Automatic analysis of human social behavior in - the - wild using multimodal streams

Cabrera Quiros, Laura

DOI
10.4233/uuid:811ba745-18e1-4dca-8321-249ba000a142

Publication date
2018

Document Version
Final published version

Citation (APA)
Cabrera Quiros, L. (2018). *Automatic analysis of human social behavior in - the - wild using multimodal streams*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:811ba745-18e1-4dca-8321-249ba000a142

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

AUTOMATIC ANALYSIS OF HUMAN SOCIAL BEHAVIOR IN-THE-WILD USING MULTIMODAL STREAMS

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
Thursday 27 September 2018 at 10 o'clock

by

Laura Cristina CABRERA QUIROS

Master of Science in Electronic Engineering,
Instituto Tecnológico de Costa Rica,
born in San José, Costa Rica.

This dissertation has been approved by the:
Promotor: Prof. dr. ir. M. J. T. Reinders and
Copromotor: Dr. H. S. Hung

Composition of the doctoral committee:

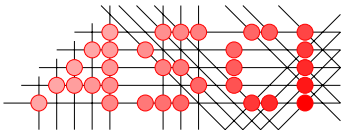
Rector Magnificus,	chairperson
Prof. dr. ir. M. J. T. Reinders,	Delft University of Technology, promotor
Dr. H.S. Hung,	Delft University of Technology, copromotor

Independent members:

Prof. dr. A. Hanjalic,	Delft University of Technology
Prof. dr. D. Gavrila,	Delft University of Technology
Prof. dr. M.A. Larson,	Delft University of Technology / Radboud University
Prof. dr. R.C. Veltkamp,	Utrecht University
Dr. P. Vogt,	Tilburg University

TEC | Tecnológico de Costa Rica

This work was supported by the Instituto Tecnológico de Costa Rica, Costa Rica.



Advanced School for Computing and Imaging

This work was carried out in graduate school ASCI.
ASCI dissertation series number 394.

Cover Designed by L. C. Cabrera-Quirós.
Printed by Proefschriftmaken.nl
ISBN: 978-94-6366-072-3

Al ser supremo,
por todas las bencidiones que me ha otorgado.

A mi madre Lidiette,
quien siempre interpuso nuestro bienestar al propio.
Mi mejor amiga, consejera y modelo a seguir.

A mis hermanos Ariel y Diego, mi trifuerza,
quienes me mantuvieron con los pies en la tierra.

A William,
quien me ayudó a levantarme cuando caía.

CONTENTS

1

INTRODUCTION

Everyday we engage in a wide variety of social interactions, from sending a text message to casually chatting with a coworker. But what defines a social interaction? How can we automatically detect, analyze or synthesize these forms of human behavior? And, once we have understood this concept, can we give computers the capacity to manage and replicate such behaviors? This thesis is one of many efforts towards answering such questions.

In this Chapter, we will first describe the ultimate goal of the automatic analysis of human social behavior using non-verbal behavior. Then, we will describe the opportunities provided by ubiquitous technologies in the analysis of social behavior in real settings, and the benefits and motivation for using multiple sensors for such analysis.

Finally, we will narrow down the field towards the specific scenario that is studied in this thesis: the **mingle events**, which are crowded scenarios where people are inherently encouraged to interact (e.g. parties). Mingle scenarios provide rich information about multiple social interactions happening dynamically and in parallel, and can be studied using ubiquitous multisensor technologies without interfering the natural behavior of the people. Note then that we will focus on face to face interactions so other types of interactions (e.g. social media) are not analyzed.

1.1. AUTOMATIC ANALYSIS OF NON-VERBAL BEHAVIOR

When it comes to face to face interactions, humans rely on verbal and non-verbal communication. Verbal communication consists of the spoken message; whereas non-verbal communications comprises all wordless forms of communication such as body language (e.g. gestures, posture) or face, eye and vocal behavior (e.g. turn taking, prosody), among others [169]. Both forms of communication play an important role in the successful transmission of the message [73].

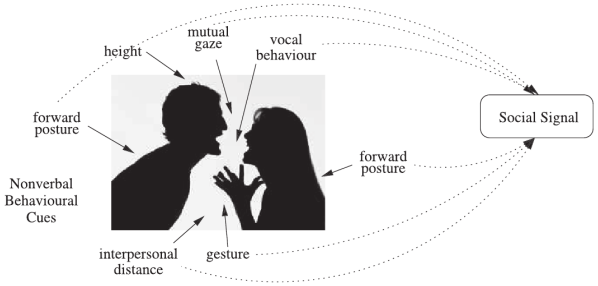


Figure 1.1: Nonverbal behavioral cues (taken from Vinciarelli et al. [170])

Take for example the image in Figure 1.1(a). Without hearing the spoken message between these two persons (if any), we immediately try to infer their interaction and relationship. What are they doing? Are they fighting? Or maybe singing? And what is their relationship? Have they just met? Or do they seem like long-time friends, siblings or a couple? Their non-verbal behavior during this interaction, also called *behavioral cues* or *social cues* in this context, give us an insight to try to answer such questions. Together, these cues become a *social signal* with an intended message [170].

While the automatic detection and comprehension of the spoken (or written) message has been studied for decades now [40], automatically detecting and incorporating the non-verbal behavior with the human message is still an interesting and unsolved problem. And this missing piece of information can dramatically change the message and meaning of the interaction. Take for example *sarcasm*, in which the written (or spoken) message might not exactly match the non-verbal communication. A sarcastic person could say *look how excited I am?* while not showing signs of excitement in its facial expression, but a neutral face instead.

Nowadays, researchers in the computing community aim to automatically detect and predict social signals and their implications, by analyzing the non-verbal behavior of the people while they interact. The understanding of the non-verbal behavior can either complement or complete the spoken message, or be used as a standalone mechanism to infer the meaning and/or characteristics of the interaction when the spoken message is missing. In this thesis we focused on applications in the latter category.

In recent years, the domain of *Social Signal Processing* (SSP) was born as an interdisciplinary niche to compile all the efforts regarding detecting, analyzing and synthesizing human social signals [169, 170]. It includes efforts in the areas of computer vision [43, 44, 79, 91, 149, 151, 166], affective computing [55, 58], Human-Computer Interactions [177] and wearable technologies [33, 106, 108], among others[1], and its impact extends within and across disciplines such as computer science, linguistics and social psychology.

[1]The reader can refer to `http://sspnet.eu/` for a compilation of works in the domain or to [170] for a comprehensive survey.

The ultimate goal of efforts in the area is to provide computers with social intelligence [169]. This refers to as the capacity to understand and manage social signals, in particular those coming from non-verbal cues, to communicate with a human in a similar way another human would do. To do so, a variety of different sensors can be used as input (e.g. video cameras, wearable sensors, microphones and proximity tags, among others). All these sensors will record the non-verbal cues of the people, which inherently are a window to their behavior, emotions and intentions [73].

To show the multidisciplinary nature of Social Signal Processing, let us take the detection of group formations as an example. This is an important first step in social interaction analysis which can be analyzed from different perspectives. Several works have addressed this problem from the computer vision perspective, focusing on scenarios of free-standing conversations with high ecological validity and in-the-wild settings, which gives rather noisy images or videos. For instance, Setti et al. [149, 151] focused on the detection of groups in still images, while Hung and Kröse [79] and Cristani et al. [43] did the detection using videos. Similarly, one could address the group detection problem through the use of wearable sensors. For instance, Cattuto et al. [33] involved up to 575 people in their analysis of person-to-person interactions using distributed RFID devices. Also, Martella et al. [105] used wearable bands to assess the proximity behavior of up to 600 people while visiting a museum. These latter efforts are more oriented to large sets of subjects (in the range of hundreds) and for a long amount of time (several hours to weeks), with less fine-grained observations. These examples illustrate different solutions that can be given to the same problem under different scenarios, and illustrate the amplitude and multidisciplinary nature that the Social Signal Processing domain can have.

Similar to group detection, works can be found about group characteristics (e.g. head and body orientation [141, 157]), group dynamics (e.g. cohesion [77], dominance [61, 78, 131], deception [133], or focus of attention [182]), interpersonal relationships (e.g. attraction [167], aggression [91]) or personal differences (e.g. personality [126, 168]), among many others works about the detection or recognition of social cues, affects or mental states; ranging from simple scenarios such as dyadic conversations or small group meetings, to complex settings with hundreds of subjects. All these examples give a hint of how diverse Social Signal Processing can be and the different facets it has, while still aiming for the same ultimate goal: giving social intelligence to computers.

1.1.1. CHALLENGES OF THE AUTOMATIC ANALYSIS OF NON-VERBAL BEHAVIOR

The capability to detect and manage the extensive variety of social signals that a human can display is rather rudimentary for computers nowadays, limited in some instances to one or two social signals at the time. For example, toolbox for the recognition of facial expressions are openly available nowadays [14]. However, these tools only take into account visual features, are affected by the presence of non-stationary poses and can not create a direct link between the expression and the persons' inner mental state. Therefore, the development of

systems with true social intelligence, that can take into account non-verbal cues, presents several complex challenges.

Humans are mostly unaware of their own non-verbal behavior. This form of communication is inherent to us, almost as a muscle memory. Even when we do not consciously orient our body towards a person given the situation, speak or remain silent to allow turn taking (so we do not to 'appear rude'), or move our head to agree to what it is being said; we do all of these actions inherently while interacting [73]. The same applies when assessing the non-verbal behavior of other people. We can infer that a person is, for example, upset but it is difficult for us to determine exactly what makes us come to such conclusion. The above also differs given the cultural background and can impact the overall 'smoothness'of the interaction [88]. Thus, for all possible instances, **there is not a clear link between a specifically set of social cues (or social signal) and the affect, feeling, mental state or intent that is being transmitted and perceived**. We can detect a smile, but is this social cue a polite, friendly or contempt response? Fortunately, insights from social sciences can provided a light about these relationships, making this field a niche for interdisciplinary collaboration.

Furthermore, **social interactions comprises different elements occurring in parallel, which makes it difficult to analyze its components individually**. During a conversation, for example, we not only speak but can also gesture with our hands and head or shift our body weight at the same time. All these movements and different social cues are registered by the sensors as one, and are semantically intertwined. Hence, one should not only try to separate these cues from unified signals, but also study the reason these cues are performed together in the first place, how does this differs from person to person and how does it affects the overall interaction.

Finally, **most human social interactions during free-standing conversations occur *in-the-wild***. We casually start a conversation with our co-workers while grabbing a coffee or, as in the example in Figure 3.3(a), during a mingle event. This brings along a technical challenge, as the sensing and data collection becomes rather complicated. For such cases, the natural behavior of the people involved in the interaction should not be disrupted or affected. Luckily, the development of new technologies makes this ubiquitous sensing possible nowadays.

1.2. UBIQUITOUS TECHNOLOGIES TO STUDY NON-VERBAL BEHAVIOR

In a matter of 40 years, microprocessors have decreased in size from micrometers (10μm with the Intel 4004) to a few nanometers (from 45 to 32 for the current Intel microprocessors), while also increasing in power efficiency, following the famous *Moore's law* [146]. This has allowed its portability, triggering among other fields the rise of mobile technologies, personal computers and embedded systems.

(a)

(b)

Figure 1.2: (a) Multiple social interactions occurring during a mingle gathering. (b) The *Chalcedony*, a smart badge used for all experiments in this thesis.

The same exponential improvement applies to wearable sensors and actuators, mostly due to the insights and developments in Micro-Electro-Mechanical systems (MEMS). This parallel phenomena, called *More than Moore's law*, is less related to size and power of the technologies and more to additional functionalities, such as sensors and actuators [86]. Thus, different types of sensors (e.g. movement, location and temperature) can now be embedded in portable devices accomplishing rather precise measurements.

The combined development of microprocessors and wearable sensors has allowed pervasive and ubiquitous technologies to rise as well, aiming to leverage these new 'cool gadges', and the algorithms that come with them to automatically interpret the world surrounding us and enhance our perception of it. Although mostly used for medical purposes, its use in other fields including Social Signal Processing was rather straightforward.

As mentioned before, one of the main challenges of analyzing human social interactions through their non-verbal behavior is the necessity to do so without interfering with the natural behavior of the people involved. This is particularly tricky for scenarios *in-the-wild*, and is the main reason for the use of these new ubiquitous technologies.

Remember the last time you were in a conference and they gave you a badge with your name? Imagine if that badge had the capacity to store new acquaintances' contact information just by standing close by, let you know where the next talk you want to see is, or even tell you if the person you are talking with is not interested in the conversation. The use of ubiquitous technologies can make these somehow utopic ideas possible. For example, Figure 3.3(b), shows the *Chalcedony*[2], a smart badge that can be used to record body movement and body proximity in a pervasive way during a networking event, thus enabling the assessment of non-verbal behavior of the participants through their movement and proxemics.

Moreover, aside from its pervasiveness, there are other main advantages for the use of

[2]The Chalcedony badges were release in 2009. Since then, more efficient devices have been created.

ubiquitous technologies. Among others benefits, there are their low costs, scalability and reproducibility, and rather easy use for both the researchers and the final users. And this is also what makes them so attractive for the study of social signals.

1.3. MULTIMODAL INFORMATION: COMBINING SENSORS

One of the main components of the work in this thesis is the leveraging of the complementarity of different modalities for the study of human social behavior (see Section 1.4 for our specific scenario). There are two main reasons for the use of complementary modalities in non-verbal behavior analysis: 1) as humans are multimodal entities they should be sensed/recorded accordingly, and 2) to overcome missing data due to noisy environments. Note that one is a semantic reason, while the other one is more practical.

Firstly, while studying the different types of social cues (see [169]) one can notice that humans are multimodal[3] entities by nature. We do not limit our communications to words, but use one or more other media to transmit our messages to others (e.g. visual gestures or change of body weight to emphasize a point while speaking). So, if we produce social signals via different channels, the sensing to such signals should be done accordingly. Therefore, to have a complete understanding of the interaction one should sense all types of social cues according to the way they are transmitted.

Fortunately, the new ubiquitous technologies previously mentioned allowed researchers to sense and collect different modalities with a wide variety of low cost sensors that can be easily embedded in wearable sensors or deployed in sensing venues without much additional work. Moreover, in some cases these sensors are already part of the infrastructure (e.g. surveillance cameras) or are embedded in objects that are part of our daily-life (mobile phones).

Secondly, due to the real nature of the recording settings, the sensing or acquisition of social signals can be compromised due to its noisy environment. For example, a person can go to the bathroom during a party and engage in a conversation with another person there. For such case, the use of video to record their interaction is frowned upon. However, other modalities that can still maintain the people's privacy such as wearable sensing of body movement or proximity, can compensate for the lack of video.

Moreover, each modality provides different information about a given action, which also relates with the scenario being studied. For example, a laugh can be more distinctive in audio than in video during a meeting, but during a noisy mingle event, video or a combination of audio with video can be a more reliable source of information. This complementarity can vary given the situation, the type of sensors or even the event itself. Thus, the modalities used for a given task can not be chosen arbitrarily. Instead, one should study the nature

[3]The term *multimodal* will be used to define information from different sources (e.g. sensor types) and not multiple modes in a distribution. Thus, a camera recording only video will be considered as unimodal wheres a camera recording video and audio will be multimodal.

of the event or interaction and determine the best sensor and modality type to record it with.

Furthermore, previous work has shown that leveraging the multimodality of social interactions further improves our comprehension of them. Alameda-Pineda et al. [2], for example, showed that combining video, audio and infrared data (for proximity) increased the accuracy performance of head and body estimation during a free-standing conversational group setting (typical study case for social interactions in the wild). They also found that when data is missing in one modality, the use of other modalities compensates for the missing information.

Specifically, in this thesis we dealt with 3 different types of modalities: **video, body movement and body-to-body proximity**. These are collected by using video cameras, wearable acceleration and wearable proximity, respectively.

1.3.1. CHALLENGES OF USING MULTIPLE MODALITIES

Although using multiple sensors has proved to provide complementary information while analyzing human social behavior, there are several challenges when using them jointly. These challenges are mostly from the practical perspective and become more strict as the scenario goes from lab environments towards natural and *in-the-wild* conditions. And although immensely important as a previous step to the data analysis, these are generally underestimated.

Mainly, these challenges are related to 1) the development of devices and systems for recording an specific event or interaction, 2) the deployment of the entire system, and 3) the synchronization and association of the different signals for the same person, during or after the event. In this thesis, we address the importance of these three challenges as an integral part of automated social behavior analysis. In fact, challenges 2 and 3 are inherently part of the proposals in Chapters 2 and 3, respectively; while overcoming the first challenge makes it possible for any multimodal sensor platform to exist nowadays.

To better explain these challenges, one could use a generic example about data collection for social interaction analysis. Imagine that we require a system that can record the video and the wearable body movement of all participants during a party to determine if they are having a good time.The main requirement that all signals from all the devices must be synchronized with each other and with the video. Also, in order to properly use the multiple modalities, we want to know who is using which device in the video.

Development of the devices
First, let us focus on the wearable devices. Although easily available at a low cost, most sensors and computing units (processors) are still commercialized to be general enough to provide solutions to a wide range of problems. As a consequence, for a specific task such as recording the body movement of people during a mingle event, we will need either a custom-made device or modify an open-source platform towards our needs. Either option

requires that the necessary modalities (sensors) are or can be embedded onto the device. We also should give it the capabilities of storing locally the data from all sensors, or communicate its data to a sink node (central computer or device). In both cases, the firmware of the device must be created or modified to give the wearable device the capacity of recording following an automated preset schedule (e.g. check all sensors work, record every Xms, store/send every minute). Note how there are several design decisions to be made just for the devices, which will affect directly the data that will be collected.

It is true that there are some alternatives to create such device from scratch, but these are mostly limited to commercial licenses[4] which are generally expensive. Moreover, even when available such commercial options will still require some level of modification to fit the exact needs of the task at hand. It is not the same recording the social interactions of people during a calm drink party, than recording them during a dance party.

Deployment of the acquisition system
Continuing with our generic example, the devices should be synchronized not only with each other but also with the camera(s) recording the event. To do so, one could connect all cameras to a main computer, that can also communicate with the devices wirelessly (if implemented), this for an online solution. Alternatively, and as can be seen in Figure 3.3(a) on the left side, one could record timestamps from the devices in the video feed, to later synchronize the signals in an offline manner. Either way, the chosen solution must align with the event needs and the possibility to deploy the chosen system as umintrusively as possible, in order to maintain the natural behavior of the participants.

Furthermore, the deployment of any acquisition system comes with other specific technical aspects. Which sampling rate should the devices and cameras use? How long must the battery in the devices and cameras last? Can all the data (both cameras and wearable devices) be stored locally or must it be sent to a sink cluster? To answer such questions one should study the event itself, the research questions at hand, what the main goal of the data collection is and what an acceptable trade-off is between the system's capabilities and the requirements of the data to be collected.

Synchronization/associations of signals
We described above the synchronization steps necessary for the signals from different sensor types. But what about the information from different people in the same signal? For example, in the video signal (e.g. Figure 3.3(a)) multiple persons are recorded at the same time.

To obtain personalized data for all participants, how can we associate a specific wearable device with the person wearing it in the video? One could associate them manually, which is in fact the approach taken by many works using video and wearable modalities these days [2, 79, 182]. However, this requires manual tracking of the people, accounting for their exit or reentry of the field of view of the camera(s), and manual association of all devices to these tracks.

[4]Only at the end of this thesis work the Open Badge by the MIT Media Lab was released [96].

Although viable for a small number of people (e.g. small meetings), this manual approach becomes unfeasible, time consuming and costly as the number of people involved in the events increases. Imagine, for example, the cost and time needed to associate 100 participants to 2 hours of video on an event recorded by 8 cameras. Not to mention that this makes online approaches rather impossible to apply.

The example described above was, in fact, **exactly what was encountered during this thesis**. The creation of such a system, including the recording of wearable proximity, its deployment and further data collection during a real mingle event lead to the creation of a new dataset, specifically for the analysis of social interaction using multiple sensors (Chapter 2). Also, an automatic approach to associate wearables to the regions in video of its wearers was conceived, and becomes a key first step for the analysis of non-verbal behavior from multiple sensors that is discussed in Chapter 3.

1.4. MINGLE EVENTS: OUR SCENARIO

In previous sections we have discussed the context in which this thesis work lies. Now, we will narrow down to the specific scenario studied. As mentioned before, this thesis focuses on mingle[5] scenarios, which are crowded in-the-wild scenarios where people are inherently encouraged to interact as part of the event itself, such as cocktails parties, poster sessions and networking events.

Figure 4.1 shows examples of different mingle scenarios, provided by previous efforts in the analysis of social interactions. First, Figure 4.1(a) presents the *Cocktail Party* dataset[182], one of the first efforts to collect multimodal data for the analysis of human behavior during this type of scenarios. Figure 4.1(b) shows the *Idiap Poster* dataset[79], a dataset designed to analyze group formation dynamics. Finally, Figures 4.1 (c-d) illustrate the *SALSA* and *MatchNMingle* datasets. The latter two dataset are the state-of-the-art multimodal resource for the analysis of human behavior during mingle scenarios. More about these datasets and their comparison to our *MatchNMingle* dataset can be found in Chapter 2.

As can be seen in these examples, mingles events present several parallel instances of free-standing conversational groups, which are naturally emerging small groups of two or more conversing people. Such spatial formations, also called F-Formations [88], dynamically form, merge, and dissolve according to the goals and desires of each person within the group.

Hence, mingle gatherings compile several unstructured face to face interactions occurring dynamic and simultaneously. Due to its social setting, these are a perfect example of social interactions occurring naturally and *in-the-wild*. Also, unlike small meetings where normally the task is assigned beforehand [66], the interactions during mingle scenarios follow

[5] *Mingle*: Move among and engage with others at a social function (definition provided by the Oxford dictionary).

Figure 1.3: Examples of mingle scenarios. (a) *Cocktail party* dataset [182], (b) *Idiap Poster* dataset [79], (c) SALSA dataset [1], (d)*MatchNMingle* dataset (our dataset)

.

the goals and desires of the participants themselves.

Moreover, these scenarios present an ideal opportunity to deploy and test the use of multimodal ubiquitous technologies for the sensing of non-verbal communications during real social events. All the above makes these types of scenarios intriguing cases of study for face-to-face social interactions occurring the *in-the-wild*, and were the main reasons to be selected as topic of this thesis.

Towards analyzing mingle events with multiple sensors

As mentioned before, we aimed to analyze real social interactions *in-the-wild* during a mingle event, without disturbing the natural behavior of the people involved. Following this statement and studying the type of event itself (as was discussed in Section 1.3) helped us select the most adequate type of sensors for a multimodal data collection.

Not all modalities are adequate to record these types of events, given its nature. For example, the audio of the event itself (e.g. the non-stationary noise of many groups of people conversing in a pub or cafe) will interfere with audio recording, making it hard to separate specific speakers without additional equipment (e.g. wearable microphones). And introducing this equipment will compromise the privacy of the participants of the event. Finally, one should remember to preserve the natural behavior of the people while interacting. This

means that the sensors chosen for recording the event must be pervasive enough while still obtaining feasible data for the analysis of the event.

As a consequence, the sensors chosen for the work in this thesis are **video cameras** and **wearable sensors** recording body acceleration and body-to-body proximity. These sensors offer the pervasiveness necessary to guarantee a natural behavior during the event and provided enough complementary information to analyze the interactions.Moreover, the use of the techniques describe in this work can be replicated rather easily in conventional devices, such as smart badges or mobile phones (the latter with some changes in the experiment design), as these sensors are embedded in most commercial devices nowadays.

1.4.1. CHALLENGES OF ANALYZING MINGLE SCENARIOS

These unstructured social scenarios are rich in information but present several challenges when processed automatically. First, as they are recorded in-the-wild, these events are susceptible to noisy or missing data. For example, people can leave the field of view of the cameras at will or devices can fail while recording the event. Both cases result in missing data, either partial or complete. Also, all sensors are subjected to noisy data acquisition. People might occlude other people in the cameras or sensors might get physically misplaced during the event (e.g. participants taking off or manipulating their smart badge). This results in noisy data that is more complex to analyze. Overcoming this challenge is the main reason of the use of multiple modalities.

Furthermore, given the nature of the event itself, several interactions occur at the same time, and each person will mostly perform more than one social cue in parallel (e.g. speaking and gesturing at the same time). Because of these two reasons there are a strong cross-contaminations, both of interpersonal and intrapersonal data. Take Figures 4.1(a) and (c), for example. Here most state-of-the-art person/object detectors will apply a bounding box to extract the region of interest (ROI) for a person, which has a high probability of also including the limb or torso of another participant.

From a Computer Vision perspective, these scenarios also present several challenges while using techniques solely-based in this field. This is mostly due to occlusions, changes in light conditions, shadows or strong changes in the appearance of the people given their position with respect to the camera. Nevertheless, video is still a rich source of information that can be complemented with other modalities.

Additionally, note for example how in Figure 4.1(d) for the people in the center of the image we can only see their head and shoulder, while the torso (and even face) of those in the border of the image (away from the camera) can be seen more clearly. Nonetheless, these appearance changes are preferable than total occlusions. This is the main reason why overhead views are preferable for recording crowded mingle scenarios, instead of elevated side views (such as those in Figure 4.1 (a,c)). The latter are more prone to strong participant occlusions, especially for those people who are farthest away from the camera.

1.5. CONTRIBUTIONS OF THIS THESIS

Overall, this thesis encompasses topics related to the leveraging of multiple modalities for the analysis of social behaviors during mingle scenarios. Nonetheless, first the reader must notice that this analysis can be performed at different levels of abstraction (see Figure 2.2), from simple raw signals to more complex concepts, depending on the task. For example, while detecting group formations is an important first step for further analysis of groups dynamics, the analysis of this task can be considered as a preliminary step (and of lower abstraction) to determine the personality of the people from the way they interact with others.

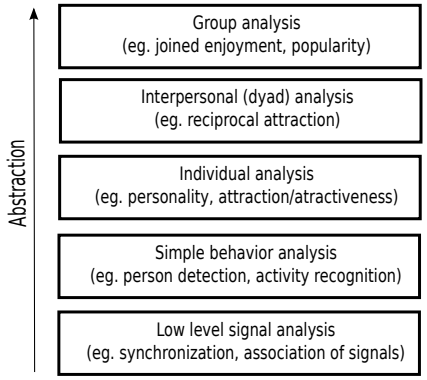


Figure 1.4: Different levels of abstraction in which the analysis of human social interactions during mingle scenarios can be made.

In summary, this thesis contributes to the scientific community with a new state-of-the-art dataset for the analysis of human interactions, which can be used in the analysis of any of these levels of abstraction. Additionally, it addresses the automatic association problem between video and wearable modalities, directly overcoming the challenges described in Section 1.3.1 related to modality association. Moreover, it tackles three different tasks within the analysis of social interactions, each at a different level of abstraction: social hand gesture detection, personality estimation and group enjoyment.

Creation of a multimodal dataset for the analysis of human interactions
As the analysis of human interactions in mingle environments using multiple sensors is a rather new topic, just a few datasets were available at the start of this thesis. Moreover, none of them were specific enough to be used as a resource to answer the research questions proposed. Hence, our *first contribution* is the design and collection of the *MatchNMingle* dataset, a multimodal resource for the analysis of social interactions in-the-wild in the form of free-standing conversational groups during mingle scenarios and of seated dyads during speed dates.[6] This dataset consists of 2 hours of data from wearable acceleration, binary proximity, video, audio, personality surveys, frontal pictures and speed-date responses

[6] *MatchNMingle* is available for research purposes at `https://matchmakers.ewi.tudelft.nl/matchnmingle/pmwiki/pmwiki.php` under an End-User License Agreement (EULA).

for 92 people, in addition to manual annotations for participants' positions, conversational groups (F-formations), and social actions (e.g. speaking, hand gesture) for 30 minutes at 20fps. This makes it, thus far, the dataset for analysis of social interactions with the largest number of participants, longest recording time and largest set of manual annotations for social actions available in this context in a real-life scenario (Chapter 2).

Automatic association of different modalities

The *second contribution* of this thesis is addressing the complex problem of associating several wearable devices with the spatio-temporal region of their wearers in video during crowded mingling events using only acceleration and proximity. This is a particularly important first step for multi-sensor behavior analysis using video and wearable technologies. Nonetheless, manually associating a specific device to a region of the video (corresponding to the person using the device) quickly becomes a challenging practical issue as the number of streams to associate per modality increases, making the correct associations harder to discriminate. Here, we address large-scale data association (69 devices compared to 3 in previous efforts) in challenging crowded environments with missing data, leveraging the use of proximity information obtained from the wearable devices and video as a spatial prior to the association process. Our experiments showed that leveraging such proximity information, even when imperfect, further helps discriminating between the streams. (Chapter 3).

Detection of gestures during conversational interactions

The *third contribution* of this work is the multimodal detection of hand gestures during free-standing conversations in crowded mingle scenarios, focusing on gestures that are inherently conversational instead of just symbolic gestures (as has been done by most state-of-the-art efforts). To do so, we use videos from overhead cameras and wearable acceleration data recorded via smart badges. Unlike the scenarios of previous works in gesture detection and recognition, crowded mingle scenes have additional challenges such as visual cross-contamination between subjects, strong occlusions and non-stationary backgrounds. To overcome these challenges we applied a Multiple Instance Learning (MIL) approach, clustering our video trajectories into *bags-of-trajectories*; and leverage the complementarity of video and wearable acceleration in a decision fusion manner. By leveraging the complementarity of video and wearable acceleration in a decision fusion manner we show improvements over the unimodal approaches. Also, we present a static and dynamic analysis to assess the impact of noisy data on the fused detection results, showing that moments of high occlusion in the video are compensated by the information from the wearables (Chapter 4).

Estimation of self-assessed personality

Our *fourth contribution* is the automatic classification of self-assessed personality traits from the HEXACO inventory during crowded mingle scenarios. Most state-of-the-art efforts addressed personality estimation in a rather controlled setup, whereas mingle scenarios have a challenging social context as people interact dynamically and freely, resulting in noisy and missing data. To address this challenging setup, we leverage the use of wearable sensors recording acceleration and proximity, along with video. We separated our features

in 3 different behavioral modality types (e.g. movement, speech, proximity) coming from 2 sensors (wearable and camera) and compared the performance of the different types on the estimation of the personality traits. Our experiments showed that features corresponding to movement dynamics, coming from both video and wearable acceleration, are a feasible alternative to address crowded and dynamic scenarios such as mingle events. Also, we found that different behavioral modalities, even when originating from the same digital modality (sensor), have complementary information that better helps the classification task. Finally, we found that although in theory movement features (e.g. variance of the acceleration) are similar when recorded from cameras compared to a worn accelerometer, in practice they have low correlations. We found that each modality records the event differently and provided complementary information (Chapter 5).

Estimation of group enjoyment

Our *final contribution* comes with the use of multiple streams to assess the group enjoyment of a live performance. Thus, we leverage tri-axial accelerometers, worn by ordinary members of the public during a dance performance, to predict responses to a number of survey answers, comprising enjoyment, immersion, willingness to recommend the event to others, and change in mood. Also, we analyze how behavior as a result of seeing a dance performance might be reflected in a people's subsequent social behavior during mingle event using proximity and acceleration sensing. To do so, we recorded the behavior of the participants during 2 rounds of mingle, one before and one after the performance. Using the movement signals of the participants while watching the performance, we were able to predict whether they enjoyed the event (balanced classification accuracy of 90%). Also, we were able to correlate salient moments of the event to participants' shared movement. These results show that even during a seated situation where there is not much movement, we still can predict someone' s enjoyment of the performance. In addition, we computed the difference between the proximity behavior for the 2 rounds, and the correlation between the reported change in mood and the difference between the variance in acceleration magnitude in the rounds. We found a similar networking behavior with respect to proximity between the two rounds but a positive correlation of 0.60 for the variation in acceleration and a self-reported change in mood as a result of watching the performance. Our analysis suggests that while the participants tend to act similarly as a group in terms of networking behavior, the quality of the interactions between rounds may have been different (Chapter 6).

2

MATCHNMINGLE: A MULTIMODAL DATASET FOR THE ANALYSIS OF MINGLE SCENARIOS

The contents of this chapter are based on the work originally published in:
L. Cabrera-Quiros, A. Demetriou, E. Gedik, L. v.d. 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**. To appear in Transactions on Affective Computing, 2018.
MatchNMingle is available for research purposes at `https://matchmakers.ewi.tudelft.nl/matchnmingle/pmwiki/pmwiki.php` under an End-User License Agreement (EULA).

2.1. INTRODUCTION

One way to study human beings as social entities is to study their nonverbal behavior (i.e. all aspects of behavior except language) while they interact. These nonverbal behaviors are a commonplace part of the everyday interaction of people, and a fundamental aspect of our daily life.

Moreover, ubiquitous technologies have allowed researchers to automatically analyze human social behavior without disturbing their natural interaction. As a consequence, specific domains such as Social Signal Processing (SSP) have emerged which seek to give computers the capacity to accurately perceive, interpret and/or display social signals and social interactions from sensors (e.g. video, audio, wearables) [169, 170]. While these endeavours have benefits for areas such as Human Computer Interaction, Affective or Social Computing, leveraging ubiquitous technologies can also be beneficial for the field of social psychology itself by providing an inexpensive and easy way to collect and analyze data from social interactions.

One of the more common forms of social interactions appears during free-standing conversational groups, which are naturally emerging small groups of two or more conversing people. Such spatial formations (known as F-Formations [88]) dynamically form, merge, and dissolve according to the goals and desires of each person within the group. These unstructured social scenarios are rich in information but present several challenges when processed automatically.

In this chapter we introduce ***MatchNMingle***, a multimodal/multisensor dataset created specifically to contribute to the efforts to overcome the challenges of the automatic analysis of social signals and interactions. This dataset consists of about 2 hours of uninterrupted recordings for 92 people, and comprises cases of conversations in free-standing groups and sitting dyads. *MatchNMingle* was collected in an indoor in-the-wild scenario, during 3 real speed date events, each followed by a mingle/cocktail party. As it was recorded during a speed date event, *MatchNMingle* also has the additional component of a romantic attraction setting. Thus, all participants in the event have an actual goal during the evening event of finding new friends or a romantic partner.

The main contributions of this chapter are:

Multimodal dataset

- We collected multimodal data (eg. acceleration, proximity and video), using wearable devices and cameras, for over 60 minutes of dynamic social interactions for **92 participants** attending one of 3 speed date events in a public pub followed by a mingle session/cocktail party.
- We leveraged the use of smart-badges and surveillance cameras to collect dynamic in-the-wild data (instead of the usual lab-setting) for the analysis of dynamic social interactions in a non-intrusive manner. Thus, this dataset has strong changes in appearance, lighting conditions, shadows and occlusions in video.

Interdisciplinarity

- We designed the data collection in a way that adheres to the standards of both the social and data sciences, and can be used by both fields.

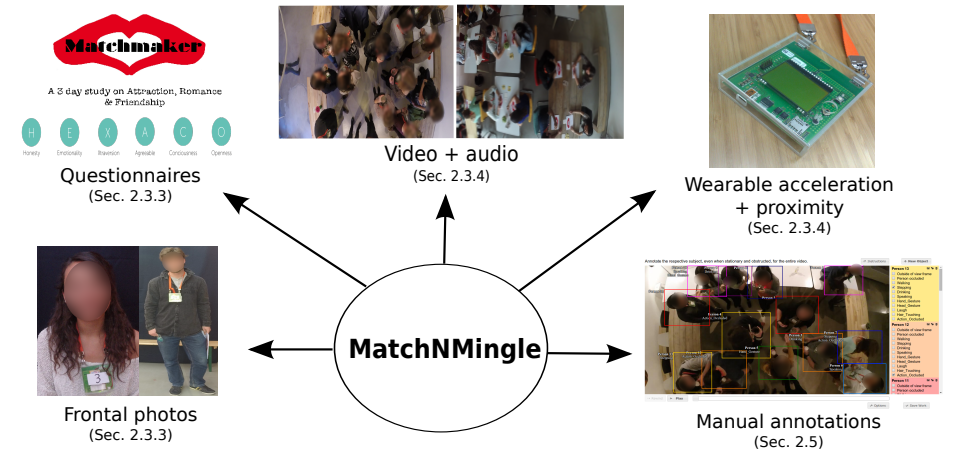


Figure 2.1: Visualization of all the modalities included in *MatchNMingle*.

Manual annotations

- We reported 30 minutes of fine-grained manual annotations of video for social actions (eg. speaking, walking, hand gesture, head gesture, hair touching) with a resolution of 20 frames per second, for over 36000 frames annotated. Additionally, we reported F-Formation manual annotations for 10 minutes of the mingle session.
- We compared crowdsourcing tools (eg. Amazon Mechanical Turk or MTurk) with trained annotators for tasks of low and high complexity, specifically position of people in video against social action annotations. This comparison shows that, although widely used, MTurk has limitations on the type of HITS that will result in high inter-annotator agreement.

Self-reported data

- We provided the HEXACO scores for personality trait (6 dimensions) and speed date responses (6 questions per date, maximum of 15 dates per participant) from all participants to be used as self-assessed ground truth on works related to personal differences or attraction preferences during a speed date event.

2.1.1. MOTIVATION FOR CREATING *MatchNMingle*

There were 4 main reasons that motivated us to create *MatchNMingle*. Firstly, we wanted to provide the research community with an open-access resource for the analysis of the non-verbal behavior during natural social interactions that captures the multimodal nature of the event by recording data with multiple sensors. We focused on cases with free-standing conversational groups, as these triggered one of the more common forms of social interactions: F-Formations [88].

Secondly, we wanted to design and record an event where the same participants were involved in 2 different natural contexts, structured sitting dyads (speed-dates) and an unstructured mingle setting, that happened one after the other. Thus, one could study the effects of one context on the other, among other open questions.

Table 2.1: Summary of all the elements included in *MatchNMingle*. Unless stated otherwise, all data is publicly available.

Sensor/Input	Modality/Survey	Details
Questionnaires	HEXACO	Scores and sub-scores for each trait.
	SOI* SCS*	
	Date Responses	All dates in the event. See Section 2.4
Hormone baseline*	Cortisol Testosterone	Collected using hair samples.
Cameras	Video	9 overhead cameras recoding both the speed dates and mingle.
	Audio	General audio from the event.
	Frontal Photos	Face(neutral/smile) + full body.
Wearable Sensors**	Acceleration	Triaxial at 20Hz for entire event.
	Proximity	Binary values at 1Hz for entire event
Manual Annotations***	Positions Social Actions	$30min$ at 20 FPS for the mingle (Social actions detailed in Section 2.5)
	F-Formations	$10min$ at 1 FPS for the mingle

*Due to privacy reasons, these elements are not publicly available.
**Due to hardware malfunction, only 70 of the 92 devices worked properly for the entire event (72 for dates). See more in Section 2.4.2.
***Position and social action annotations were performed by 8 different annotators. More details in Section 2.5.

Thirdly, we intended to study the effects of initial romantic attraction on non-verbal behavior, based on self-reports of people that were not already acquainted. In particular, our aim was to capture data of the moments when a pairbond might begin, and to present the data in such a way it would allow for fruitful research regarding romantic attraction to be conducted.

Finally, we seek to trigger the collaboration of social and data scientists, by collecting a dataset that follows the specifications of (and can be used by) both fields.

2.1.2. WHAT IS INCLUDED IN MATCHNMINGLE?

A comprehensive summary of all the elements include in *MatchNMingle* is shown in Table 2.1. Unless specified directly, all the data is publicly available. [1] Also, Figure 2.1 presents a visual summary of all the modalities of the dataset. Details about each component/sensor type of the dataset can be found in Section 2.3.

Similar to previous efforts ([1, 43, 79]), participants' positions and F-Formations were manually annotated. But most importantly, *MatchNMingle* also provides manual annotations for social actions (eg. speaking, hand gesture) for 30 minutes at 20 FPS, making it the first dataset for automatic analysis of free-standing conversational groups to incorporate the annotation of such cues in this context.

Each day event (from 3) consisted of a speed dating round ($3min$ date with participants of opposite sex) immediately followed by a mingle party of about an hour where participants could interact freely following their own desires and intentions. Details of data collection procedures can be found in Section 2.3.5.

[1] *MatchNMingle* is available for research purposes at `https://matchmakers.ewi.tudelft.nl/matchnmingle/pmwiki/pmwiki.php` under an End-User License Agreement (EULA).

MatchNMingle is, to the best of our knowledge, the dataset with the largest number of participants and longest recording time, that is publicly available in the context of free-standing conversations. Also, it is the only dataset with manual annotations for social actions for this context. In addition, the data were collected in a specific social context (where we would expect attraction to occur) but it is not limited by it as a wide range of types of social interactions also occurred (eg. friends coming together to the event). Finally, *MatchNMingle* is the first dataset for the automatic analysis of first encounter interactions within romantic settings that is publicly available.

2.1.3. POSSIBLE USES OF MATCHNMINGLE

Although *MatchNMingle* was created for the analysis of social interactions in the wild, possible uses of the dataset are not limited to this specific domain. Figure 2.2 shows the different levels of abstraction, from raw signals to more complex concepts, in which analysis can be done using *MatchNMingle*. Hence, research about simpler components within the interactions (eg. activity recognition, people detection/tracking with high camera perspectives or group detection) can also benefit with the use of the dataset.

Overall, *MatchNMingle* was created as an exploratory resource so it can be used to answer multiple research questions in different research domains, including (but not limited to) Ubiquitous computing, Affective Computing, Social Signal Processing (SSP), Computer Vision-Pattern Recognition and Social Psychology. A suggestion for the reach of these areas (by no means definitive) is also presented in Figure 2.2.

