden voice. Ahmed, I'm a fan of how you dress and do research. Hamdi and Gorkem, thanks for making my adaptation period as easy as possible and thanks for all the guidance. Alexandar, thanks for introducing me to the gym after my 20s and thanks for enduring my distasteful jokes about your nationality. Speaking of it, Christian your weird remarks are always amusing and Thies, it's always nice to see the mushroom guy pop-up in the building. Veronika, our talks, in our former office or over a beer anywhere, always had a positive effect on me and thanks for all the BBQ! Jesse, your work ethics and analytical thinking is impressive. I guess that's how you manage to endure our drunken ramblings without drinking yourself :). Amogh come to Delft

more often and don't forget to invite me to good electronic parties in Amsterdam. Tom, only person who can drown my voice with his own. To receiving more rude comments regarding the level of our voices in trains! Tamim, you introduced me to football after more than 10 years and put up with my sub-par skills. Arlin, I should come to one of your shows one day. Wenjie, I'm sad that I still haven't got to use your one and only model yet. Yazhou, I wish you a life as active as your learning. Saskia, the amount of thanks I can extend is not enough for you, you were always there for all of our questions and problems. Ruud, Bart and Robbert, the unsung heroes of the group. Thomas, your knowledge in everything, from tech to video games to beers, is fascinating. Osman, the second person who can endure our drunken antiques and my mentor in football. Taygun, the NLP expert and a fellow CRPG fan. David, before coming here I read some of your papers. Considering the citation counts, I thought I'll be meeting a snob person, however you were the exact opposite if it. Thanks for all the inspiration and the book recommendations. Silvia, your positive aura always makes me happy and I'm glad to have you around. Wish to meet Luna one day! Bob, I was looking forward to all of your lab talks since you don't get to listen to a legend that often. Jan, CV expert, dad joke machine and sci-fi enthusiast. Hope Delft as a city treats you well. Bioinformatics Tom, thanks for replacing me in buying the borrel equipment. Christine, keep up your positive stance in life and please don't play football anymore. Joana, next time I'll see you with the dog, I'm definitely petting him/her. Seyran, your comments in coffee talks always opened new perspectives for me. Stephanie, the first of the second generation. Hope I'm not bothering you with my weird noises in the office too much and I'm keenly waiting to meet Dobby. Bernd, your different point of views in our discussions opened my eyes to new possibilities. Stavros, the sports guy of the group. Thanks for organising all these events. Soufiane, it is always nice to hear about your travels. Yancong, I will always remember your sprints in football matches. John, the original body builder. Yunqiang, Jose, Ramin, Yeswanth, Chirag, and Ziqi, you are the most recent additions but you are already an integral part of the group.



Laura, Wouter and Sally, the usual suspects. I can't thank you enough guys. You were there for me all throughout this journey, the goods and the bads. Having you with me made this journey much more bearable. Laura, my sister from another mister. I was really lucky to have a bright and cool colleague, and more importantly, a friend like you. Our weird similarities, all of our conquests in different cities, countries and even continents, extended Thursday nights, these are all the memories I will remember forever. I'm glad you are not returning to Central America yet but please don't forget about us in Eindhoven. Wouter, the golden boy. Since I need to keep this somehow formal, I cannot enumerate all the farming, piracy or baking related phrases and concepts I have learned from you. You definitely expanded my horizons. Thanks for showing me from the beginning that Dutch can be spontaneous and giving. Alex, the Tex-Mex in flesh. From the time that Thomas asked me to show you around to this day, you were in all the good stories I remember. You searching for me in my own house, the night we waited for Cristian's train to arrive, you getting angry at a wall ornament in Klooster (YOU

SEE THIS GUY?) – all good memories and I'm waiting to add more to them.

Claudio, the Italian. I hope that we get to collaborate again and I will have more chances to serve you pastirma after a night out. Gwenn, thanks for accepting my last minute requests of paper revisions, regardless of where you are. Andrew, thanks for providing us with the view of a social scientist and all your help in data collection.

Of course, I need to mention my Turkish Gang that I spent most of my time with during these four years. Onursal, my bromance, a true intellectual and memleketlim. To more concerts, to more raki, to more craft beers and to more in depth discussions! Argun, the most handy person I've ever met. Just be positive to yourself as much as you are positive to us. Bas (honorary Turkish gang member) who turned from Cansin's boyfriend into a real friend of mine. I was honored to be your best man. Burak, the Muhtar. You are the father of the group and you always acted like one. Cansin, the motherly figure. I'm still craving for that amazing soup you made. Alper, the most carefree person of us all. Keep the attitude and I hope we can visit a Greek island together next year, like we planned. Tilbe, the social butterfly and my drinking buddy. Klooster or Bouwpub, you always answered the call and came running. Rose (second honorary Turkish gang member), I hope we can visit Ikaria some time together, preferably your and Burak's retirement house. Emre, I already miss our crazy nights in Rotterdam.