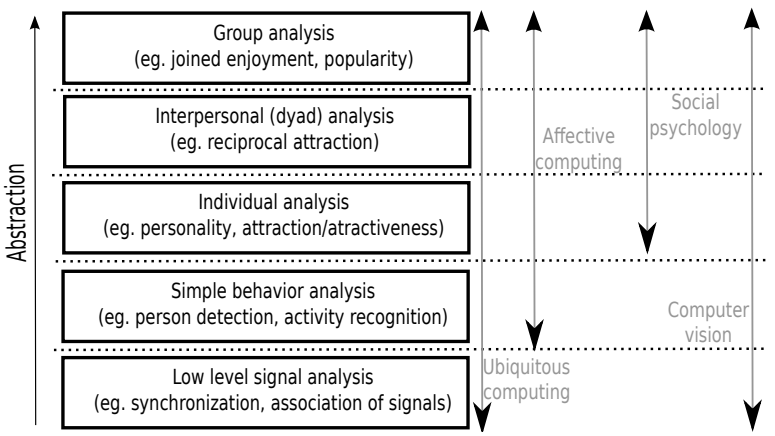


Figure 2.2: Levels of abstraction while studying social interactions in which *MatchNMingle* can contribute as a new multimodal resource. A suggestion for the reach of scientific areas (no definitive) is also presented.

Moreover, there are 4 key novel aspects of *MatchNMingle* that can trigger new and exciting research: 1) its annotations that are focused on the social context instead of everyday activities or spatial descriptions (eg. body/head orientations), 2) its romantic setup, 3) the high number of multiple groups forming and splitting dynamically, which allows better generalization in topics such as group dynamics, and 4) the possibility to study the relationship between 2 different settings (sitting dyads and conversational groups) with the same

people, and its relation to the interests of each participant (eg. attraction).

The first point allows to analyze the social component of the interaction in a more deeper level, which was not possible (without additional annotation work) with other datasets. Secondly, as the first dataset to present publicly available sensor data and responses of a free speed dating event, *MatchNMingle* provides a key resource to analyze the relation between non-verbal behavior and attraction/attractiveness. Thirdly, although other works have recorded mingle scenarios, MatchNMingle is the first to collect data of spontaneous interactions for such a high number of people (92 compared to 18) in such a fine-grained time resolution. This reflects directly in the number of groups and their dynamic behavior. Finally, the two different but consecutive settings (structured and unstructured) allows the study of the people's behavior within changing scenarios. For example, for the analysis of attraction, one could study the relationship between the matches in the dating part and the group formations in the mingle.

Can we understand the link between non-verbal behavior and the person's intentions or desires during a freely occurring social interaction? And, can we detect this automatically? Perhaps these are the ultimate questions that researchers might aim to answer with *MatchNMingle*.

Thus far, our research team used this dataset for diverse topics with various levels of abstraction (in increasing order): multimodal data association [28], gesture detection [30], speaker detection from wearable devices [67, 68], personality estimation [27] and perceptions of attractiveness [48]. Still, many possibilities for using MatchNMingle as an unimodal or multimodal resource are largely untapped.

We strongly believe that there is even wider range of possibilities and open questions that can be answered using *MatchNMingle*, and hope that the presentation of this dataset will encourage collaboration and scientific inquiry.

2.2. RELATED WORK

We will focus on related datasets that allow 1) analysis of free-standing conversational groups and 2) speed date events. Different communities have made efforts to detect, track and analyze groups and face-to-face interactions using mobile phone technologies. However, we do not refer to works that addressed the problem on a large scale and with a broader view than a fine-grained analysis (e.g. over the course of weeks), and therefore considered these outside of scope. For a survey on sensing using mobile phone and its use during social interactions (among others), refer to [179].

Also, although participants are seated during the Speed Dates in our dataset, we will not refer to works in analysis during seated conversations (e.g. meetings) as we will focus specifically on works about speed dates. For more detail on group analysis during sitting conversations, refer to [66].

2.2.1. FREE-STANDING CONVERSATIONAL GROUPS

Most efforts on the analysis of free-standing conversational groups have focused on group detection (or F-Formation detection), and the use this information for further analysis of the group. Thus, we follow a similar approach in this review.

For a summary, Table 2.2 shows a numerical comparison of datasets oriented specif-

Table 2.2: **Free-standing conversational groups and face-to-face interaction.** Numerical comparison with other datasets.

Dataset	Number of people	Total time (minutes)	Numb. annotated frames		Maximum group size(***)	Scenario (context)
			F-Formations	Social Actions		
Cocktail [182]	6	30	320	0	6	Mingle in a lab environment
CoffeeBreak [43]	10	-	120	0	-	Outdoor mingle in social event
Big Game [74]	32	30	600	0	4	Indoor quiz game in teams
Idiap* [79]	50	360	82	0	5	Indoor poster session
Salsa* [1]	18	60	1200	0	7	Indoor poster session+Mingle
MatchNMingle*	**92**	120	**4200****	**36000**	8	Indoor SpeedDate event+Mingle

*Dataset is (or will be made) publicly available.
**Every second for 10 minutes of the mingle + every frame during the speed dates.
***Obtained from F-formation annotations provided by each work.

Table 2.3: **Free-standing conversational groups and face-to-face interaction.** Sensor comparison with other datasets.

Dataset	Sensors			
	Video	Audio	Wireless	Accel.
Cocktail [182]	X			
CoffeeBreak [43]	X			
Big Game [74]	X		X	
Idiap* [79]	X			
Salsa* [1]	X	X	X	X(**)
MatchNMingle*	X	X	X	X

*Dataset is (or will be made) publicly available.
** SALSA provides processed accel., instead of raw triaxial.

ically to f-formation detection and free-standing conversational groups, while Table 2.3 compares these in terms of modalities.

VISION-ONLY DATASETS

The *cocktail party* dataset, published by Zen et al. [182], was one of the first datasets designed specifically for the analysis of free-standing conversational groups. This dataset consists of a mingle involving 6 people recorded by multiple cameras and was used to explore the relation between people's proxemics, their visual attention, and their personality.

The *CoffeeBreak* dataset by Cristani et al. [43] has been used in several works on the detection of people's position and orientation in images, and in detection of F-Formations [149, 150]. This dataset consists of the free interactions of 10 people. Hung and Kröse proposed the *IDIAP poster* dataset [79], which consist of a poster presentation, to which 50 people attended, recorded from above and was used by the authors for F-Formation detection using dominant sets.

WIRELESS COMMUNICATION DATASETS

Along with video, works using wireless communication have had a significant impact in group detection. For example, Cattuto et al. [33] collected data from wearable RFID devices, worn by 25 to 575 individuals during different social gatherings. As they stated, most efforts at the moment either: 1) scale to millions of mobile devices but provide no information about face-to-face interactions, or 2) collect rich data on face-to-face interactions under lab conditions, at high cost on deployment. The aim of their dataset was to achieve

a balance a between scalability of device's deployment and resolution, while monitoring social interactions.

The sensing platform on [33] was later used by Isella et al. [82] to collect face-to-face interaction data for more than 14 000 attendees at a Science Gallery, and a conference. In this work, the authors focus on a deep analysis and a comparison of each event, in terms of its context.

Martella et al. [106] collected data from wearable devices recording proximity from 137 participants during a IT conference. Thus, they could detect group formations using dynamic proximity graphs. Similarly, Atzmueller et al. [10] collected data from 77 RFID tags used by participants during an introductory freshman week. They analyzed the face-to-face interactions of participants, and investigated the relation between spatial and social networks, and gender homophily. Matic et al. [108] collected proximity data from 24 participants, wearing a mobile phone in a known place, in order to detect social interactions through proxemics obtained from RSSI values. Unlike *MatchNMingle*, here the participants were instructed to (randomly) talk with each other, so these interaction were natural up to a certain point.

These datasets use a large number of devices, but share the disadvantage of not having an actual ground truth for the group formations. Thus, there is no accurate way of assessing if the interactions detected by the devices indeed have a social component, from an F-Formation perspective.

MULTIMODAL/MULTISENSOR DATASETS

The main advantage of current multimodal/multisensor datasets is that they provide the high scalability of wireless communication approaches for proximity, while also having video and/or audio to either use as ground truth or as complementary source of information.

Using this approach, Hung et al [74] provided the *Big Game* dataset, which consists of 32 subjects playing a quiz game in teams. This dataset was initially used to classify social actions (eg. speaking), and later used to detect conversing groups from wearable acceleration, using video as ground truth [75]. Also in [108], Matic et al. collected another set of data in a multimodal approach including accelerometers and proxemics (RSSI values). Although recordings were provided for 7 working days, only 4 subjects (officemates) participated in the collection of the data, which was later used to detect social interactions.

The SALSA dataset by Alameda-Pineda et al. [1] is the work that most closely resembles the MatchNMingle dataset. SALSA consists of recordings of 18 previously acquainted participants during a poster session, followed by a mingle, similar to ours. They collected video from multiple cameras, wearable acceleration, and IR-based proximity using a commercial version of the sociometer [39]. In addition, they gathered information about personality traits using the Big-5 [50], and annotated participant's position, head/body orientation, and F-Formations.

Compared to SALSA, *MatchNMingle* has over 5 times the number of participants (92) and double the recording time. This results in a more dynamic scenario where people change groups more regularly (see Section 2.4.3), and a large distribution of groups sizes is observed. This allows a better study of group dynamics (eg. formation, merging, splitting) and the reasons behind it. This high number of people, while compare to 18 in SALSA, allows to better regularize the learning of group behaviors.

In addition, in *MatchNMingle* the participants were never assigned a specific role and all social interactions are natural and spontaneous, whereas in SALSA they do have a role for the poster session part of the dataset. Similar to SALSA, for *MatchNMingle* a personality trait survey was also collected. However, instead of the Big-5 survey, we collected the HEXACO inventory (100 items) as it has been shown to better capture the multi-dimensional nature of personality (see [8] for review and Section 2.3.3 for more).

But, the main difference between SALSA and MatchNMingle (in the context of free standing conversations) is the depth and detail of the **manual annotations** collected for *MatchNMingle*, which are based on social constructs (see Section 2.5 for more details). Thus, manual annotations were incorporated for social actions (or behavioral cues) such as speaking, hand gestures, and hair touching (cue associated with flirting and important in the context of a speed date [116]), making it the first dataset to incorporate such annotations in this context. Thus, our intention is to provide the research community with labels that are truly associated with social behavior, in addition to the usual spatial labels such as position and orientation. These types of labels will help answer open questions in the domain of social interactions by examining the data at a higher level of abstraction (eg. social cues instead of spatial-temporal positions or actions).

Also, for SALSA the people's position and head/body orientation were manually annotated and used to automatically predict F-Formations using the method proposed by Cristani et al. [43]. On the contrary, for *MatchNMingle* all positions and F-Formations are manually annotated directly. This provides additional resources for training people detectors from a top down perspective, as currently all models are trained from elevated side views in less crowded scenarios.

Notice that all the above holds while comparing only the mingle segment of *MatchNMingle* to SALSA. But *MatchNMingle* also incorporates a speed date segment, which is compared to other efforts in the next subsection.

2.2.2. SPEED DATES

Speed-dating events have been used in the social sciences for the study of romantic attraction, as they allow for a balance of experimental control and ecological validity. During these events, participants meet potential romantic partners for 3-4 minutes, after which they each indicate (yes/no) if they would like to meet their partner again after the event.

Data collected during such events is rich, and allows for the application of sophisticated analytic techniques (e.g. Kenny's Social Relations Model [90]). Each participant meets with a number of interaction partners, which allows for data to be collected on a large number of interactions using a relatively small sample. In addition, each participant is evaluating while simultaneously being evaluated, yielding data from both perspectives.

Social science researchers have collected various forms of unimodal and multimodal data to test various hypotheses in speed-dating studies, including photos ([23, 42]), video ([127, 162]) and audio ([81, 109]). These studies employed ratings of media given by participants or trained raters, with the exceptions of [81], who transcribed interactions and subjected the transcripts to text analysis software, and [109] who transcribed interactions and extracted features from the audio, both for a qualitative analysis.

So, despite the number of speed-dating studies, few have leveraged the potential of ubiquitous technologies to examine and predict the outcomes of these interactions, or to

Table 2.4: **Speed dating events.** Numerical and sensor comparison with other datasets.

Dataset	Numb. dates	Time Date (min)	Sensors			
			Video	Audio	Wireless	Accel.
Madan et al. [101]	57	5		X		
SpeedDate Corpus [85]	991	4		X		
Veenstra and Hung [167]	64	5	X			
MatchNMingle*	674	3	X	X	X	X

*Dataset is (or will be) publicly available.

assess how speed dates unfold. Table 2.4 compares all these efforts to *MatchNMingle*.

First, Madan et al. [101] and Pentland [125] presented one of the first data collections specifically used for the automated analysis of speed dates. They collected audio data of 57 5-minute speed-dates, and correlated the 4 measures of vocal social signaling proposed by Pentland [125] to levels of attraction and friendship. Jurafsky et al. [85] created the *SpeedDate* Corpus, which consists of spoken audio of 991 4-minute speed dates, collected with a shoulder-worn audio recorder. In order to collect this corpus, they held 3 speed-date sessions (such as ours). This corpus has been used by Jurafsky et al. [85] to detect whether the speaker is awkward, friendly, or flirtatious, and by Ranganath et al. [134, 135] to investigate the difference between intention and perception during speed dates.

To the best of our knowledge, Veenstra and Hung [167] is the only work for which video features are extracted and used to predict the outcome of speed dates. They collected video for 64 5-minute speed dates with 16 participants (8 females), and predicted physical attraction (from self-reported surveys) and the intent of exchange contact information using movement-based features from video.

In *MatchNMingle*, we considerably increased the number of modalities which recorded the event, aimed for a larger number participants and added surveys to assess their predisposition regarding social conduct or personality (see Section 2.3.3).

2.3. DATA COLLECTION FRAMEWORK

The *MatchNMingle* dataset was collected during 3 events over the course of three different weeks, each consisting of a speed-dating session, followed by a mingle which resembled a cocktail party. In this section, we describe the framework of our data collection. [2]

2.3.1. VENUE

A local cafe/bar/restaurant was chosen as an ecologically valid venue for the events. In addition, it was chosen because 1) it was located in the center of the dormitory campus, 2) the building had a large, separate room outside of the dining area that could be used for taking photos and preparing the registration (see Section 2.3.5), and 3) because staff allowed researchers to reconfigure the dining area to suit the needs of the study.

For the speed-dating portion of the study, tables were arranged in several rows with opposite sex interaction partners facing each other. For the mingling portion of the study, the tables were re-arranged to create a rectangular area for participants to enjoy drinks while freely socializing.

[2]The Ethics Committee of the Faculty of Psychology and Pedagogy of the VU University Amsterdam (Vaste Commissie Wetenschap en Ethiek van de Faculteit der Psychologie en Pedagogiek: VCWE) approved the study, and it was registered under VCWE-2015-037.

2.3.2. PARTICIPANT RECRUITMENT

Participants were recruited from a university campus. The goal was to recruit approximately 30 participants per event, 15 of each sex. Researchers posted fliers around campus and dormitory buildings, made in-class announcements, promoted the events on social media, and recruited participants from their personal social networks. To be a possible candidate, participants had to be 1) single, 2) heterosexual and 3) between 18 and 30 years old.

As compensation, apart from the possible outcome of the speed date event itself, all participants were given in return €10 and 2 free drinks during the event. From prior data collection experience, we have found that this type of compensation increases the interest of potential participants.

Participant registration was conducted via an online survey. The survey screened for relationship status (single / not single), sexual orientation (heterosexual, homosexual, bisexual, other), and age (18-30). Here, they also filled questionnaires to test individual differences (see Section 2.3.3).

In addition, the initial survey screened for medicinal and recreational drug use, recent emotional events, and hair length for the purposes of hormone sampling. Although collected, due to the sensitivity of the information, all the latter can not be made publicly available. However, it is worth stating that these surveys and hormone baselines were also collected for all participants during the events.

2.3.3. OFFLINE DATA COLLECTION

QUESTIONNAIRES

In order to test individual differences among participants, the initial online registration survey included 1) the HEXACO personality inventory [5], 2) the brief Self Control Scale (SCS) [158] and 3) the revised Sociosexual Orientation Inventory (SOI) [123]. [3] Only those who filled these questionnaires were allowed to participate in the events.

Collecting self-assessments of participants' personality facets allows for the comparison of various traits expected to affect social outcomes. For example, studies have shown a corelation between people's attraction and personality traits [12, 161]. Within a mating and/or interaction context, inclusion these self-assessments could allow researchers to see how these predict or affect behavior during the mingle, and/or speed-dating outcomes.

The HEXACO personality inventory measures personality along 6 dimensions: Honesty-humility, Emotionality, eXtraversion, Agreeableness, Conscientiousness, and Openness to experience. We chose the HEXACO rather than the more frequently used 5 factor models such as the Big-5 or the Five Factor Model (FFM). While the Big-5 and HEXACO are both derived from the same lexical studies (see [8] for review), the six-dimensional HEXACO model has been shown to more optimally capture the data in cross-cultural replications [7], and to outperform the FFM in both self-ratings (i.e. when participants complete the inventories about themselves) and in observer ratings (i.e. when participants complete the scale about another individual [6].

Briefly, the HEXACO and five factor models are related in a number of ways: 1) extraversion and conscientiousness are the most similar among all the dimensions to their five factor counterparts, 2) agreeableness and emotionality in the HEXACO are rotated

[3] Due to privacy issues and the sensitivity of the information, only the HEXACO inventory is publicly available. For the SOI and SCS, please contact the authors for possible collaborations.

versions of their five factor counterparts, with traits related to anger loading on HEXACO Agreeableness instead of Big-5 Neuroticism, and traits relating to sentimentality loading on HEXACO Emotionality instead of Big-5 Agreeableness, and 3) terms such as honest, sincere, fair etc. that load on Big-5 Agreeableness are the separate dimension of HEXACO Honesty-Humility instead (see [8] for review).

In addition, each scale in HEXACO can be further separated into facet-level scales (e.g. Social Self-Esteem, Social Boldness, Sociability and Liveliness are part of the eXtraversion domain). This survey consists of 100 questions [4] which are answered on a scale from 1 (strongly disagree) to 5 (strongly agree).

The brief Self Control Scale was designed to assess dispositional self-control and consists of 10 items such as *I am good at resisting temptation*. Each of these items can be rated in a scale of 1 (*not at all like me*) to 5 (*very much like me*).

Finally, the revised Sociosexual Orientation Inventory assesses attitudes, behaviors, and desire for unrestricted sexual relationships, with 9 items such as *Sex without love is ok*. Sociosexuality has been characterized as an individual's attitude, desire, and behavior regarding sexual relationships: specifically, unrestricted individuals have been shown to have more short-term sexual encounters, consider uncommitted sexual relationships positively, and engage in more flirtatious behavior [12, 123]. Similarly to the other surveys, this could be answered in a scale of 1 to 5.

FRONTAL PHOTOS

Before each event (or during the intermission) three frontal photographs were taken of all participants: 1) neutral facial expression, 2) smiling facial expression and 3) full body (see Figure 2.1). We collected also these as prior research has shown that facial attributes, such as facial height-width ratio [163] or closeness of a person's face to the mean face [71] correlate with perceived attractiveness.

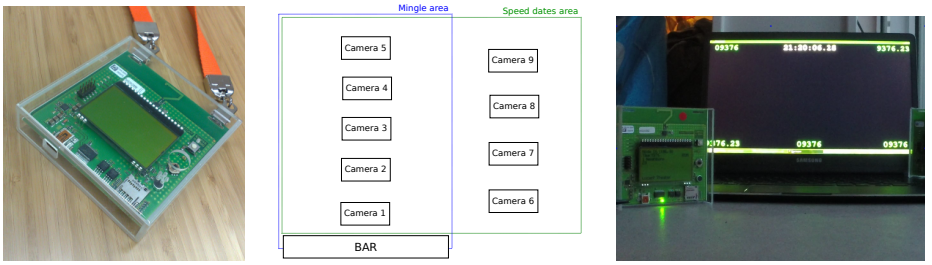
HORMONE BASELINES

Researchers collected a total of 3 hair samples from each participant for the purposes of gathering hormonal baselines.Strands of hair on the lower back of the head (posterior vertex) were cordoned off with a string and cut as close to the scalp as possible. They were cut into ~3mm diameter, with 3cm lengths from the point closest to the scalp, as in prior research [121]. Results obtained reflected approximate 3 month averages for each of the measured hormones. As hormone baselines have been shown to affect behavior in various contexts, these baselines were collected to test variance in popularity and selectivity over the course of the speed dates. Due to privacy issues and the sensitivity of the information, these baselines can not be made public. Please feel free to contact the authors for possible collaborations.

2.3.4. ONLINE DATA COLLECTION

We sensed the entire area of the event through: 1) wearable devices recording triaxial acceleration and proximity, and 2) video cameras arranged in surveillance style, facing downwards from the ceiling. All the data collected by these sensors is synchronized to a global time. In addition, after each speed date all participants filled a match booklet with their impressions.

[4] http://hexaco.org/



(a) (b) (c)

Figure 2.3: (a) Custom-made wearable device. (b) Distribution of top view cameras on the venue area. (c) Cameras and devices' synchronization.

WEARABLE DEVICES

As they arrived at the venue, participants were given a device that was to be hung around the neck, emulating a badge similar to those used in conferences, and to be worn for the duration of the event. These badges (see Figure 2.3(a)) are custom-made wearable devices designed specifically for applications on social interaction and group dynamics analysis [53]. They record triaxial acceleration at 20Hz with a maximum range of ±2G. Also, these devices can detect each other using wireless radio communication. Thus, each device broadcasts its unique device identifier (ID) every second to all neighbor devices within a distance of about 2-3 meters. The reception of this ID by the devices nearby is considered a binary proximity detection. This way, each device can create and locally store a binary proximity graph of its neighbors every second. This communication also allows the devices to synchronize to a global time-stamp. Refer to [53] more for technical details.

VIDEO CAMERAS

Top-view, video of the event area was captured using 9 different GoPro Hero 3+ cameras, which were configured to a resolution of 1920 x 1080 (16:9), a sample rate of 30 fps and a ultra wide field of view. Also, each camera recorded audio (due to privacy issues, audio for each person using microphones could not be used). Figure 2.3(b) shows the camera's distribution on the venue. The GoPro Remote Control was use to ensure synchronize the cameras. An additional camera recorded a screen showing the global timestamp from the wearable devices, as seen in Figure 2.3(c). Thus, we can synchronize cameras and wearable devices. The main reason for using top views is to reduce at a maximum the interpersonal occlusions, which is higher in side views for this type of crowded scenes.

For the first portion of the event, the 9 cameras are arranged so each of the 15 tables for the speed dates are captured by at least one of the cameras. For the second portion of the event, the tables are set aside to create a rectangular space for the mingle. For this area, 5 cameras recorded the mingle with some overlap between the cameras. Figure 2.4(a) shows snapshots from 4 of the cameras recording during the speed dates. These snapshots correspond to cameras 6 to 9 on Figure 2.3(b). In Figure 2.4(b) are shown snapshots from 5 cameras (1 to 5 from Figure 2.3(b)) during the mingle session. Notice how our event has different illuminations, shadows, occlusions, and a crowded environment (during the mingle), making the data challenging to analyze using methods solely-based on computer vision.

(a)

(b)

Figure 2.4: Snapshots for (a) the speed date session (cameras 6 to 9) and (b) mingle session (cameras 1 to 5). The speed date snapshots correspond to the first day event, while the mingle ones correspond to the last day.

SPEED DATE RESPONSES

During each date, participants completed a questionnaire in the form of a booklet, designed to resemble materials from a commercial speed-dating event. The booklet format was used so that the participant could hold one end upright, preventing their interaction partner from seeing their responses. After each date, participants indicated whether they would like to meet their interaction partner again (yes/no); a "match" occurred when both participants answered "yes" to this question. In addition, participants indicated how much they would like to see their interaction partner again (low = 1, high = 7), and how they would rate them as a short term sexual partner (low = 1, high = 7), and a long term romantic partner (low = 1, high = 7). Participants received an email following the event, with photos of the faces of their *matches*. They then indicated which of their matches they would like their contact details sent to.

2.3.5. DETAILED COLLECTION PROCEDURE

As participants entered the venue, researchers checked their registration and assigned them an anonymized participant number. They were then provided a wearable device showing their participant number (to facilitate the process of completing the match booklet questionnaire). Women and men were separated during the entire preparation process to ensure

that their first encounter occurred during the speed dates. During the preparation process, researchers collected photos of one group (either the men or the women), while collecting hair samples of the other group (for hormone baselines). During the break after the 7th speed date, the groups were reversed so that hair samples and photos could be taken. For example, if during registration photos were taken of the women and hair samples were taken from the men, during the break photos were taken of men and hair samples were taken from women. After the speed dates, there was a second break where any remaining hair samples and photos were collected. Participants of opposite sex remained separated during all breaks.

The first part of the event was the **speed dates**. Each participant had an approximately 3-minute date with a participant of the opposite sex, followed by approximately 1 minute to fill their match booklet. Once completed, all participants of the same sex were asked to move to the next seat. For the rotation process, we alternated the sex that was asked to move so as to prevent confusion and regulate first impressions. This rotation was repeated until this portion of the event was complete. Approximately half-way through the speed-dating session (after the 7th date), we introduced a pause to reduce the effect of fatigue on participants' impressions.

For the second part of the event the participants were asked to **mingle freely** within the area limited for this purpose. This area was limited to ensure high spatial density of people during the mingle. Participants were not instructed in any way, and could move through, leave and re-enter the mingle area at will. During this part of the event, soft and/or alcoholic drinks were provided (2 for free with the option of purchasing more) in the bar or by request to one of our team members. Snacks were also available for purchase. Sadly, the bar's staff were not members of our experiment, so a detailed number of alcoholic drinks ingested by participant could not be collected.

2.4. THE *MatchNMingle* DATASET

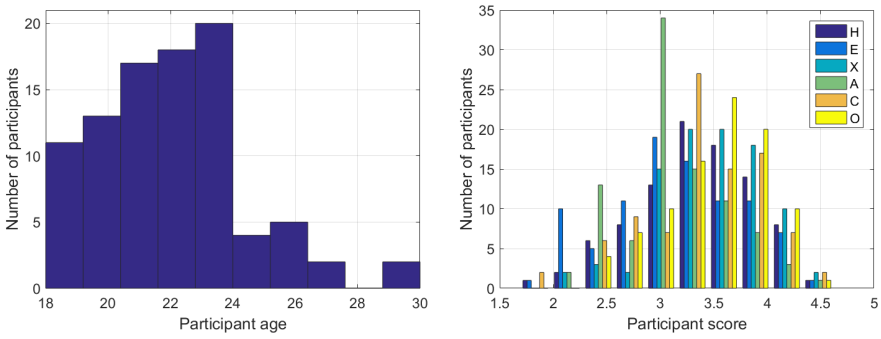
We had a total of 92 single, heterosexual participants (46 women: 19-27 yrs., M = 21.6, SD = 1.9; 46 men: 18-30 yrs., M = 22.6, SD = 2.6) divided on 3 events. From these, 16 men and women attended the first event, and 15 men and women attended the second and third events.

Due to hardware malfunction, some of the devices failed to record the event partial or totally. In total, we collected sufficient information for 72 wearable devices during the speed dates and 70 during the mingle session. These correspond to 28 devices (26 for the mingle) from Day 1, 22 from Day 2 and 22 devices from Day 3. The number of failing devices that were assigned to female participants were 4 for Day 1, 3 for Day 2 and 3 for Day 3.

2.4.1. PARTICIPANT STATISTICS

As introduced in Section 2.3.2, our participants were mostly students that were not acquainted before the event. Figure 2.5(a) shows the proportion of participants of different ages (mean=22.09, std=2.34). Similarly, Figure 2.5(b) shows the proportion of participants with a similar score on each of the personality traits on the HEXACO inventory.

Here we report the Cronbach's α coefficient, widely used to test the internal reliability



(a) (b)

Figure 2.5: General statistics of our participants. (a) Proportion of participants for different ages. (b) Proportion of participants with similar personality trait scores on the HEXACO inventory.

of scales [45]. By convention, 0.65 is considered sufficient, and 0.8 is considered good in terms of reliability. The higher the internal consistency, the more interpretable the scores [64]. For the HEXACO inventory, the coefficients were 0.81 for Honesty, 0.87 for Emotionality, 0.84 for Extraversion, 0.82 for Agreeableness, 0.83 for Conscientiousness and 0.77 for Openness to experience. The coefficient for the Self Control Scale was 0.89, the global SOI score was 0.87, and the subscales for attitude, behavior, and desire were 0.87, 0.82, 0.89 respectively. Examination of internal reliability revealed that the correlation between behavior and attitude $r = 0.23$, $p < 0.05$, was weaker than with desire $r = 0.40$, $p < 0.01$, and that desire and attitude showed a stronger correlation $r = 0.40$, $p < 0.01$.

Selectivity and popularity variables were computed using the "yes" and "no" responses to the question "Would you like to see this person again?". For the selectivity variable, the responses given by each participant were summed, and for the popularity variable, the number of responses received by each participant were summed. To account for the differences in the number of dates attended by each participant, we divided the number of responses by the number of dates that participants attended. Independent sampled t-tests were then conducted with 1000 bias accelerated bootstrapped samples to test for sex differences.

Men ($M = .60$, $SD = .25$) said "yes" slightly more often than women ($M = 0.49$, $SD = .23$), $t(90) = 2.295$, $p < 0.05$, $d = 0.48$. As such, men were slightly less selective than women. Men ($M = 0.49$, $SD = 0.20$) also received slightly fewer "yes" responses than women ($M = 0.60$, $SD = 0.22$), $t(90) = -2.607$, $p = 0.011$, $d = -0.54$. As such, women were slightly more popular than men.

Sex differences might also be expected in reported sociosexuality among participants. An independent samples t-test was conducted with 1000 bias accelerated bootstrapped samples. It showed that men (M= 46.43, SD =11) reported a more unrestricted SOI than women (M= 33.44, SD = 13.56) $t(90) = 5.051$, $p < 0.001$, $d = 1.05$ indicating a large effect. We then tested the individual subscales using the same procedure, expecting that men would report greater scores on all subcales: this was true for desire $t(90) = 6.39$, $p < 0.001$, $d = 1.35$ (large effect) and attitude $t(90) = 3.929$, $p < 0.0001$, $d = 0.83$ (moderate) effect, but not behavior $t(90) = 0.87$, $p = 0.413$. No sex differences were found in scores on the SCS $t(90) = 0.537$, $p = 0.593$.

Table 2.5: Number of answers to the question *Do you want to see this person again?* during the speed dates. Total of date interactions = 674.

Day	Match	Women yes	Men yes	Both no	Total
1	70 (31.25%)	45 (20.09%)	57 (25.45%)	52 (23.21%)	224
2	61 (27.11%)	42 (18.67%)	65 (28.89%)	57 (25.33%)	225
3	79 (35.11%)	39 (17.33%)	74 (32.89%)	33 (14.67%)	225

Table 2.6: Summary of dates and date interaction for participants carrying a functional wearable device. Original data: 674 date interactions (Day 1= 224, Day 2=225, Day 3=225)

Day	Females		Males		Date Interactions*
	Num. Partic.	Dates	Num. Partic.	Dates	
1	13	182 (81.3%)	15	210 (94%)	195 (87%)
2	12	180 (80%)	10	150 (67%)	120 (53%)
3	12	180 (80%)	10	150 (67%)	120 (53%)
All	37	542	35	510	435 (65%)

*Date interactions where the 2 have a functional device.

2.4.2. SPEED DATES STATISTICS

For clarity, when addressing the speed dates we treat *date* as the information from a **single** person during a 3 minute date, and *date interaction* as the interaction between 2 participants during a 3 minute speed date. Thus, during a date interaction we will have 2 dates, one for each participant.

During the speed dates, each participant had a 3 minute date interaction with all other participants of the opposite sex. Thus, for Day 1 each participant had 14 dates, and 15 for each participant in days 2 and 3.

In total, we collected 674 date interactions (Day 1= 224, Day 2=225, Day 3=225) which gives us 1348 match booklet sets of answers, one for each date (6 answers per set/date). From these, half correspond to female responses.

SPEED DATES MATCHES

Table 2.5 summarizes the answers for the question *Do you want to see this person again?*. Notice that a *match* only happens when the two participants answer positively to this question. Figure 2.6 shows the score distribution for the all the responses in the match booklet. From these we can see that most participants' impressions were more inclined towards a friendship relationship with their dates.

WEARABLE ACCELERATION

Due to hardware malfunctioning, wearable acceleration was not recorded for some of the date interactions, either for both or one participant. Table 2.6 summarizes the number of dates recorded using wearable acceleration per gender for each day, and the number of date interactions of each day for which both participants were using a functional device. This table also reports the percentages of the dates and dates interactions that were successfully recorded.

Overall, we found a distinctive difference between participants' movements (intrapersonal difference), and between the movement of the same participant for different dates (interpersonal difference). The first might relate to personal characteristics (eg. personality) whereas the second might be related to the interaction between 2 specific persons (eg.

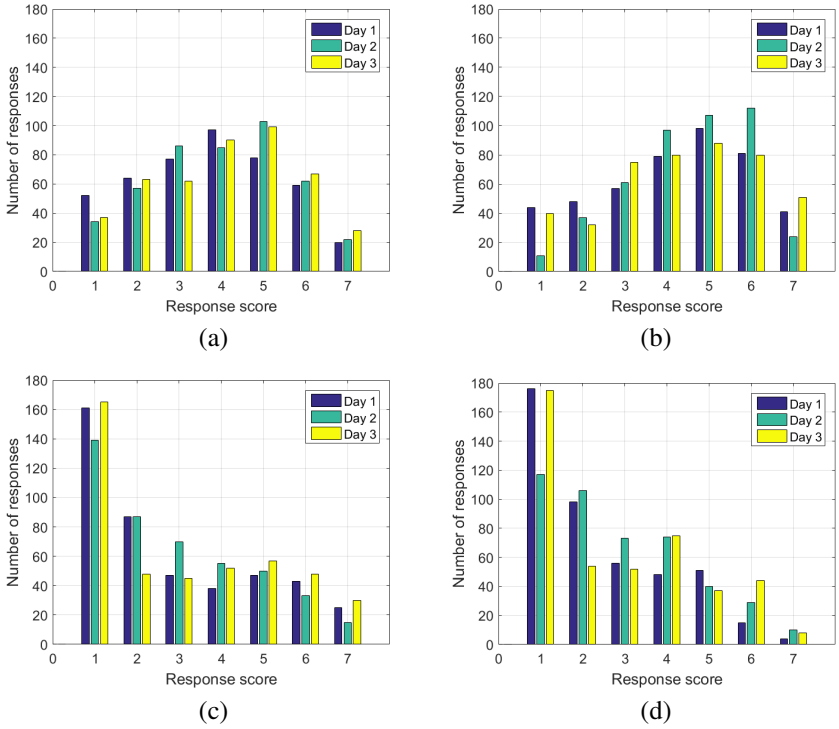


Figure 2.6: Distribution of score responses for all dates divided by day. All scores are within a range of 0-low and 7-high. (a) How much would you like to see this person again? (b) How would you rate this person as a potential friend? (c) How would you rate this person as a a short term sexual partner? (d) How would you rate this person as a long term romantic partner?

attraction). *MatchNMingle* gives us the possibility to further study these open questions, which we first approach in Section 5.5.

2.4.3. MINGLE STATISTICS

In total, 70 participants had working devices during the mingle segment (Day 1=26, Day 2=22, Day 3=22) and over 45 minutes of free mingle were recorded for each day (Day 1=56, Day 2=50, Day 3=45).

WEARABLE ACCELERATION

For each participant with a functional device, we calculate its mean movement during the mingle session and normalize it across all participants. Thus, participants had a mean normalized movement of 0.43, with a deviation of 0.18 between participants. In addition, when separating female and male participants we found that males have a mean normalized movement value of 0.77 while females had a mean value of 0.76, with a deviation between participants of 0.17 and 0.22, respectively. Note that while there are significant differences between participants' movements, this effect is not significant given their gender.

PROXIMITY INFORMATION

For each day of the event, the proximity information was calculated for all participants. Notice that there is a maximum of $N-1$ interactions per day (N = number of participants per day). Also, these are the proximity estimations from the devices, and not the annotated ground truth for the F-Formation.

The mean number of people interacted with per participant was 26.5 ± 3.8. The person who interacted with the fewest number of people interacted with 15 persons, while the participant with the most interactions had 33 different neighbors throughout the event. The mean longest interaction over participants was 23 ± 5 minutes.

We also evaluated the accuracy of the proximity detections by comparing it to the ground truth annotations for the F-Formations (see Section 2.5) in a pairwise fashion every second. The metrics were calculated on the data from all days. We obtained Recall and Precision scores of 90.1% and 33.0%, respectively. This is mainly due to the use of radio communication to detect proximity. We further discuss this issue, which is due to the omnidirectional proximity detection, in Section 2.7.

2.5. MANUAL ANNOTATIONS FOR *MatchNMingle*

In this section, we describe the process of manually annotating *MatchNMingle*. We first describe the social cues present in social contexts and the contribution of *MatchNMingle* to them. Then, according to the analysis of social cues, we present the social actions (eg. actions performed specifically as part of a social cue) selected for annotation, and the motivation for them. We then present a detailed description of the tool, and the process used for collecting the manual annotations. Also, we analyze and compare the performance of annotators hired through an online crowd-sourcing platform to the performance of trained annotators. We do this comparison for both a simple and a complex type of task (or HIT), such as people's position on video and social actions, respectively. Finally, we present the statistics of the annotations collected.

Our analysis highlights the care required for generating rich annotations of social behaviour at short time scales. Unlike many publicly available datasets that rely on crowd sourcing to label the data, our results show that, for more complex HITs, the label quality was insufficient and required a more intense training phase with annotators that is not possible within the current setup of Mechanical Turk.

2.5.1. SOCIAL CUE CATEGORIES

Efforts in activity recognition tend to focus on the detection of daily activities such as walking, or biking, among others ([16, 41, 137]). In contrast, when addressing social interaction scenarios, one will encounter human behavior that depends on the social context. These are more complex to analyze than basic daily activities, due to the large variations between the behavior of different individuals for the same class (person independence) as illustrated by [68].

Also, unlike daily activities where people tend to perform one action at a time, social actions tend to overlap. For example, during a conversation a speaker accompanies their vocalized speech with head and hand gestures . Hence, these actions are not mutually exclusive but instead are complementary and/or correlated. These two aspects should also

be considered during the annotation.

The most important social cues for the judgment of social constructs (attraction, personality, etc.) have already been categorized in social psychology and used extensively by the computing community. These categories are 1) physical appearance, 2) gestures and posture, 3) face and eye behavior, 4) vocal behavior, and 5) space and environment (see Vinciarelli et al. [169] for a more exhaustive explanation). How *MatchNMingle* contributes to these categories is detailed below. Note that some will require explicit manual annotation, others are implicitly included in the data collected (i.e. further processing of the raw data is required), and others cannot be addressed by *MatchNMingle* due to the nature of the event (eg. facial behavior analysis).

PHYSICAL APPEARANCE

This category includes characteristics such as height, attractiveness, and body shape. For *MatchNMingle* , the height is obtained explicitly via self-report and implicitly from frontal photos for the body shape and attractiveness.

GESTURES AND POSTURE

All hand and head gestures belong to this category (including visible laughter), along with the posture (e.g. head and body orientation) and shift of posture of the body (eg. shifting weight from one leg to the other). One of the **main contributions** of *MatchNMingle* is the expansion of the amount and type of gesture behavioral cue, compared to works on activity recognition. Specifically, we annotate all hand and head gestures performed by our participants. As the definition of 'gesture' is widely debated [73], we treated gesture (hand or head) as any intentional or unintentional movement of the hand or head. These annotations can be later sub-categorized according to each researcher's needs.

FACE AND EYE BEHAVIOR

All facial behavior, including eye gaze, are included in this category. Since we use overhead cameras, the analysis of facial or eye behavior is not reliable to annotate for. However since head pose and body pose can be observed, they could be used as a proxy for gaze direction ([11, 37, 157]). We also have static images of the participants' faces in the frontal photos (neutral and smile), which could be used for the analysis of physical facial attributes.

VOCAL BEHAVIOR

This accounts for all non-linguistic verbal behaviors (eg. prosody, turn taking, silence). As the audio recorded in our dataset is not person specific but recorded from the cameras, cues from this category cannot be extracted from audio directly. However, by annotating speaking from video one can imply cues such as turn-taking.

SPACE AND ENVIRONMENT

This category desscribes distances between interacting people and where they place themselves relative to each other in the environment they are in. In *MatchNMingle*, we recorded proximity from the wearable devices to account for the space cue. In addition, by annotating the people's position and using camera calibration techniques, one could infer the distance between participants from video. In the case of the environment, we intentionally created an event in a real venue (for ecological validity) with 2 different contexts (free-standing groups and sitting dyads).

2.5.2. ANNOTATIONS FOR POSITIONS, F-FORMATIONS AND SOCIAL ACTIONS

After the aforementioned analysis per social cue, we decided to annotate the following 8 social actions: **1) Walking, 2) Stepping, 3) Drinking, 4) Speaking, 5) Hand Gestures, 6) Head Gesture, 7) Laugh,** and **8) Hair Touching**. These actions are strongly linked with the social context of our events. In addition, we annotate the **spatial position** for all participants during the entire segment selected (see next subsection) and 10 minutes for the F-Formations[5]. Note that some of these are also part of daily activities (eg. walking). In addition, we annotate separately from hand gestures for hair touching due to the romantic attraction context, as this particular gesture is distinctive during flirting situations [116]. Also, the action for *walking* represents a long spatial displacement, whereas *stepping* is used for changes of posture in a limited space.

Compared to *MatchNMingle*, SALSA [1] is the only other dataset which considers the social context for automatic analysis of conversational groups, and annotates accordingly. Nonetheless, the number of social cues considered from the above categories in SALSA is limited to spatial constructs (eg. head and body orientation (posture)), audio statistics (min, max, average, variance, standard deviation from audio energy) and turn taking (vocal behavior), which are significantly less social cues than those in *MatchNMingle*.

2.5.3. THE ANNOTATION PROCESS

We randomly selected a 30-minute segment of the mingle session for each day. Segments were only restricted to be 5 minutes after the beginning and 5 minutes before the end of the mingle, to eliminate the possible effects of acclimatization, and to maximize the density of participants and the number of social actions that could occur in the whole scene.

Each day, the 3 cameras with the highest concentration of subjects were selected for annotation[6]. These were enough to ensure that all participants had annotations for at least 75% of the time, with the exception of 5 participants that were outside the mingle area (eg. going to the bathroom or leaving the event). This was possible as all 5 cameras during the mingle had overlapping coverage (see Figure 2.4(b)).

The Vatic tool proposed by Vondrick et al. [171] was used for manual annotations of the positions and social actions. This tool was designed for crowd-sourcing annotations in Amazon's Mechanical Turk (MTurk) and has an interface similar to a video-player (see Figure 2.7). With Vatic, the annotator can create a new object (which type depends on the final application), follow it through time and give it attributes from a checklist. Vatic also interpolates between frames for both position and attributes, so the annotation of every single frame is not necessary. Although mostly used for tracking tasks, a simple modification of the tool allowed us to also include the social action annotations as attributes. Also, the F-Formations were annotated directly from a video showing the participant's number.

Using this tool, our annotators had to 1) manually track all the people in the video and 2) annotate the 8 selected social actions for each of them. To do so, they had to create a bounding box for each person in the video, either at the first frame or the first time they appeared in the video. Each bounding box included check box es for each of the 8 social

[5] A detailed description of times is provided with the dataset.

[6] A previous version of the dataset using the annotations of only 2 cameras per day has been used in past works. Both versions are available for reproducibility purposes. Refer to our website for details.

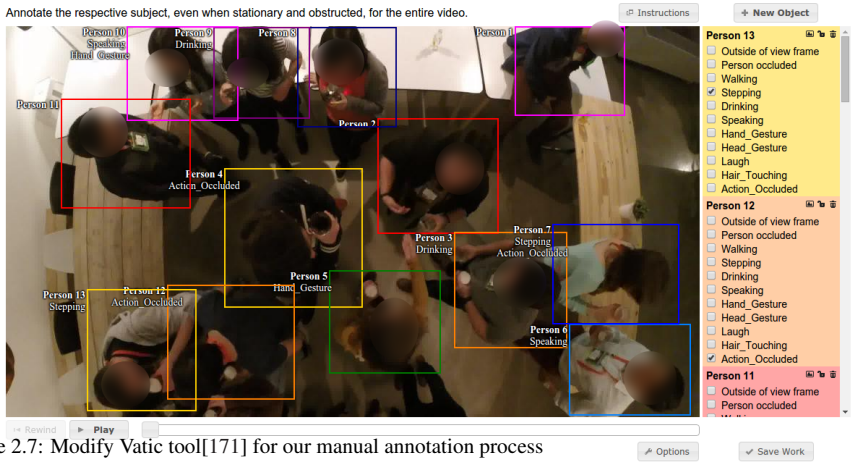


Figure 2.7: Modify Vatic tool[171] for our manual annotation process

actions, as can be seen in on the right in Figure 2.7. The annotators were instructed to check the box for an action once it had begun, and to uncheck it when the action was completed. More than one action could be selected in parallel. In addition, checkboxes for *person occluded*, *action occluded* and *outside of view frame* were also included. The first allows to specify occlusions between people, while the second one allows to explicitly give a confidence on the annotations (eg. person giving the back to the camera). The latter is used when the person leaves the field of view.

Each annotation task was divided into smaller tasks or HITS of a length of 2 minutes. For 30 minutes of recordings, using 3 cameras per day we had a total of 135 HITS. Finally, to select only good workers, we applied a non-paid practice run for all workers only once. This consisted of a comparison against a *gold standard*, which was annotated by an expert. Only those workers that passed the practice run were allowed to do paid HITS. The goal of such gold a standard was to guarantee that the workers could apply the instructions provided.

2.5.4. COMPARING PERFORMANCE OF CROWD-SOURCING WITH ON-SITE ANNOTATORS

The annotation process was initially intended to be conducting solely using crowd-sourcing, specifically Amazon's Mechanical Turk (MTurk), as it provides access to low-cost workers and task completion in a relatively short time. Nonetheless, we still needed to ensure the quality of the annotations as they are part of the publicly available dataset.

To evaluate the quality and feasibility of the process we focus on *objective* and *subjective* measurements. Specifically, we evaluate the Fleiss'-Kappa coefficient [7], and the relation of time/cost, respectively. Experience has shown us that a pilot test is a good practice for estimating these.

Thus, we first annotated only 12 participants in data from one camera for a 2-minute interval (or 1 HIT) as a pilot test. This interval was annotated by two sets of annotators: 1) workers hired through Amazon's Mechanical Turk, which are called MTurk workers; and 2) by personally hired annotators, called from now on *on-site* annotators, which were trained

[7] Widely used to assess multiple inter-annotator agreement.

by an expert via a video. The Vatic tool was not altered between the 2 groups of annotators and the guide provided to them stayed the same, in written (GIFs were also added for clarity) and video form, respectively. The only difference between the two training process is that for MTurk workers it was the responsibility of each worker to read the guide, whereas the *on-site* annotators received the same guide in video form via email (not a face-to-face meeting) and were not allowed to proceed until they had watch the entire video.

Also, we separated the annotation into two phases: 1) people's position, and 2) social actions. In the first phase, a single individual annotated the positions of all visible people. In the second stage, the position's were provided and all social actions were annotated. Note that the level of complexity for these two phases is different; tracking a person's position in video is rather simple compared to annotating its social actions.

In summary, the pilot test for the social action annotations using MTurk workers showed considerable inter-annotator disagreement and an overhead in time per completed task, which was not the case for the people's positions. The details of this comparative experiment, separated by type of task, are presented below.

SIMPLE HIT (PEOPLE'S POSITION)

For each set of workers (MTurk and on-site) we had 3 different annotators for this type of task. They were asked to follow all 12 participants in the same 2-minute video interval (or HIT) using bounding boxes with the Vatic tool. Both groups did so by accessing the tool via a web address.

In both cases, we calculated the mean across annotators of the overlapping ratio of all bounding boxes annotated for the same participant during each HIT. For a given time t, this overlapping ratio corresponds to the intersection of two bounding boxes over the area of the bounding box with the minimum area of the sets at that time, or (for 2 annotators):

$$r(t) = \frac{Area(BB_1(t)) \cap Area(BB_2(t))}{min(Area(BB_1(t)), Area(BB_2(t)))} \tag{2.1}$$

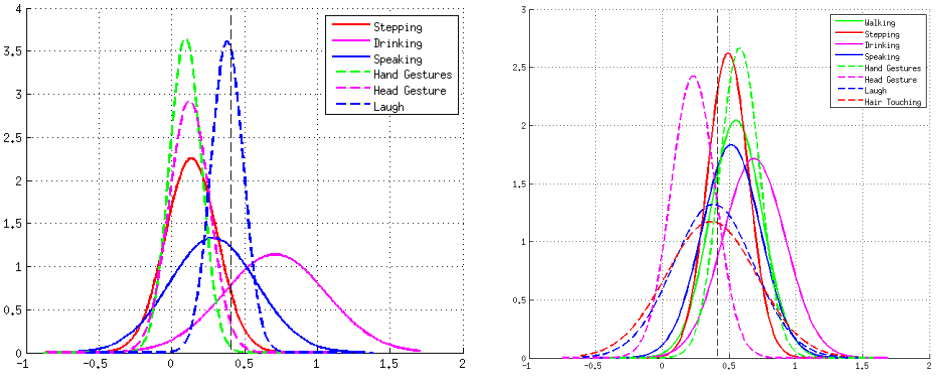
We used the minimum area as denominator in Eq. 2.1, instead of the union (Jaccard index), as some annotators account for the entire body while others annotated only for the head and torso of the people in the video. In both cases, the bounding box correctly followed the person.

Averaging the overlapping ratios in time ($1/T \sum_{t=1}^{T} r(t)$) gives us a single value for comparison. MTurk workers had a mean overlapping ratio of 0.8446, while the on-site workers had a ratio of 0.9289 for the same participants followed (not significant variance in both cases). Notice that both ratios are high enough to be acceptable as inter-annotator agreement for the positions. Hence, both types of annotators are adequate for a simple task as manually tracking a person in video.

Additionally, it took around one day for all 6 HITS to be selected and completed by MTurk workers. One HIT had to be rejected and repeated as the worker submitted an empty HIT. Around the same amount of time was required by the on-site workers to complete the test pilot for positions.

COMPLEX HIT (SOCIAL ACTIONS)

For the pilot tests for this task, the 8 social actions presented in Section 2.5.1 were annotated for 12 participants for an interval of 2 minutes. This was done using 3 different trained



(a) (b)

Figure 2.8: Distribution of inter-agreement Fleiss'-Kappa coefficient (k) over participants annotated for 3 annotators. (a) MTurk workers. (b) On-site trained annotators. Dotted vertical line represents the inter-agreement threshold for moderate agreement. *Walking* and *Hair Touching* are excluded for the MTurk workers as they ignored these classes (no labels).

annotators, each annotating all participants. In the case of the Mturk set of workers, the 12 participants were separated into 12 HITS to reduce the workload for a total of 36 HITS. Although a MTurk worker could do more than one HIT, it was not permitted for the same person to do a HIT for the same participant twice, so all participants were annotated by different workers. Also, only workers that had passed the gold standard practice run were allowed to participate[8]. All on-site annotators passed the this run at the first try.

Figure 2.8(a) shows the distributions of the Fleiss'-Kappa coefficient for all actions over participants (eg. one different k value per participant annotated) in the MTurk pilot test, while Figure 2.8(b) shows the same for the on-site annotators. As can be seen on Figure 2.8(a), there was strong disagreement between MTurk workers for almost all annotations, whereas for the on-site annotators the difference varied more depending on the type of social action. In fact, the classes *Stepping, Hand Gesture* and *Head Gesture* have a mean agreement coefficient below or equal to 0.1 for the MTurk workers and, although present in the segment provided, the classes *Walking* and *Hair Touching* were completely ignored.

Moreover, some of the actions annotated by on-site people also have a low inter-annotator agreement, as can be seen in Figure 2.8(b). More specifically, the social cues of laughing and hair touching lie at the threshold between fair and moderate agreement, while head gestures are considered to have only fair agreement. These differences between actions may be due to the subjectivity of the annotations. Thus, while actions like drinking are rather evident, head gestures have more subtleties that can produce disagreement between annotators.

In addition, when checking the statistics from MTurk for the workers, we noticed that 11 workers attempted the practice run but failed it and quit. 18 HITS had to be repeated as the worker submitted an empty job which was rejected. Finally, the time necessary to complete the set of 36 possible HITS available was 10 days, whereas it only took one day for the on-site annotators. All the above also resulted in an overhead during the annotation process, as the performance of the workers (empty versus completed HITS) had to be assessed and

[8]The practice run is passed if the worker accomplishes 70% overlap with the gold standard for a 1-minute HIT.

additional HITS submitted accordingly, which was not the case for the on-site annotators. Thus, the benefits in costs and time that are generally provided by crowdsourcing tools did not apply to this task.

Main finding: The experiment in this section has shown that not all types of annotations tasks can be done using crowd-sourcing tools. In some instances, the complexity of the task results in low inter-annotator agreement (with differences in the Fleiss'-Kappa coefficients of up to 0.4) and time/cost overhead when using crowd-sourcing. As a result, it requires hiring and closely training the annotators (called *coders* in psychology) to ensure rich annotations. This also demonstrates the importance of coders in social psychology, and how their work and insights should be better reflected in the computing community for these type of efforts.

2.5.5. SOCIAL ACTION STATISTICS

Following the results of the pilot test, we decided to delegate the social action annotations only to trained annotators. Thus, 8 different annotators were hired and given an introductory video where they were instructed on how to annotate social actions by an expert. They all passed the gold standard on the first try. To further ensure inter-annotator agreement, 2 additional 2-minute intervals (one for each remaining day of the event) were annotated by 3 of the annotators divided randomly. The agreement for this 2 additional HITS resemble the results in Figure 2.8(b), which are overall fair agreements scores.

The 30-minute segment of the mingle session for each day was then annotated for these trained annotators. We divided each segment into 2-minute intervals to create the HITS and used the Vatic tool, as described in Section 2.5.3. The only difference is that, in these HITS each annotator was asked to annotate the social actions for ALL participants in the video. This was done for simplicity and to unify the annotation process for the annotators. The assignment of the HITS to the annotators was done randomly.

Two versions of MatchNMingle

There were 2 circles of annotations, a preliminary one where 2 cameras per each day were annotated, and a second annotation circle where an additional camera per day was annotated. From these, we created 2 versions of the dataset, in order to maintain reproducibility: **version 1** for which 2 cameras per day are annotated, and **version 2** for which 3 cameras per day are annotated. For version 2, we can ensure that all participants had at least 75% of annotated data, with 5 exceptions of people that left the mingle area (eg. bathroom or leaving the event early).

For the version 1 of the dataset, 35 of the 92 participants are visible in one of the annotated cameras for the entire 30 minutes, 48 for 90% of this time, 58 for 80%, 68 for 70%, 70 for 60%, 74 for 50% and At a resolution of 20 FPS, the 30 minutes annotated correspond to 36000 samples. We calculated the percentage of this time interval in which each social was annotated as occurring. Thus, the participants were walking $0.98 \pm 1.15\%$ of the time, 8.89 ± 6.76 were Stepping, 3.19 ± 2.61 Drinking, 22.62 ± 14.38 Speaking, and 1.96 ± 2.92 Laughing. In the case of gestures, hand gestures where annotated for a 17.14 ± 11.12 percent of the time while head gesture occurred $7.03 \pm 5.41\%$. Finally, Hair touching was registered 2.02 ± 2.87 of the time.3 are not visible at all (either captured by one of the other cameras or outside the mingle area).

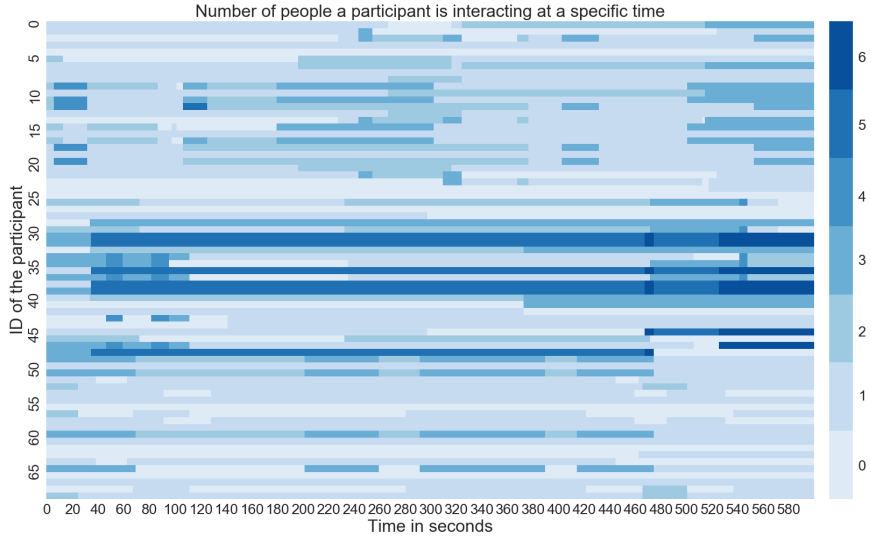


Figure 2.9: Number of people each participant is interacting with at a specific time.

For the version 2 of MatchNMingle, 51 of the 92 participants are visible in one of the annotated cameras for the entire 30 minutes, 69 for 90% of this time, 82 for 80%, 88 for 70%, 88 for 60%, and 89 for 50%. Only one participant was missing for more than 80% of the time, as he left the event early. The participants were walking $1.35 \pm 1.42\%$ of the time, $11.75 \pm 9.64\%$ were Stepping, $4.09 \pm 2.86\%$ they were Drinking, $27.87 \pm 12.75\%$ Speaking, and $2.32 \pm 3.13\%$ Laughing. In the case of gestures, hand gestures were observed for 23.14 ± 11.57 percent of the time while head gesture occurred $12.75 \pm 10.02\%$. Finally, Hair touching was registered $2.58 \pm 3.26\%$ of the time. The percentages were obtained by calculating the mean and standard deviation of the total time all participants performance the action, and reflect that for these segments our participants are mostly engaged in conversations without much walking.

2.5.6. F-FORMATION ANNOTATION

As stated before, the F-Formations were annotated directly from a video showing the participant numbers. The annotations were made every second and for an interval of 10 minutes. Figure 2.9 shows the number of people each participant is interacting with during this interval. Note that there are more or less stable group sizes, as people are possibly engaged in conversations. Nonetheless, there are strong variations between participants, and variations through time for the same person when they leave, form or merge groups. Note that Figure 2.9 does not represent unique groups.

2.6. EXPERIMENTS USING *MatchNMingle*

In this sections, we provide some examples of state-of-the-art methods applied to *MatchNMingle*. These examples complement the research questions addressed in Chapters 3 (multimodal data association), 4 (gesture detection) and 5 (personality estimation), demonstrating

Figure 2.10: Examples of participants with clear faces. Left from right: participants 2, 9, 21, 41 and 52.

Figure 2.11: Snapshots of pose detection using OpenPose [32] for 2 of our cameras.

the range of possibilities of *MatchNMingle* as a multimodal resource.

2.6.1. FACE AND POSE ESTIMATION

We applied 2 state-of-the-art methods commonly used as preprocessing steps for behavior analysis: OpenFace [14] for face detection and OpenPose [32] for pose estimation. For face detection, Table 2.7 summarizes the mean confidence (and deviation) and success from the face detection by OpenFace for 12 of our participants. These participants were selected as the best case scenario, where their faces are visible from the human perspective (see Figure 2.10). Similarly for the pose estimation, Figure 2.11 shows a snapshot of the performances of OpenPose for our dataset.

Table 2.7: Mean and deviation of confidence in face detection provided by the OpenFace tool by our best case scenario subjects (see Figure 2.10)

Participant		2	9	12	14	17	21	40	41	45	52	83	92
confi-	Mean	0.02	0.13	0.00	0.02	0.00	0.01	0.00	0.02	0.00	0.00	0.00	0.35
dence	Std	0.02	0.25	0.00	0.03	0.00	0.07	0.00	0.01	0.01	0.00	0.00	0.15
success	Mean	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Std	0.02	0.28	0.00	0.03	0.00	0.06	0.00	0.01	0.01	0.00	0.00	0.01

We hypothesize that these discouraging performances are due to the inherent difference between the data for which each method was created (or trained with) and our dataset, particularly in terms of camera perspective. This is a straightforward conclusion after seeing Figure 2.11, where mostly the people on the top of the image are detected. The adaptation of these methods to a scenario like ours is discussed in the limitations of this dataset in Section 2.7, and in Chapter 7 as future extension of this thesis.

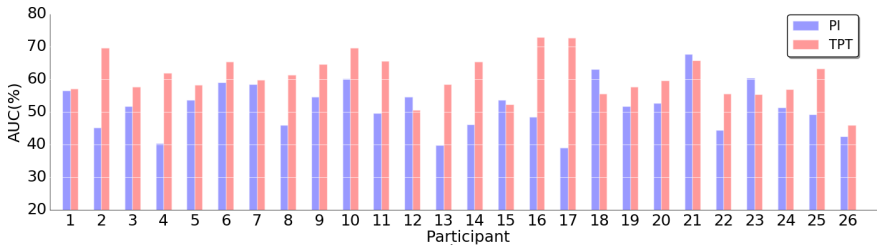


Figure 2.12: Performance in terms of AUC for speaking status detection. PI: Traditional person independent setup. TPT: Transductive Parameter Transfer

2.6.2. SPEAKER DETECTION

Here, we present a simple baseline experiment for speaker detection from wearable acceleration. To do so, we use the acceleration data and speaking annotations provided in *MatchNMingle* to train and compare two different machine learning models. The aim of the experiment is to detect speaking status from acceleration data, which is by definition challenging as 1) the connection between speech and acceleration is not theoretically well defined and 2) individual differences while speaking. To do so, we use the method proposed by Gedik and Hung [68], and compare this to a logistic regressor.

To perform a comparative baseline fairly, we selected only those participants that have 10-minutes of consecutive labels. This way, we do not account for acclimatization issues and we have a balanced amount of labels per participant. By acclimatization, we refer to the warming up period of the mingle where people are entering and introducing or re-introducing themselves to one another. Unlike [68] where a leave-one-subject-out methodology was applied, we explicitly divided the event days, using day 1 for training, and days 2 and 3 for testing. Under the aforementioned restrictions, the subset for this experiment included data from 17 participants in the first day for training, whereas the test set included data from 26 participants from the second and third days.

For the training set, the percentages of the speaking intervals for participants differed between 4% and 59%, with the mean and standard deviation of 32% and 15%, respectively. Similarly, the percentages for the test set was between 6% and 66%, with mean and standard deviation of 27% and 16%. These percentages show that we were able to capture varying amounts of speech in our dataset, even for a fairly comparative subset.

We have used the same features and feature extraction setup explained in [68] and obtained a 70-dimensional feature vector per each 3s window. In order to obtain the speaking label for each window, we used majority voting. Given that we hypothesized that body movements accompany speaking are highly person specific, we compared the performance of two different machine learning models: the binary class L2 penalized logistic regression classifier, and Transductive Parameter Transfer Method (TPT), in a similar manner to [68]. As the performance metric, Area Under Curve (AUC) was selected, since it provides a better estimation of the actual performance in case of imbalance. Results obtained for each participant in the test set by both methods can be seen in Figure 2.12.

As can be seen from Figure 2.12, TPT outperformed the traditional setup in the majority of the cases. With TPT, we obtained an average AUC of $61 \pm 6\%$ where it was $51 \pm 7\%$ for the traditional person independent setup. A paired one tailed t-test between the performances

resulted in a p-value of smaller than 0.01, showing high significance. When analysed individually for each participant, it can be seen that TPT provided better results for 21 out of 26 participants. Also, in all cases, TPT provided a performance better than random (50%), whereas the traditional person independent setup failed to do so in many cases. The results obtained with this data is on par with those in [68].

2.6.3. ATTRACTION DETECTION FROM MOVEMENT

Speed-dating events offer an ecologically-valid context to study initial interactions. Although the use of automated techniques has shown promise [70], behavioral assessments are often gathered by human raters in the social sciences when analyzing nonverbal behavior. Instead, here we attempt to provide baselines for the automatic classification of attraction scores using movement-based features. Thus, we emulate the features presented by Veenstra and Hung [167], which use movement as a link to arousal to predict the participant's responses during each date. Recall that after each date, participants indicated whether they would like to see their interaction partner again (yes/no) as well as how much (7-point Likert scale). They were also asked to rate their date as a short term sexual partner, and as a long term romantic partner or as a friend (7-point Likert scale).

We treated the attraction detection task as a binary classification problem, separating it into the classes *See Again*, *Romantic*, *Sexual*, or *Friendly*. To do so, we use a the median of each class as threshold to convert the Likert scale to a binary class. Given the 7-point Likert scale ranging from 1 (low) to 7 (high), the medians for each category are: 2 for Sexual and Romantic, 4 for SeeAgain and 5 for Friendly.

We have wearable acceleration and booklet responses from 10052 dates (542 and 510 from females and males respectively). We treat each date as a 4-dimensional feature vector. We extracted the mean and variance of the magnitude of acceleration ($abs = \sqrt{x^2 + y^2 + z^2}$) for each date. In addition, we calculated the variance over a 1 seconds sliding window with a shift of 0.5 seconds. Two additional features were extracted from this variance over a sliding window: the mean and the variance, leading to a total of 4 basic features. These feature emulate as close as possible those in [167].

We chose a logistic regressor as classifier to avoid overfitting and applied a 10-fold cross-validation. For significance, we applied a paired one-tailed t-test to the performance values of our baselines and the random baseline classifier (most-frequent). Thus, we obtained a mean F-score for the folds of 0.55 ± 0.11 for the SeeAgain class, 0.65 ± 0.12 for Romantic, 0.61 ± 0.14 for Sexual and 0.41 ± 0.141 for the Friendly class. All these values are statistically significant with respect to the baseline ($p < 0.01$).

Furthermore, [167] showed that separating males from females can further improve the performance of the 'SeeAgain' class. This is also a normal practice in works for attraction estimation [85, 125]. Applying this separation, we obtained a similar finding with a F-score of 0.60 ± 0.14 for males and 0.42 ± 0.11 for females (both with $p < 0.001$).

Note that only using movement-based features from the accelerometers to measure arousal, emulating what was done by Veenstra and Hung [167] in a simple classification setup, provides rather acceptable F-scores for the two classes related to attraction (Romantic and Sexual interest). Then by further investigating the relation between non-verbal behavior and attraction, one could use other resources given by *MatchNMingle* to increase this performance.

2.7. LIMITATIONS OF *MatchNMingle*

Perhaps the most straightforward limitation is the number of devices that malfunctioned during the data collection: a total of 20 out of 92 (22%) for the speed dates and 22 (24%) for the mingle. This level of malfunctioning devices is unfortunately one of the disadvantages of working in real life scenarios. Thus, some of the people interacting during the mingle, and for which there are annotations, do not have wearable acceleration and proximity. This malfunction also affected the date interactions (see Table 2.6).

Nonetheless, 70 working devices (72 for the dates) is still the largest number of devices recording an event of this nature. Also, we have more than 67% of the devices working per day for the dates. Furthermore, one could always address the problem considering missing data in one modality as proposed by Alameda-Pineda et. al. [2].

Also regarding the devices comes the distinction of using radio instead of IR communication for the proximity detection, as is done in SALSA [1]. The main difference between the 2 approaches is that, while IR communication is directional (in the form of a cone pointing forwards), radio communication is omnidirectional thus detecting devices in mostly all directions[9].

This is the reason why the recall reported in Section 2.4.3 (90%) is high while the precision is rather low (33%). Our wearable devices detect neighbors in mostly all directions, which is a rather high number in such crowded environments and where there are many distinct F-Formations that are spatially close. Fortunately, this also allows our devices to detect people in the same group which are standing next to each other (eg. see people standing next to the tables in Figure 2.4(b)), but also wrongly detects as neighbors people from different groups standing close. In contrast, IR communication tends to have high precision values but low recall as they detect strictly face-to-face interactions but miss the detection of people standing next to each other if the F-Formation is not strictly a circle or if the group has too many participants. Thus, while our detection can be later refined to detect the true conversing groups, the detection of these groups gets lost in the data collection when using IR.

The audio recorded also has its limitations, as it was recorded by the cameras instead of personal microphones (also discussed in Section 2.5.1). Thus, the audio available consists of recordings of the global audio of the events, with a high noise level due to the inherent nature of the events.

Finally, due to the perspective of the cameras and the *in-the-wild* nature of the events, face detection and pose estimation are still open and interesting problems for this type of data. As saw in Section 2.6, OpenFace [14] and OpenPose [32] were applied to this data, with discouraging results which are mainly due to the camera perspective. The limitations of this dataset for the analysis of facial cues was also discussed in Section 2.5.1. Nonetheless, a top view perspective was necessary as side views are more prone to participant occlusions, especially for crowded scenes such as this, and for those people who are farthest away from the camera.

Nevertheless, we must emphasize that this dataset represents an opportunity to study those cases where clean data from each participant is unavailable, and where the nature of the event itself makes it difficult to capture some standard elements in other scenarios (eg.

[9] The human body has proved to be a natural damping for radio communications [72], thus limiting detections of people standing back to back.

facial expressions). Thus, while the recording of the participants' faces is straightforward in a meeting setup for example, this is not the case for mingle scenarios.

2.8. DISCUSSION

For the purposes of this thesis, the use of *MatchNMingle* was limited to answer research questions regarding cue detection (Chapter 4) and personality estimation (Chapter 5), and analyze the role of multimodality in these estimation tasks. The dataset was also a resource to adress tasks in lower levels of abstraction, specifically the association of streams of different modalities using similarity matching (Chapter 3). These examples show how the same dataset can be used to analyze different types of research questions, all within the same scenario. However, due to the exploratory nature of *MatchNMingle*, they are still a rather restrictive use of the true potential of the dataset.

Hence, the use of *MatchNMingle* for future works on automatic multimodal analysis of social interactions in-the-wild is rather straightforward and encouraged. This is the main reason why the dataset was made openly available, to act as a multimodal resource for such analysis without limiting the researchers to imposed research questions.

Also, it should be discuss the main finding of this Chapter regarding the use of crowdsourcing for annotations of social context. The results in Section 2.5.4 show that, while helpful with simple tasks such as annotating the position of the people in images, Mechanical Turk (MTurk) annotators failed to provide quality annotations for complex tasks such as social actions. Even when provided with the same conditions for the annotations, the inter-annotator agreements and time to control the productivity between MTurk and on-site workers differed dramatically.

One should wonder about the reason behind this difference. We hypothesize that the main reason behind it is the type of tasks that are historically predominant on crowdsourcing platforms, generally of a simple nature (e.g. *Does this picture has a dog? Yes/No*). Thus, workers are used to tasks that are simple and time efficient. This is not the case for tasks regarding social actions, which are time consuming, repetitive and have high subjectivity. This demonstrates the importance of coders (trained annotators) in social psychology, and how their work and insights should be better reflected in the computing community for these type of efforts.

3

AUTOMATIC ASSOCIATION OF MULTIPLE WEARABLE DEVICES WITH ITS WEARER ON VIDEO

The contents of this chapter are based on the work originally published in:
L. Cabrera-Quiros and H. Hung. **Who is where?: Matching People in Video to Wearable Acceleration During Crowded Mingling Events**. Proceedings of the ACM International Conference on Multimedia (ACM MM), 2016.
L. Cabrera-Quiros and H. Hung. **A hierarchical approach for associating body-worn sensors to video regions in crowded mingling scenarios**. Submitted to IEEE Transactions on Multimedia and under revision.

3.1. INTRODUCTION

Past efforts in human behavior analysis have proved that fusing modalities (eg. video and audio or video and wearable sensors) increases the performance of recognition and classification of a wide variety of tasks, such as dominance [84], leadership [145] or cohesion [77]. Thus, each modality contributes to a different element of the event and acts as a complementary source of information. In addition, the use of multiple modalities had shown to be a suitable alternative to deal with challenging scenarios, including group gatherings [13, 154].

Wearable devices are a modality that has been used considerably in mingle scenarios, due to its versatility. Nonetheless, although the use of wearable sensors as a complementary source of information has many advantages, manually associating a specific device to a particular region of the video (corresponding to the person using the device) quickly becomes a challenging practical issue as the number of streams to associate per modality increases, making the correct associations harder to discriminate.

In fact, when using other modalities along with video, the majority of works either i) manually associate video to the other modalities [2, 75], or ii) avoid the problem entirely by using only one source in the other modalities (eg. only one wearable device or microphone) [145, 154].

In this chapter, we present a solution to tackle this problem by associating the time series signals from wearable accelerometers to the acceleration streams extracted from video flows. Thus, we aim to associate each device with the spatio-temporal region of video of its wearer. This association is particularly challenging for mingle scenarios, as people's social behavior in these events (unlike simple actions like walking or running) do not tend to have a predictable and easily distinguishable pattern; and as the number of people increases. To do so, we use wearable devices recording wearable acceleration hung around the neck, and overhead cameras.

We address large-scale data association in challenging crowded environments by proposing a novel method based on an extension of the Hungarian algorithm. Thus, we leverage the use of proximity information from the wearable devices and video as a spatial prior to the association process. Using the proximity, we can subdivide the association problem to areas in the real world sharing the same spatial-social context. Thus, we could applied a *divide and conquer* strategy, by associating the streams within all possible group combinations from different modalities, and then selecting the optimal group-to-group association.

A preliminary version of the method was presented in [28]. However, this version did not accounted for missing data or uneven number of streams in each modality. Instead, for groups with unequal instances in the modalities (eg. more devices than people in video) the streams remaining after the group-to-group association were discarded, making our method able to handle uneven streams globally but not optimized for such case. In addition, all experiments were done with a limited number of participants (19), all which have complete time series.

As a consequence, a modified version of the method was later presented in [29]. In this new version we improved over the aforementioned aspects and present several additional novel contributions:

• We modify our method to account for unequal groups of streams and streams with missing data. Thus, our method is now optimized to handle any combination of streams, dynam-

ically accounting for uneven numbers of streams both globally and in the group-to-group associations.
• We increase the number of streams to be associated to 69 in each modality. In addition, these streams have also missing data in different proportions (given the behavior of each person), which makes them a more suitable example of cases in real scenarios.
• We include a more comprehensive evaluation of our method. We address issues related with understanding the association process such as evaluating the impact of the number of participants and the period over which observations are accumulated on the accuracy of the association, the effects and errors introduced by the group-to-group matching and assess the impact on the association performance of missing streams, either partially (missing data) or completely (missing streams).
• We further analyze qualitatively if shared social actions (eg. shared gestures or laughter) have any impact on the association process, as we hypothesized that due to mimicry these could become failure cases for our method.

3.2. RELATED WORK

Several works have used information from video and wearable sensors for a wide range of tasks such as human action/activity classification [13, 154] or group detection [2], among many others. However, very few have addressed the challenging task of automatically associating the video pixels or regions with the additional sensor modalities, such wearable sensors. Although many works exists on video-to-audio association [3, 76, 164], which is generally called *speaker diarization*, we will only refer to works about association of video with wearable devices, as other modalities are outside the scope of this paper. For more details in audio-video association, see [3].

When associating wearable acceleration with video, previous works can be divided in 2 main approaches: 1) pixels-to-device association and 2) region-to-device association. In the former, each device is associated to the set of more similar pixels in terms of a given similarity measurement (eg. Euclidean distance). As they have several more streams in one modality, such as the pixel trajectories of all people moving, these approaches tend to use 3D and orientation measurements in both modalities, which allows them to be more discriminative. In a region-to-device association, the set of pixels is previously clustered by a defined technique such as manual annotation, image-based segmentation or object tracking, among others. Then, each region of interest is associated with their corresponding device. Our work is an example of the latter.

Rofouei et al. [142] and Wilson and Benko [177] are examples of works using a pixels-to-device association. They proposed similar methods to match the 3D acceleration of a smartphone (also using its gyroscope) to the set of pixels with the higher similarity in a video recorded with a Kinect, which also recorded depth information. Thus, constructing the real 3D world coordinates from the Kinect and knowing its position w.r.t the real world, they mapped all the devices to these real world coordinates. To measure the similarity, their methods are based on an euclidean distance minimization between both streams. Bahle et al. [13] proposed a similar association of pixels-to-device, but limiting the pixels to those regions on the joints detected by the Kinect. They also used a 3D reconstruction of the real world and the Dynamic Time Warping (DTW) distance as similarity measure.

Although these methods essentially match acceleration streams like ours, their solutions

are oriented to the interaction with a display using mobile phones. Hence, they do not consider a high number of devices and the implications that this could have in the association process with video. In addition, they reported problems with fast movements and during moments when the device was not moving.

For region-to-device association, the closest works to our own is Texeira et al. [159]. They presented an approach based on Hidden Markov Models to identify and localize moving smart phones (by their accelerometers and magnetometers) in a camera network. To do so, they modeled the association as a missing data problem where a person's behavior is observed twice, once by the camera and once by the wearable device, but the link between the 2 modalities is unknown. They proposed a solution that could ultimately work for more than one device, but in their experiments have one single person walking under the network of cameras. This unique stream is later divide into 5 and each is treated as a different participant. This solution seems to be a suitable option to 'generate' more participants, but they do not address the challenges of occlusion resulting from a crowded scene making their solution infeasible for mingling groups. In addition, unlike in our case, the streams that they generate do not have any interaction between each other in real life, which makes the dataset they used not a nice representation of a mingle scenario, with its possible consequences in the matching process.

Other works in region-to-device association include Shigeta et al. [152], Plotz et al. [129] and Nguyen et al. [120]. The methods proposed by Shigeta et al. and Plotz et al. first detected the moving areas in the video and associate these to a corresponding device within a set of 5 and 3 devices, respectively. As their acceleration signal are not synchronized in time, unlike our case, they used the peaks in the Normalized Cross Correlation (NCC) between the acceleration signal and the region in the video to detect the proper alignment between the signals. Thus, once the peaks are found they choose the matches between devices and video using a greedy assignment.

These methods are feasible for a small number of devices but when this number increases the discrimination between the streams is harder to perform in a greedy manner, as we will prove later in this work (see Section 3.5.2), and the NCC starts to fail while providing the correct alignment. Also, both methods, are limited to moving objects.

Compared to these works (including Texeira's), our approach proposes a considerable increase in the number of accelerometers to be associated, where we show improvements in performance over the state-of-the-art methods even when matching over 60 video and wearable acceleration streams using a hierarchical grouping approach.

To the best of our knowledge, thus far we are the first to consider the association of video with multiple wearable devices in such large and crowded scenarios, considering missed streams (devices or video missing) and streams with incomplete data. In addition, we propose to solve the association problem in a much more challenging context where people's behavior can not be as easily characterized as simple actions like standing and walking and it is harder to discriminate between people's movements.

Figure 3.1: Changes in appearance of 2 of our participants through the mingle event

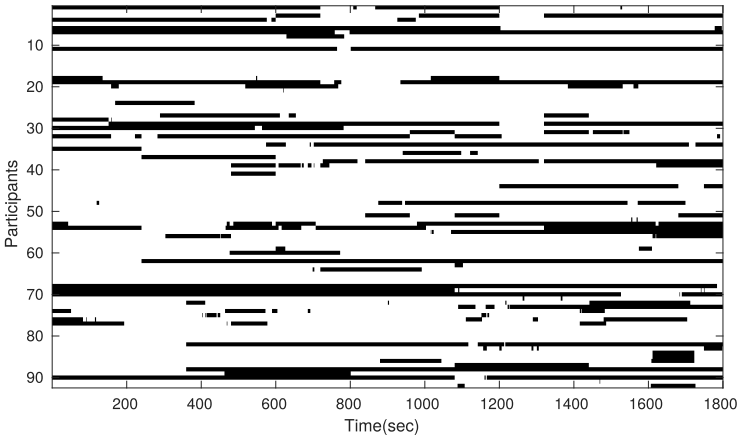


Figure 3.2: Participants visibility status. Black = participant missing.

3.3. OUR ASSOCIATION METHOD

3.3.1. REAL MINGLING SCENARIO DATASET

For these experiments, we use the version 1 of the *MatchNMingle* dataset (see Chapter 2), where it was collected video and wearable acceleration for 92 participants during 3 separated group gatherings. Due to hardware malfunctioning, 22 of the devices did not record data during the event leaving 70 functioning devices. One more device was excluded due to incomplete data for a total of **69 participants**.

As explained in Chapter 2, each person wore a custom-made wearable device hung around the neck which recorded triaxial acceleration at 20 Hz, and overhead video was captured using 5 different GoPro Hero 3+ cameras that covered the whole mingling area with some overlap. Finally, 8 different social actions (Walking, Stepping, Drinking, Speaking, Hand Gesture, Head Gesture, Laugh and Hair Touching) were annotated for each participant (when visible) every second.

COMPLEXITY OF OUR DATASET FOR THIS TASK

Since our event was recorded during a real mingle event, all the participants had the liberty to move around and leave the mingle area at will. They were recorded by different cameras, with different light conditions and strong appearance changes given their position w.r.t. the camera. For example, Figure 3.1 shows the changes in appearance of 2 of our participants. Moreover, Figure 3.2 shows the visibility status of the 69 participants, under the 2 cameras with the higher concentration of people (version 1 of the dataset), for an interval of 10 minutes chosen randomly from the mingle segment. This is the same 10 minute interval

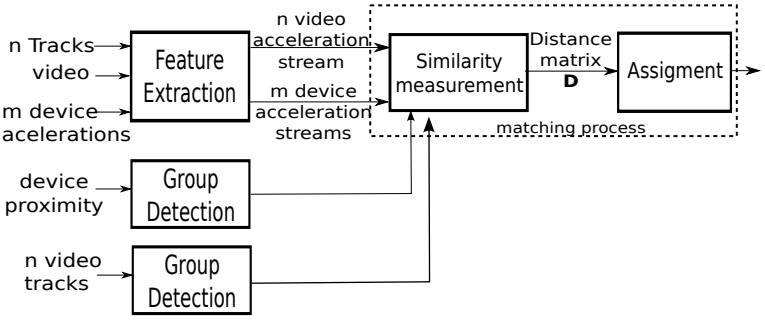


Figure 3.3: Overview of our approach.

that will be used later in our experiments. The visible times are not necessarily consecutive and that, in fact, some were missing for the entire interval.

From the entire 69 participants, 22 are visible for the entire 10 minutes while there is not video data for 14 people. The rest of participants are visible for a variable proportion of the time, summarized in Table during the 10 minutes. Table 3.1 summarizes the number of participants visible for at least a given amount of time (X). Here, the subset were all people are under the FoV for the entire time (last column) will be our *ideal subset*, while the set with all 69 participants is our *entire set*. Thus, only a 31.9% of the streams are complete for our entire set while 20% of the streams are entirely missing.

Table 3.1: Number of participants visible for at least an X amount of time

Minimum time X under FoV (minutes)	1	2	3	4	5	6	7	8	9	10
Number of participants	54	53	52	51	50	46	46	37	33	22

3.4. OUR APPROACH

Our approach is summarized in Fig. 3.3 and detailed below.

3.4.1. FEATURE EXTRACTION

WEARABLE DEVICES

For the wearable devices, a single acceleration stream for each device is obtained using the magnitude of the 3 axes. Using magnitudes allow us to compare the device's acceleration to the video without knowing the orientation reference between the two modalities in the real world. To eliminate the influence of gravity, each axis is first normalized using its mean and variance over the entire observation time.

VIDEO

Each device stream must be assigned to a specific person in the video. As stated before, in this work we do not intend to perform a pixel-to-device association, but rather associate each device with a region containing a person. Hence, all those regions of interest (or bounding boxes), which include a person with a device, are first extracted. Then, we

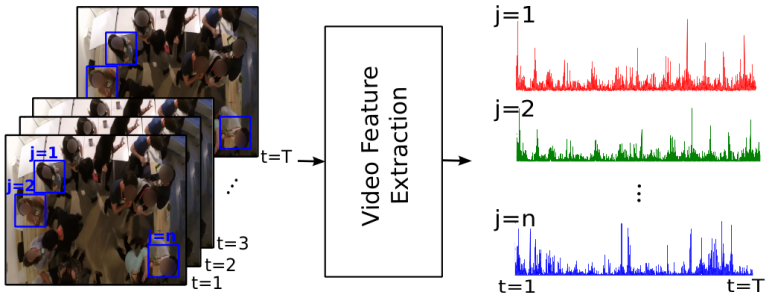


Figure 3.4: Feature extraction from video for 3 example tracks (subjects). Output: speed stream for each participant for interval of length T.

concatenate the bounding boxes over time for each person, to generate a track or tube (area of interest over time) for that person (see Fig. 3.4).