My Valentini. Thank you for being with me in the hardest parts of this journey. Knowing that you are with me, physically or mentally, kept and keeps me going. I'm lucky to find you and have you. To more years together!

Finally, I would like to thank my family. Kuzen, I'm glad our first real trip together was in Japan. Father, you always believed in me, no matter what. You supported me in all aspects of my life and this thesis is made possible by you. I'm extremely lucky to have a cool, understanding and loving father like you. I love you. And of course my mother, the woman who made me the man I'm today. All the thanks and the love in the world is not enough for you.

# Curriculum Vitæ

## Ekin Gedik

Ekin was born on October 29th 1988 in Ankara, Turkey. In 2010, he received his Bachelor's degree (with honors) in Computer Engineering from Middle East Technical University (METU), Ankara. Between 2010 and 2013, he worked as a teaching and research assistant in METU, while pursuing a Master's degree. During this period, he worked on various research projects that mainly focused on automatic analysis of satellite imagery and collaborated with leading defense and telecommunication firms of Turkey. In September 2013, he obtained his Master's degree with high honors.

In April 2014, he relocated to the Netherlands and started to work as a PhD candidate in the Pattern Recognition and Bioinformatics Group at Delft University of Technology, under the supervision of Dr. Hayley Hung. Ekin's position was partly funded by Commit/ and Delft Technology Fellowship. His research mainly focused on automatic analysis of human behaviour in real life scenarios through wearable sensing.

Currently, Ekin is a post-doctoral researcher in Pattern Recognition and Bioinformatics Group at Delft University of Technology.

# List of Publications

## Journals

1. **E. Gedik**, L. Cabrera-Quiros, C. Martella, G. Englebienne, and H. Hung, *Towards analyzing and predicting the experience of live performances with wearable sensing*, IEEE Transactions on Affective Computing, 2018.

2. **E. Gedik** and H. Hung, *Detecting Conversing Groups Using Social Dynamics from Wearable Acceleration: Group Size Awareness*, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2(4), 2018.

3. L. Cabrera-Quiros, **E. Gedik**, H. Hung, *Multimodal self-assessed personality estimation during crowded mingle scenarios using wearable devices and cameras*, Submitted to IEEE Transactions on Affective Computing, Under Revision.

4. L. Cabrera-Quiros, A. Demetriou, **E. Gedik**, L. van der Meij, and H. Hung, *The matchnmingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates*, IEEE Transactions on Affective Computing, 2018.

5. **E. Gedik** and H. Hung, *Personalised models for speech detection from body movements using transductive parameter transfer*, Personal and Ubiquitous Computing 21(4), 2017.

## Book Chapters

1. H. Hung, **E. Gedik**, and L. Cabrera-Quiros, *Complex Conversational Scene Analysis Using Wearable Sensing*, Chapter 11 of the book *Multi-modal Behavior Analysis in the Wild: Advances and Challenges*, Elsevier, X. Alameda-Pineda, E. Ricci, N. Sebe, 2018.

## Conferences

1. L. Cabrera-Quiros*, **E. Gedik***, and H. Hung, *Estimating self-assessed personality from body movements and proximity in crowded mingling scenarios*, Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016.[1]

2. **E. Gedik**, *Are you (not) entertained? estimating the state of a crowd in an event using wearable sensors*, Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct, 2016, Doctoral School.

3. **E. Gedik** and H. Hung, *Speaking status detection from body movements using transductive parameter transfer*, Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct, 2016.

[1] *Authors contributed equally.

4. K. Schellekens, E. Giaccardi, H. Hung, and L. Cabrera-Quiros, **E. Gedik**, C. Martella *Impact of connected objects on social encounters*, 4th Participatory Innovation Conference, 2015.

5. C. Martella*, **E. Gedik***, L. Cabrera-Quiros*, G. Englebienne, and H. Hung, *How was it?: exploiting smartphone sensing to measure implicit audience responses to live performances*, Proceedings of the 23rd ACM International Conference on Multimedia, 2015. [2]

[2]*Authors contributed equally.