The annotations provided by the dataset (see Chapter 2. While this are manual annotations, we found that using the SPOT tracker [184] gave us similar results, with a mean overlapping ratio between participants of 0.9006 (1 equals the highest) and a deviation of 0.0632 for 10 participants randomly selected in the 10 minute segment.

The annotated ROI are concatenated to form the position tubes for each person, we proceed to treat each tube as follows. First, we extract dense optical flow for the entire video. Then for a given bounding box, which belongs to a tube, we take the magnitude of all the flow vectors and then compute the mean for those with a magnitude greater than zero. In this way, we obtain a vector of mean flow magnitudes for a given tube over the entire video of length T (where T is the number of frames in the video interval). This is used to represent the velocity of movement for that person between two consecutive frames. This approach allows us to consider the influence of fine grained movements such as gestures or laughter as well as movement of the entire body. Figure 3.4 shows a graphical representation of this process. Finally, we compute the acceleration vector from the speed using finite difference approximation to obtain a measurement comparable to the acceleration in the devices.

After we extract the acceleration streams from the video and wearable accelerometers, we proceed to treat each stream (video and device) as follows. First, we normalize the maximum value of all streams to one, so a comparison between video and wearable acceleration can be made. Next, we apply a sliding window calculating the variance over each stream. Using this instead of the raw acceleration will give us a better representation of the activity levels of the people [80]. Additionally, it has been proved in activity recognition using wearable acceleration that working with raw acceleration values can present difficulties due to recording noise, among others factors [94].

3.4.2. SIMILARITY METRICS

Both video and acceleration streams are noisy because they capture only partially the behavior of a person. Since the device is hung around the neck, movements from the torso are strongly captured by this modalitiy. However, energetic gesturing in the video will not necessarily be directly translated into similarly energetic movement in the body. Therefore, we need measurements to assess how similar 2 streams are and not if they are equal. Different

metrics are compared to quantify the affinity between the acceleration streams from video and the devices: covariance (COV), Dynamic Time Warping Distance (DTW) and Mutual Information (MI). These metrics are widely used to assess affinity between streams [47]. Notice that the similarity metrics only consider those sections where there is information for both modalities (for the improved version). Hence, for the covariance and the mutual information, sections with missing data in one or both modalities are ignored, and weighted given the length of the complete stream. For the Dynamic Time Warping distance, all sections of complete data are treated as separately streams and the overall distance is calculated by taking the mean distance of all segments weighted by their length.

3.4.3. ASSIGNMENT METHODS

We consider the matching process to be an assignment problem, where m elements of a set M (device streams, in our case) need to be associated with n elements of a set N (video streams), by fulfilling a given function or constraint. The distances matrix $\mathbf{D}_{ij}$, of size $m \times n$, is formed by the pairwise distances between all possible combinations of m acceleration and n video streams, where

$$\mathbf{D}_{ij} = d(i,j), i \in \{1 \cdots m\} \text{ and } j \in \{1 \cdots n\} \tag{3.1}$$

and d is one of the similarity metrics in Section 3.4.2.

WINNER-TAKES-ALL (GREEDY) ASSOCIATION

State-of-the-art methods ([142, 152, 177]) use a greedy approach where the element in $\mathbf{D}_{ij}$ that has the highest value determines the assignment. The corresponding column and row are removed from $\mathbf{D}_{ij}$ and the assignment process is repeated. This relies on a strong correlation between the sensor data for a given device and its corresponding video stream. This is the baseline that we compare our proposed method with.

HUNGARIAN METHOD

Although the winner-takes-all method is a reasonable baseline, it does not consider that there is likely to be noise in both sensor streams. Hence, it may not be able to distinguish one possible assignment from the other. This is particularly problematic as the number of streams increases. In this case, trying to optimize the assignments globally may help.

The Hungarian method [25] computes a solution for the linear assignment problem by optimally matching the elements m and n, based on a global optimization of $\mathbf{D}_{ij}$. For this assignment problem, given the matrix of distances $\mathbf{D}_{ij}$, the aim is to find the global cost c that minimizes

$$\begin{aligned} &\min \sum_{i=1}^{m} \sum_{j=1}^{n} d(i,j) * w(i,j) \\ &\text{s.t.} \sum_{i=1}^{m} w(i,j) = 1, && j = 1,2,...,n \\ &\sum_{j=1}^{n} w(i,j) = 1, && i = 1,2,...,m \\ &w(i,j) = 0,1 \end{aligned} \tag{3.2}$$

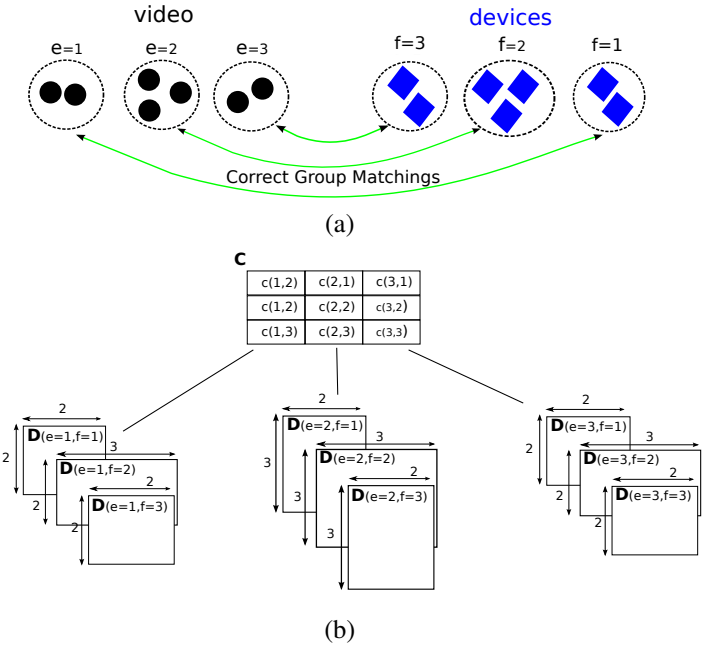


Figure 3.5: Example of assignment method. (a) Devices and Video streams representations. The dotted circles show the group detection. (b) Our proposed *Hierarchical Hungarian* method using the streams and clusters from (a).

where $w(i,j)$ is the binary weight for matrix $\mathbf{W} \in \{0,1\}^{m \times n}$ for the element (i,j). Thus, $w(i,j) = 1$ if the two pairs are associated, the method will choose the pairs of elements with the lowest total pairing cost and the elements of sets M and N can only be paired once. Several solutions exist to solve this problem [124]. Notice that Eq. 3.2 is defined such that it holds for those cases where $m \neq n$. Thus, our association is not limited to an equal number of streams on each modality.

HIERARCHICAL HUNGARIAN METHOD

As the number of streams to be associated increases there is a higher probability of finding 2 or more people with similar streams. Hence, the observation period needs to be longer to increase the chances of discriminating between them. Computationally speaking, however, it is desirable for a potential real-time application of this work, to be able to rely on shorter time intervals to make the association. Although, if the streams are too short, we will not have enough observable behavior for the distance metric to be discriminative enough.

By initially sub-dividing the problem based on the local spatial neighborhood in each sensor, we hypothetically could improve the numbers of correctly associated streams. Therefore, we propose an extension to the original Hungarian method by performing the assignment procedure in a hierarchical manner using a divide-and-conquer strategy where all the streams are subdivided into groups in each modality. This reduces the problem initially to a smaller size represented by the number of groups in each modality. We propose to generate the groups by clustering based on their proximity over a particular time interval (described later in Section 3.4.4). This further reduces the assignment problem from a global to lo-

cal assignment problem, which exploits the local-spatial and social context of the mingling gathering.

So, for this new assignment methods, the n video and m accelerometer streams are clustered into p groups for the acceleration and q groups for the video streams. Then, $p \times q$ different distance matrices are generated; one for each possible group combination (e, f) where indices $e \in \{1 \cdots p\}$ and $f \in \{1 \cdots q\}$. For each of these matrices, the corresponding stream assignment is calculated. So, within each group-to-group matching, the possible stream combinations are now reduced to $n'_e \times m'_f$, where n'_e and m'_f are the number of elements in the e^{th} and f^{th} device and video groupings, respectively.

The cost $c(e, f)$ of each group-to-group assignment is then obtained by Eq. 3.2. These costs are allocated in a new matrix $\mathbf{C}$, which represents the costs of assigning the elements within each possible group combination e and f. Note that each cost $c(e, f)$ must be normalized by dividing by the number of total costs that were used in each assignment so $\mathbf{C}(e, f) = c(e, f) / \min(m'_f, n'_e)$. For example, when comparing a group of 3 streams against a group of 2, only 2 costs from the 3×2 matrix are used for the final assignment, whereas comparing 2 groups of 3 streams we will have a summation of 3 costs from the 3×3 matrix.

Finally, the Hungarian algorithm is applied to matrix $\mathbf{C}$ to find the optimal group-to-group assignment. The stream assignment for that specific group-to-group pairing is then chosen. An example of our Hierarchical Hungarian assignment procedure is illustrated in Figure 3.5.

Notice that due to a mismatch in the number of streams on each modality (eg. missing people) or incorrectly matched groups, some streams can be left without associating. To account for these cases, we modify our original method. First, all similarity metrics used in the assignment already account for missing data, as explained in Section 3.4.2.

In addition, from an early submission presented in [28] we noticed that when the groups are wrongly matched some streams are left unassigned even when is one or more streams left in each modality. Thus, we improved our method and made it more resilient against wrongly matched groups or incorrect group detections in either modality, as we must take the proximity prior as noise and imperfect information. To do so, when there is one or more streams in *both* modalities, we performed a final assignment with the remaining streams, without grouping, treating all the streams as singletons (group of one person). Those streams remaining without associating after this final step are treated as missing people in one of the modalities ($m \neq n$ in Eq. 3.1).

3.4.4. GROUP DETECTION

The group detection is performed independently per sensor type using mainly the same clustering algorithm based in maximal cliques [79], treating the proximities of the participants as graphs. The difference between the 2 modalities lies in the process to create such graph. For each sensor type, this process and the clustering is described below. We choose maximal cliques as this clustering method has prove to be an accurate approximation for conversational groups [79], following the same scheme as the behavior presented by people during group forming.

CLUSTERING VIDEO STREAMS

To create the graph for cluster the video streams, we use the tracks extracted for each of the participants, focusing in the position of the center of the track in each frame of the video. Thus, for each frame, an affinity matrix $\mathbf{A}$ is created, which defines a symmetric distance between person i and j

$$\mathbf{A}_{ij} = -e^{\frac{d_{ij}}{2\sigma^2}} \tag{3.3}$$

where d_{ij} is the Euclidean distance in the image plane between the centroids of the bounding boxes for person i and j and σ is the width of the Gaussian kernel. In our experiments, σ was set to 150 pixels, as this was an approximate value for group distance given the image size and resolution of the camera. This threshold was selected by learning the mean area of coverage of all our participants in video for the entire dataset.

Then, we apply the group detection algorithm that extracts clusters as maximal cliques in edge-weighted graphs [79]. This is an iterative procedure that optimizes the group clustering based on the notion of a dominant set. If we have a graph G with each node representing the centroid of a person's bounding box and the affinity between people to be the edges, we can consider a representation of the closeness of a subset S of the graph as follows. We define a measure called the average weighted degree of a vertex $i \in S$ with respect to set S as $k_S(i) = \frac{1}{|S|}\sum_{j\in S} a_{ij}$. The relative affinity between node $j \notin S$ and i is defined as $\phi_S(i,j) = a_{ij} - k_S(i)$, and the weight of each i with respect to a set $S = R \cup \{i\}$ is defined recursively as

$$w_S(i) = \begin{cases} 1 & \text{if} |S| = 1 \\ \sum_{j\in R} \phi_R(j,i)\, w_R(j) & \text{otherwise} \end{cases} \tag{3.4}$$

$w_S(i)$ measures the overall relative affinity between i and the rest of the vertices in S, weighted by the overall affinity of the vertices in R. Therefore to find the cliques in the graph $w_S(i) > 0, \forall i \in S$. For every graph, only one maximal clique can be identified at a time and a peeling strategy is employed where the same conditions are repeatedly applied to the remaining sub-graph until no more cliques remain.

Finally, the cliques selected per frame are combined into a single set of groupings q for the entire video segment using majority voting. Thus, groups with the same set of participants are counted for the entire segment of recordings and the ones with the exclusive majority are selected.

Luckily, in a mingle scenario the people tend to stay in the same group for long intervals of time, making this selection method feasible. For example, for our event 17% of the participants stayed in the same group for the entire 10 minutes, 20% joined only 2 groups, 11% joined 3 and 17% joined 4 groups (total of 65%). Only 17% joined 6 or more groups. Notice that these statistics includes merging groups and excludes singletons (people alone). Thus, 2 groups of 2 people joining counts as 2 groups joined, even if they stayed with a same person during the entire event.

CLUSTERING DEVICES

As stated in Section 5.3, each of the wearable devices outputs a dynamic binary proximity graph, which is later refined to eliminate false neighbor detections using the method proposed by Martella et al. [106]. Thus, for each time sample which is recorded at the same sample rate as the video (20Hz against 20fps) a proximity graph is created between the participants. To refined false neighbor detections, they apply a density-based clustering to

group all the neighbor detections in time, by comparing the graphs of consecutive times. This method leverages the bursty nature of the proximity graphs, meaning that the correct neighbor detections tend to appear sequentially together in time and the false detections tend to be isolated (see [106] for more details).

Finally, the maximal cliques are identify from the proximity graphs, to obtain p sets of fully connected nodes, using the same maximal clique methods as with video. Here, d_{ij} in the affinity matrix **A** from Eq. 3.3 is created with the binary values from the proximity graphs.

3.5. EXPERIMENTAL RESULTS

For our experiments, we selected a 10 minute interval chosen randomly in the middle of the mingle event. For all 69 people with functioning devices, we extracted our wearable acceleration streams (see Section 3.4.1) using a sliding window of 50 samples with a shift of one sample for which we calculated the variance. This window length (equals to 2.5 seconds) gives enough time for an human action to fully develop.

As explained in Section 5.3, not all participants were present under the FoV of the cameras for the entire interval. So, the video acceleration streams were extracted for these 69 participants were video data was available. If their video data was incomplete, the acceleration stream was set to zero for those times only for practical purposes. This is done for purposes of a further comparison with our old method (see Section 3.5.4). Nonetheless, as explained in Section 3.4.2, these sections are not take into account for the creation of our distance matrix with our new approach.

In general, we will treat as true positives (TP) all the pairs of streams that were associated correctly. Thus, our association accuracy will be number of true positives over the total number of streams to associate in the modality with less streams, or $acc = \frac{TP}{min(m,n)}$. Notice that, as well as Eq. 3.2, this considers a different number of streams on each modality. Also, in those cases with K-folds (eg. leave out experiments), the mean accuracy will be the equal to acc_{fold}/K.

For the association including grouping (see Section 3.4.3), the accuracy will be equal to the number of true positives that were correctly associate within a group matching that was also correctly associated. Also, TP_{group} will be used to denote those groups that were correctly matched and $acc_{group}(e,f)$ as the association accuracy within a given group pair (e,f).

3.5.1. COMPARING BETWEEN DISTANCE METRICS (WITHOUT GROUPING)

First, we compare the metrics in Section 3.4.2. Our intention is to assess the impact of each metric on the original linear assignment problem without applying our hierarchical approach just yet. To do so, we used our *ideal subset* (22 people, as seen Table 3.1) where there is not missing data which represents an ideal scenario and our entire set of 69 participants. For both sets we used the entire segment of 10 minutes. Also, for the participants with missing video the acceleration streams from video were set to zero.

Table 3.2 summarizes the results for the association of both sets. For both sets, all similarity metrics (using greedy or Hungarian) outperform the random baseline. Using the

Table 3.2: Association accuracy without grouping for the ideal subset (22 participants) and the entire set (69 participants).

	Accuracy (%)					
	Greedy			Hungarian		
	MI	COV	DTW	MI	COV	DTW
Ideal subset	36.36	63.64	18.18	22.73	**86.36**	77.27
Entire set	11.59	37.68	11.59	13.04	**46.38**	36.23

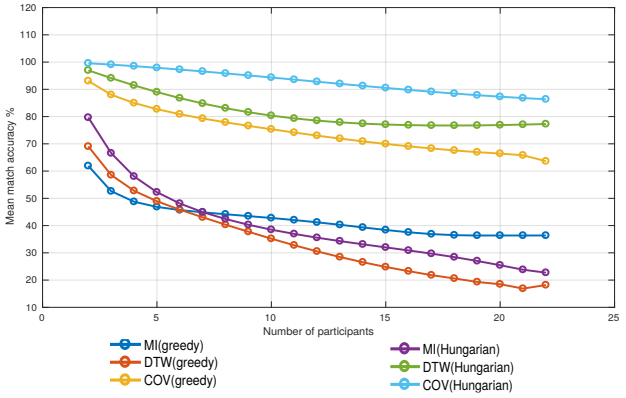


Figure 3.6: Mean association accuracy without grouping for the different number of participants using the ideal set as base (10 minutes).

covariance (COV) as a similarity metric gives the best association performance for either the greedy or the Hungarian assignment. The DTW seems to work well when combined with the Hungarian assignment, suggesting that this metric generates similarity values which are close together while the COV gives more discriminative values. Hence, a global optimization is necessary for the DTW but not for the COV.

We also found a significant difference in the association accuracy between our ideal subset (22 people with only clean data) and our entire set (69 people with missing data). This difference could be explained by one of 3 factors (or a combination of them): 1) different number of participants, 2) quality of the data (clean versus missing) or 3) an social aspect within the observations. These 3 aspects will be further developed in the next subsections.

3.5.2. EFFECTS OF THE NUMBER OF STREAMS AND THE INTERVAL LENGTH ON THE ASSOCIATION PROCESS

In this section we analyze the impact on the association accuracy of the number of participants to be associated and the observation length when extracting the acceleration streams. To do so, we use only our *ideal set* as base to maintain clean conditions (eg. no missing data in the streams).

First, for the analysis of the impact of the number of participants on the performance, we run associations with different number of participants. On each run, were $N \in \{2,...,22\}$ participants (ideal set has a total of 22 streams), we leave out k different participants iteratively ($k = 22 - N$) considering all possible K tuples. We then calculated the mean accuracy obtained by each association with N streams. Figure 3.6 summarizes these results.

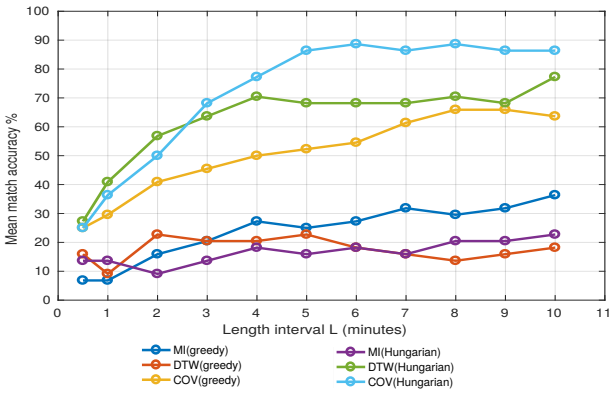


Figure 3.7: Association accuracy without grouping for the different length intervals using the ideal set (22 participants).

We can see how there is only an accuracy difference of about 13% between the sets of 22 and 2 participants, even when the number of participants was increased by a factor of 10. This was possibly due to the long interval (10 minutes) that was used for the association, and its is related to the strong trade-off between streams to associate (participants) and observation time when the association is done without grouping. Given that shorter observation intervals are preferred and supported by the results in the last rows of the Figure 3.6, we opted for a group-to-group assignment (Section 3.5.4).

Now, to analyze the impact of the length of the observation time, we gradually decrease this interval for the extraction the acceleration streams and calculated the association accuracy. Given that different parts of the interval can have different actions/events, we calculated the association using a sliding window of length L and shift it by $L/2$ and then report the mean value with its deviation over all the intervals. Figure 3.7 shows these results. Here, at least an observation time of 5 minutes is needed to accurately associate more than 80% of the 22 streams.

3.5.3. IMPACT OF MISSING PEOPLE IN VIDEO

As seen in Table 3.1, the 69 participants in our dataset stayed under the field of view of the cameras for different intervals of time. This implies that some acceleration streams from video will be partially or totally missing. Nevertheless, our method can also work in such cases, as can be seen in the formulation in Equation 3.2. The following is the empirical proof of this claimed. Our experiments only consider missing video streams, but the insights found will also applied for missing streams from the wearable devices.

For each subset of participants in Table 3.1 (number of people under the field of view of the cameras for at least a given amount of time X), we applied the greedy and Hungarian association assignments without grouping. Those streams with less information than 10 minutes, for all subsets, were filled with zeros for the missing parts. Figure 3.8 summarizes the association accuracy for these subsets.

Similar to the results on Table 3.2, the combination of the COV as similarity metric and the Hungarian assignment has the best performance. Notice how the overall association

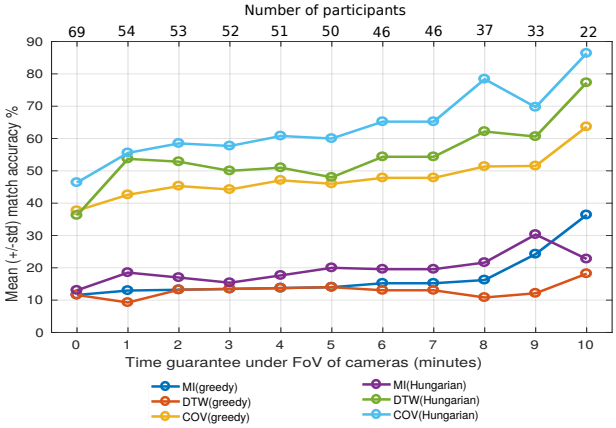


Figure 3.8: Association accuracy without grouping for the different subsets of participants in Table 3.1(10 minutes interval).

accuracy decreases as the data becomes more incomplete and the number of participants increases.

3.5.4. EVALUATION OF GROUP-TO-GROUP ASSIGNMENT

After analyzing the different components that can influence the association, we now introduce our Hierarchical Hungarian method which applies grouping. Table 3.3 summarizes the results for the association accuracies using this method. As well as in Figure 3.8, for these associations we selected those participants that were under the field of view of the cameras for at least a given amount of time X, an calculated the association accuracy for these different subsets. This table also includes the total number of groups involved in the association. The last 3 columns of Table 3.3 represents the accuracy with an ideal grouping. This means that the group formation of the participants (both in video and in the devices) are used and the correct group-to-group assignments $\{(e = 1, f = 1),\ (e = 2, f = 2),\ (e = 3, f = 3)\}$ are known. Thus, the the overall accuracy will be the mean accuracy for all $acc_{group}(e, f)$.

In addition, Figure 3.9 show the association accuracy against the number of participants (as in Table 3.1) for the 3 different metrics and the random baseline. Notice that both Table 3.3 and Figure 3.9 have sets with missing data.

Overall, all approaches are better than a random baseline (see Figure 3.9). Furthermore, our Hierarchical method over-performs all other approaches when using the covariance as metric. Moreover, when analyzing the TP_{group} (groups correctly matched) of each association one can see that the association errors come from incorrectly matched groups in different modalities. For example, in the second row of Table 3.3 we see that from 25 groups (in each modality, 50 in total) our method correctly matched 19, resulting in an accuracy of 81.82%. If all groups were correctly associated (ideal case), we can obtained a 100% accuracy using this metric, as seen in the second to last row. This implies that better algorithms to detect and match groups will improve our method. However, the correct group detection in each modality is not the main goal of this work. Nonetheless, we proved that using group detection as a prior, even when defective, increases the association performance.

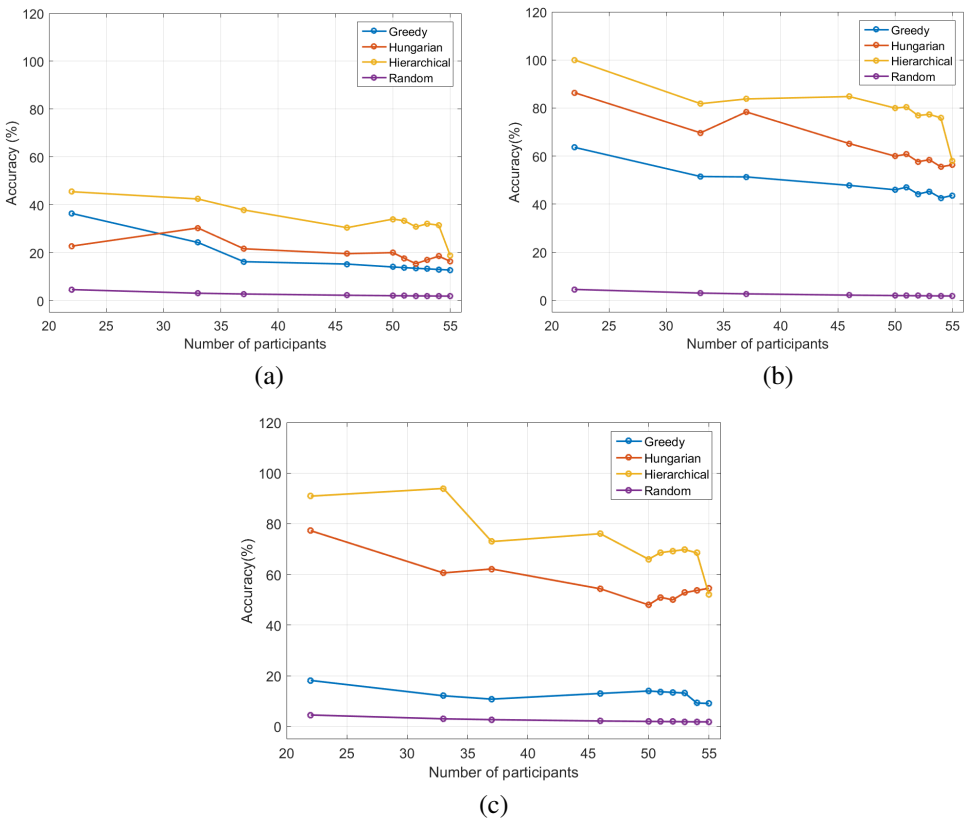


Figure 3.9: Accuracy of stream association for our proposed method (Hierarchical), the state-of-the-art (Hungarian and Greedy) and the random baseline using (a) MI, (b) COV and (c) DTW as similarity metrics.

Comparison of hierarchical method (missing versus complete data): To further evaluate the difference between our 2 implementations, one accounting for missing data and one none, we proceed to compare the results here presented to those obtained using the method presented in [28]. As explained in Sections 3.4.2 and 3.4.3, our method was optimized to account for missing data, either completely (missing streams) or partially (streams with missing data).

Table 3.4 summarizes the results for both methods using the covariance (COV) as metric, as this gave us the best results for both methods. Here, the sections with missing data in the streams were set to zero in order to use our old method. Nonetheless, our improved method account for this sections differently as was explained in Section 3.4.2.

The results for the complete set (22 people with no missing data) are rather similar with each other, and to what was presented in [28]. Here, as there is no missing data, the matrices for both methods are the same, which leads to the same result for the group-to-group assignment. The difference between the two is due to unmatched singletons, which

Table 3.3: Association accuracy and number of correctly associated groups using the Hierarchical Hungarian method for the different subsets of participants in Table 3.1(10 minutes interval).

Num. Partic.	Num. groups** ($min(p,q)$)	TP_{group}			Accuracy (%)					
					Hierarchical Hungarian			Ideal Hungarian(*)		
		MI	COV	DTW	MI	COV	DTW	MI	COV	DTW
22	16	9	16	14	45.45	100.00	90.91	68.18	100.00	90.91
33	25	11	19	23	42.42	81.82	93.94	78.79	100.00	93.94
37	28	10	22	20	37.84	83.78	72.97	81.08	100.00	89.19
46	34	10	27	25	30.43	84.78	76.09	69.57	100.00	91.30
46	34	10	27	25	30.43	84.78	76.09	69.57	100.00	91.30
50	37	12	27	23	34.00	80.00	66.00	72.00	100.00	92.00
51	38	12	28	25	33.33	80.39	68.63	72.55	100.00	92.16
52	38	13	29	25	30.77	76.92	69.23	73.08	100.00	92.31
53	38	14	29	26	32.08	77.36	69.81	73.58	100.00	92.45
54	39	14	30	26	31.48	75.93	68.52	74.07	100.00	92.59
69	40	10	31	26	18.84	57.97	52.17	57.97	88.41	81.16

*Results using the ground true groups and their correct matching (all manually annotated).
**A singleton is also treated as a group if output by the group detection as such.

Table 3.4: Comparison of our improved hierarchical method to our previous version presented in [28] using the covariance (COV) as metric.

Method	Num. Participants							
	22	33	37	46	52	53	54	69
Acc. [28]	81.82	48.48	48.65	52.17	46.15	45.28	44.44	26.09
Acc. New	100	81.82	83.78	84.78	76.92	77.36	75.93	57.97

remained after choosing an uneven group-to-group matching (each modality grouped the streams differently) and were obliviate by our previous method.

In contrast, the results between both methods differ significantly as missing data is introduced. These differences are due to a combination of the way the values in the similarity matrices are calculated, and singletons omitted (and so unmatched) after an imperfect group-to-group association. For example, for the case were only 60% of the streams is guaranteed (46 participants) 22 streams have complete data while the rest have different proportions of missing segments. While these segments are omitted while calculating the similarity matrices by our new method, they remained as zeros for our previous version resulting in different values in the similarity matrices **D**, and subsequently generating different values in the matrix of costs **C**. Moreover, the latter can even result in a different group-to-group assignment. Nonetheless, as seen by these results, the improvements done to our hierarchical method account for such cases and maintain the functionality of our method for missing data.

3.5.5. ASSOCIATION VS. SOCIAL CONTEXT

The results obtained so far show that, although the length of the interval, the number of participants and amount of missing data have a significant impact on the accuracy, there are some confusions that cannot be totally explained by the aforementioned and detailedly described parameters.

We hypothesize that such confusions are due to the role of social context and in this section we intend to analyze this aspect further. To do so, we used social actions anno-

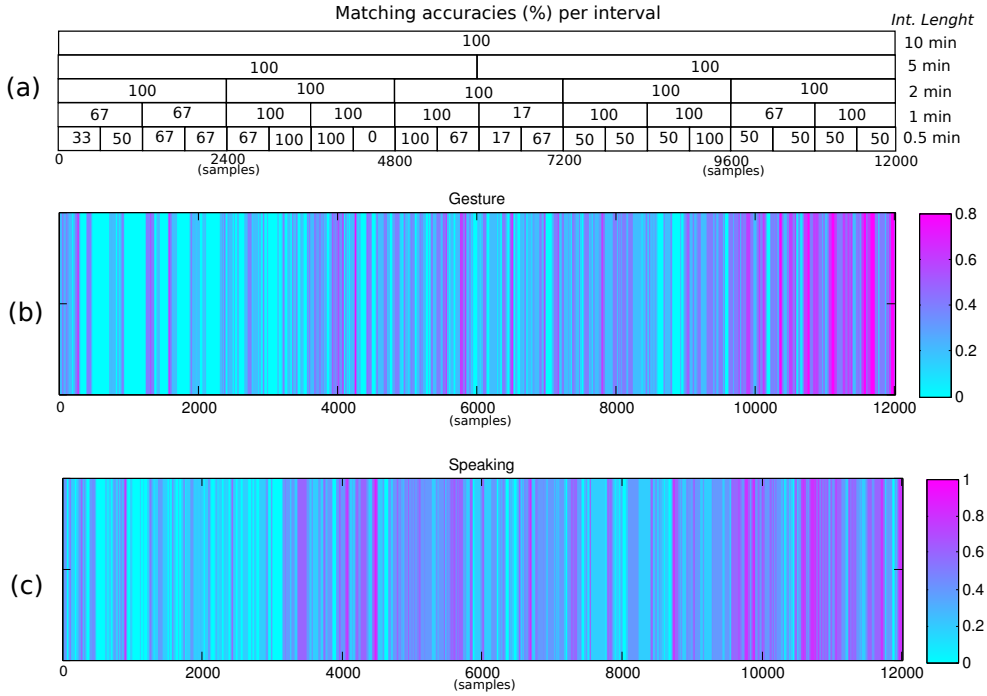


Figure 3.10: Analysis of impact the impact of social actions in the association (better seen in color). (a) Matching accuracies for selected 6 participants under different lenght intervals at different times. (b) Normalized density of *hand gestures* for all participants (1 equals all participants gesturing). (c) Normalized density of *speaking* for all participants (1 equals all participants speaking).

tations provided with the *MatchNMingle dataset* and specify on Section 5.3. From all 8 social actions provided, we focus in hand gestures and speaking which are more related to conversational aspects of the context.

Figure 3.10(a) shows the percentage of correct associations over time for 3 pairs of participants (6 people). These person stayed together for over an 90% of the 10 minute interval, so they have a high number of shared social actions. We selected 6 people as this is within the higher number of people interacting in the same group.

To obtain this figure, we took into consideration different interval lengths over time, so we can see the association performance for these 6 people for different times and observations lengths. Thus, the block in the far bottom right represents the accuracy percentage for the last interval of 0.5 minutes (600 samples) within our 10 minutes. Similarly, the top block represents the accuracy performance using the entire 10 minutes. Moreover, Figures 3.10(b) and 3.10(c) represent the normalized density over time of the actions of hand gestures and speaking, respectively. With these figures, one can see graphically the correlation between social actions and the percentage of mismatches.

It can be seen from Figure 3.10, specifically at the right side of all figures, that when there is a higher density of hand gestures and speaking (which are inherently associated with body movement [68]) the short intervals (bottom blocks of Figure 3.10(a)) present a consistent lower association percentages compared to those where the occurrence of so-

cial actions is relatively low. This implies that shorter intervals with a high concentration shared of social actions become a failure case for our method. This also relates with the trade-off between observation time and number of participants discussed in Section 3.5.2. Nonetheless, Figure 3.10(a) also shows that even these cases can be compensated with longer observation intervals, to allow the method to properly discriminate between people interacting.

Furthermore, when analyzing the mismatches per individual we found that most mistakes are due to people talking to each other. So, for 2 people interacting actively (eg. speaking and gesturing), our method switch their assignment, even within the same group. This also might explain why our hierarchical method is better than a normal Hungarian, as people moving at the same time but in different groups are not considered for an association.

3.6. DISCUSSION

Comparing between distance metrics

As it was seen through Sections 3.5.1, 3.5.2, and 3.5.3; when applying our method without grouping, the performances of the association vary significantly with the metric and assignment used. Mainly, the Mutual Information (MI) performed poorly regardless of the assignment method, the Dynamic Time Warping distance (DTW) was competitive when using the Hungarian approach only, and using the covariance (COV) as metric gave us the best results for both assignment methods (greedy and Hungarian). This summary is better seen in Table 3.2, and Figures 3.6, 3.7, and 3.8.

We hypothesize that the difference between the DTW and the COV lies on the local and global nature of the computation for each metric, respectively. The final goal of the DTW is to warp one stream to the other optimally in time. Thus, the comparisons between the streams are performed locally up to some degree. In contrast, the COV takes into account the stream globally, even computing implicitly the expected values of each entire stream separate and jointly[1]. From Figure 3.7 we can see that the separation between the DTW and COV becomes smaller as the interval length for the observation reduces. For such cases, as the number of samples on each streams reduce, the two metrics start measuring similar distances. This also supports this global versus local hypothesis. This analysis shows that not only the assignment with a global optimization is important. Also, a metric that computes the distances in a global manner is a better option for computing the distance matrices, specially for longer intervals of time. This might also explain why the DTW works only for the Hungarian method (a global optimization) but fails when using the greedy association (local).

A particularly interesting result is the low accuracy achieved when using the Mutual Information (MI) as similarity metric, as it is generally used for synchrony between streams. Even for a real scenario where the signals are more noisy (such as in our entire subset) the covariance and DTW distance are able to cope with the noise up to limit whereas the MI cannot. We hypothesis that this relates to what the MI is measuring in essence. This metric is more complex than just a measurement of similarity and, unlike the covariance and DTW distance, it is designed to also account for those moments of inverse synchrony. Hence, it might not be adequate for the association of the streams.

[1] COV(X,Y)=E[XY]-E[X]E[Y], where X and Y are the 2 streams.

Impact of missing data

Note that as the number of people increases in the subset of Figure 3.8, the number of missing data also increases. Although there is an influence of the number of participants on the accuracy decrease for this table, we believe this is strongly related to the missing data in the sets used as we gave a rather long observation interval. Nonetheless, for a subset of 51 people only present under the FoV of the cameras for 4 minutes (from the total 10 minutes), the normal Hungarian method is still capable of matching correctly 60% of the streams. This can be further improved using our hierarchical method, as we see in Section 3.5.4.

Impact of group-to-group association on final performance As mentioned before, the results presented in Table 3.3 suggests that the association between streams is constantly accurate when applied in a hierarchical manner. Nonetheless, it is the association of groups which then becomes inaccurate. This could be mainly due to two reasons: inaccurate group detection or high number of groups. Firstly, the group detections are still imperfect as it was mentioned before. For example, one modality could detect 3 people as member of a group in the wearables but exclude one of them in video. Thus, better group detection techniques could improve the association performance. Secondly, as the number of groups to match increases the association becomes harder, similar to discriminate larger numbers of streams without any grouping. A practical solution might be to further apply a hierarchy, now grouping the conversational groups into spaces (e.g. rooms). For this, one should know the position of each camera and have a sniffer node for the devices on each space.

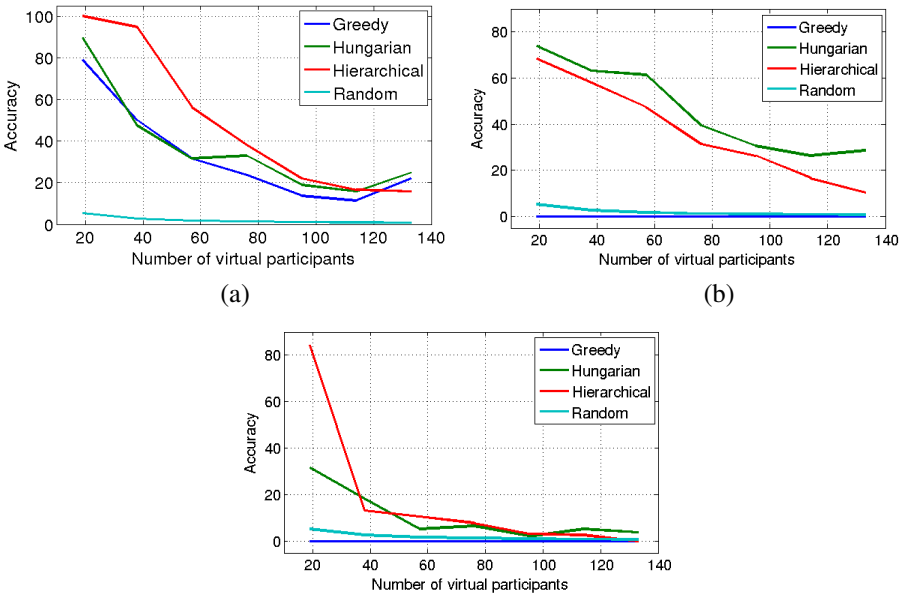


Figure 3.11: Accuracy of the stream association for our previous Hierarchical method, the state-of-the-art (Hungarian and Greedy) and a random baseline (Random) using (a) COV, (b) DTW and (c) MI, using virtual streams.

ADDENDUM

On the previous work presented in [28] we proved our hierarchical apporach using only 19 participants, with no missing data. We do so using our hierarchical method without any modification to account for missing data or uneven streams in both modalities.

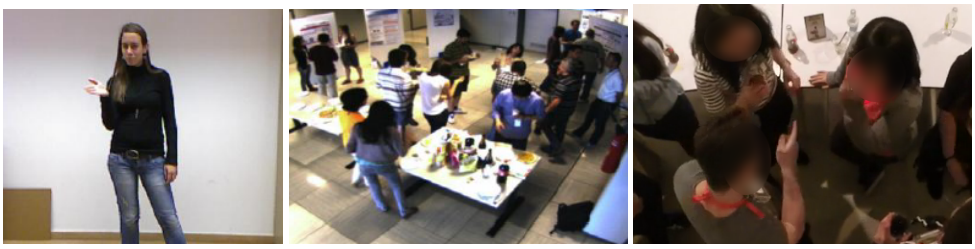
To stress test the number of streams to associate, we used *virtual streams*, similarly to Teixeira et al. [159], by splitting the total time interval of the streams into smaller intervals and treating each one as if they were occurring at the same time. From our 19 original streams, we created subsets with 38, 57, 76, 95, 114 and 133 virtual subjects, increasing significantly the number of streams we could associate. This leads inevitably to a reduction in the length of the streams, which was a trade-off to endure in this submission. Figure 3.11 shows the assignment accuracy against the number of participants for the greedy, Hungarian and hierarchical Hungarian assignment methods, for the 3 metrics.

Although outperforming the other methods, the accuracy for our Hierarchical Hungarian method decreases with a higher number of participants. We hypothesized that as the number of groups increases but the observation time remains low, the method will now have problems to discriminate between the groups instead of streams. The results shown in Figure 3.9 suggest this previous hypothesis to be true, as the curves in this Figure have a smaller decay than those in Figure 3.11. For the experiments of Figure 3.9 the number of participants increases but the observation time stays the same (only with missing data within).

4

DETECTION OF CONVERSATIONAL HAND GESTURES USING MULTIMODAL STREAMS DURING CROWDED MINGLE SCENARIOS

The contents of this chapter are based on the work originally published in:
L. Cabrera-Quiros, D. Tax and H. Hung. **Gestures in the wild: detecting conversational hand gestures in crowded mingle scenes using bags of video trajectories and wearable acceleration**. Submitted to IEEE Transactions on Multimedia and under revision.



(a) (b) (c)

Figure 4.1: Different scenarios for hand gesture detection. (a) Symbolic gesture [60], Conversational gestures: (b) Salsa Dataset [1] and (c) Our scenario.

4.1. INTRODUCTION

Hand Gestures constitute one of the key elements of face to face interactions. As described by Adam Kendon [89](1): *"Willingly or not, humans (...) [communicate] their intentions, interests, feelings and ideas by means of visible bodily action."* During a conversation there is a high probability of observing conversational hand gestures, and their analysis can provide further insights about the interaction itself [92, 110].

However, when it comes to detecting and recognizing gestures, nowadays efforts are not oriented to the gestures that we all perform on a daily basis, which are mostly inherently conversational as described by [89] and [110]. Instead, most works are focused on scenarios where the person is only performing symbolic gestures[1] that are clearly visible. For example, over the last five years the gesture recognition Chalearn challenge [59, 60, 130, 172] has provided datasets of over 40000 videos of one person at a time performing sign language gestures in front of a Kinect (see Figure 4.1(a)). For this challenge, the clips are either trimmed to fit the gesture (isolated gestures) or have only up to 5 symbolic consecutive gestures [130].

Unfortunately, this does not reflect the majority of real life situations where gestures are used. These works, although interesting for other applications (eg. HCI), only addressed a subset of the wide variety of gestures a human can perform, and this subset report a consistent and discriminative pattern [92] (eg. 'hello' in sign language is always the same). Also, they present rather stationary backgrounds with a single subject, without any cross-contamination between subjects.

On the contrary, datasets for the analysis of social interactions *in-the-wild* should maintain ecological validity. For this reason, these types of scenarios have a crowded nature as the people come together to form conversational groups. To better capture these events a top view is preferred (see Figure 4.1(c)). Side or elevated view tend to have higher amounts of occlusions, particularly for those people away from the camera (see Figure 4.1(b)).

The scenario studied in this paper are the *crowded mingle events*, which are scenarios where people are inherently encourage to interact in a real setting (eg. parties). Thus, they provide a perfect example of the use of hand gestures during conversation within a social context in an *in-the-wild* scenario. Nonetheless, we hypothesize that our method can be applied in other cases with related contexts.

From the visual modality perspective, mingle scenarios have four main challenging

[1] Symbolic gestures are those with a specific meaning (eg. thumbs up).

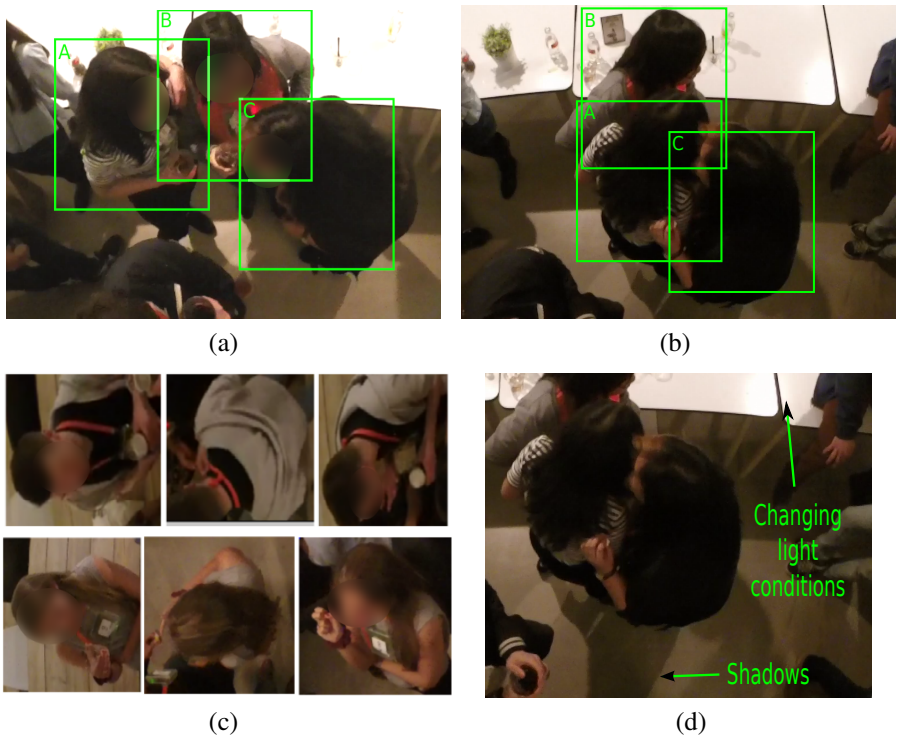


Figure 4.2: Challenges while analyzing mingle scenarios with video (see better in color). (a) Interpersonal cross-contamination, (b) strong interpersonal (A to B) and intrapersonal (A to herself) occlusions, (c) strong appearances variations for two subjects, and (d) Non-stationary background.

differences when compared to the symbolic gesture scenarios: 1) cross-contamination between subjects, where using a bounding box as it is done by most works in object/person detection could include two or more people in one person's box; 2) strong occlusions, making it impossible to see the subjects in some cases; 3) strong changes in appearance for the same subject; and 4) non-stationary backgrounds, which are affected by the position of the subjects, lighting conditions or shadows. A visual depiction of these challenges is shown in Figure 4.2. Thus, the detection of gestures in such scenarios must be addressed considering the presence of noisy data.

Fortunately, previous works on free-standing conversational groups have shown that wearable sensing alternatives can provide additional information when learning with video [2]. Thus, each modality can provide different information to understand the event, relying on one modality when the other is missing or leveraging their complementarity.

In this chapter we detect conversational hand gestures in crowded and strongly occluded scenarios using dense trajectories from video and wearable acceleration. We do so using the MatchNMingle dataset (see Chapter 2). To the best of our knowledge, thus far we are the first to address the problem of gesture detection during crowded scenes.

Overall, we leverage the use of video trajectories and wearable sensors to detect gestures in crowded scenes in a window-based approach. Thus, our method identifies if a time interval (window) of length w contains a gesture or not. To do so, it uses as inputs 1) RGB

video and 2) the triaxial acceleration on the wearable device of each participant.

The main contributions of this chapter are: (i) unlike previous works, we addressed the detection of conversational hand gesture *in-the-wild*, using a dataset collected during a real mingle event with strong interpersonal occlusions, (ii) we propose to use a multiple instance learning approach (MILES [38]) representing gestures as bags of trajectories to overcome subject cross-contamination and to become robust against noisy backgrounds in video, (iii) we leverage the MILES instance classification capability to analyze which dense trajectories (time and space) are more representative for a gesture in video, (iv) we combine video and wearable acceleration in a decision-level manner leveraging the complementarity between modalities, particularly for cases where occlusions in video are too strong to have a clear view of the person performing the gesture, showing improvements over unimodal approaches; and finally (v) we analyze the impact of noisy data (eg. strong occlusions) of the participants on the overall performance, both static and dynamically in time.

4.2. RELATED WORK

Most work on the detection and recognition of gestures focused on cases where there is a clear view of the person performing a symbolic gesture, generally from the front. This also includes works for Human Computer Interaction (HCI).

In general, for all these works the process is quite similar: 1) pose estimation or use of its skeleton if available, and 2) gesture detection or recognition. Some works also include a previous hand segmentation step in this pipeline. Overall, these steps are possible for these works as their view of the person is rather clear. However, this condition does not apply for crowded settings. Firstly, the pose estimation is rather complicated due to strong changes in appearance (see Figure 4.2(c)). These works, although relevant for its specific goal (eg. using gestures to interact with a computer), do not address the same inherent problem as our work and most do not share the same challenges (eg. strong occlusions due to natural interactions). For a comprehensive review of the domain, please refer to any of the past Chalearn Challenges about gesture recognition [59, 60, 130, 172] and to [136] for a review on gesture recognition for HCI applications.

Here we focus instead on related works with scenarios similar to ours, where there is a conversation between 2 or more people and the goal is to detect or recognize gestures within a social context. As an example, Xiong et. al. [181] and Xiong and Quek [180] presented their work on the analysis of gestures during conversations, the first for the analysis of the symmetric behavior of gestures and the second for their frequency. Although oriented to conversational gestures (as ours), these works are based on the dataset collected by Quek et. al. [132], which has a rather clear side view of the speaker and only 30 seconds of video.

Similarly, Marco-Ramiro et. al. [103, 104] addressed the detection of conversational gestures during seated encounters. The first work focused on the detection of upper body monocular motion (including hands), while the latter used such features to look for adaptors (eg. unintentional gestures, generally performed while fidgety) or beat gestures in the context of an interview. Also, Cerekovic et. al [36] detected the rapport between people and virtual agents using as one of their features the hand gesture activity of the people while interacting with the agent.

Unlike other works, these latter efforts addressed the wide range of human gestures during conversations. Nonetheless, as the scenarios permitted it (seated interview) they used

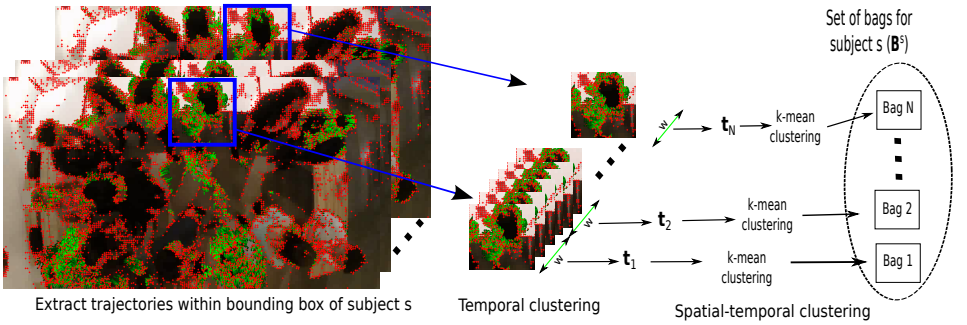


Figure 4.3: Process of clustering in space and time to create bags of trajectories.

a rather clear front view of the participants while interacting, so they do not present the same additional challenges regarding the visual perspective (eg. subject cross-contamination) as our mingle scenario.

To the best of our knowledge, we are thus far the first to address the problem of gesture detection during crowded free-standing conversational scenes, emphasizing the importance of the social context of the gestures and its impact in the challenges for the detection of gestures.

4.3. OUR APPROACH

As mentioned before, our method identifies if a time interval (window) of length w contains a gesture or not. To do so, it uses as inputs 1) RGB video and 2) the triaxial acceleration on the wearable device of each participant.

Next, we explain the process of fusing both modalities (Section 4.3.3) in the decision level by using the posterior probabilities of two unimodal classifiers (one per modality). Each of these classifiers is also explained in detail. Also, we explain the process of extracting and clustering dense trajectories from video to be treated as bags, and the multiple instance learning method applied for their classification (Section 4.3.1); and the extraction of features and classification from the wearable devices (Section 4.3.2)

4.3.1. VIDEO CLASSIFICATION

Figure 4.3 summarizes the process of clustering trajectories in space and time to create our *bags of trajectories* in video for each subject. This process consists of the following steps:

Extraction of Dense Trajectories

Firstly, we extract our trajectories employing the method of dense trajectories proposed by Wang et al. [175]. These have proven to be an efficient representation for human activity recognition [165, 175, 176]. The dense trajectories are extracted for the entire frame using a length L of 20 frames (default).

Using the bounding boxes for each participant on each frame we create a voxel (as seen in Figure 4.4) following the participant over time. Thus, we reduce the number of trajectories to those around or from each participant by selecting only those inside this

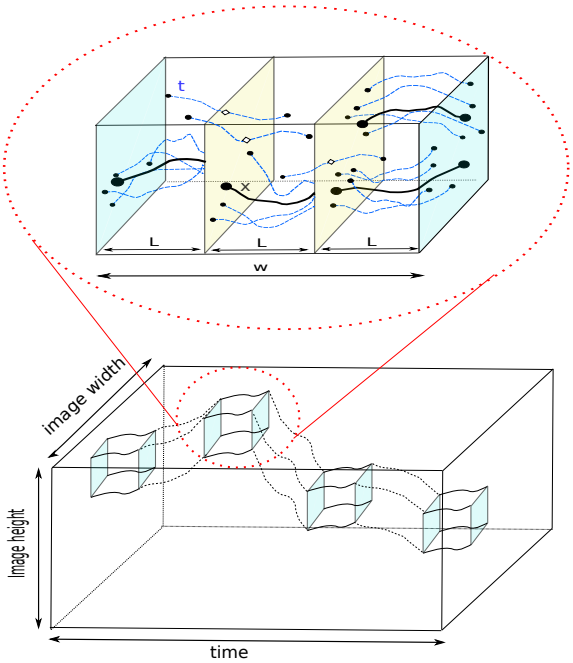


Figure 4.4: Clustering of trajectories for a bag. Dense trajectories (in blue) are extracted for a sliding window of size w and clustered in time-space into prototypes (black trajectories). These prototypes become the instances of our bags.

voxel. This selection also accounts for trajectories that start outside of the voxel but enter it, and those that start within the voxel but drift out.

For bounding box extraction one can use any existing tool for this purpose [140, 156], however we use ground truth annotations to avoid further contaminations. We leave the analysis of the impact of the people detection error on the overall gesture detection for future research.

Notice that, due to the crowdedness of the scene, bounding boxes of different subjects can heavily overlap. A bounding box can therefore contain trajectories of both the subject of interest and of 'background' subjects. Fortunately, our multiple instance learning (MIL) approach will account for this duality.

We then have a set of trajectories for the voxels of each subject, which are then clustered in time to create our bags using **sliding windows**.

Selection of trajectories for a bag

First, let us define $\mathbf{B}^s$ as the set of bags (positive and negative) created using the bounding boxes for subject s, where $s = \{1, ..., S\}$ and S is the total number of subjects. A bag from this set is then $\mathbf{B}^s_j$, where $j = \{1..., N^s\}$, and N^s is the total number of bags possible for subject s.

For a given window $\mathbf{B}^s_j$ we 1) select the trajectories corresponding to this bag (temporal clustering), and 2) cluster the trajectories within a bag (spatio-temporal clustering). The latter will be explained in the next subsection.

The set of trajectories $\mathbf{t}_j^s$ corresponding to bag $\mathbf{B}_j^s$ is then determined by the position of the sliding window in time, as seen in both Figures 4.3 and 4.4, and its size w. $\mathbf{t}_j^s$ are to those trajectories within the voxel for participant s (previously selected from the entire set) that have at least an 80% overlap in time with the window used for bag j. These trajectories are represented in blue in Figure 4.4.

Note that the sliding window does not necessarily have to be the same length L as the trajectories (as presented in Figure 4.4), or the same shift. Also, the trajectories can start at any point within the sliding window, but will always have a size L. This size is fixed to L to avoid drift, as explained by Wang et al. [175]. So, if $w > L$ there will be trajectories in the bag that are only partially within the window.

It is important to emphasize that although the bounding boxes for each subject s were used, **the trajectories inside these boxes do not necessarily belong to subject s**. Instead, they could also represent the background or other subjects. This is the main motivation for using our MIL approach (more in Section 4.3.1).

Clustering of trajectories within a bag

A final clustering is important to create less noisy trajectories and more representative trajectory prototypes. For practical purposes, it also results in a more efficient memory usage, without losing information. This is a common practice in works using dense trajectories [165, 176], as these descriptors tend to be redundant for similar local-temporal instances.

To cluster the trajectories within a bag, we use k-means clustering. This way, the trajectories for each bag $\mathbf{t}_j^s$ are clustered into the k most representative prototypes for the bag ($\mathbf{x}_j^s$). These prototypes are represented in Figure 4.4 in bold, and become the instances of our bags ($\mathbf{B}_j^s = \{\mathbf{x}_j^s\}$).

The label of each bag (y_j^s) is set using the annotations provided by the dataset, which are made every frame. To select a single value we use majority voting.

Finally, we create the set of bags $\mathbf{B}^s$ (positive and negative) for subject s by applying the procedure described above for a window, then sliding it and repeating for the entire video, as seen in Figure 4.4. Thus, each window becomes a bag.

MULTIPLE INSTANCE LEARNING

As stated before, due to the crowded nature of our scenes we opt for a multiple instance learning approach using bags of trajectories. Thus, each bag $\mathbf{B}_j^s$ will consist of *good* trajectories (corresponding to subject s) and *bad or noise* trajectories (other subjects or shadows and other background artifacts).

As our Multiple Instance Learning (MIL) approach we use Multiple Instance Learning via Embedded Instance Selection (MILES)[38]. Overall, MILES classifies a bag by considering both contributing information (e.g. trajectories of subject s in our case) and opposing information (e.g. trajectories from other subjects or background). It does so by creating a *concept* in an embedded space and comparing all instances to this concept. Instances close to the concept will have a higher contribution (see more on the explanation of Eq. 4.3).

Thus, unlike other MIL approaches where at least one positive instance in a bag automatically converts it into a positive bag, MILES does not have this restriction. This also allows us to assess the role of individual instances in the classification of a bag (see Section

4.4.2 for an analysis about this).

More specifically, MILES maps each bag into a feature space (IF_c) defined by the instances in the training set using bag to instance similarities. The bags are then classify in this space, depending on how close the instances within the bag are to the concept defined by the instances in the training.

Let us define $\mathbf{B} = \{\mathbf{B}^1, \mathbf{B}^2, ..., \mathbf{B}^S\}$, as the set of bags for all participants. $\mathbf{B}_a$ is then a bag of this set $\mathbf{B}$, where $a = \{1..., A\}$ and A is the sum of the total number of bags for all S subjects.

For a given bag $\mathbf{B}_a$ the measure of similarity between this bag and all other instances in the training set[2] (disregarding their bag) is calculated by

$$s(\mathbf{x}^k, \mathbf{B}_a) = \max_b \exp\left(-\frac{||x_{ab} - \mathbf{x}^k||^2}{\sigma^2}\right) \tag{4.1}$$

where $\mathbf{x}^k$ is the set of instances in the training and x_{ab} is a given instance b within bag $\mathbf{B}_a$. Thus, bag $\mathbf{B}_a$ is embedded into a space of similarities defined as

$$\mathbf{m}(\mathbf{B}_a) = [s(\mathbf{x}^1, \mathbf{B}_a), s(\mathbf{x}^2, \mathbf{B}_a), ..., s(\mathbf{x}^{n_a}, \mathbf{B}_a)]^T \tag{4.2}$$

where n_a is the total number of instances in the training set. This results in the matrix representation of all training bags in the embedded space (IF_c) : $\mathbf{m}(\mathbf{B}) = [\mathbf{m}(\mathbf{B}_1), ..., \mathbf{m}(\mathbf{B}_A)]$.

The creation of this distance matrix, which is directly dependent on the number of bags (A) and instances in the training set (n_a), can be memory consuming. This is one of the main reasons behind the spatio-temporal clustering within the bags using k-means.

On this representation a (sparse) linear classifier is then trained. The classification of new bags is done by:

$$y = \text{sign}(\sum_{k \in I} w_k^* s(\mathbf{x}^k, \mathbf{B}_{new}) + b^*) \tag{4.3}$$

where I is the subset of instances with non-zero weights ($I = \{k : |w_k^*| > 0\}$). Note that instances with contributing information will have positive weights w_k^*, while those with opposing information will have negative weights. For more details, please refer to [38].

4.3.2. WEARABLE ACCELERATION CLASSIFICATION

Each wearable device (one per subject) recorded the triaxial acceleration at 20Hz. For each, we also calculate the magnitude of the acceleration ($|accel| = \sqrt{x^2 + y^2 + z^2}$); resulting in 4 different time series for which we can extract features using a sliding-window approach, similarly to the video.

Then, for each window we extract features that have proven to be efficient to analyze human actions from wearable acceleration [68]. These features are mainly statistical and spectral, where the statistical features focused on mean and variances from each axis and from the magnitude, and spectral using the power spectral density (PSD). All features are concatenated to obtain a 70-dimensional feature vector per window, and then classified using a logistic regressor (see Section 4.4.1 for details). To ensure that each of these windows

[2]The separation of $\mathbf{B}$ into train and test set is addressed in Section 5.5.

Table 4.1: Summary of results using unimodal classifiers and their fusion in a decision-level. Mean AUC (± deviation) of folds.

Classifier	Wearable Device	Video (Baseline)	Video (MILES)	Fusion
AUC	0.65 ± 0.08	0.61 ± 0.07	0.67 ± 0.09	$\mathbf{0.69 \pm 0.10}$

can be directly compared to the bags in the visual classification for the decision-level fusion, they were extracted for the same time intervals. This will also be important for the decision fusion described in the next subsection.

4.3.3. DECISION FUSION CLASSIFIER

Both unimodal classifiers (video and wearable) provide a posterior probability, which can be used in a decision fusion classifier. This classifier has only 2 features, which also makes it suitable for direct visualizations using scatter plots.

We opt for a decision fusion instead of a early fusion approach as we aim to maintain a constant (and fair) feature space. Thus, while MILES will map each bag to a embedded space (IF_c) defined by its instances, this can not be applied to the features from the wearable acceleration.

As the MILES classifier bases its output on the sum in Equation 4.3 instead of a proper probability, we applied Platt scaling [128] to obtain the probability distribution of the classifier from video, as is generally done for these type of classifiers such as SVM.

4.4. EXPERIMENTAL RESULTS

We now proceed to evaluate our classifiers, both separately and combined using decision fusion. A summary of the results is presented in Table 4.1. A detailed explanation of each classifier is now presented.

4.4.1. WEARABLE ACCELERATION CLASSIFICATION

We selected a window size (w) of 60 samples (3 seconds) with no overlap for the classification using the wearable acceleration. Empirical tests, performed always separating training and test data, shown that these values are optimal for our task. For these windows, we extracted for each participant in our dataset (70 in total) the features described in Section 4.3.2. As a classifier we selected a linear logistic regressor and used a leave-one-subject-out cross-validation strategy.

We obtained a mean AUC of 0.65 ± 0.08 for the 10 minute interval. This result is similar to what has been found in the past for the detection of other social actions using wearable acceleration [68].

4.4.2. VIDEO CLASSIFICATION

For each participant we extracted their set of bags of trajectories $\mathbf{B}^s$ following the process described in Section 4.3.1. Same as for the wearable acceleration, we selected a window size (w) of 60 samples (3 seconds) and no overlap. Hence, for a segment of 10 minutes we obtained a maximum of 200 bags per participant. Some of the participants had less bags, as they left the field of view of the camera for different intervals of time during this interval.

Then, we proceeded to evaluate our MILES approach using leave-one-subject-out. However, this procedure results in a training set of around 13,500 bags with a total of around 270,000 instances per training fol (exact number depends on the subjects), even when applying the k-means clustering described in Section 4.3.1. As consequence, the creation of the matrix of distance $\mathbf{m}(\mathbf{B})$ for the MILES becomes significantly consuming in terms of memory and computing time.

To overcome this issue, we implemented an efficient training (in practical terms) where we randomly sample the training set so the optimization of $\mathbf{m}(\mathbf{B})$ is manageable, while maintaining samples from all subjects and enough information for classifier regularization. To find the optimal value for this trade-off (enough samples versus memory limitations) we train with a maximum number of bags ranging from 100 to 5,000. This experiment showed us that at 1,500 bags the results start saturating to a maximum and adding more samples has no evident benefit. Hence, we chose this value for all our following experiments.

Also, our data has a strong imbalance between positive and negative bags for most subjects, resulting in a strongly imbalanced training set. To overcome this we do the sampling in a stratified manner. In contrast, for the test set we used the entire set of bags $\mathbf{B}^s$ for the subject left out.

We used the AUC as evaluation metric instead of the accuracy to account for the imbalance in our samples. For the classification we used the MILES implementation in PRTools [57]. Applying this methodology we obtained a mean AUC of 0.67 ± 0.09 with 68 subjects (2 had issues for the segment selected so were discarded). In addition, using the same training samples given to the MILES, we trained a Fisher classifier which is used as a baseline comparison for the video.

ANALYSIS OF BAG INSTANCES

We are interested in analyzing qualitatively which instances in the bags, which correspond to specific trajectory prototypes, contributed the most in the MILES classifier. Our intention is to assess whether the instances chosen by the classifier correspond in fact to trajectories of the correct subject, and that trajectories for changes in the background or other subjects are ignored.

For this, we leveraged the instance classification capacity of the MILES algorithm. Thus, for the classification of a given new bag $\mathbf{B}_i$ with instances $\mathbf{x}_{ij}$, $j = 1, ..., n_i$ (n_i equal to the total number of instances in the bag), we can define which instances contributed the most for the classification of the bag. To measure this contribution level of the instances we use the following weight:

$$g(\mathbf{x}_{ij^*}) = \sum_{k \in I_{j^*}} \frac{w_k^* s(\mathbf{x}^k, \mathbf{x}_{ij^*})}{m_k} \tag{4.4}$$

where $\mathbf{x}^k$ corresponds to the instances in the training set, I_{j^*} corresponds to the subset of all instances in the bag for which there is a maximum similarity with one of the instances in the training set ($\exp(||\mathbf{x}_{ij} - \mathbf{x}^k||^2/\sigma^2)$) and whose weights are $|w_k^*| > 0$; and m_k is the number of instances in I_{j^*}. Hence, Eq. 4.4 determines the contribution of $\mathbf{x}_{ij^*}$ on the classification of the bag $\mathbf{B}_i$. For more details, please refer to [38].

Figure 4.5 shows example cases of instance importance used by the MILES classifier. From left to right we have the original image, the original dense trajectories, and those

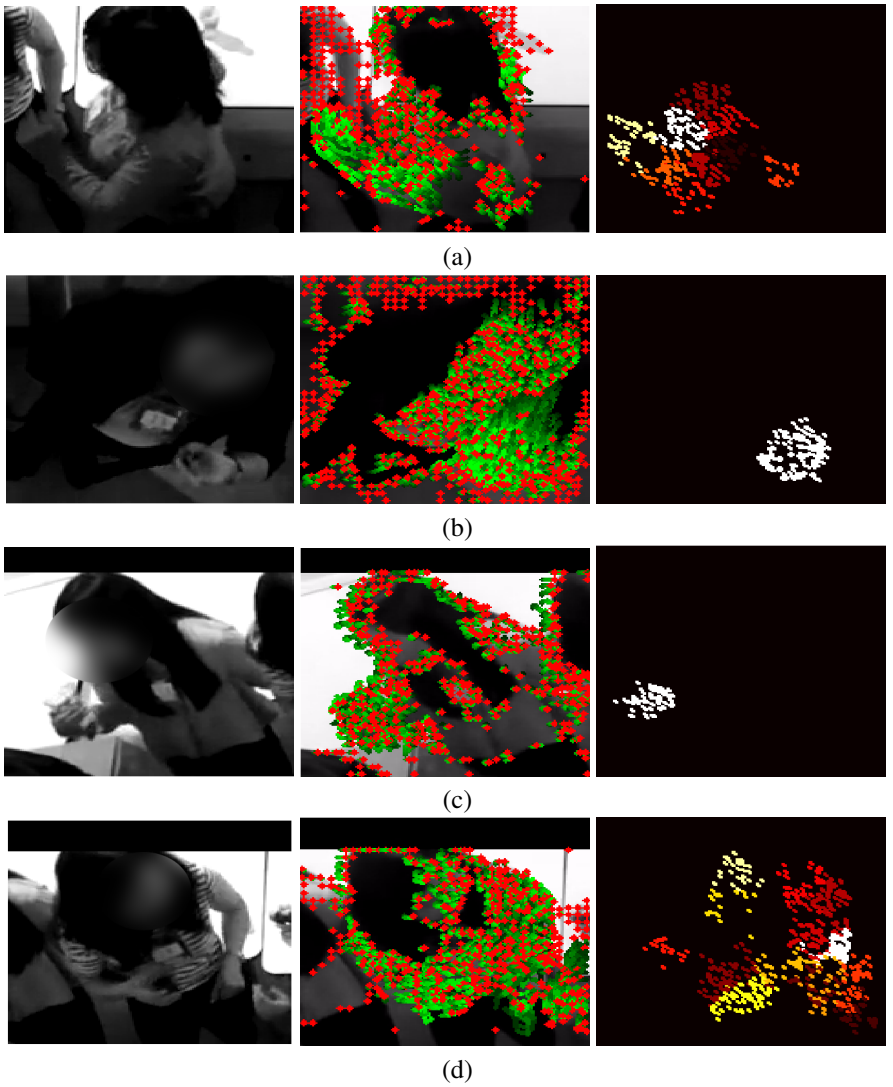


Figure 4.5: Analysis of instances with higher contribution in the MILES classifier. From left to right: Original image, original dense trajectories, trajectories considered as of high contribution. Different colors in right images represent different clusters (see section 4.3.1). (a) MILES focusing on trajectories corresponding to the hand/arm area (b) Background trajectories (eg. shadows) beign omitted. (c) MILES handling cross-contamination of subjects. (d) Failure case, MILES also considers gesture from another participant.

trajectories considered by the MILES classifier to be of high contribution following Eq. 4.4. To be considered as 'high', the absolute value of the instance contribution had to be higher than 0.0001 (to filter noise). Different colors in the images showing trajectories with high levels of contribution (right images) represent different clusters. Thus, if a prototype trajectory (see Section 4.4.2) was selected as having a high contribution, all the trajectories in that cluster are shown.

First, in Figure 4.5(a) we can see an example of MILES having high contribution levels for trajectories corresponding to the arm an hand region of the person while a gesture is performed. This shows that the MILES classifier is learning the correct representations. Figure 4.5(b) shows that background trajectories are ignored while trajectories of the subject's hands are important.

Figure 4.5(c) shows the case where the MILES approach handles cross-contamination of subjects, assigning high contributions to the trajectories belonging to the subject owning the bounding box and no contribution to the trajectories from to the person causing the cross-contamination.

Finally, Figure 4.5(d) shows a failure case where trajectories corresponding to another subject (cross-contamination) are given high contribution. This case in particular was interesting to observe, as during those segments the two subjects were engaged in a conversation, and the subject causing the cross-contamination was also gesturing. Thus, the MILES learns correctly that these trajectories were corresponding to a gesture but fails to discriminate the subject observed.

4.4.3. DECISION FUSION

As stated before, to leverage the complementarity of both modalities we performed a decision level fusion using the posterior probabilities of both unimodal classifiers. We chose decision fusion as the bags in MILES are embedded into a different space than the features from the acceleration, so an early fusion approach is not appropriate.

Similar to the experiments in the past sections, we applied a leave-one-subject-out cross-validation, this time using Fisher's linear classifier to avoid overfitting.

This experiment gave us a mean AUC of 0.69 ± 0.10. So, as we hypothesized, the use of the complementarity between modalities increases the performance of the detection while compared to the unimodal approaches (0.67 and 0.65 for video and acceleration, respectively).

Nonetheless, as seen in the deviation, there is still a high variability between the participants' results. Thus, we now proceed to analyze the levels of noisiness of the data, due to the cross-contamination between subjects and their occlusions, as a possible cause. This will be done in a static (in time) and dynamically manner.

STATIC ANALYSIS OF NOISY DATA IN VIDEO

We first aimed to analyze the impact of the noisy data using a static comparison. Thus, we separated our set of subjects in 2 subsets: 1) the *clean* and 2) the *noisy* subjects. To select the subjects belonging to the *clean* subset we computed the overlapping ratio on each frame for all subjects and then computed their mean over time. For the visibility constraint, we used the annotations given by the dataset for *out of view*. We also added as *out of view* those moments for which the subject is too close to the borders of the frame and is only partially visible. A subject is part of the *clean* subset if their mean overlapping ratio is lower than 0.4 and is visible for 60% of the time. On the contrary, will be part of the *noisy* subset.

Table 4.2 summarizes the results while training with one of the subsets (or the entire set) and testing with another, using a leave-one-subject-out cross-validation. When training for a given subset we only used those subjects within the subset; leaving out the subject only if this is part of the subset selected for testing. For this table, the first row corresponds

Table 4.2: Analysis of the impact of data complexity in gesture detection. Mean AUC (± deviation) using leave-one-subject-out and different subsets for training and testing.

Train (sub)set	Test (sub)set	Mean AUC ± std		
		Video	Wearable	Fusion
Entire	Entire	0.67 ± 0.09	0.65 ± 0.08	0.69 ± 0.10
Entire	Clean	0.67 ± 0.10	0.63 ± 0.07	0.68 ± 0.11
Entire	Noisy	0.69 ± 0.08	0.66 ± 0.09	**0.71 ± 0.09**
Clean	Entire	0.68 ± 0.09	0.65 ± 0.09	0.66 ± 0.10
Clean	Clean	0.66 ± 0.10	0.67 ± 0.09	0.66 ± 0.12
Clean	Noisy	0.70 ± 0.08	0.63 ± 0.11	0.66 ± 0.07
Noisy	Entire	0.68 ± 0.09	0.66 ± 0.10	0.69 ± 0.10
Noisy	Clean	0.67 ± 0.10	0.66 ± 0.11	0.68 ± 0.11
Noisy	Noisy	0.69 ± 0.08	0.70 ± 0.07	**0.71 ± 0.09**

to the results presented in the previous subsections.

First, we can see in these results that the fusion column has in most cases a higher performance than the video classification. The exception are the experiments trained with the clean subset which suggests that the information added by the wearable devices in this case is redundant as the MILES has learned clear examples of gestures, and the information from the wearable is redundant. This might also be the reason of the performance for the training with the clean subset and the test with the noisy one.

Also, note that overall the classifier using the wearable acceleration information performs similarly to the video. And these values are also below the fusion except for the clean subset as training. This suggests, as that the wearable devices are not affected by the vision noisy states, and instead different factors cause the differences between the subsets. We hypothesize that these difference might be due to interpersonal differences, as suggested by [68].

DYNAMIC ANALYSIS OF NOISY DATA IN VIDEO

Our aim with such comparison is to determine if exits a correlation between those moments with high occlusion between participants (*noisy states*) and the confidence of the MILES classifier.

Similar to the static analysis, for each participant we calculated the overlapping ratio at each frame. In addition, we computed the distance of each person's bounding box center to the closest image border and normalize it (border = 1, center of image = 0). The sum of these two on each frame give us a *ratio of occlusion* for each participant in time. Note that this ratio is not normalized as you can have a person in the border and occluded, for which ratio of occlusion > 1.

Figure 4.6 shows the error analysis in time for a subject chosen at random. Aside from the ratio of occlusion, these plots show the confidence of each classifier and the error for the MILES and the decision classifier. The confidences are calculated with the normalized absolute distance of the sample to the hyperplane for the logistic regressor, and with the normalized sum of weights for the bag for the MILES (see Equation 4.3).

First, in Figure 4.6 we can see how for those time with high occlusion ratio the MILES classifier makes error while the decision fusion compensates for these. In these intervals, although MILES is having trouble with confidence due to the noisy state, the wearable acceleration based classifier maintains or increases its confidence and allows the fusion classifier to correctly classify such moments.

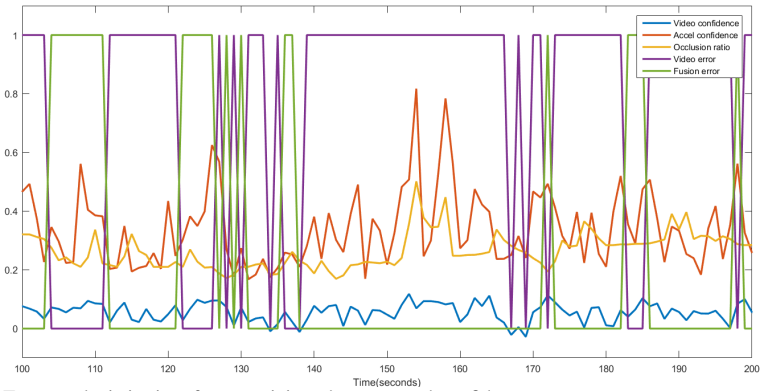


Figure 4.6: Error analysis in time for a participant's errors and confidences.

4.5. DISCUSSION

Perhaps one of the most interesting findings of this chapter was the dynamic (in time) complementarity of the modalities when one of them has a low confidence. As saw in Figure 4.6, when the occlusion ratio of the video was high, meaning that person was severely occluded, the confidence of that classifier was relatively low compared to the wearable acceleration classifier. For those segments the video classifier tend to make errors, which is expected due to the high values of occlusion for that person. Nevertheless, the fused classifier overcomes for those moments leveraging the information and confidence of both unimodal classifiers.

Note also how the confidence of the wearable acceleration classifier also is slightly affected by the occlusion periods, although not as much as the video. We hypothesize that this might be due to the moment of the interaction that is causing the occlusion. Thus, although does not affect in such as extreme way the wearable data as the video, it causes a certain effect nonetheless (e.g. person moving to let someone else pass through during a crowded scene).

Another interesting finding comes with the analysis of the instances which have high contribution for the MILES. Specifically, for the failure cases. As saw in Figure 4.5(d), some instances of this example were assigned as of high contribution to the MILES even when they belong to trajectories of a different participant.

Nonetheless, we notice that most of these failure cases in fact corresponded to moments were the bounding box captures 2 or more people while strongly engaged. Thus, the high contributions attributed to a different person occasioning the cross-contamination are either due also to gesturing or as a reaction in a different manner (e.g. head nodding) to the other person.

This suggests that the MILES learns a representation of gesture but fails to discriminate the correct participant, or further extrapolates to movements and gestures that are close related to gesturing, such as a synchrony reaction to what the other person said.

5

ESTIMATION OF SELF-ASSESSED PERSONALITY USING MULTIMODAL STREAMS DURING CROWDED MINGLE SCENARIOS

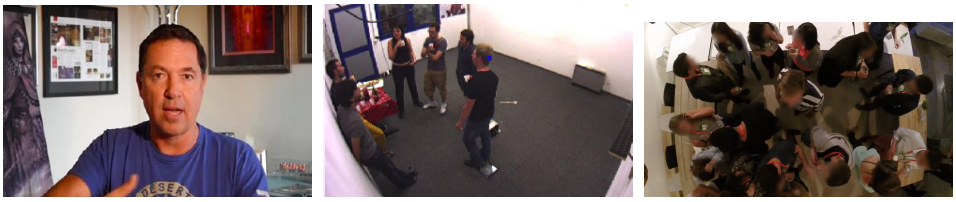
The contents of this chapter are based on the work originally published in:
L. Cabrera-Quiros, E. Gedik and H. Hung. **Estimating self-assessed personality from body movements and proximity in crowded mingling scenarios**. Proceedings of the ACM International Conference on Multimodal Interactions (ICMI), 2016.
L. Cabrera-Quiros and H. Hung. **Multimodal self-assessed personality estimation during crowded mingle scenarios using wearables devices and cameras**. Submitted to IEEE Transactions on Affective Computing and under revision.

5.1. INTRODUCTION

Different modalities have been used to analyze and estimate personality traits, with audio-visual approaches being predominant among the works [168]. In addition, the estimation of such traits has been addressed in different types of scenarios including meetings ([65, 126]), video logs (VLOGS) or self-presentations ([18, 21, 130]), radio broadcasts ([114, 115]) or social media ([35]), among other situations.

Nonetheless, most of the aforementioned efforts tend to share the same characteristics: 1) data of a single person can be easily differentiated from the rest, and 2) they do not have much missing data. For example, works using VLOGS (such as the Chalearn challenge [130]) have a clear view of the participant's faces and unique speech. In contrast, other scenarios do not allow the acquisition of clear, personalized and high quality data without specialized equipment.

One of such scenarios are mingle events. These are intriguing scenarios from the social signal analysis perspective [169] due to their dynamic nature and also comprise a wide range of social interactions and the formation of free-standing conversational groups which also triggers research in group dynamics [1, 106].



(a) (b) (c)

Figure 5.1: Example snapshots of typical scenarios for personality estimation. (a) VLOG taken from the Chalearn Challenge 2016 [130], (b) mingle event taken from [182], (c) a more crowded mingle event (our dataset, more in Chapter 2).

In this Chapter, we focus in the estimation of *self-assessed* personality traits from the HEXACO inventory [4] during crowded mingling events using wearable sensors and video cameras in a noninvasive manner.

Compared to other scenarios where the estimation of personality has been addressed, crowded mingle events are harder to analyze using audio-visual modalities. For example, during meetings or VLOGS settings the audio and frontal video for each participant is generally recorded separately, as can be seen in Figure 5.1(a). Hence, for these scenarios the camera has a clear view of a single participants' face and its speech is unique or can be robustly separated, providing rather clean data from these 2 modalities.

In contrast, mingle scenarios are crowded events where obtaining clean data from computer vision techniques is hard due to occlusion problems, changing light conditions and challenges with people re-identification. In addition, mingle scenarios present ambient noise due to the event itself that makes harder to record good quality audio for each person without customized equipment (eg. microphones).

We focus on the estimation of personality traits a during mingle scenario, leveraging wearable devices, sensing acceleration and proximity, and video cameras recording the event from above (see Figure 5.1(c) as an example). Using these types of sensors also

allows our method to be unobtrusive and to scale rather easily to a higher number of people. Furthermore, we focus on a crowded scenario, including up to 56 people freely interacting for 30 minutes.

The main contributions of this chapter are: 1) we leverage the use of wearable devices and overhead video cameras to estimate self-assessed personality traits during a crowded mingle event, 2) we compare the impact of different modality types on the estimation of the different personality traits as we hypothesize that each modality captures the event differently, 3) we study the impact of fusing different modality types in the estimation performance of each trait and, 4) we analyze the impact of the speaker detection in the overall performance of personality estimation.

While a previous version of this work was presented in [27], it only contemplated the modalities from the wearable devices, and a set of 69 participant (from now on called *wearable-only* set) was used. See more in the Addendum of this chapter for our previous work. To further extend it, the video modality was later included and comparisons made against the modalities from the wearable devices (eg. acceleration, proximity and speaking status). In addition, we analyze the correlation between feature types, and the impact of a speech detection stage and the visibility of the participants in the cameras (missing data) on the classification performance. To do so, we use a subset of 56 subjects (*wearable+video* subset), which have both video and wearable data (see Section 5.3.1).

5.2. RELATED WORK

As the amount of works on personality analysis and estimation is extensive, we only focus on works estimating *self-assessed* personality (meaning that the participants filled an inventory/survey to score in their own personality traits). Nonetheles, many efforts have been made in automated third-party attribution-based personality recognition (or impression of personality), or focused on personality estimation in social media, which are beyond the scope of this paper. A comprehensive review of the related personality computing literature can be found in [168]. Also, specific workshops and challenges (also using impressions) such as the Mapping Personality Traits Challenge and Workshop (MAPTRAITS) [34] or the Chalearn Looking at People Challenge [130] have encouraged researchers in the automatic analysis of personality.

Within the domain of automated self-assessed personality estimation, works can be grouped mainly in small meetings and mingle scenarios. First, an example of the meeting setting, Pianesi et al. [126] 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. This work was extended by Lepri et al. [98]. Both works used the corpus which was first introduced by Mana et al. [102].

Closely to a meeting setup, Batrinca et al. [18] presented a method to analyze 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 part of the call while the main segment for non-verbal cue extraction was a monologue.

Thus, the closest work to our own was presented by Zen et al. [182], were they proposed a classification method to recognize Extraversion and Neuroticism (from the Big-

Five) using proximity related features extracted from multiple cameras in a considerably less crowded mingle event than ours (see Figure 5.1(b)). 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 56 people. Also, their proximity features are based on distances while ours rely on binary neighbor detection (see Section 5.4.1).

5.3. MINGLE DATA

For this section we use the version 1 of *MatchNMingle* [26] (see Chapter 2). During 3 different speed dating events, each followed by a mingle session, between 30 and 32 different people interacted freely, with a total of 92 participants for the 3 events. From these, 69 have functional devices and became our *wearable-only* set. Only 56 have both video and wearable information, becoming our *wearable+video* subset (detailed explanation of why in Section 5.3.1). A 30 minutes segment from the mingle was selected where the number of people interacting was maximized.

As can be recalled, 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 [97]. These will be used as labels.

5.3.1. SUBSET DUE TO CAMERA LOW VISIBILITY

As stated in Chapter 2, only 2 of the 5 cameras have annotations (including positions) for the entire 30 minutes in the version 1 of the dataset. Hence, those participants outside the field of view of the camera are treated as *not visible*. Also, due to the dynamic nature of the event itself, some participants are not captured by any of the cameras at some points (eg. going to the bathroom).

We analyzed the video data to extract a subset of participants that allows a fair comparison between modalities (eg. participants with video and wearable data). For this subset, all participants should be under the FoV of one of the cameras for at least half of the time (15 minutes). This time is not necessarily continuous. Hence, we ensured that there is a representative amount of data for each participants' video, even with missing data. Thus, we are left with the final *wearable+video* subset of **56 participants** that have both a working device and are visible at least 50% of the time.

5.4. EXTRACTION OF NON-VERBAL CUES

Firstly, a summary of all our non-verbal cues is shown in Table 5.1, separated by the modality type and the sensor they come from (wearable or camera). Thus, from the 3 digital modalities at our disposal (wearable acceleration, proximity and video) we extracted 5 behavioral modality types: 1) Speech (S), 2) Movement from wearable (W), 3) Movement from wearable while Speaking (WS), 4) Proximity (P) and 5) Movement from video (V). In the next subsections, a detailed description of the preprocessing on each sensor and the extraction of each global feature is presented.

Table 5.1: Summary of our features divided by modality type: W=Mov. from wearable, S=Speaking, WS=Mov. from wearable while Speaking, P=Proximity and V=Mov. from Video. (S.T.= Speaking turns.)

	Feature	Type	Sensor
1	mean of accel. magnitude var. per window	W	Wearable
2	var. of accel. magnitude var. per window		
3	maximum length of S.T.	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 lenght of S.T.		
10	mean of accel. magnitude var. per window for S.T.	WS	
11	var. of accel. magnitude var. per window for S.T.		
12	mean size of group interacted with	P	
13	largest size of group interacted with		
14	total number of people interacted with		
15	mean of total number of zeros of OF magnitude	V	Video
16	mean of mean OF magnitude var. per window		
17	var. of mean OF magnitude var. per window		
18	mean of mean OF magnitude from distribution (low)		
19	mean of mean OF magnitude from distribution (medium)		
20	mean of mean OF magnitude from distribution (high)		

5.4.1. WEARABLE DEVICES

For the wearable devices we grouped our cues, which originated from 2 different sensors or *digital* modalities (triaxial acceleration and proximity), in 3 *behavioral* modality categories and their combination: body movement energy (W), speaking turns (S), body movement during speaking turns (WS) and proximity (P).

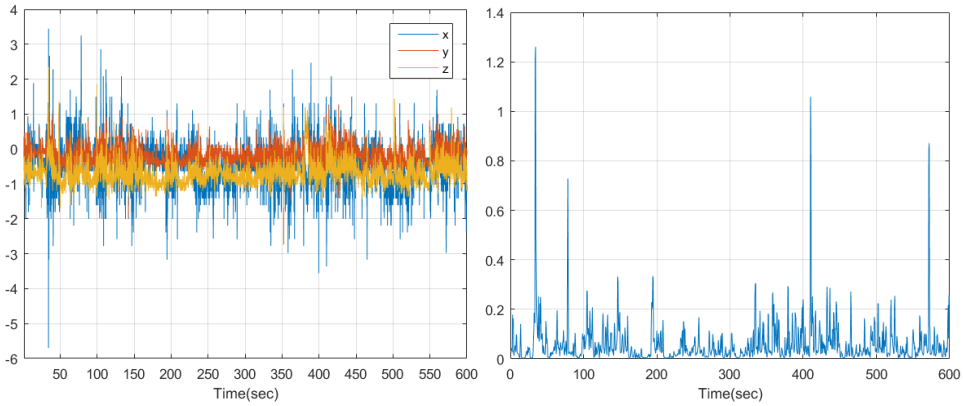
BODY MOVEMENT ENERGY (W)

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 window with a 2s shift. A graphical representation of this process is presented in Figure 5.2. This give us a better representation of *movement energy* over time than the raw acceleration magnitude, as can be seen in Figure 5.2(b).

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 over all windows. These features are 1 and 2 in Table 5.1.

SPEAKING TURNS (S)

Building on prior findings that people's speaking status is representative of their personality [18, 126, 168], 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 [89, 111]. We have used a novel transfer learning



(a) (b)

Figure 5.2: (a) Original raw acceleration of a wearable device (after filtering effects of the gravity). (b) Body movement energy resulting from preprocessing (described in Section 5.4.1).

method, Transductive Parameter Transfer (TPT) [183], which 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 unlabeled 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 $P_i^s \neq P_j^s$. In the notation used, s always corresponds to source datasets while t corresponds to the target one.

For our case, i is the index for the different subjects in the source data. Thus, P_i^s represents the distribution marginalized for that specific individual, as seen in Figure 5.3(a). 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 a Linear SVM as in [183], 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 5.1.

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

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

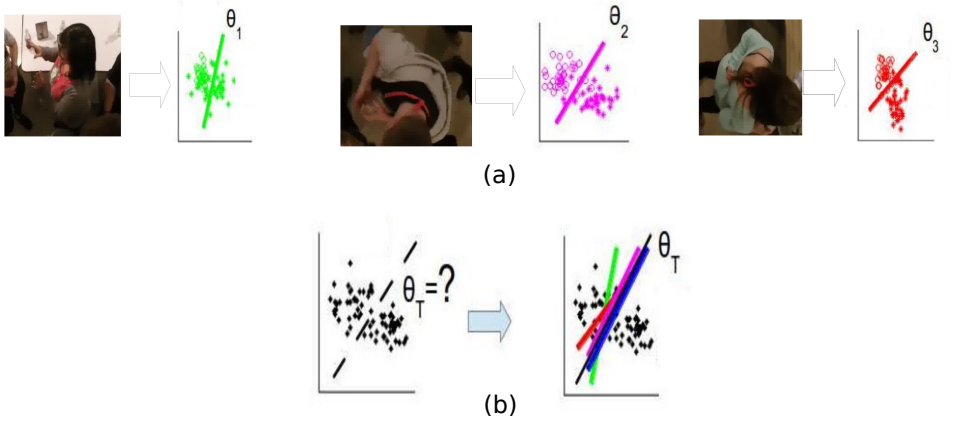


Figure 5.3: Transductive Parameter Transfer (TPT). (a) Marginal distributions and parameters for each subject. (b) Using regression to obtain parameters in the target distribution for new subject. Partially taken from [183]

2. **Learn the relation between the marginal distributions $\mathbf{P_i^s}$ and the parameter vectors $\theta_\mathbf{i}$ using a regression algorithm:** Training set $T = \{X_i^s, \theta_i\}_{i=1}^N$ is formed by samples X_i^s and parameters θ_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 θ needs to be learned. Assuming that elements in θ may be correlated, we have employed Kernel Ridge Regression [118], instead of the independent Support Vector Regressors used in [183]. Since we need to define the similarities between distributions X_i^s instead of independent samples, we employ an Earth Mover's Distance [144] based kernel. EMD kernel is computed as:

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

   In Equation 5.2, $EMD(X_i, X_j)$ corresponds to the minimum cost needed to transform X_i into X_j. The user defined parameter γ 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 θ^t. With θ^t known, we can infer the labels for the target dataset.

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 [74]) that are expected to be representative of speech related body movements. 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 labeled data of 18 participants as sources, we obtained speaking turns for all 56 participants during 30 minutes. The time interval used for these 18 participants is not the same as the 30 minutes used for our experiments. As stated in

Section 3, the labels for the speaking status of these 18 participants (sources) are obtained by manual annotation using the video.

Finally, derived features were extracted from the speaking turns (see Table 5.1). We create 7 global features from the speaking turns, which have the reference numbers 3 to 9 in Table 5.1. These features are the maximum, mean, variance and total of length of speaking turns, and the same for non speaking turns. In addition, we create 2 additional multi-modal behavioral features (WS), which combines the movement and the speaking turns (reference numbers 10 and 11 in Table 5.1). These are the mean and variance of the movement energy only in those windows with detected speaking.

PROXIMITY (P)

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 neighbor. From these binary detections, a dynamic (in time) binary proximity graph can be generated for each participant. To eliminate false neighbor detections, the method proposed in [106] was applied.

Then, 3 features (ref. numbers 12, 13 and 14 in Table 5.1) were calculated for each participant from the proximity graphs: mean size and 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 neighbors if they detect each other for more than one minute in the graphs.

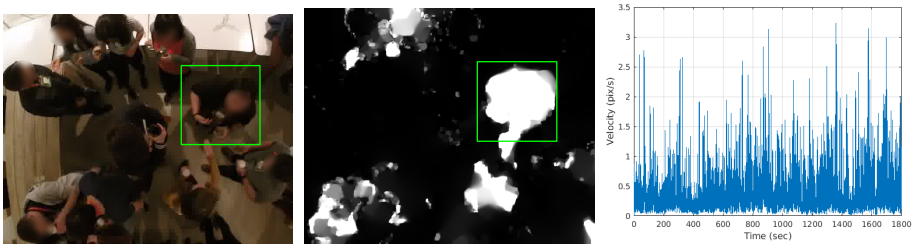
5.4.2. VIDEO CAMERAS

MOVEMENT FROM CAMERA (V)

First, we extract the dense optical flow of the entire video frame using the Farneback's algorithm. Then, we obtain the position of each participant in each frame using a bounding box and extract the magnitudes of the flow vectors within this box, as seen in Figure 5.4(b). Then, a single movement value per participant per frame is calculated using the mean value of the magnitudes within the bounding box. This is done for all frames in the video, which are later concatenated. Hence, we can represent the movement of the participant in video with a single time series (Figure 5.4(c)). Notice that we use magnitude of the flow vectors instead of the Cartesian values, as the participants always have a relative frame of reference (eg. their orientation changes with respect to the camera).

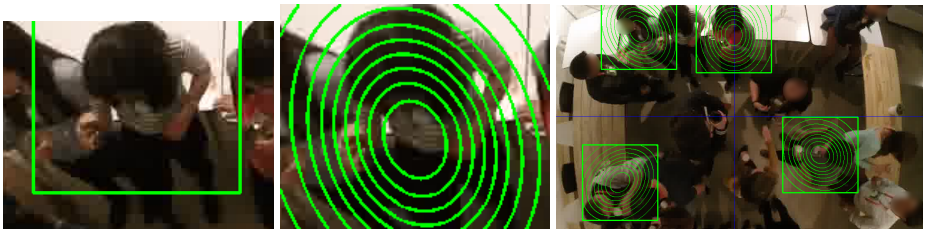
Additionally, we separate the magnitude values in 3 bins (low, medium and high), using the third percentile. Thus, we obtain 3 additional time series per participant. We do this to further analyze the impact of the type of event (high versus low acceleration variation) on the detection of personality.

Finally, to obtain our global features we calculate the mean (in time) total number of zeros in the optical flow vectors, and the mean variance of all time series (entire and 3 separated by bins). We also include the variance for the entire time series. This give us a total of 6 global features for the video modality (V).



(a) (b) (c)

Figure 5.4: Extraction of speed time series from video. (a) Participant's location in video. (b) Magnitude of dense optical flow for an entire frame. (c) Speed time series for one participant extracted from the mean magnitude of its optical flow.



(a) (b) (c)

Figure 5.5: Correction of outside flow vectors. (a) Example of third-party movement capture by the bounding box. (b) Weight correction using a multivariate Gaussian function. (c) Examples of multivariate Gaussian functions given the position w.r.t the camera

5.4.3. COMPENSATING SUBJECT CROSS-CONTAMINATION IN VIDEO

As can be seen in Figure 5.5(a), sometimes the bounding box with the participant's location captures movement that does not correspond to the participant itself. Hence, instead of using the raw optical flow, we apply a multivariate Gaussian function centered at the bounding box location as a weighting factor to compensate for the extracted flow vectors that possibly do not belong to the participant (eg. borders):

$$f(X,\mu,\Sigma) = \frac{1}{\sqrt{|\Sigma|(2\pi)^2}} e^{(-\frac{1}{2}(x-\mu)^T\Sigma-1(x-\mu))} \tag{5.3}$$

where $\mu = [\mu_x, \mu_y]$, μ_x and μ_y are the center of the bounding box, and Σ is the covariance matrix.

The aim of Equation 5.3 is to adapt to the form of the person given its position in the image plane, and give it a higher weight to the flow vectors located in the center of the bounding box where, we hypothesize, the person is truly located. Thus, μ_x and μ_y control the position of the box and the covariance matrix Σ its orientation. More specifically, given:

$$\Sigma = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix} \tag{5.4}$$

we define Σ_{XX} as a quadratic function of the position for the bounding box with respect to the image plane, or $f(x) = a * \mu_x^2 + b * \mu_x + c$. The same applies for Σ_{YY}, using μ_y. Finally, Σ_{XY} and Σ_{YX} are define as a function of 2 variables given by $f(x, y) = d * (\mu_y - w/2) * -(\mu_y - h/2)$, were w and h represent the width and height of the image. Here, a, b, c and d are constants that depend of the resolution of the image.

Figure 5.5(c) shows how the form of the Gaussian given by Eq. 5.3 changes depending on the position in the image plane of the participant with respect to that of the camera. Thus, if the person is directly under the camera Eq. 5.3 produces a more symmetric distribution, whereas Figure 5.5(b) shows the distribution required for the weighting of a person located in the top right position of the image plane.

5.5. EXPERIMENTAL RESULTS

As summarized in Table 5.1, we divided our set of features in 5 behavioral modality types: 1) Speech (S), 2) Movement from wearable (W), 3) Movement from wearable while Speaking (WS), 4) Proximity (P) and 5) Movement from video (V). In the next subsections we compare and analyze the complementarity of these feature types, both with a correlation analysis and during classification.

In addition, we analyze the impact of the speech detection in the overall performance of the personality estimation by comparing it to the speech annotations provided by the MatchNMingle dataset.

5.5.1. FEATURE CORRELATION ANALYSIS

Figure 5.6 shows the correlations for our final 20 features (summarized in Table 5.1). First, we can see 5 clusters in this figure that correspond to each modality type: 1-2 for W, 3-9 for S, 10-11 for WS, 12-14 for P and 15-20 for V.

Some of the correlations results are as expected. For example, we can see how the features related to speaking turns from the set S (3 to 5) are inversely correlated to those related to non-speaking turns (6 to 8).

Nonetheless, there are some interesting results. The feature for low distribution values of OF magnitude (18) does not correlate strongly with any of the other features, not even those in the same modality set V. This might be due to remaining noise in the video (after our filtering described in Section 5.4.2) most likely captured by this feature.

Another interesting result in this figure are the correlations between the features of mean and variance movement from the wearable W (1 and 2) and the video V (16 and 17). The absolute values for these correlations are low (around 0-0.2). An explanation for such low values can be that, as we hypothesized, each modality might be providing complementary information about the person's personality.

5.5.2. COMPARISON OF BEHAVIORAL MODALITY TYPES

Once we have seen the correlation between all the features, we proceed to analyze the impact of each modality type separately. To do so, we trained 5 different binary classifiers using only those features for the given set (W, S, WS, P or V). We used a L1 penalized logistic regressor (to reduce possible overfitting) and applied a 10-fold cross-validation.

Note that, as we only have one sample per subject (i.e. a 20-dimensional vector), using

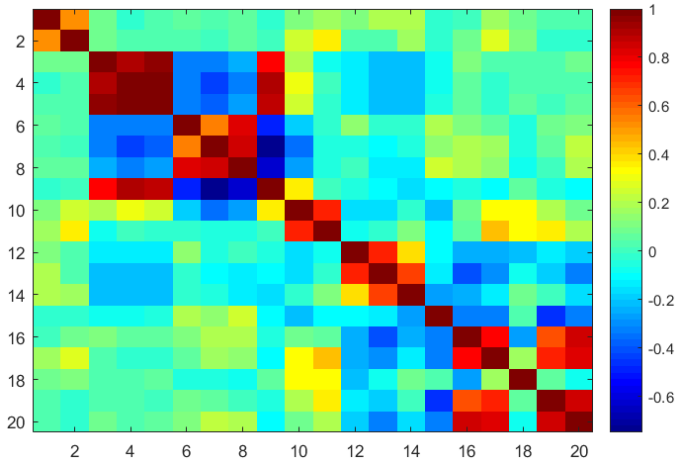


Figure 5.6: Correlation between all features in Table 5.1 (better seen in color).

Table 5.2: Comparison of behavioral modality types. Mean accuracy (± deviation per fold) of classification for the modality types (Table 5.1). Bold values represent the best performance per trait.

Trait	Modality type				
	W	S	WS	P	V
H	0.47±0.10	0.59±0.23	**0.64±0.20**	0.48±0.12	0.55±0.21
E	**0.46±0.05**	0.37±0.19	0.43±0.10	0.39±0.16	0.43±0.11
X	0.54±0.13	0.36±0.13	0.53±0.13	**0.55±0.22**	0.42±0.16
A	0.35±0.21	0.50±0.10	0.46±0.12	**0.63±0.25**	0.45±0.10
C	0.45±0.11	**0.58±0.12**	0.60±0.17	0.48±0.15	0.54±0.20
O	0.57±0.11	0.54± 0.17	0.44±0.05	**0.63±0.20**	0.56±0.19

a 10-fold crossvalidation is valid in our case without contamination between the train and test set.

To create binary labels from our trait scores, we used the median value for each trait as threshold with the higher values in the positive class. We do this for each fold, so the median is calculated only for the training set. The mean (± deviation per fold) median values were 3.40±0.05 for H, 3.18±0.02 for E, 3.56±0.01 for X, 3.12±0.03 for A, 3.34±0.06 for C and 3.5±0.01 for O. This procedure resulted in balanced class distributions.

Table 5.2 summarizes the mean accuracy and standard deviation within the folds for each trait and modality type set. Most of these results are equal or below the random baseline, with some exceptions.

Notice also that the accuracy values differ given the trait and the modality set. This shows that there is not a general modality set that would work for all traits and that each feature type has a different impact given the personality trait. Furthermore, one of the few accuracies over the random baseline is the Openness to experience (O) using the Proximity (P) features. This correlates with what has been found in previous research [182], which supports that proximity features are a good indicator for this trait.

Table 5.3: Complementarity of behavioral modality types from all sources (wearable + video). Mean accuracy (± deviation per fold) of classification for different combinations of modality types (Table 5.1). Bold=best result for the trait. Underline=best result while including video.(**p< 0.01 and *p< 0.05 for t-test comparison with classifier assigning labels at random.)

Modality set combination	Performance per trait					
	H	E	X	A	C	O
W	0.47 ± 0.10	0.46 ± 0.05	0.54 ± 0.13	0.35 ± 0.21	0.45 ± 0.11	0.58 ± 0.11
S	0.59 ± 0.23	0.37 ± 0.19	0.36 ± 0.13	0.50 ± 0.10	0.58 ± 0.12*	0.54 ± 0.17
WS	0.64 ± 0.20*	0.43 ± 0.10*	0.53 ± 0.13	0.46 ± 0.12	0.60 ± 0.17*	0.44 ± 0.05
P	0.48 ± 0.12	0.39 ± 0.16	0.55 ± 0.22	0.63 ± 0.25	0.48 ± 0.15*	0.63 ± 0.20**
V	0.55 ± 0.21	0.43 ± 0.11*	0.42 ± 0.16	0.45 ± 0.10	0.54 ± 0.20	0.56 ± 0.19*
W-S	0.51 ± 0.14	0.39 ± 0.16	0.38 ± 0.13	0.36 ± 0.19	0.58 ± 0.14	0.44 ± 0.17
W-WS	0.59 ± 0.15	0.44 ± 0.09	0.47 ± 0.14	0.44 ± 0.21	0.53 ± 0.13**	0.50 ± 0.11
W-P	0.47 ± 0.10	**0.46 ± 0.05***	**0.56 ± 0.25**	0.49 ± 0.29	0.42 ± 0.13	**0.69 ± 0.17***
S-WS	0.49 ± 0.20	0.37 ± 0.17	0.37 ± 0.13	0.41 ± 0.10	**0.65 ± 0.11****	0.47 ± 0.15
S-P	0.63 ± 0.23	0.41 ± 0.10	0.46 ± 0.17	0.53 ± 0.27	0.58 ± 0.14	0.51 ± 0.14
WS-P	0.57 ± 0.13	0.44 ± 0.16	0.51 ± 0.22	**0.64 ± 0.16***	0.56 ± 0.20	0.62 ± 0.22
W-V	0.56 ± 0.12	0.36 ± 0.13	0.43 ± 0.14	0.41 ± 0.11	0.56 ± 0.19	0.62 ± 0.22
S-V	0.53 ± 0.20	0.39 ± 0.14	0.39 ± 0.15	0.40 ± 0.10	0.59 ± 0.25	0.53 ± 0.22
WS-V	0.63 ± 0.24	0.39 ± 0.12	0.46 ± 0.14	0.41 ± 0.09	0.51 ± 0.19	0.51 ± 0.19
P-V	0.56 ± 0.18	0.41 ± 0.09	0.48 ± 0.19	0.52 ± 0.17	0.59 ± 0.14	0.60 ± 0.17
W-S-WS	0.58 ± 0.19	0.39 ± 0.16	0.38 ± 0.19	0.37 ± 0.17	0.57 ± 0.17	0.45 ± 0.16
W-S-P	0.49 ± 0.15*	0.37 ± 0.16	0.43 ± 0.18	0.41 ± 0.24	0.47 ± 0.10	0.52 ± 0.10
W-WS-P	0.53 ± 0.17	**0.46 ± 0.05**	0.55 ± 0.26	0.59 ± 0.20	0.48 ± 0.13	0.65 ± 0.13
S-WS-P	0.53 ± 0.19	0.40 ± 0.15	0.42 ± 0.20	0.56 ± 0.24	0.63 ± 0.15	0.50 ± 0.11
W-S-V	0.53 ± 0.13	0.44 ± 0.06	0.39 ± 0.15	0.43 ± 0.10	0.58 ± 0.22*	0.63 ± 0.25
W-WS-V	**0.71± 0.15****	**0.46 ± 0.05***	0.39 ± 0.09	0.39 ± 0.13	0.55 ± 0.21	0.57 ± 0.18
W-P-V	0.55 ± 0.19	0.39 ± 0.16	0.49 ± 0.22	0.48 ± 0.22	0.55 ± 0.19	0.68 ± 0.20*
S-WS-V	0.54 ± 0.26	0.37 ± 0.17	0.37 ± 0.18	0.39 ± 0.11	0.60 ± 0.17**	0.49 ± 0.24
S-P-V	0.50 ± 0.13	0.33 ± 0.16	0.46 ± 0.19	0.51 ± 0.16	0.58 ± 0.19	0.50 ± 0.09
WS-P-V	0.61 ± 0.20	0.41 ± 0.09	0.42 ± 0.17	0.58 ± 0.19*	0.57 ± 0.15	0.61 ± 0.19
W-S-WS-P	0.56 ± 0.14	0.34 ± 0.16	0.47 ± 0.21	0.43 ± 0.21	0.61 ± 0.12**	0.52 ± 0.10
W-S-WS-V	0.63 ± 0.09*	0.37 ± 0.16	0.36 ± 0.10	0.41 ± 0.13*	0.57 ± 0.17	0.54 ± 0.21
W-S-P-V	0.43 ± 0.09	0.41 ± 0.10	0.46 ± 0.18	0.43 ± 0.16*	0.50 ± 0.15	0.65 ± 0.16
W-WS-P-V	0.66 ± 0.12	0.43 ± 0.10	0.52 ± 0.26	0.50 ± 0.23	0.57 ± 0.15	0.61 ± 0.19*
S-WS-P-V	0.50 ± 0.21	0.37 ± 0.16	0.50 ± 0.19	0.53 ± 0.18	0.56 ± 0.18	0.54 ± 0.15**
All	0.56 ± 0.09	0.39 ± 0.12	0.46 ± 0.18	0.41 ± 0.15**	0.58 ± 0.14	0.60 ± 0.18

5.5.3. MODALITY COMPLEMENTARITY

We now proceed to evaluate the complementarity of our 5 different modality types for the binary classification of personality traits. For this purpose, we trained different classifiers with the different combinations of the modality types using early fusion. Similar to Section 5.5.2, we selected a L1 penalized logistic regressor with a 10-fold cross-validation, and use the median to create binary labels from the personality scores.

Table 5.3 presents the mean accuracy and deviation for selected combinations. Here, the first 5 rows are the same as the columns of Table 5.2, which we replicated for an easy comparison. For each trait, the best result is in bold and the best result when the video type (V) is used is underlined, for further comparison between sources. This table is also separated given the number of modality types combined (double line), and given the presence of the video modality (bottom of each sub-block). As seen in Table 5.3, our best results corresponds to the traits of Honesty (H) and Openness to experiences (O).

Table 5.4: Impact of speech detection in the personality estimation. Mean accuracy and deviation of classification for the S and WS types using the Ground Truth and the estimated (TPT) speech. (Bold=best result for the trait in this table, underline=ground truth has better result).

Trait	Ground Truth		Estimated (TPT)	
	S	WS	S	WS
H	0,53±0,24	0,43±0,09	0.59±0.23	**0.64±0.20**
E	0,40±0,11	0,42±0,15	0.37±0.19	**0.43±0.10**
X	0,49±0,17	0,42±0,14	0.36±0.13	**0.53±0.13**
A	0,53±0,20	**0,59±0,16**	0.50±0.10	0.46±0.12
C	0,50±0,17	0,40±0,17	0.58±0.12	**0.60±0.17**
O	0,53±0,21	**0,60±0,13**	0.54±0.17	0.44±0.05

5.5.4. IMPACT OF SPEECH DETECTION ON THE PERSONALITY ESTIMATION

Finally, we intended to assess the impact of the speech estimation status in the personality estimation performance. As recall, for obtaining our global speech features we relied in a transfer learning approach which extracted binary speech status from wearable acceleration (see Section 5.4.1). Now, we intend to compare the estimation of the personality traits using this speech estimation with the speech ground truth.

To do so, we used the annotated binary speech status provided by the MatchNMingle dataset [26]. These annotations were done manually by trained annotators every frame at 20 FPS. To perform a fair comparison against the speech estimation, we applied the same window to the annotations as we applied to the speech detection using TPT (see Section 5.4.1). Thus, we obtained binary time-series (speaking/no speaking) with the same number of samples that can be directly compared to those extracted from the TPT method.

From these ground truth time-series, we extracted our 7 global features for speaking (3 to 9 in Table 5.1) using the same process as described in Section 5.4.1. In addition, we use these streams to calculated the global features for movement while speaking (10 and 11 in in Table 5.1). The procedure was the same as in Section 5.4.1, when using the TPT estimation.

Table 5.4 shows the comparison of the accuracy, for each trait, when using the ground truth speech and the speech estimated using our TPT (same as in Table 5.2). Similarly to previous sections, these results were obtained using a 10-fold cross-validation and a L1 penalized logistic regressor.

5.6. DISCUSSION

The best and worst performing traits

As seen in Table 5.3, our best results corresponds to the traits of Honesty (H) and Openness to experiences (O). For the trait of Openness to Experiences (O) we already obtained an acceptable result ($63\% \pm 2\%$) using only proximity-based features. In addition, we can see across Table 5.3 that combinations of this modality (P) with the types movement (W) and video (V) tend to give the better results, even when only combining 2 types. For instance, combining these 3 modality types gives a 68% accuracy (deviation of 20%). However, the best result for this trait is obtained when combining only proximity and movement ($0,69\% \pm 0,17\%$). Thus, it appears that the modality of movement (either video or wear-

able) added complementary information to that in proximity for this trait. Furthermore, one should notice that the movement from wearables and movement from video are, from the technical perspective, recording similar features (eg. mean movement). Nonetheless, the results of each combined separately with proximity differed, with a 69% against a 60% respectively.

For the trait of Honesty (which is our best result overall) it appears that the type of movement while speaking (WS) gives the most information, complemented by the proximity and movement from video types, as can be seen in the different combinations of these in Table 5.3. Furthermore, for this trait we obtained ($71\% \pm 15\%$) when combining the modality types of movement (W), movement while speaking (WS) and movement from video (V). To properly link these modalities' impact in the trait one most study the nature of the trait. People with low Honesty scores '*will flatter others to get what they want, (...) and will feel a strong sense of self-importance*[1]. This might get reflected better in WS, P and V as all these modalities take into account an interaction with someone (movement from video is more likely to come from gestures while interacting than from torso movement).

Notice also that the experiment using only the modality sets from the wearable devices are similar to those presented in our previous work using only the wearables [27] (see Addendum). Nonetheless, here we used a different subset of participants (*wearable* versus *wearable+video* sets) for a fair comparison with video. Thus, the results might vary with respect to our previous work but the general insights maintained.

Except for the trait of Honesty, all the best results per trait are obtained when combining only 2 modalities (second sub-block of Table 5.3). Also, these combinations do not include video. Nonetheless, they all included a type of movement, with M-P having the best result for 3 traits (Emotionality still performs under the random baseline). Thus, these results suggest that movement features are a feasible indicator to assess personality of people during crowded and in-the-wild scenarios.

The trait of Extraversion (X) shows some of the lower performance, with the best result been $55\% \pm 22\%$ accuracy while using only the set for the proximity modality (Table 5.2). This modality has also proved in previous efforts to be a good indication for extraversion [182]. Nonetheless, one should wonder about the low results for this trait in Tables 5.2 and 5.3, which are in most cases not even over the random baseline. We firstly hypothesized that the modalities of Speech (S), Proximity (P) and movement from video (V) could be a good indicator for this trait, as they record elements that will help to detect an extrovert person: speaking, proximity to others and hand gestures [4, 182]. However, apart from proximity, these modalities provided performances below the random baseline when evaluated independently. This was also the case when combined. We can only infer that the hand-crafted features described in Section 5.4 are not representative enough to distinguish between extroverts and introverts in such a complicated setup as are the mingle scenarios.

Another interesting insight comes with the fact that fusing all modalities does not guarantees the best result. In fact, as can be seen in Table 5.3, none of the traits has a better performance when combining all modality types (last column). Thus, each trait is reflected differently in the modality types and this is in consequence relevant for the classification. For example, we have already discussed that features from the proximity set are a good indicator for Openness to experience (O) and that this increases when combined with the set

[1] Taken textually from `http://hexaco.org/`

of features from video. Nonetheless, when also combining these 2 sets with the modality sets of movement (W) or Speech (S) decreases the performance, even when compared to the result for the Proximity set only (Table 5.2). This shows that the type of modality itself has an important role in the estimation of the different traits and that this is not only related to more information.

The impact of video

With respect to video, we first must emphasize that all the results including video only have partial information, as the subjects selected were chosen to account only for those participants that were at least 50% of the time under the cameras. This means that, although some participants have complete video data for the entire 30 minutes, some participants have the worst case of only 15 minutes of video.

Nonetheless, Table 5.2 shows that movement from video (V) alone gives results over the random baseline for the traits of Honesty (H), Conscientiousness (C) and Openness to Experiences (O). Furthermore, Table 5.3 now shows that adding the modality of video (V) to any of the other modalities improves the overall results for these 3 traits. This suggests that the features selected for video can deal with missing data in video (until a certain point) as they are meant to be accumulative over time and still give complementary information for the other modalities. Also, when adding movement from video we also obtained acceptable results (underline in Table 5.3) even with missing data for some participants. Thus, we hypothesize that obtaining more data from video might further improve these results. However, when intended to increase the visibility threshold for the participants from 50% to 80% to at least, we reduce the number of participants from 56 to 22, which will not be representative enough to generalize.

Impact of speech detection on the personality estimation

Firstly, we can see that all the best results (for Table 5.4 only) correspond to the modality set of movement while speaking (WS). This is interesting as the set itself is a multimodal representation, taking into account the status in one modality (speech) to filter the information on the other (movement). Thus, also in this experiment we can see how the complementarity between modalities increases the performance of the estimation.

Furthermore, perhaps the most intriguing result in this table is that using the ground truth from the speech status does not necessarily implies a better performance in the estimation of personality traits. As can be seen in Table 5.4, only 5 results (those underlined in the table) are better when using the ground truth speech. One could hypothesized that improving the performance in an early stage (speech detection, in this case) will have a positive impact in the final estimation (personality). However, this experiment proves that this is not necessary the case.

For this case in particular, we hypothesize that the better performances while using the estimation for speech instead of the ground truth are due to the method that was used to detect the speech from the wearable acceleration. Our TPT method relies in the assumption that we move when we speak and hence we can use this to approximate the speaking status. Nonetheless, it is quite possible that this is not necessarily the case for all events. Thus, the speech estimation might also be taking into account movement from other components of the interaction (eg. gestures or fidgeting movements), which can be informative for the

estimation of the personality traits.

ADDENDUM

Our previous work was presented in [27], and contemplated only the modalities from the wearable devices. This means, only the features 1-13 in Table 5.1 were used. Also, it used the *wearable-only* set (69 participants), as this was the total number of working devices available.

First, we calculate the correlations between the features and the traits, which are summarized in Table 5.5. In this Table, only the comparisons between features and traits with a significant value are summarized.

Table 5.5: 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

Then, we proceed to classify the traits in a binary way. The labeling was also obtained by using the median value for each item and using it as a threshold, with higher values in the positive class. We selected the logistic regressor as our classifier, and we used 10-fold cross validation. The accuracies obtained for each feature type are presented in Table 5.6 and with different feature combinations are provided in Table 5.7.

Table 5.6: Mean accuracy (%) ± std. error for original features. 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).

	Different Features			
	M	**S**	**MS**	**P**
H	59 ± 22	$66 \pm 17^{**}$	$68 \pm 17^{**}$	44 ± 12
E	47 ± 7	43 ± 13	52 ± 3	52 ± 3
X	52 ± 12	46 ± 9	48 ± 12	53 ± 15
A	54 ± 9	52 ± 10	54 ± 8	55 ± 14
C	46 ± 19	49 ± 19	57 ± 13	46 ± 8
O	58 ± 1	56 ± 5	58 ± 1	$\mathbf{69 \pm 17}^{*}$

Table 5.7: Mean accuracy (%) ± std. error for different feature combinations. 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	**M+MS**	**M+P**	**S+MS**	**S+P**	**MS+P**	**M+S+MS**	**M+S+P**	**M+MS+P**	**M+S+MS+P**
H	$62 \pm 20^{*}$	$\mathbf{69 \pm 15}^{**}$	47 ± 20	58 ± 16	57 ± 14	$62 \pm 14^{*}$	58 ± 18	61 ± 22	$63 \pm 13^{**}$	$62 \pm 18^{*}$
E	48 ± 12	48 ± 7	52 ± 3	45 ± 13	46 ± 13	52 ± 3	48 ± 10	46 ± 13	52 ± 3	52 ± 3
X	51 ± 4	48 ± 10	59 ± 17	46 ± 13	50 ± 12	$\mathbf{60 \pm 12}^{*}$	49 ± 7	51 ± 7	$\mathbf{61 \pm 14}^{*}$	54 ± 9
A	53 ± 15	55 ± 6	56 ± 15	53 ± 17	58 ± 18	59 ± 15	$62 \pm 10^{*}$	53 ± 12	54 ± 20	$\mathbf{65 \pm 14}^{*}$
C	52 ± 16	55 ± 13	42 ± 19	56 ± 12	53 ± 13	50 ± 13	$66 \pm 15^{**}$	55 ± 14	49 ± 16	$\mathbf{69 \pm 15}^{**}$
O	55 ± 9	53 ± 9	63 ± 17	58 ± 1	66 ± 14	60 ± 19	53 ± 13	48 ± 17	65 ± 18	56 ± 19

6

MEASURING IMPLICIT AUDIENCE RESPONSES TO LIVE PERFORMANCES USING MULTIPLE MODALITIES

The contents of this chapter are based on the work originally published in:
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 ACM International Conference on Multimedia (ACM MM), 2015. (*These authors contributed equally to this article).

6.1. INTRODUCTION

Art and cultural events such as dance, art exhibitions, and concerts have an inherent value [63], which some studies have shown are correlated with a perceived quality of life [113] as well as self-rated health levels [119]. Driven by the high reward that such information would have for the art and cultural industry, the aim of this paper is to investigate ways of automatically measuring the response to the experience of art and cultural events as a means of enhancement for both consumers and practitioners.

In this Chapter, we target the task of quantifying people's experience and automatically predicting factors related to watching a modern dance performance. We show experimental results based on **two different public modern dance performances**. However, unlike prior works that have used biosignals or physiological sensing [62, 153], we hypothesize that behavior, as sensed by sensors that one would find in smart phones, could also be used to sense affective responses to a performance. Our experiments show that using more pervasive sensors opens up huge possibilities for implicit affective sensing on a large scale in the wild.

By working closely for at least 2 years with Holland Dance (HD)[1], an organization whose role is to promote dance in The Netherlands, we have identified some key challenges to measuring audience response:

The limits of survey responses: Organizers of live performances are always interested to gauge audience opinions about the performance they organize — if they enjoyed a performance or enjoy dance performances in general, they are more likely to recommend it to others, thus sustaining the popularity of the art form. Such responses must also be obtained after the performance and do not therefore capture the spontaneous response of the performance audience to specific moments. Note that sentiment about a performance can also be assessed via social media but again requires audience members to actively participate in putting forward an opinion publicly [160].

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 [62, 153] or live performances [174] have tended to rely on physiological or brain activity measurements. While such signals are considered fairly reliable, equipment that is able 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 dance performance, so measuring implicit and measurable responses from pervasive sensors are likely to be even more challenging.

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 moments 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 but still incomplete information about someone's experience that need to be manually processed. Going further, interviews can also be used, but provide more unstructured responses for those few who are willing to spend more time reflecting on their experience. They are, therefore, limited to an even smaller subset of an entire audience.

Quantifying the impact of a Performance on our Social Lives To our knowledge, most works focus on measurements obtained solely during the performance, measuring direct

[1] http://www.holland-dance.com

responses to it. However, the value of a performance can stretch beyond this period to its affect afterwards. For instance, it could affect ones mood, serve as a topic of stimulating discussion over drinks, leading to positive feelings about the entire performance and socialising experience. In an ideal case, its effects should last beyond the performance itself, perhaps even providing lasting memories that are recalled collectively by friends. This is perhaps the most difficult challenge but, if answered, even in part, would provide a broader metric to quantify the value of arts and cultural events.

More concretely, we make the following novel contributions; we show that (i) even when people are sitting and watching a live dance performance, they spontaneously react it via body movements that can be captured from a standard acceleration sensor, (ii) moments of common spontaneous bodily reaction correspond to memorable events in the performance as reported by survey responses relating to the performance, (iii) their 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) 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.

6.1.1. THE LUCENT AND HDF DATASETS

Additionally to the *MatchNMingle* dataset (see Chapter 2), for this Chapter we collected two more datasets to assess the enjoyment of the attendees to a dance performance. First, the *Lucent* dataset consists of the wearable acceleration data of 32 participants watching a dance performance for which we address the detection of enjoyment directly (see Section 6.3).

Furthermore, we also aimed to answer the research question: "Can we quantify the influence of a performance on the audience even after the performance is over?". Thus, for the *HDF* dataset apart from the performance itself we also recorded the wearable acceleration and proximity during the interactions during a mingle (similar to the *MatchNMingle* dataset) before and after the dance performance, for up to 35 participants. See Section 6.4 for more.

6.2. RELATED WORK

Traditional methods to investigate the response of an audience to a live dance performance make use of self-reports, such as surveys and interviews [24, 138]. Digital technologies can overcome the limitations of surveys and interviews, by giving more direct and fine-grained insights into the response of an audience. For example, the explosion in popularity of the social media, e.g. Twitter, and mobile computing has broadened the borders of a live performance, as fans comment and post information and opinions live to the online community [19].

Other rather less pervasive technologies can also overcome the granularity issues of surveys using sensors. For example, work in neuroaesthetics use fMRI scanning to relate viewer responses to the aesthetics of the performance [22, 31, 46]. Moreover, tracking of

eye gaze from video has been used when trying to distinguish novice from expert observers of dance [155]. Finally, physiological 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 [95]. Similarly, GSRs have been used also to measure the response to other types of live performance, such as comedy [174] and movies in a cinema [62].

These attempts show an increasing interest in quantifying the experience of arts and cultural events, such as live dance performances. Unlike these approaches, we advocate the use of pervasive sensors which are readily available in smartphones, which enable less obtrusive measurements and on a massive scale, compared to those obtained via physiological sensing. In particular, in this work, we focus on using acceleration and proximity sensors to measure people's reactions to live performance, which have been used thus far to measure very different phenomena.

Specifically, most work that consider accelerometers and people have addressed the problem of activity recognition of daily activities such as walking, running, sitting, climbing the stairs [93], daily household activities including eating or drinking, vacuuming or scrubbing, lying down [15], or identifying modes of transportation taken [139]. There is a trend moving towards the detection of medically relevant events, such as fall detection [56, 185], 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. Few works do exist 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 [108], Hung et al. [74] used body movements to predict socially relevant actions with a device hung loosely round the neck or for detecting conversations[75]. 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 abovementioned 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 Englebienne and Hung [58] 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. [17] 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 whilst the user was holding it were used to gauge responses. Using a multimodal approach they were able to predict the rating of users to movies they watched on the tablet. However, in this case, the user sat alone to watch the movies and was not inhibited by the social norms usually adhered to in an auditorium.

Proximity sensors have been used to study the interactions between individuals with approaches more similar to complex network analysis. Cattuto et al. [33] have used wearable sensors to analyse social interactions in crowded social settings, by means of proximity data collected through RFIDs. Martella et al. [107] used data collected through a series of wearable proximity sensors to identify the different communities attending a multi-disciplinary ICT conference. Roggen et al. [143] and Wirz et al. [178] 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 arts and cultural events.

6.3. DIRECT RESPONSES TO A PERFORMANCE

To inspect whether it is possible to predict responses to a performance by using data collected with wearable sensors, we have conducted an experiment in an actual dance performance. We start this section by explaining the characteristics of the performance and the resulting dataset we obtained. Then, we describe the features we used in the data analysis and classification experiments, and present the qualitative analysis and classification experiments we did based on questionnaire responses.

6.3.1. DATA COLLECTION (THE LUCENT DATASET)

We organized a data collection experiment during a dance performance. This event consisted of almost an hour and a half of performance without intermission. It mainly consisted of dancing but also included monologues by the performers in Italian, while the music was mainly based on live cello arrangements but also included pre-recorded songs. Using triaxial acceleration and IR cameras (for additional data verification), we recorded 41 participants watching the performance. The accelerometers were located in a custom-made device hung around every participant's neck (the Chalcedony devices used through this thesis). These devices recorded at 20Hz and were synchronized to a global time obtained by communicating through a wireless network. However, due to hardware malfunctions, only 32 accelerometers recorded data. In addition, the performance was recorded using a GoPro Hero +3 to analyze salient moments (i.e. favorite moments that were reported by the participants).

To evaluate the experience, a questionnaire was filled in by all 41 participants after the performance. Each questionnaire had 12 questions, where each group of three questions aimed to measure one aspect of the experience. The four aspects were "enjoyment", "recommendation (to a friend)", "immersion" and "mood changes". Each participant evaluated these aspects of the performance on 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 [148]. For the task of immersion, we selected involvement questions from the Igroup Presence Questionnaire [147]. For recommendation we used items from O'Brien's questionnaire [122]. Each of these questions were 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 describing 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. From the total number of participants, 32 responded with the Dutch questionnaire and 9 to the English one.

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

6.3.2. FEATURE EXTRACTION

We used the variance of the accelerometer readings, which is expected to act as a proxy for the physical activity level of the participants. Our assumption was that both subtle as well as more expansive movements of the the participants is related to the experience of the event. We expect participants with different evaluations of the event to have different movement patterns throughout the event, especially during salient parts of the performance. We calculate the variance in a sliding window of 2 seconds with 1 second shift, which corresponds to 40 samples for each window with a shift of 20 samples. This window size is carefully selected to capture the subtle and short variations in motion while still preserving a fine time scale and is empirically proven to perform well.

We extract features from an interval of ~79 minutes, starting just before the first piece, when all participants are seated and ending when the final piece of the performance finishes. With this we obtain 4705 different variance values for each axis. Before calculating the variance along each axis, each axis is normalized by computing the z-score to remove interpersonal differences. We also calculate the variance of the magnitude, resulting in 4 different variance values for each interval. For the qualitative analysis, we only use the variance of the magnitude. For classification, we treat the variance values of each axis as well as the magnitude as our features, resulting in a 18820 dimensional feature vector for each participant. This feature choice restricts the representation to be temporally dependent. Therefore we may not capture cases where two participants liked (or disliked) the event during different parts of the performance. With another formulation of the problem, like using bag-of-words or a multiple instance learning approach, temporal dependence can be avoided. Although such cases are quite probable, in this experiment we assume that participants with similar evaluations of the performance tend to respond similarly during salient parts of it.

6.3.3. DATA ANALYSIS

The variance in magnitude signals from all the participants were compared against each other to create a pairwise co-occurrence measurement over time using Mutual Information (MI). These signals were calculated over a sliding window (size of 60 samples) shifted by one sample, resulting in a vector of co-occurrence over time between two participants. The mean mutual information at each time interval was calculated, allowing us to evaluate the collective response of the participants to the performance over time. We hypothesized that

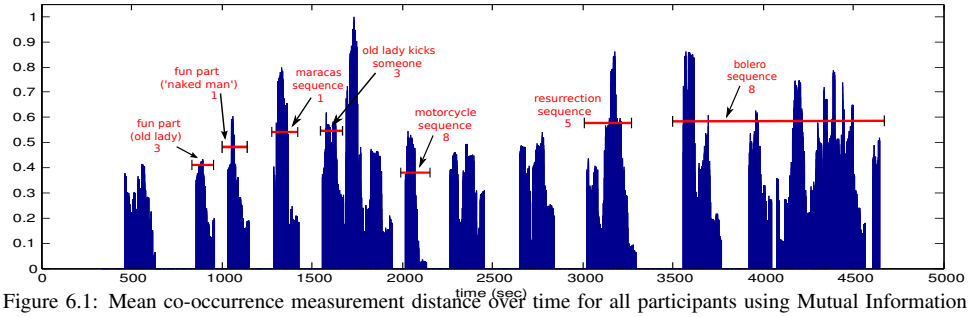


Figure 6.1: Mean co-occurrence measurement distance over time for all participants using Mutual Information (MI). Salient moments are highlighted in red with number of appearance

salient moments would correspond to a high MI among all participants. These moments were chosen using a Otsu threshold [69] on the values of the computed signal. Figure 6.1 shows these salient moments captured by points where the mutual information goes beyond the threshold (blue), as well as the frequency of the reported favorite moments (red) that appeared in the free text survey responses. The time stamps were generated by manually identifying the time period(s) where the reported favorite moments occurred. Notice that the two moments declared as favorites for the majority of participants (*motorcycle* and *bolero finale*) are captured.

This shows that memorable moments for people during these events can be captured by their coordinated movements, as they share the experience.

Furthermore, the role of music during the performance is also interesting and we want to understand its effect, if any, on the collective behavior 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 normalized 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 main factor stimulating coordination between our participants. For this reason, we decided that the song changes would not be useful for the following classification experiments and have focused solely on acceleration data instead.

With this preliminary qualitative information, the next subsection describes our classification experiments using the four questionnaire tasks mentioned in Section 6.3.3.

6.3.4. CLASSIFYING EXPERIENCE

Labelling samples

We use a standard pattern recognition approach to automatically predict the responses of participants. We start by labelling our participants according to the questionnaire answers they gave. For each group of three questions, we obtained one numerical value by averaging and rounding the three answers. This way, we obtain four different labels for our participants, where each label corresponds to the one of the tasks. We divided the participants into two classes for each task corresponding to a "positive" and "negative" report on

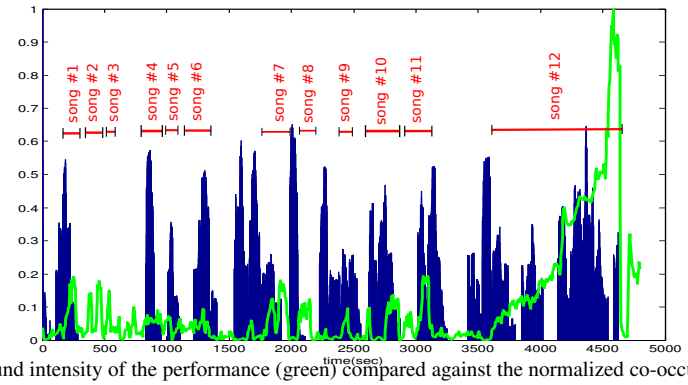


Figure 6.2: Sound intensity of the performance (green) compared against the normalized co-occurrence measurement calculated by MI (blue).

their experience of the performance. Participants whose averaged answer to any group of questions was below 5 was placed in the negative class for that task, meaning this participant, either did not enjoy the event, would not recommend the performance, did not feel immersed throughout the performance or did not think the performance uplifted their mood. The positive class thus contains participants who gave positive responses to the questions.

In this study, we mainly focus on enjoyment and recommendation, since these two are questions with clearer indications, but still provide results for immersion and the mood changes as we believe they can help in obtaining a general understanding of the performance's effects on the participants. After defining the classes for each task, our class distributions were not always balanced. For "enjoyment" and "recommendation", the majority of participants (26) gave positive answers. 22 participants thought "the performance affected their mood positively". Therefore, for these tasks, the problem becomes challenging for the smaller negative classes. These imbalance in class distributions also affected our choices of performance measure; in addition to accuracy, we also provide the balanced accuracy so each classes contributes equally to the final measure. The distribution for the "immersion" task is more balanced with 17 participants in the positive class.

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. More precisely, our features are the variance of acceleration along each axis and of acceleration magnitude, extracted with the aforementioned setup. We selected a Linear SVM as our classifier. Since the number samples is limited, we chose to use a model with few parameters. For evaluating the performance of our method, we used leave-one-out cross validation, training with 31 samples and with the remaining one is used for testing. The hyperparameters of the SVM are also selected using cross validation on the training set.

As stated in Section 6.3.2, representing the features using the whole performance would require classification on a 18820-dimensional feature vector for each participant. Since we do not expect all intervals to be equally informative and to avoid the curse of dimensionality, we decided to use a filtering approach which selects the informative intervals before feature

extraction.

To do so, we selected a Dynamic Time Warping (DTW) distance computed over a window (sizes were set from 20 to 60 samples) with a shift size of 1 sample. Similar to the mean MI co-occurrence vector used in Section 6.3.3. Our assumption here is that if we select the intervals where the average DTW distance between each pair is significantly higher than the rest, we should end up with time intervals that 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 coorespond to intervals where the intra-class distances are maximised. We hypothesise that using this metric provides better discrimination between classes compared to using mutual information where moments of high mutual information could also correspond to moments where the mean DTW is low and both classes would be almost indistinguishable. Empirical results using a threshold on the mean MI supported this claim, with performance scores significantly lower than the proposed method for the majority of tasks.

Using the same setup explained in Section 6.3.3, we detect the intervals where pairwise DTW distances are significantly higher than the rest and extract variance features for each person from these intervals only. The number of remaining intervals after filtering depends on the window size selection. In our experiments, where we have used windows of 20 to 60 samples, the number of selected intervals ranged from 86 to 830. Finally, after interval selection and feature extraction, we perform further dimensionality reduction by applying principal component analysis (PCA) to the feature vectors. We keep the principal components which preserve the 99 percent variance of features and use them for training and testing our model. This resulted in 12 and 19 resulting feature dimensions.

Results

The performance obtained by the proposed method, with different window size selections when using the thresholded DTW distance to pre-filter the salient intervals, are presented in Table 6.1. This table also includes the performance scores obtained without interval selection. The statistically significant results between those using interval selection and those using whole event are specified with a sign '*'.

Table 6.1: Performance scores obtained with different methods

Method \ BAcc | Acc(%)	Enjoyment	Recommend	Immerse	Mood
DTW IS(20 Sample)	60 | 66*	61 | 78	63 | 63	60 | 63*
DTW IS(40 Sample)	90 | 94**	70 | 81	53 | 53	55 | 66*
DTW IS(60 Sample)	78 | 84**	71 | 84	49 | 50	48 | 63*
Whole Event	36 | 38	74 | 78	51 | 50	38 | 38

(→ p<0.05) (** → p<0.01)*

The performance obtained without interval selection, which are reported in the final row of Table6.1, are unsatisfactory in general. Any task other than predicting recommendation has an accuracy lower than or equal to the proposed method, regardless of the window size. We should note that we did also apply PCA to the feature vectors for the non-filtered method. However, these scores showed that the extracted principal components were still affected by variance features extracted from many non-informative intervals, supporting our claim of interval selection is necessary.

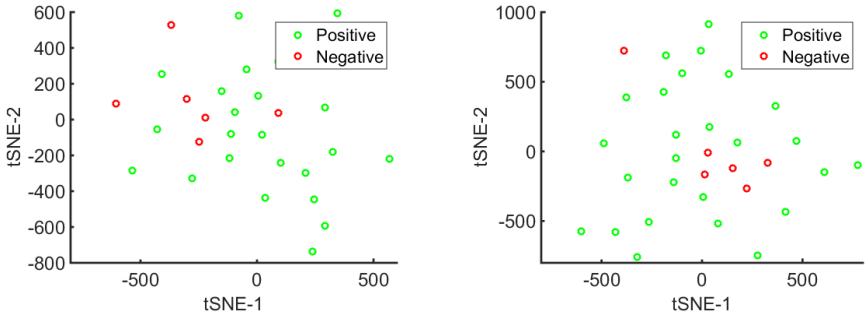


Figure 6.3: Non-linear embedding of feature vectors for the enjoyment class for (a) the whole event and (b) using interval selection with a 40 sample window.

The performance obtained with different window sizes have some interesting implications. For the tasks of immersion and mood, we obtained the highest performance with a window of 20 samples, corresponding to one second. With the increasing window size, our performance for these tasks dropped below random. We can see that this same window size is not optimal for the enjoyment and recommendation tasks. This could suggest that some tasks are shorter in time scale than others. However, we need more data to draw solid conclusions about such implications. It should be also noted that for the mood task, our proposed method always provides a significantly better result than the whole interval method, regardless of the window size.

Another interesting implication can be seen in the performance scores for "recommendation". For this task, the highest performance score is obtained when the whole event was used. Supporting this idea, the performance for this task increases as the window size increases. We should note that using a wider window may not always guarantee that more intervals are selected for classification. None of the results presented for the recommendation task showed significant improvement over the whole event baseline. However, the recommendation accuracy is very high for the whole event, suggesting that the behaviour of people who are interested in recommending dance performance to others could be distinguished from those who did not tend to recommend dance performances to others, and that this was independent of any particular moments during a performance.

Finally, we can see that for enjoyment, one of our most important tasks, we obtain relatively high performance, with the highest score of 94% accuracy and 90% balanced accuracy. With a window size of 40 samples, only one sample from each class is misclassified, resulting in high precision and recall scores for both classes. This result is significantly better than the whole interval method ($p>0.01$). To further investigate how the interval selection affects the distribution of samples, we visualise the filtered set of feature vectors of each participant using t-SNE [49], a non-linear embedding technique. The resultant two-dimensional embedding shown in Figure 6.3 illustrates that interval selection allows data points in the negative class to be clustered together in the feature space.

We also experimented with computing DTW distances on the raw accelerometer magnitude signal, instead of the variance. This experiment resulted in performance scores that

were worse than random for "enjoyment", "immersion" and "mood". For "recommendation", we obtained a balanced accuracy score of 85 percent. This result is quite interesting, since the highest performance we obtained for "recommendation" in Table 6.1 was the case where the whole event was used. However, in general it can be said that this experiment empirically supported our claim that the variance in acceleration is a valid feature for our experiments, both as a feature and for the interval selection using the thresholded DTW distance.

6.3.5. FURTHER ANALYSIS OF SALIENT MOMENTS

This section aims to further explain the salient moments of the performance, now relating these with the classes identified in Section 6.3.4. For space reasons, we focus on the enjoyment task as it has given us the best performance. 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, the same unified similarity measurement as described in Section 6.3.3 was calculated and the salient moments were obtained using the Otsu threshold level. Since the goal is to asses the similarity of people within the same class, pairs of different classes are left out.

Figure 6.4 shows the measurements of mean MI over time for both classes in the enjoyment task (where the negative class was plotted below the positive class). Notice how the two moments considered as favorites for the majority of participants (*motorcycle sequence* and *bolero finale*) 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 1 and 2. This reaffirms that specific acts or sequences in a performance (or movie) can have a significant impact in the final assessment of enjoyment.

Furthermore, Figure 6.5 shows the mean DTW distance for members within the 'Enjoy' class (blue), the 'Not Enjoy' class (green) and all pairs in opposing classes (red). 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 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 [173].

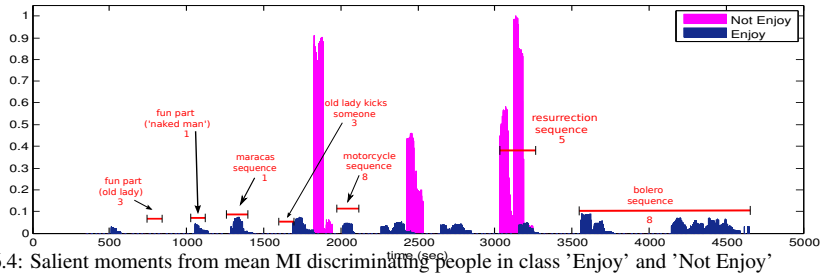


Figure 6.4: Salient moments from mean MI discriminating people in class 'Enjoy' and 'Not Enjoy'

6.4. IMPACT OF PERFORMANCE ON SOCIAL BEHAVIOR

Section 6.3 provided interesting insights into how the response to live performances can be measured with pervasive sensing. However, while working with HD, we came across a

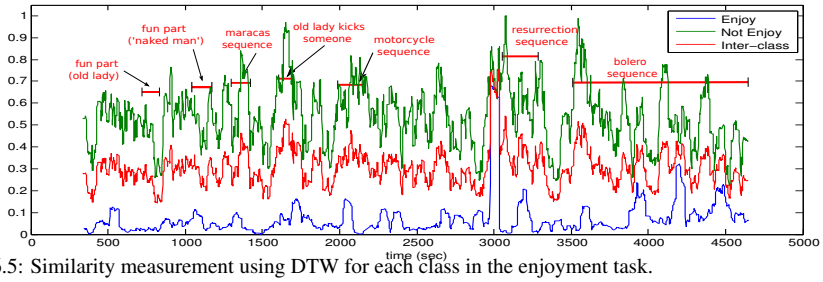


Figure 6.5: Similarity measurement using DTW for each class in the enjoyment task.

different perspective on the problem. Can we quantify the influence of a performance on the audience even after the performance is over?

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 drink afterwards. We hypothesized that people's social behavior (as measured through proximity and acceleration) could also be affected by watching a dance performance. As a business model HD was already co-organizing networking events around dance performances together with two local networking organizations. 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.

To investigate this hypothesis, we decided to investigate whether we could measure differences in how people socialized during the event. Although a networking event is not exactly the same as the more casual ways that people might attend live performances socially, we believe this initial investigation provides a feasibility study for larger scale less controlled studies in the future.

6.4.1. DATA COLLECTION (HDF DATASET)

We measured mingling behavior during two networking sessions, one right before a dance performance and another right after it. With HD and regional networking groups, we co-organized a networking event with 48 volunteers. The same sensing devices (Chalcedony) described in Section 6.3 were used. An example snapshot of the mingling data is shown in Figure 6.6.

Similar to the past datasets, we used proximity sensors as proxies for face-to-face social interactions, together with accelerometer data as described above. Each device used is equipped with a wireless radio, which we used to broadcast 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. Since the device lies on the front of the torso, the radio transmission is shielded by the body, hence restricting the proximity sensing mostly towards the front of the individual. The device logs each detection on the on-board storage along with their timestamps. We used an energy-efficient MAC protocol [54] to allow the devices to communicate their IDs and detect each other's proximity. Because of the unreliability of the wireless medium, we used a density-based filtering technique to increase the sensitivity of the signal for detecting face-to-face proximity [106].

To measure whether people were interacting or not, we processed the proximity detections collected by the devices as follows. For each pair of individuals, we computed the

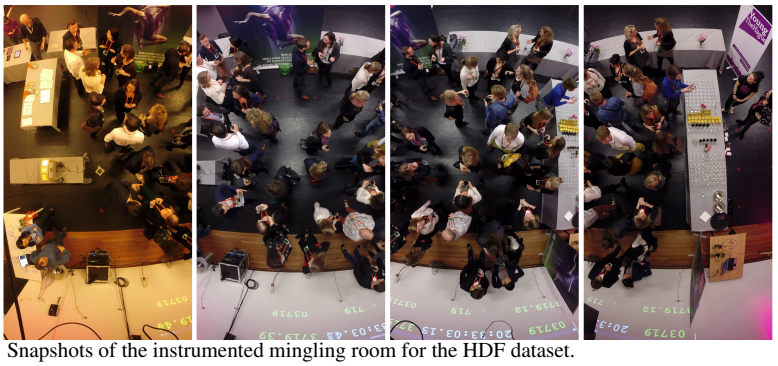


Figure 6.6: Snapshots of the instrumented mingling room for the HDF dataset.

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 non-overlapping time intervals 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 60s to indicate interactions.

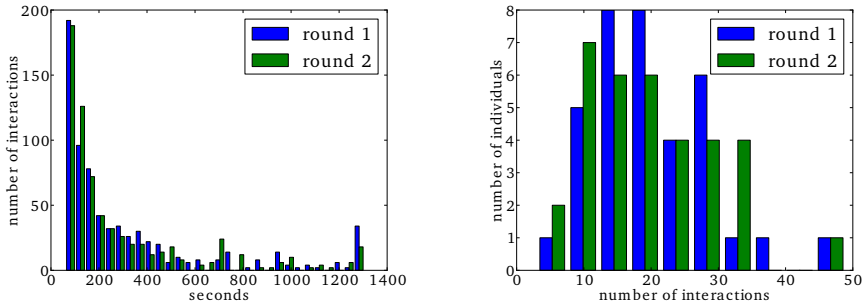
Two months after the experiment, we published sensor analysis results for the volunteers and asked them to answer a survey about their experience of the dance performance and the networking event.

6.4.2. RESULTS

Since some attendees did not attend the entire event, only the data from 35 of the participants was available for analysis. We first hypothesized a difference in the length or the number of interactions between the two sessions. For example, one could imagine that individuals would interact in longer conversations, or with more people. In Figure 6.7(a) we present the distribution of the length of the interactions for the two sessions (from here on referred to as round 1 and round 2) across all the individuals. In both rounds shorter interactions are predominant. Note that as drinks were served at the bar, during both rounds often individuals 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 6.7(b) we present the distribution of the number of distinct interactions for round 1 and round 2 across all the individuals.

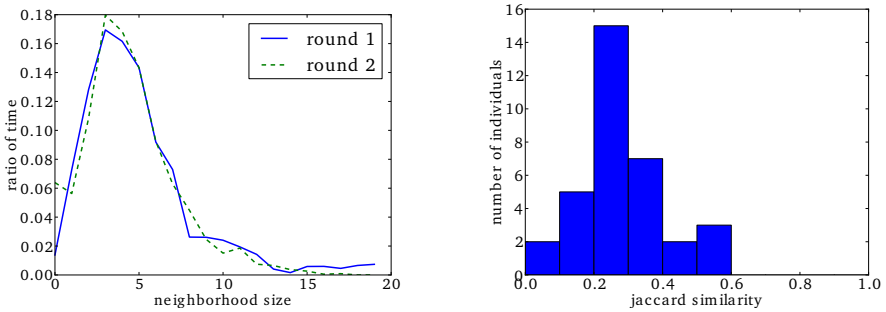
A second difference we hypothesized was in the size of conversational groups. For example, people could be engaged in conversations involving more people, or conversely more one-to-one conversations, perhaps to discuss the content of the performance. We define a *neighborhood* as the set of nodes a sensor a detects at a given moment in time, i.e. the individuals in physical proximity of the individual wearing sensor a. In Figure 6.8(a) we present the distribution of neighborhood 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



(a) (b)

Figure 6.7: (a) Distribution of the lengths of the interactions during the two rounds. (b) Distribution of the number of interactions for round 1 and round 2 across all the individuals.



(a) (b)

Figure 6.8: (a) Relative amount of time sensors detected a certain number of other sensors (at a specific moment in time). (b) Distribution of the jaccard similarity across the individuals.

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 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 6.8(b) 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).

Acceleration

The image emerging from the pure proximity measurement is that of an ordinary mingling event. Overall, these results indicate that the volunteers, as a group, applied a consistent pattern in their mingling behavior during the two rounds, a pattern that they used, however, to target different conversational partners between the two rounds. The measurement pictures a socializing 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 behavior 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 as a result. This correlation supports our hypothesis that the mood change can be linked to implicit behavior as measured by acceleration, though we would like to verify this with a larger dataset later. In conclusion, the results suggest that while individuals acted similarly as a group in terms of networking behavior captured by the proximity sensors, the quality of those interactions seemed different between the two sessions, as captured by the accelerometers.

6.5. DISCUSSION

Our experiments show that using sensors commonly available in smart phones, we are able to leverage group dynamics in the audience to predict an audience's experience of a dance performance and that their implicit responses to the performance can automatically highlight salient moments in it.

In addition, to our knowledge, thus far it is the first time that the context surrounding a performance has also been analysed. Our results suggest that measureable changes in social behaviour before and after watching a dance performance are correlated with an audience's perceptions of how a performance can affect their willingness to socialise as well as their mood. Given that the problem of automatically analysing and quantifying the experience of audience members of live performances is a difficult and multi-faceted domain to conduct experiments, we believe that our first steps show much potential for further devising automated methods to enrich live artistic performances via implicit responses.

6.5.1. OPPORTUNITIES

Our experiments shed light on the huge and under-explored potential of linking implicit responses from pervasive sensing to augment digital signals that already connect with live performance such as verbal expressions of sentiment via social media. Given the potential of acceleration and proximity to be measured pervasively, we identify a number of key areas in which this information, when coupled with multimedia systems could be of real societal benefit.

Enriching the experience of a live performance Live performances tend already to have much multi-media data associated with them such as advertisements of the event on the Internet, video or audio recordings from smart phones, the associated social media, critical reviews of a performance in news or blog posts, attendees and their associated social media profiles, to name but a few. However, in all cases, online responses to an event requires an active and declarative response by audience members. One could easily imagine an audience member's experience of an event could be further enriched by seeing whether

their responses to the performance matched that of others in the audience.

Live performance recommendation By using the acceleration signal generated by the sensor data, we were able to capture spontaneous responses. Our experiments show that the temporal characteristics of the variation in acceleration were similar for people who enjoyed the performance and sufficiently different for those that didn't. An immediate question following this would be whether such response signatures could be used for recommendation via collaborative filtering – if person A who responded similarly at particular moments of a live performance to a person B also like other performances that person B enjoyed?

Benefitting performers One could easily imagine that the salient moments in the performance that were identified automatically in our experiments could be used to verify or highlight key moments of audience response. While our analysis was performed afterwards offline, one could also imagine that such information could be provided in real time to the performers while they are performing. Moreover, the implicit responses from multiple audience members show that even if they do not report certain salient events in the performance to be memorable, their implicit responses can still provide supporting evidence for more sparse explicit survey responses.

Benefitting organisers of live performances For many, going to a live performance involves both the experience of the performance as well as the social event surrounding going to the performance. The national dance organisation that we worked with is aware of this and needs quantitative proof of its benefit. Importantly, from both events that we organised, volunteers paid to take part so they saw inherent value in it. Both experiments that we carried out show that dance performances can have an effect on how people behave and that these responses reflect positive experiences to the performance. This suggests that further studies should be carried out to investigate what triggers people to recommmend dance performance in general and using pervasive sensing provides a realistic means of doing this.

6.5.2. OPEN CHALLENGES

Based on our experiences, it seems that ordinary members of the public are interested in being measured about their responses to a dance performance and having their sensed responses given back to them. However, working with real events with members of the public in dynamic and uncontrolled environments leads to a number of open challenges.

From controlled to uncontrolled large scale measurement First, to obtain large scale measures, we would need to use people's personal smart phones to record their behaviour. Like in our experiments, better responses can probably be gained from hanging their phones around their neck rather than keeping it in a pocket or handbag. Aside from requiring a special pouch to hang the phone around the neck (which could be easily made at low cost and large scale), we are not able to prevent, for example, tampering with the phone during the performance. Moreover, to investigate the role that live performances could have in improving people's social lives, we must also be able to measure their behaviour before and after the performance, which requires further collaboration by audience members unless some other method of incentivisation is provided.

Handling low numbers of explicit responses Implicit responses to a live performance can only be better understood when coupled with survey responses. However, as reported by HD, people tend not to answer surveys about a performance voluntarily unless they have

extreme views about it. There are two ways of considering the problem. One perspective is to address how to more easily obtain even light declarative forms of sentiment about a performance (e.g. 'like' vs. 'dislike') – perhaps socialised incentives could lead to more willingness to report such information. Another perspective would be to consider label propagation techniques to estimate the reaction of the larger unlabelled data.

Subjectivity of responses Perhaps one of the biggest challenges remains in the fact that even if a set of people enjoyed a performance, it is highly probable that they enjoyed different parts. There is no guarantee that responses to the dance performance will be the same for everyone, particular if performances are unstructured or more abstract. Fortunately, the performance that we analysed was less abstract, having quite specific components that could be easily referred to. It remains an open question as to how more abstract performances would be responded to and whether they would as easily analysed.

7

DISCUSSION AND FUTURE WORK

In this thesis, mingle scenarios have been analyzed for 3 tasks in different levels of abstraction (see Figure 2.2 in Chapter 1): gesture detection, personality estimation, and enjoyment estimation. In addition, a novel multisensor dataset for the analysis of human social behavior in crowded scenes was created. Also, a method for the automatic association of video and wearable devices for large group events was proposed, as a necessary first step for further multimodal analysis. In the following subsections, remaining insights of this thesis and future directions in which this work can be extended are discussed.

7.1. SENSING *in-the-wild* SCENARIOS

One of the key aspects in this thesis was the sensing of people interacting in a real scenario without disturbing their natural behavior, thus guaranteeing ecological validity. To do so, ubiquitous technologies in the form of cameras and wearable devices were used throughout this thesis. The creation of *MatchNMingle* (Chapter 2) was an important effort oriented to study social behaviors in real in-the-wild scenarios, by taking the recording space of mingle events from a laboratory or university facility to a real event in a public restaurant.

Nonetheless, one should be critical about the use of the sensing technologies. Were the sensors used in this work truly ubiquitous? Although emulating closely the wearing method of a conference badge, one could argue that our wearables are not yet entirely ubiquitous under the true definition of ubiquitous technology ("development of computing in which computers are sufficiently small and inexpensive to be embedded frequently in everyday objects", taken from Oxford academy). This statement applies for most state-of-the-art smart badges nowadays, both prototypes and commercial [100]. The reason behind this is its current size, weight, and functionality protocols. Our Chalcedony badge, for example, was always seen as an additional/external element to the events. We even saw the case of a device being stored in a purse during one of the events in Chapter 6. Hence, unlike other technologies (e.g. mobile phones or fitness trackers), these wearable devices can be considered pervasive but not yet ubiquitous.

Imagine a fully instrumented conference venue, using true ubiquitous technologies such as small devices embedded in a paper conference badge. It will have no 'lights' or exces-

sive weight that can distract you but still holds the same (or more) functionalities of the Chalcedony devices described here. Current technologies in embedded systems and microelectronics in general are still limited in realizing such an idea, but are steadily evolving towards that end. Here is where the challenges in the ubiquitous domain lie.

Nevertheless, one should always think about the type of activity one is aiming to record, the placement of the sensor to do so, and the implications that a new location for such placement might have on the methods and algorithms. Take our association method in Chapter 3 for example. Other types of applications might opt for a sensor placed on a wrist instead of hung around the neck. For such cases the performance for our association will probably be lower than that stated in Chapter 3, mainly due to the difference in the movement captured by the sensor. While for our case we capture movement from the torso, which is a reflection of the whole body, a sensor on the wrist will focus on movement from the hand. The same applies for the gesture detection (Chapter 4) but in an inverse condition, as a sensor placed on the wrist might be more discriminative for hand gestures than the one used in our experiments. Other works have shown this issue as well [9], emphasizing the importance of the sensor placement for a given task. This critique, however, does not necessarily imply that our methods can not be modified to fit new sensor placement locations.

7.2. MODALITY COMPLEMENTARITY AND SIMILARITY

Although the multiple modalities in this thesis were mostly used in a complementary manner, one can notice that the association process described in Chapter 3 is based on a similarity matching approach. This dichotomy of complementarity and similarity of the streams can be perceived at first hand as contradictory. Nonetheless, there is a key difference while approaching the two that might explain it: the length of the observation intervals.

On the one hand, we have tasks leveraging complementarity (e.g. gesture detection in Chapter 4). Here, the observation intervals are relatively short, ranging from 1 to 3 seconds. On the other hand, for the association of modalities in Chapter 3 the intervals of observations used were of 10 minutes or longer. We also showed that short intervals were not discriminative enough to properly associate the streams as the number of people increased. Furthermore, when comparing streams for similarity purposes we treated the signals as time series, comparing them directly without any previous feature extraction. In contrast, in those chapters where complementarity was exploited we extracted more descriptive features (focusing on the variance of movement) from the times series.

Nonetheless, the exact relationship between the length of the observation windows and this complementary/similarity nature is still unknown. Is this just a matter of length in the time series or there are other aspects affecting this phenomena? The effect of the shared social actions by members in a group on the association task described in Chapter 3 indicates the latter. Thus, the effect of social interactions in the personal streams should also be accounted for.

In this line of research, this work can be further extended to properly analyzing what makes two streams, coming from two different sensors but recording the same action or event, be both correlated and complementary at the same time. Perhaps, an approach oriented to information theory can be a good starting point, aiming to understand the mutual information shared by the streams and how this changes given observation times, preprocessing of other phenomena (e.g. shared actions such as hand shaking).

7.3. RELATIONSHIP BETWEEN THE DIFFERENT LEVELS OF ABSTRACTION

When comparing the different chapters in this thesis, one can notice that each one addressed a different level of abstraction of analysis. We presented a low level signal analysis (stream association in Chapter 3), an analysis at the level of social cue detection (gesture detection in Chapter 4), an analysis of an individual's traits (personality in Chapter 5) and group analysis (group enjoyment in Chapter 6).

Most findings in these works show that although a task can be originally located in a specific level of analysis, these levels tend to intertwine given the nature of the problem as a whole. In this way, an event of higher abstraction (e.g. a social shared experience) seems to affect other more simple tasks at lower levels of abstraction in the analysis.

Take for example the association of acceleration streams presented in Chapter 3. This task can be categorized as a preliminary step for social behavior analysis, and as a low level signal analysis. Nevertheless, it was also shown in Chapter 3 that a better understanding of the shared social actions of people while interacting can further explain those moments when the matching of the streams confuse people in the same group.

Due to this, further work on the analysis of social behavior should always have a global picture of the problem, no matter the level of analysis of the specific task at hand. Tasks as simple as action recognition will be affected by the social component of the events. Thus, they should be analyzed accordingly, considering all the variables that might (or not) affect our current task. For example, are the gestures of a person independent or are they dependent on the conversation he or she is having? We saw in Chapter 4 that occlusions were strongly present in our data, and that gestures from other people cause low confidence and errors in the video classifier. Both these issues are due to the conversation the person is having. So perhaps these issues will be fixed by improving our gesture detection method in Chapter 4, so it also considers the levels of engagement of the persons while they interact, or even uses the gestures from other people in the group to better predict a specific person's gestures.

7.4. USE OF UNCONVENTIONAL MODALITIES FOR BEHAVIORAL ANALYSIS

Due to the crowded and noisy nature of a mingle event, these types of scenarios are hard to analyze for social human behavior with what are considered 'conventional' sensors or modalities. For example, while in other scenarios, using a camera to record facial expressions is somehow straightforward, in crowded in-the-wild events (such as the one presented in Chapter 2) this is not the case. The faces of all participants are not always visible. So, what happens in these cases? We must adapt.

One way of doing so is to use what can be considered unconventional sensor types while analyzing social human behavior, such as wearable devices recording fine-grained acceleration. However, there are two main issues that researchers should consider while doing so: 1) these sensor might record the effects of an action or interaction during an event instead of the action/interaction itself, and 2) the features for such tasks might need further analysis.

First, let us discuss the first point. As mentioned before in Section 7.3, the levels of abstraction on the analysis tends to intertwined. Thus, it is not strange to think that a social behavior, interaction or action (e.g. people conversing in a group) is somewhat represented in the raw signals captured by the sensors, such as the movement signals themselves for our scenario. As a consequence, perhaps in some cases it is wiser to not analyze an interaction or action directly, but its effects in the signals.

Take for example the gesture detection in Chapter 4. For this work we did not apply the 'normal' process for assessing gestures, meaning extracting the position of the hands and/or the skeleton of the participants and use this additional information. Instead, we looked for the effect of gestures on the overall movement of the participants. To do so, we leveraged the movement information from the wearable accelerometer (acceleration) and video (dense trajectories without segmentation between body parts). Thus, we never truly look for the hands or how they moved in either modality, but instead look for its effects on the movement of all the participant's body. The personality estimation in Chapter 5 follows the same principle, measuring the effect of changes in personality from the variance in movement of the participants, instead of directly assessing speaking activity (which we estimate from movement as well), prosody or gaze behavior as has been done for most prior works [168].

Similarly, some other works have already leveraged this premise to detect speech [67] or conversing groups [75] directly from body worn accelerometer signals by looking at the people's shared movements or synchrony. So, this idea can be further extended to overcome the messy and uncontrolled nature of crowded in-the-wild settings.

Regarding point 2 listed earlier, there is an issue with the features which is more of a limitation. Generally, specific sensor types and features have been studied extensively for affective and social behavior analysis. Face analysis, for example, has several state-of-the-art features (e.g. level of activation of different action units) which are based on knowledge from social science and affective computing. Thus, they are meant to address problems in these particular domains. However, their use is limited to rather controlled setups.

Other modalities have not been used for applications specifically oriented to social behavior analysis, much less during in-the-wild scenarios, until very recently. As an example, wearable acceleration was largely used in the past decades for recognizing activities of daily living (ADL). But only until recently have been used for the automated analysis of social behaviors. Thus, the commonly used features in wearables are based on ADL task. However, are these also the best option to address social behavior analysis? This is still an open question. Through this thesis, we use variance of acceleration to extract our features, instead of the raw acceleration. This variance of movement (applied using sliding windows) proved to be a better indication of salient events of movement than the raw acceleration, and was motivated by literature on the relation of movement and arousal [20, 117]. Similarly, future research should also focus on the motivation for their features. And their motivation should be based on knowledge from fields such as social psychology, to better address problems in the Social Signal Processing domain overall.

7.5. TACKLING MISSING AND NOISY DATA

Another effect of working with real scenarios is the presence of missing data. This statement applies for recordings of hours just such as ours (Chapter 2 or [2]) and to efforts

encompassing days of data [83]. Researchers often choose to simply discard samples, reducing considerably their training/test set. And even when present, the data can be noisy due to the acquisition process or the event/activity itself.

In Chapter 2, we showed that the MatchNMingle dataset also presented a significant amount of missing data in one or more modalities, particularly in its first version. This is mainly due to the free and dynamic nature of the event, where people can leave the mingle area at will (e.g. go to the bathroom) or due to device malfunction. In addition, due to the crowded nature of the mingle event, the video has multiple subject cross-contaminations which generate noisy data for each participant. As a consequence, in Chapters 3, 4 and 5 we also dealt with both missing and noisy data.

Thus, in Chapter 3 we saw how our hierarchical association method needed to be modified to account for intervals of missing data, and an additional matching step was needed to compensate for uneven numbers of streams in both modalities. Similarly, not all the video streams used in the personality detection presented in Chapter 5 are complete. Instead, we selected those participants that had at least 50

Moreover, in Chapter 4 we dealt directly with noisy data. The basic difference between missing and noisy data is that for one the sensor failed to provide information, while the latter comprises data where each person's data is affected by an external factor (e.g. occlusion in video). Thus, the use of Multiple Instance Learning (MIL) for the gesture detection in Chapter 4 is mainly based in the challenging noisy data in video due to subject cross-contamination.

Overall, working with missing data brings first a practical challenge as researchers need to account for those sections where data is unavailable. This is also the case for multimodal approaches, where either only complete samples can be used or the method should be capable of working with one or more missing modalities. Moreover, working with missing data also represents a conceptual challenge. Is the remaining information enough to generalize when partial data is missing? The results in Chapter 5 (personality estimation) suggested that this might be the case, until a certain level. Also, the experiments on Chapters 2 and 5 gave a hint to the impact of missing data on the association and estimation tasks.

A straightforward extension of this work then comes with the reconstruction of the missing data exploiting the complementarity of the modalities, instead of just an analysis of the impact of this missing data on the prediction tasks. This is particularly important for fields such as Affective Computing or Social Psychology which aim to give a meaning to the results and impact of the different modalities, and relate these to human behavior. A gradually increasing number of efforts are addressing the problem of missing data in tasks related to behavior analysis [2, 83]. Following the premise of modality complementarity, these works aimed to reconstruct the missing modalities (or labels) using those samples with complete information. To do so, they either applied matrix completion, high dimensional k-means clustering or autoencoders; which leverage the correlations between the features to reconstruct missing data. Other works have formulated the problem using transfer learning [51, 52], treating a type of features as source and those missing as a target domain. Nonetheless, there is still open questions about the complementary nature of different modalities. As we wondered before, when are these modalities complementary? When are they correlated? This will inevitably affect the reconstruction of missing data. Hence, this problem is still inherent to works dealing with real life applications with several open questions.

7.6. ANALYZING TOP-VIEWS FOR MINGLE SCENARIOS

One of the different aspects of the MatchNMingle dataset (Chapter 2) with respect to other works is the use of top view cameras instead of side or elevated views. This is because side views are more prone to participant occlusions, especially for crowded scenes such as ours, and for those people who are farthest away from the camera.

However, the use of a different view comes with its implications. Firstly, the use of state-of-the-art method for common tasks, such as head and body pose or orientation detection, is not straightforward. As discussed in Chapter 2, these methods are generally trained or designed for other type of views, mainly where parts of the head and torso are visible. Thus, they failed for a different domain, such as top views. In addition, the changes of appearance for top views are stronger than in other views, specifically depending on the position of the person with respect to the camera.

A straightforward next step to overcome these challenges is to include examples of top views in the training set for the detection models. This could allow them to generalize to all types of views, up to a certain degree defined by the samples given. Another possible extension is the use of a 3D mapping, which can convert these images to real world coordinates and then perform the detection in this coordinate system.

REFERENCES

[1] X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrinca, E. Ricci, B. Lepri, O. Lanz, and N. Sebe. SALSA: A Novel Dataset for Multimodal Group Behavior Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2015.

[2] X. Alameda-Pineda, Y. Yan, E. Ricci, O. Lanz, and N. Sebe. Analyzing free-standing conversational groups: a multimodal approach. *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2015.

[3] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals. Speaker Diarization: A Review of Recent Research. *IEEE Trans. on Audio, Speech and Language Processing*, 2012.

[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] M. C. Ashton and K. Lee. Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review : An Official Journal of the Society for Personality and Social Psychology*, 11(2), 2007.

[6] M. C. Ashton and K. Lee. The prediction of Honesty – Humility-related criteria by the HEXACO and Five-Factor Models of personality. *Journal of Research in Personality*, 42:1216–1228, 2008.

[7] M. C. Ashton, K. Lee, M. Perugini, P. Szarota, R. E. 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.

[8] M. C. Ashton, K. Lee, and R. E. D. Vries. The HEXACO Honesty-Humility, Agreeableness, and Emotionality Factors: A Review of Research and Theory. *Personality and Social Psychology Review*, 18(2):139–152, 2014.

[9] L. Atallah, B. Lo, R. King, and G. Yang. Sensor placement for activity detection using wearable accelerometers. *International Conference on Body Sensor Networks (BSN)*, 2010.

[10] M. Atzmueller, T. Thiele, G. Stumme, and S. Kauffeld. Analyzing group interaction and dynamics on socio-behavioral networks of face-to-face proximity. *Proc. Intern. Joint Conf. on Pervasive and Ubiquitous Computing (UbiComp): Adjunct*, 2016.

[11] S. O. Ba and J.-M. Odobez. Recognizing visual focus of attention from head pose in natural meetings. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1), 2009.

[12] M. D. Back, L. Penke, S. C. Schmukle, K. Sachse, P. Borkenau, and J. B. Asendorpf. Why Mate choices are not as Reciprocal as We Assume: The Role of Personality, Flirting and Physical Attractiveness. *European Journal of Personality*, 2011.

[13] G. Bahle, P. Lukowicz, K. Kunze, and K. Kise. I see you: How to improve wearable activity recognition by leveraging information from environmental cameras. *IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 2013.

[14] T. Baltrušaitis, P. Robinson, and L. P. Morency. OpenFace: an open source facial behavior analysis toolkit. *IEEE Winter Conference on Applications of Computer Vision*, 2016.

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

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

[17] 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. 2013.

[18] L. M. Batrinca, N. Mana, B. Lepri, F. Pianesi, and N. Sebe. Please, tell me about yourself: automatic personality assessment using short self-presentations. *International Conference on Multimodal Interaction (ICMI)*, 2011.

[19] L. Bennett. Patterns of listening through social media: online fan engagement with the live music experience. *Social Semiotics*, 22(5), 2012.

[20] N. Bianchi-Berthouze.

[21] J. I. Biel and D. Gatica-Perez. The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Transactions on Multimedia*, 15:41–55, 2013.

[22] 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(2), 2012.

[23] R. Bowers, S. Place, P. M. Todd, L. Penke, and J. B. Asendorpf. Generalization in mate-choice copying in humans. *Behavioral Ecology*, 2012.

[24] A. S. Brown and J. L. Novak. *Assessing the intrinsic impacts of a live performance*. WolfBrown San Francisco, CA, 2007.

[25] B. Burkard, M. D.'Amico, and S. Martello. *Assignment problems*. 2009.

[26] L. Cabrera-Quiros, A. Demetriou, E. Gedik, M. v.d. L, 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. *Submitted to IEEE Transactions on Affective Computing and under revision*, Attach as complementary material., 2017.

[27] L. Cabrera-Quiros, E. Gedik, and H. Hung. Estimating self-assessed personality from body movements and proximity in crowded mingling scenarios. *International Conference on Multimodal Interaction (ICMI)*, 2016.

[28] L. Cabrera-Quiros and H. Hung. Who is where? Matching People in Video to Wearable Acceleration During Crowded Mingling Events. *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2016.

[29] L. Cabrera-Quiros and H. Hung. A hierarchical approach for associating body-worn sensors to video regions in crowded mingling scenarios. *Submitted to IEEE Transactions on Multimedia and under revision.*, 2018.

[30] L. Cabrera-Quiros, D. Tax, and H. Hung. Gestures in-the-wild: detecting conversational hand gestures in crowded scenes with bags of video trajectories and body worn acceleration. *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2018.

[31] B. Calvo-Merino, C. Jola, D. E. Glaser, and P. Haggard. Towards a sensorimotor aesthetics of performing art. *Consciousness and cognition*, 17(3), 2008.

[32] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[33] 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*, 2010.

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

[35] F. Celli, E. Bruni, and B. Lepri. Automatic Personality and Interaction Style Recognition from Facebook Profile Pictures. *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2014.

[36] A. Cerekovic, O. Aran, and D. Gatica-Perez. Rapport with virtual agents: What do human social cues and personality explain? *IEEE Transactions on Affective Computing*, 2017.

[37] I. Chamveha, Y. Sugano, Y. Sato, and A. Sugimoto. Social Group Discovery from Surveillance Videos: A Data-Driven Approach with Attention-Based Cues., 2013.

[38] Y. Chen, J. Bi, and J. Z. Wang. MILES: Multiple-Instance Learning via Embedded Instance Selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2006.

[39] T. Choudhury and A. Pentland. Sensing and Modeling Human Networks using the Sociometer. *IEEE Intern. Symposium on Wearable Computers*, 2003.

[40] G. G. Chowdhury. Natural language processing. *Annual Review of Information Science and Technology*, 2003.

[41] D. Cook, K. D. Feuz, and N. C. Krishnan. Transfer learning for activity recognition: A survey. *Knowledge and information systems*, (3), 2013.

[42] J. C. Cooper, S. Dunne, T. Furey, and J. P. O. Doherty. Dorsomedial Prefrontal Cortex Mediates Rapid Evaluations Predicting the Outcome of Romantic Interactions. *Journal of Neuroscience*, 2012.

[43] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, A. Tosato, A. Del Bue, G. Menegaz, and V. Murino. Social interaction discovery by statistical analysis of F-formations. *British Machine Vision Conference (BMVC)*, 2011.

[44] M. Cristani, R. Raghavendra, A. D.Buea, and V. Murino. Human behavior analysis in video surveillance: A Social Signal Processing perspective. *Neurocomputing*, 2013.

[45] L. J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, 1951.

[46] 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, 2011.

[47] 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*, 2012.

[48] A. Demetriou. Rose Colored Lenses: The Role of Testosterone and Cortisol in Mate Assessment. Master's thesis, VU University Amsterdam, 2015.

[49] L. der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

[50] J. M. Digman. Personality structure: Emergence of the five-factor model. *Annual review of psychology*, 1990.

[51] Z. Ding, M. Shao, and Y. Fu. Latent low-rank transfer subspace learning for missing modality recognition. *Proceedings of the Conference on Artificial Intelligence (AAAI)*, 2014.

[52] Z. Ding, M. Shao, and Y. Fu. Missing modality transfer learning via latent low-rank constraint. *IEEE Transactions on Image Processing*, 2015.

[53] M. Dobson. *Low-power epidemic communication in wireless ad hoc networks*. PhD thesis, Amsterdam: Vrije Universiteit, 2013.

[54] 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.

[55] W. Dong, B. Lepri, F. Pianesi, and A. Pentland. Modeling Functional Roles Dynamics in Small Group Interactions. *IEEE Transactions on Multimedia*, 2013.

[56] C. Doukas, I. Maglogiannis, P. Tragas, D. Liapis, and G. Yovanof. Patient fall detection using support vector machines. *Artificial Intelligence and Innovations 2007: from Theory to Applications*, 2007.

[57] P. Duin R.P.W. Juszcak, P. Paclik, E. Pekalska, D. de Ridder, and D. M. J. Tax. PRTools, A {M}atlab Toolbox for Pattern Recognition, mar 2017.

[58] G. Englebienne and H. Hung. Mining for Motivation: Using a single wearable accelerometer to detect people's interests. *International workshop on Interactive multimedia on mobile and portable devices (IMMPD)*, 2012.

[59] S. Escalera, X. Baró, J. Gonzalez, M. A. Bautista, M. Madadi, V. Ponce-López, H. Escalante, J. Shotton, and I. Guyon. ChaLearn Looking at People Challenge 2014: Dataset and Results. *Workshop at the European Conference on Computer Vision (ECCV)*, 2014.

[60] S. Escalera, J. Gonzalez, X. Baró, M. Reyes, O. Lopes, I. Guyon, V. Athitsos, and H. Escalante. Multi-modal gesture recognition challenge 2013: dataset and results. *International conference on multimodal interaction (ICMI)*, 2013.

[61] S. Escalera, R. M. Martinez, J. Vitria, P. Radeva, and M. T. Anguera. Dominance detection in face-to-face conversations. *Conference on Computer Vision and Pattern Recognition Workshops (CVPR)*, 2009.

[62] J. Fleureau, P. Guillotel, and I. Orlac. Affective Benchmarking of Movies Based on the Physiological Responses of a Real Audience. 2013.

[63] D. Fujiwara, L. Kudrna, and P. Dolan. Quantifying and Valuing the Wellbeing Impacts of Culture and Sport. Technical report, UK Department of Culture, Media and Sport, Apr. 2014.

[64] R. M. Furr and Bacharach R. *Psychometrics An Introduction*. SAGE Publications Ltd, Thousand Oaks, CA, second edition, 2014.

[65] D. Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: A review. *Image and Vision Computing*, 2009.

[66] D. Gatica-Perez, O. Aran, and D. Jayagopi. Small Group Analysis. *Social Signal Processing. Cambridge Univ. Press*, 2017.

[67] E. Gedik and H. Hung. Speaking status detection from body movements using transductive parameter transfer. *Proccedings of the International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp): Adjunct*, 2016.

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

[69] R. Gonzalez and R. Woods. *Digital Image Processing*. Prentice Hall, 2008.

[70] K. Grammer, M. Honda, Juette A., and A. Schmitt. Fuzziness of nonverbal courtship communication unblurred by motion energy detection. *Journal of Personality and Social Psychology*, 1999.

[71] K. Grammer and R. Thornhill. Human (Homo-Sapiens) Facial Attractiveness and Sexual Selection - the Role of Symmetry and Averageness. *Journal of Comparative Psychology*, 108(3), 1994.

[72] P. Hall, Y. Hao, V. I. Nechayev, A. Alomain, C. C. Constantinou, C. Parini, M. R. Kamarudin, T. Z. Salim, D. T. M. Heel, R. Dubrovka, A. S. Owadall, W. Song, A. Serra, P. Nepa, M. GalIo, and M. Bozzetti. Antennas and Propagation for On-Body Communication Systems. *IEEE Antennas and Propagation Magazine*, 2007.

[73] R. Hinde. *Non-verbal communication*. Cambridge University Press, 1972.

[74] H. Hung, G. Englebienne, and J. Kools. Classifying Social Action with a Single Accelerometer. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, 2013.

[75] H. Hung, E. G., and L. Cabrera-Quiros. Detecting Conversing Groups with a Single Worn Accelerometer. *ACM International Conference on Multimodal Interaction (ICMI)*, 2014.

[76] H. Hung and D. Gatica-Perez. Estimating Dominance In Multi-Party Meetings Using Speaker Diarization. *IEEE Transactions on Audio, Speech and Language Processing*, 2011.

[77] H. Hung and D. Gatice-Perez. Estimating Cohesion in Small Groups Using Audio-Visual Nonverbal Behavior. *IEEE Transactions on Multimedia*, 2010.

[78] H. Hung, Y. Huang, C. Yeo, and D. Gatica-Perez. Associating Audio-Visual Activity Cues in a Dominance Estimation Framework. *Computer Vision and Pattern Recognition Workshop on Human Communicative Behaviour*, 2008.

[79] H. Hung and B. Kröse. Detecting F-Formations as Dominant Sets. *ACM International Conference on Multimodal Interaction (ICMI)*, 2011.

[80] T. Huynh and B. Schiele. Analizing features for activity recognition. *Joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies*, 2005.

[81] M. E. Ireland, R. B. Slatcher, P. W. Eastwick, L. E. Scissors, E. J. Finkel, and J. W. Pennebaker. Language Style Matching Predicts Relationship Initiation and Stability. *Psychological Science*, 2011.

[82] 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*, 2011.

[83] N. Jaques, S. Taylor, and A. S. R. Picard. Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction. *International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017.

[84] D. 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*, 2009.

[85] D. Jurafsky, R. Ranganath, and D. McFarland. Extracting social meaning: identifying interactional style in spoken conversation. *Proc. of Human Language Technologies: The Annual Conf. of the North American Chapter of the Association for Comp. Linguistics*, 2009.

[86] A. B. Kahng. Scaling: More than Moore's law. *IEEE Design & Test of Computers*, 2010.

[87] K. Kalimeri, B. Lepri, and F. Pianesi. Going beyond traits: multimodal classification of personality states in the wild. *International Conference on Multimodal Interaction (ICMI)*, 2013.

[88] A. Kendon. *Conducting interaction: Patterns of behavior in focused encounters*, volume 7. CUP Archive, 1990.

[89] A. Kendon. *Gesture: Visible Action as Utterance*. Cambridge University Press, 2015.

[90] D. A. Kenny. Interpersonal Perception: A Social Relations Analysis. *Journal of Social and Personal Relationships*, 1994.

[91] J. Kooij, M. Liem, J. Krijnders, T. Andringa, and D. Gavrila. Multi-modal human aggression detection. *Computer Vision and Image Understanding*, 2016.

[92] R. M. Krauss, Y. Chen, and P. Chawla. Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? 1996.

[93] J. R. Kwapisz, G. M. Weiss, and S. A. Moore. Activity recognition using cell phone accelerometers. *ACM SIGKDD Explorations Newsletter*, 12(2), 2011.

[94] O. D. Lara and M. A. Labrador. A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys and Tutorials*, 2013.

[95] C. Latulipe, E. A. Carroll, and D. Lottridge. Love, hate, arousal and engagement: exploring audience responses to performing arts. 2011.

[96] O. Lederman, D. Calacci, A. MacMullen, D. C. Fehder, F. E. Murray, and A. Pentland. Open Badges: A Low-Cost Toolkit for Measuring Team Communication and Dynamics. *International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction and Behavior Representation in Modeling and Simulation*, 2016.

[97] K. Lee and M. C. Ashton. Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research*, 2004.

[98] B. Lepri, A. Mana N. an Cappelletti, F. Pianesi, and M. Zancanaro. Modeling the Personality of Participants During Group Interactions. *International Conference on User Modeling, Adaptation, and Personalization*, 2009.

[99] B. Lepri, R. Subramanian, K. Kalimeri, J. Staiano, F. Pianesi, and N. Sebe. Connecting meeting behavior with extraversion - a systematic study. *IEEE Transactions on Affective Computing*, 2012.

[100] K. Lyytinen and Y. Yoo. Issues and challenges in ubiquitous computing. *Communications of the ACM Ubiquitous Computing*, 2002.

[101] A. Madan, R. Caneel, and A. Pentland. Voices of Attraction. 2004.

[102] N. Mana, B. Lepri, P. Chippendale, A. Cappelletti, F. Pianesi, P. Svaizer, and M. Zancanaro. Multimodal Corpus of Multi-Party Meetings for Automatic Social Behavior Analysis and Personality Traits Detection. *Proceedings of the workshop on Tagging, mining and retrieval of human related activity information*, 2007.

[103] A. Marcos-Ramiro, D. P. Perez, M. Marron-Romera, and D. Gatica-Perez. Capturing Upper Body Motion in Conversation: an Appearance Quasi-Invariant Approach. *International Conference on Multimodal Interaction (ICMI)*, 2014.

[104] A. Marcos-Ramiro, D. Pizarro-Perez, M. Marron-Romera, L. Nguyen, and D. Gatica-Perez. Body Comunicative Cue Extraction for Conversational Analysis. *IEEE International Conference on Face and Gesture Recognition (FG)*, 2013.

[105] C. Martella, M. Cattani, and M. v. Steen. Exploiting Density to Track Human Behavior in Crowded Environments. *IEEE Communications Magazine*, 2017.

[106] C. Martella, M. Dobson, A. van Halteren, and M. Van Steen. From proximity sensing to spatial-temporal social graphs. *IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 2014.

[107] C. Martella, A. van Halteren, M. van Steen, C. Conrado, and J. Li. Crowd Textures as Proximity Graphs. 2014.

[108] A. Matic, V. Osmani, and A. Maxhuni. Multi-modal mobile sensing of social interactions. *Intern. Conf. on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, 2012.

[109] D. A. McFarland, D. Jurafsky, and C. Rawlings. Making the Connection: Social Bonding in Courtship Situations. *American Journal of Sociology*, 2013.

[110] D. McNeill. *Hand and Mind: what gestures reveal about thought*. The University of Chicago Press, 1992.

[111] D. McNeill. *Language and gesture*. Cambridge University Press, 2000.

[112] M. Mehl, S. Gosling, and J. Pennebaker. Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, 2006.

[113] A. C. Michalos and P. M. Kahlke. Arts and the perceived quality of life in British Columbia. *Social indicators research*, 96, 2010.

[114] G. Mohammadi and A. Vinciarelli. Automatic personality perception: Prediction of trait attribution based on prosodic features. *IEEE Transactions on Affective Computing*, 2012.

[115] G. Mohammadi, A. Vinciarelli, and M. Mortillaro. The voice of personality: Mapping nonverbal vocal behavior into trait attributions. *Proceedings of the International workshop on Social signal processing*, 2010.

[116] M. M. Moore. Nonverbal courtship patterns in women: Context and consequences. *Ethology and Sociobiology*, 6(4):237–247, 1985.

[117] F. Mueller, S. Agamanolis, and R. Picard. Exertion interfaces: sports over a distance for social bonding and fun. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2003.

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

[119] 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*, (0), 2014.

[120] L. T. Nguyen, Y. S. Kim, P. Tague, and J. Zhang. IdentifyLink: User-device linking through visual and RF-signal. *ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, 2014.

[121] G. Noppe, Y. B. De Rijke, K. Dorst, E. L. T. Van Den Akker, and E. F. C. Van Rossum. LC-MS/MS-based method for long-term steroid profiling in human scalp hair. *Clinical Endocrinology*, 83(2):162–166, 2015.

[122] 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(1):50–69, 2010.

[123] L. Penke, P. M. P. Todd, A. P. Lenton, and B. Fasolo. How Self-assessments can guide human mating decisions. *G. Geher & G. F. Miller (Eds.), Mating Intelligence: New insights into intimate relationships, human sexuality, and the mind's reproductive system.*, 2008.

[124] D. W. Pentico. Assignment problems: A golden anniversary survey. *European Journal of Operational Research*, 2005.

[125] A. Pentland. Social Dynamics: Signals and Behavior. *ICDL, IEEE Press*, 2004.

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

[127] S. S. Place, P. M. Todd, J. Zhuang, L. Penke, and J. B. Asendorpf. Judging romantic interest of others from thin slices is a cross-cultural ability. *Evolution and Human Behavior*, 2012.

[128] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 1999.

[129] T. Plötz, C. Chen, N. Y. Hammerla, and G. D. Abowd. Automatic synchronization of wearable sensors and video-cameras for ground truth annotation - A practical approach. *International Symposium on Wearable Computers*, 2012.

[130] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera. ChaLearn LAP 2016: First Round Challenge on First Impressions-Dataset and Results. *Computer Vision–ECCV Workshops*, 2016.

[131] E. Principi, R. Rotili, M. Wöllmer, S. Squartini, and B. Schuller. Dominance Detection in a Reverberated Acoustic Scenario. *International Symposium on Neural Networks*, 2012.

[132] F. Quek, D. McNeill, R. Bryll, S. Duncan, X. F. Ma, C. Kirbas, K. E. McCullough, and R. Ansari. Multimodal Human Discourse: Gesture and Speech. *ACM Transactions on Computer-Human Interaction*, 2002.

[133] N. Raiman, H. Hung, and G. Engliebienne. Move, and I will tell you who you are: Detecting deceptive roles in low-quality data. *International Conference on Multimodal Interaction (ICMI)*, 2011.

[134] R. Ranganath, D. Jurafsky, and D. McFarland. It's not you, it's me: detecting flirting and its misperception in speed-dates. *Proc. of Conf. on Empirical Methods in Natural Language Processing*, 2009.

[135] R. Ranganath, D. Jurafsky, and D. McFarland. Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates. *Computer Speech & Language*, 2013.

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

[137] N. Ravi, N. Dandekar, P. Mysore, and M. L. Littman. Activity recognition from accelerometer data. *Procedings of the conference on Innovative applications of artificial intelligence (AAAI)*, 2005.

[138] M. Reason and D. Reynolds. Kinesthesia, empathy, and related pleasures: An inquiry into audience experiences of watching dance. *Dance Research Journal*, 42(02), 2010.

[139] S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava. Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks (TOSN)*, 6(2):13, 2010.

[140] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You Only Look Once: Unified, Real-Time Object Detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[141] E. Ricci, J. Varadarajan, R. Subramanian, S. R. Buló, N. Ahuja, and O. Lanz. Uncovering interactions and interactors: Joint estimation of head, body orientation and f-formations from surveillance video. *International Conference on Computer Vision (ICCV)*, 2015.

[142] M. Rofouei, A. D. Wilson, A. J. B. Brush, and S. Tansley. Your Phone or Mine? Fusing Body, Touch and Device Sensing for Multi-User Device-Display Interaction. *ACM Conference for Human-Computer Interaction (CHI)*, 2012.

[143] 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.

[144] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision (IJCV)*, 40(2):99–121, 2000.

[145] D. Sanchez-Cortez, O. Aran, M. Schmid, and D. Gatica-Perez. A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Transactions on Multimedia*, 2012.

[146] R. R. Schaller. Moore's law: past, present and future. *IEEE Spectrum*, 1997.

[147] T. Schubert, F. Friedmann, and H. Regenbrecht. The experience of presence: Factor analytic insights. *Presence*, 10(3):266–281, 2001.

[148] E. W. K. 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(8):1484–1494, 2012.

[149] F. Setti, H. Hung, and M. Cristani. Group detection in still images by F-formation modeling: A comparative study. *International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 2013.

[150] F. Setti, O. Lanz, R. Ferrario, V. Murino, and M. Cristani. Multi-scale f-formation discovery for group detection. *International Conference on Image Processing (ICIP)*, 2013.

[151] F. Setti, C. Russell, C. Bassetti, and M. Cristani. F-Formation Detection: Individuating Free-Standing Conversational Groups in Images. *PLoS ONE*, 2015.

[152] O. Shigeta, S. Kagami, and K. Hashimoto. Identifying a moving object with an accelerometer in a camera view. In *IROS*, 2008.

[153] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *Affective Computing, IEEE Transactions on*, 3(1), 2012.

[154] S. Stein and S. Mckenna. Combining embedded accelerometers with computer vision for recognition food preparation activities. *ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, 2013.

[155] C. Stevens, H. Winskel, C. Howell, L. Vidal, J. Milne-Home, and C. Latimer. Direct and indirect methods for measuring audience reactions to contemporary dance. 2009.

[156] R. Stewart, M. Andriluka, and A. Y. Ng. End-to-end people detection in crowded scenes. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[157] R. Subramanian, Y. Yan, J. Staiano, O. Lanz, and N. Sebe. On the relationship between head pose, social attention and personality prediction for unstructured and dynamic group interactions. *Proc. of the 15th ACM on Intern. Conf. on multimodal interaction*, pages 3–10, 2013.

[158] J. P. Tangney, R. F. Baumeister, and A. L. Boone. High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality*, 72(2), 2004.

[159] T. Teixeira, D. Jung, and A. Savvides. Tasking networked CCTV cameras and mobile phones to identify and localise multiple persons. *ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, 2010.

[160] M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment in Twitter Events. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY*, 62(2), 2011.

[161] N. D. Tidwell, P. W. Eastwick, and E. J. Finkel. Perceived, not actual, similarity predicts initial attraction in a live romantic context: Evidence from the speed-dating paradigm. *Personal Relationships*, 2013.

[162] T. Vacharkulksemsuk, E. Reit, P. Khambatta, P. W. Eastwick, E. J. Finkel, and D. R. Carney. Dominant, open nonverbal displays are attractive at zero-acquaintance. *Proc. of the National Academy of Sciences*, 2016.

[163] K. Valentine, N. P. Li, L. Penke, and D. I. Perrett. Judging a Man by the Width of his Face: the Role of Facial Ratios and Dominance in Mate Choice at Speed-dating Events. *Psychological Science*, 2014.

[164] F. Vallet, S. Essid, and J. Carrive. A Multimodal Approach to Speaker Diarization on TV Talk-Shows. *IEEE Transactions on Multimedia*, 2013.

[165] J. C. van Gemert, M. Jain, E. Gati, and C. G. M. Snoek. APT: Action localization Proposals from dense Trajectories. *British Machine Vision Conference (BMVC)*, 2015.

[166] S. Vascon, E. Mequanint, M. Cristani, H. Hung, M. Pelillo, and V. Murino. A Game-Theoretic Probabilistic Approach for Detecting Conversational Groups. In *Asian Conference on Computer Vision (ACCV)*, 2014.

[167] A. Veenstra and H. Hung. Do They Like Me? Using Video Cues to Predict Desires during Speed-dates. *IEEE Intern. Conf. on Computer Vision Workshops (ICCV Workshops)*, 2011.

[168] A. Vinciarelli and G. Mahammadi. A survey of personality computing. *IEEE Transactions on Affective Computing*, 2014.

[169] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 2009.

[170] 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*, 2012.

[171] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation. *Intern. Journal of Computer Vision (IJCV)*, 2012.

[172] J. Wan, S. Escalera, X. Baro, H. J. Escalante, I. Guyon, M. Madadi, J. Allik, J. Gorbova, and G. Anbarjafari. Results and analysis of chalearn lap multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenges. *ChaLearn LaP, Action, Gesture, and Emotion Recognition Workshop and Competitions: Large Scale Multimodal Gesture Recognition and Real versus Fake expressed emotions, ICCV*, 2017.

[173] 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.

[174] C. Wang, E. Geelhoed, P. Stenton, and P. S. Cesar Garcia. Sensing a live audience. 2014.

[175] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action Recognition by Dense Trajectories. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 3169–3176, Colorado Springs, United States, jun 2011.

[176] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense Trajectories and Motion Boundary Descriptors for Action Recognition. *International Journal of Computer Vision*, 2013.

[177] A. Wilson and H. Benko. CrossMotion:fusing device and image motion for user identif., tracking and device association. *International Conference on Multimodal Interaction (ICMI)*, 2014.

[178] M. Wirz, D. Roggen, and G. Tröster. A Methodology towards the Detection of Collective Behavior Patterns by Means of Body-Worn Sensors. 2010.

[179] K. W.Z., Y. Xiang, M. Y. Aalsalem, and Q. Arshad. Mobile Phone Sensing Systems: A Survey. *IEEE comm. surveys and tutorials*, 2013.

[180] Y. Xiong and F. Quek. Hand motion gesture frequency properties and multimodal discourse analysis. *International Journal of Computer Vision*, 2006.

[181] Y. Xiong, F. Quek, and D. McNeill. Hand Gesture Symmetric Behavior Detection and Analysis in Natural Conversation. *International Conference on Multimodal Interaces (ICMI)*, 2002.

[182] G. Zen, B. Lepri, E. Ricci, and O. Lanz. Space speaks: towards socially and personality aware visual surveillance. *International workshop on Multimodal pervasive video analysis (MPVA)*, 2010.

[183] G. Zen, E. Sangineto, E. Ricci, and N. Sebe. Unsupervised domain adaptation for personalized facial emotion recognition. In *International Conference on Multimodal Interaction (ICMI)*, 2014.

[184] L. Zhang and L. Van Der Maaten. Structure Preserving Object Tracking. *IEEE Conputer Vision and Pattern Recognition (CVPR)*, 2013.

[185] 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*, 2006.

SUMMARY

The automated analysis of human non-verbal behavior during crowded mingle scenarios is part of the newly emerged domain of Social Signal Processing (SSP). This specific line of research aims to develop computational methods to automatically understand social interactions in-the-wild, while facing the many challenges inherent with the noisy nature of mingle scenarios.

While most works about the analysis of social interactions are focused on structured and task-driven setups such as small group meetings, mingle scenarios consist of free-standing conversational groups that dynamically form, merge and split aligning with the participants' intentions and desires.

Data collected in structured scenarios is rather clean, whereas mingle scenarios have frequent and heavy subject cross-contamination as well as missing data due to the inherent crowded and dynamic nature of the events, with people mingling freely. The goal of this thesis is to leverage multiple modalities for the analysis of social interactions during crowded mingle scenarios, to overcome these challenges.

The approach taken in this thesis is to record mingling events with overhead cameras and wearable sensors recording body acceleration and proximity, to be minimally intrusive and to scale rather easily to a higher number of people. We focused on different tasks for the understanding of social interactions such as automatic association of multiple modalities, detection of social hand gestures, personality estimation, and group enjoyment.

We show that the use of multiple modalities improves the performance of our classification tasks and the understanding of social interactions, compared to unimodal approaches. This was particularly important when data in one of the modalities was noisy or completely missing.

SAMENVATTING

De automatische analyse van menselijk non-verbaal gedrag tijdens drukke informele ontmoetingen maakt deel uit van het nieuwe domein van Social Signal Processing (SSP). Deze onderzoeksrichting heeft als doel computationele methoden te ontwikkelen om automatisch realistische sociale interacties te begrijpen. Hier zijn veel uitdagingen mee gemoeid die voortkomen uit de ruizige natuur van sensor data verzameld tijdens deze drukke ontmoetingen.

Gebruikelijk is om sociale interacties te bestuderen in gestructureerde of doelgerichte scenario's. Hierbij worden echter de participanten beperkt in hun bewegingsvrijheid of worden vaak kleine groepen mensen bestudeerd. Tijdens drukke informele ontmoetingen is er echter een grote dynamiek waarbij groepen zich vormen, samenvoegen en splitsen naar gelang de behoeften van de deelnemers.

In gestructureerde scenario's is het gemakkelijker om onvervuilde data te verzamelen van elke participant. Tijdens drukk informele ontmoetingen is dit echter veel lastiger. Visuele data verzameld van de ontmoetingen bevatten vaak meerdere deelnemers vanwege de drukte van de ontmoetingen. Hierdoor kan sensor data van verschillende participanten met elkaar vermengen. Verder kan informatie over participanten geheel ontbreken, omdat de mensen zich vrij kunnen bewegen. Het doel van dit proefschrift is om sociale interacties tijdens drukke informele ontmoetingen te analyseren en de problemen hierbij op te lossen door meedere modaliteiten te meten.

Om dit te bewerkstelligen gebruiken we camera's bevestigd aan het plafond en draagbare sensoren die bewegingen van het lichaam registreren en nabijheid tussen participanten meten. Deze aanpak is onopvallend voor participanten en schaalt gemakkelijk naar grote groepen mensen. We richten ons op het ontrafelen van sociale interacties op verschillende niveaus van abstractie, bijvoorbeeld automatische associatie van verschillende typen sensoren, het detecteren van handgebaren, persoonlijkheidsinschatting en het schatten van gezelligheid van een groep.

Met de analyses in dit proefschrift tonen we aan dat het gebruik van meerdere modaliteiten de prestaties van onze classificatietaken verbeterd, in vergelijking met methoden die slechts individuele modaliteiten gebruiken. Dit was met name belangrijk als een modaliteit ruis bevatte of als participanten geheel ontbraken.

CURRICULUM VITAE

Laura was born in 1988 in San José, Costa Rica. She received her *Licenciatura* as Electronic Engineer in 2012 from the Instituto Tecnólogico de Costa Rica. In 2014, she received her Master degree (with honors) as Electronic Engineer with emphasis on embedded systems, also from the Instituto Tecnólogico de Costa Rica.

She worked as a junior lecturer at the Instituto Tecnólogico de Costa Rica from 2012 to 2014, and was part of the Bounce Imaging start-up company from 2013 to 2014. While working at the Instituto Tecnólogico de Costa Rica, she received a full scholarship from the Costa Rican government to pursue her postgraduate studies abroad. Thus, she started in 2014 as a PhD candidate at the Pattern Recognition and BioInformatics Group at Delft University of Technology, under the supervision of Dr. Hayley Hung.

Laura is currently a postdoctoral researcher at the Pattern Recognition and BioInformatics Group at Delft University of Technology, in the Netherlands. Her main interests are the use and fusion of wearable sensing and computer vision techniques for applications mainly oriented (but no limited) to the analysis of human social behavior.

LIST OF PUBLICATIONS

Journals

1. **L. Cabrera-Quiros**, D. Tax and H. Hung. *Gestures in-the-wild: detecting conversational hand gestures in crowded scenes with bags of video trajectories and body worn acceleration*. Submitted to IEEE Transactions on Multimedia. Under Revision.

2. **L. Cabrera-Quiros**, E. Gedik and H. Hung. *Multimodal self-assessed personality estimation during crowded mingle scenarios using wearables devices and cameras*. Submitted to IEEE Transactions on Affective Computing. Under Revision.

3. **L. Cabrera-Quiros***, A. Demetriou*, E. Gedik, L. v.d. 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*. To appear in IEEE Transactions on Affective Computing, 2018.

4. **L. Cabrera-Quiros** and H. Hung. *A hierarchical approach for associating body-worn sensors to video regions in crowded mingling scenarios*. Submitted to IEEE Transactions on Multimedia. Under Revision.

5. E. Gedik, **L. Cabrera-Quiros**, C. Martella, G. Englebienne and H. Hung. *Analyzing and Predicting the Experience of Live Performances with Wearable Sensing*. Submitted to IEEE Transactions on Affective Computing. Under Revision.

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 International Conference on Multimodal Interaction (ICMI), oral presentation, 2016.

2. **L. Cabrera-Quiros**. *Towards multimodal analysis of human behavior in crowded mingling scenarios using movement cues from wearable sensors and cameras.* Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp): Adjunct, 2016. Doctoral School.

3. **L. Cabrera-Quiros** and H. Hung. *Who is where?: Matching People in Video to Wearable Acceleration During Crowded Mingling Events*. Proceedings of the ACM International Conference on Multimedia (ACM MM), 2016.

4. C. Martella*, E. Gedik*, **L. Cabrera-Quiros**[*1], G. Englebienne and H. Hung. *How was it?: exploiting smartphone sensing to measure implicit audience responses to live performances*. Proceedings of the ACM International Conference on Multimedia (ACM MM), Brave New Ideas paper, 2015.

5. K. Schellekens, E. Giaccardi, D. Day, H. Hung, **L. Cabrera-Quiros**, E. Gedik and C. Martella. *Impact of connected objects on social encounters*. Participatory Innovation Conference (PIN), 2015.

6. H. Hung, G. Englebienne and **L. Cabrera-Quiros**. *Detecting conversing groups with a single worn accelerometer*. Proceedings of the International Conference on Multimodal Interaction (ICMI), 2014.

[1] *Authors contributed equally in this paper.

ACKNOWLEDGEMENTS

At the beginning of the PhD, I was given a book called 'Sink or swim, mastering your PhD'. I remember it as a curious moment: could it really be that hard? Turns out yes. Here, I would like to thank all the people that help me keep swimming through it.

First and foremost, I must thank the Instituto Tecnológico de Costa Rica, which gave me the opportunity and financial support to pursue my doctoral studies. I must express my appreciation to Dr. Julio Calvo Alvarado, Ing. Luis Paulino Méndez Badilla, Dr. Roberto Pereira Arroyo, Ing. Francisco Navarro Henríquez, Ing. Milton Villegas Lemus, Ing. Arys Carrasquilla Batista and all members of the Electronic engineering, Computer engineering, Mechatronics engineering, and Scholarships departments. Without your support and good disposition, this work would not have been possible.

To my promotor. Marcel, thank you for giving a crazy electronics engineer from Costa Rica the opportunity to join a pattern recognition group, with all that entails. From the challenging discussions I had with you, I have learned to debate in an amazing Dutch style. I hope I inherit at least half of your resolution, pragmatism and amazing managing skills.

Hayley, my boss (I know, you hate that word) and daily supervisor. It was a true pleasure to become one of your 'children' for the past 4 years. It was evident from the start that you truly cared for my professional and personal development. I have learned a ton from you, from breathing deeply and taking a walk before answering to a reviewer, to the beauty behind the concept of a research question. All the pep talks and discussions about becoming a true member of academia will always stay with me. So, I think it is safe to say that you have successfully turned me to the dark side. And while I might still miss a 'd' at the end of some words when in a rush, I will never forget your patience towards me when doing such mistakes.

I would like to thank all the people in the Pattern Recognition and Bioinformatics group: Ekin (the hipster), Sally (freedom!), Alex (the bodybuilder), Tom (the loud Dutch), Wouter (the 14 year old), Christian (the Gulden Draak drinker), Wenjie (the keyboard guy), Stephanie (Hayley's new kid), Yanxia (my Zumba buddy), Bernd (all kindness), Jesse (the self-designated driver), Veronika (the social butterfly), Marieke (the accordionist), Gorkem (the stereo-man), Hamdi (advise giver), Yazhou (the active learning expert), Taygun (the word master), Osman (office and house neighbor), Seyran (our motherly adviser), Ahmed (our role model), Stravos (the witty sense of humor), Arlin (the band girl), Christine (no filter, amazing!), Tom (quite at first, then really funny), Joana (the microphone thrower), Silvia (our beer connoisseur), Soufiane (world cup winner), Sjoerd (our singer), Tamin (the soccer player), John (our other bodybuilder), Marco (the life of the party), David (the MIL genius), Bob (the dad of the group), Emile (who directed me to my PhD), Jan (infinite pa-

tience), Thomas ('Belgium is better'), Saskia (our superhero secretary), Robbert (the server keeper), Ruud and Bart (the gadget facilitators). You all welcomed me with open arms and taught me so much. Having in the same place such a mixture of backgrounds, professions, languages and cultures opened my eyes towards a broader and more beautiful world. I am going to miss the coffeetalks, the Thursday drinks, the social events, the deep learning jokes and even the complaining about certain food company during lunch.

To the usual suspects: Ekin, Wouter, Sally. I did not believe it possible to find another person with my same weird sense of humor, let alone three. The valleys of the PhD (which were plenty) became enjoyable thanks to all the Thursday night laughs. Ekin, my twisted sibling. You were not only my officemate, but a great colleague to collaborate with (when you were not talking really loud), travel companion during conferences and overall brother from another mother. You, looking for a beer in the middle of a mountain in Japan. That memory not only describes you fully but will probably always stay with me. Wouter, thank you for all the honest chats and for listening to my rants about how lame academia can be, all over a nice glass of whiskey. I believe that, for better or worse, we share the same utopic view of science. Let's hope some of that remains through the years. You still suck at Halo though. Sally, the 'American' (I refuse to use that word without quotes!). You always find a way to look at the *Brightside*, and that is contagious. You were the *somebody that told me* to *smile like you mean it*. It was funny to see how much, *for reasons unknown*, our lives resembled each other *when we were young* (see what I did there...?). That might explain why we get along and can be blunt with each other, without end up playing sicario.

A special mention for the usual suspects goes to Marco. Even though you were not one of the minions, you were not shy when joining our crazy endeavors. Thank you for the honest and serious talks when needed, and also the easygoing and somewhat disturbing ones. 'Elevators' now has a different meaning in my head, thank you very much.

Claudio and Gwenn. I met you two within 3 days of arriving to the Netherlands and somehow we managed to collect a dataset during a real event that day. This smooth and easy collaboration with you triggered data collections and publications that for the untrained eye might seem as if we wing it (I deny all charges) but they always worked. I blame Claudio's Italian ease towards life and Gwenn's almost saint-like predisposition.

Starting the second year, we decided to collaborate with social psychologists. I am not going to lie, I was scared. Luckily, we stumbled with Andrew and Leander. Your help, insights and astonishingly different way of seeing a social event had a big impact on my research. Because of all your hard work, we managed to collect an amazing dataset which makes me proud and, let's be honest, allowed me to finish the PhD. Bam!

Martha. Helping with the course you co-lecture with Hayley allowed me to stay in touch with the teacher within me. From you, I learned patience, positivism, empathy and an everlasting faith towards life and other people's dreams and capacities. I will never forget a bag of chocolates giving to me when it was most needed. Thank you.

A mi grupo de ticos en Holanda. Johan, Carmen, Martha, Andrea, Andres, Miguel, Victor, Ana, Thomas, Adolfo, Pame, Juan (a los tres), Michael, Marlen, Maricruz. Muchísimas gracias gente! Tal vez sin saberlo, ustedes me ayudaron a tener un pedacito de la casa en Europa y no morir de mal de patria. Agradecimiento especial para Johan y Carmen, que con un par de birras y consejos sacados de Google me ayudaron a salir de uno de esos valles que trae el PhD.

A mis supervisores y mentores: Pablo Alvarado, Paola Vega y Johan Carvajal. Ustedes me impulsaron primero a hacer mi maestria y luego, por que no, a seguir con un doctorado. Creo que es justo decir que sin su apoyo y consejos nunca hubiera siquiera considerado la idea. Infinitas gracias por tener fe en mí.

A mi pequeña gran familia: Mama, Ariel y Diego. Desde que les dije que me iba a vivir a Europa nunca recibí otra cosa más que apoyo incondicional, aún cuando las cosas no se veían bien. La persona que soy se define en mucho por haberlos tenido a mi lado. Esta tesis es tanto su logro como el mío. Los quiero.

William, mi novio y mejor amigo. Ya fuera en las buenas o en las malas, siempre te mantuviste cerca. Nunca imagine que en el compañero de trabajo de mi primer curso en el TEC encontraría a mi compañero de vida, pero doy gracias a Dios por eso. Vos le das balance a mi locura, me impulsas a salir de mi zona de comfort, a ser la mejor versión de mi misma y a nunca rendirme. Terminar esta tesis no me hubiera sido posible sin tener una persona tan especial como vos a la par.