# Complex conversational scene analysis using wearable sensors

Hung, Hayley; Gedik, Ekin; Cabrera Quiros, Laura

**DOI**
10.1016/B978-0-12-814601-9.00019-5

**Publication date**
2019

**Document Version**
Final published version

**Published in**
MULTIMODAL BEHAVIOR ANALYSIS IN THE WILD: ADVANCES AND CHALLENGES

**Citation (APA)**
Hung, H., Gedik, E., & Cabrera Quiros, L. (2019). Complex conversational scene analysis using wearable sensors. In X. Alameda-Pineda, E. Ricci, & N. Sebe (Eds.), *MULTIMODAL BEHAVIOR ANALYSIS IN THE WILD: ADVANCES AND CHALLENGES: Advances and Challenges* (pp. 225-245). (Computer Vision and Pattern Recognition Series). Academic Press. https://doi.org/10.1016/B978-0-12-814601-9.00019-5

Chapter 11

# Complex conversational scene analysis using wearable sensors

Hayley Hung[*,†], Ekin Gedik[*], Laura Cabrera Quiros[*,‡]

[*]*Delft University of Technology, Intelligent Systems, the Netherlands* [†]*CWI, Distributed and Interactive Systems, the Netherlands* [‡]*Instituto Tecnológico de Costa Rica, Electronic Engineering Department, Costa Rica*

**CONTENTS**



## 11.1 INTRODUCTION

In recent years, researchers have shown that it is possible to automatically detect complex social phenomena, often using predominantly or only nonverbal behavior; from dominance [15,20,21,26,39,43], functional roles [12,28,54], deception [7,40], cohesion [24], attraction [29,30,41,49], interest [16,36,48,52], influence [38,44], rapport [19,46], to friendship [53].

Most research on social behavior analysis has focused on the analysis of conversational scenes in predominantly pre-arranged settings of typically no more than 6 people. Little is known about the function and nature of

**Multimodal Behavior Analysis in the Wild.** https://doi.org/10.1016/B978-0-12-814601-9.00019-5

■ **FIGURE 11.1** Example snapshots of a mingling event. Taken from [27].

social interactions during complex conversational scenes that is, mingling behavior that occurs during social networking events (e.g. see Fig. 11.1). However attending such face-to-face mingling events has been linked to career and personal success [51]. The long term goal of such an analysis of complex conversational scenes is threefold:

- to be able to generate a social network describing the relationships between attendees of an event and how cohesive the group is;
- to quantify the interaction by considering how the social signals expressed and coordinated during the conversations can be indicative of the experience of a person and therefore how much potential influence there is within the network. This goes beyond more commonly used proxies of social relationship such as the frequency of interaction, where interactions are considered binary values [10,11,13,14];
- or ultimately to predict future behavior such as the chance that someone starts a personal or business relationship with people they first encountered at a particular social event.

One of the reasons that it is challenging to observe these types of events is that they require the coordination of scores of people together to capture such phenomena. Observing and analyzing such behavior is both hard for social scientists and computer scientists because the setting is so complex; What aspects of the behavior indicate how involved a person is in the conversation or that they are interested or engaged in a conversation? What are the patterns of behavior that determine when or how groups split or merge?

Usually when analyzing smaller groups such as for meeting analysis, findings from social science provide a framework for further analysis from computer science. For instance, we can use measures proposed in social science to build automated methods to perceive social phenomena (e.g. estimating dominance using the visual dominance ratio [21]). However, in the case of complex conversational scenes, little is known about how people behave because it has been too difficult to observe at a scale appropriate

for social scientists, the technical challenges for automatically analyzing behavior in these scenes are also difficult to handle. For instance, since the scenes are so crowded, relying on video alone is unwise due to the high level of occlusion. Amalgamating this with wearable sensors can be one way to mitigate this problem, such as for improving head and body pose estimates using audio and infrared sensing from a body worn sensing device [1,2].

Concretely, complex conversational scene understanding has to overcome the following challenges.

- Scene noise: Crowding causes frequent visual occlusions, audio is contaminated by surrounding chatter, the sheer density of human bodies reflect and distort proximity and localization signals (e.g. from radio or wifi).
- Uncontrolled scenario: People move dynamically from group to group based on their own individual goals. Conversational groups or the length of an interaction are not prearranged. Multiple conversations occur simultaneously and conversation partners can change dynamically.
- Multi-sensor fusion: exploiting wearable sensors allows us to mitigate the data association problem by linking all digital sensor information to the wearer. However, we multiply the sensor problem from one of capturing the entire social scene in a single camera to requiring one device per person in the scene.

In this chapter, we propose a break from conventional approaches to the first step in doing such a behavior analysis by detecting social actions (e.g. speaking). Detecting individual speaking status is important for generating derived turn-taking features that are foundational for further analysis of more complex social constructs such as dominance [21,25,26], cohesion [24,35], and attraction [49].

## 11.2 DEFINING 'IN THE WILD' AND ECOLOGICAL VALIDITY

Let us first discuss some definitions. The term 'in the wild' has been much used in recent years to refer to research analyzing human behavior in uncontrolled settings. It has been used frequently in relation to a video-based analysis of facial expressions recorded outside of laboratory conditions. The expectation is that the facial behavior will be unposed and therefore more spontaneous and true to real life. One might question where the 'wild' aspect of this example sits. Conceptually, we could say that it lives within the conditions in which the data is collected outside of the lab, so presumably in uncontrolled/uneven lighting conditions, varying pixel resolution and frame rates. Aside from the recording conditions, there needs to be truthfulness to

the behavior of the person being recorded; that is, the behavior should not be posed or fake.

When we consider human social behavior, we cannot avoid being influenced by social science. And when experimental psychologists conduct experiments, one of the issues they consider is ecological validity. The reason for introducing such a term in this chapter is perhaps most evident in its definition. Ecological validity describes the extent to which an experimental setting and task is true to real life. That is, it tells us the extent to which the location and the task fit with our expectations of what occurs in the expected ecology of someone's everyday life.

We can further understand the idea of ecological validity by considering a typical example. Suppose that an experiment is conducted to study whether people are able to follow emergency instructions provided in a leaflet if a building was burning down. One could approach this task by asking people to fill out a survey where they are shown the instruction leaflet and then asked to imagine what they would do in such a situation. This setting has low ecological validity as the respondents have to imagine what they might do and the task of reading the instruction leaflet is not done in the setting it was intended to be.

The experimenters could improve the ecological validity by using a state-of-the-art virtual reality system to simulate the burning building and then observe how participants react. In this case, the ecological validity may be higher but one might still doubt how realistic the experiment really is. In the optimal case, one might consider actually measuring the behavior of people as a real building is burning down. However, it is unlikely that such an experiment would get ethical approval and this would be so realistic that perhaps the sensing required to actually measure the genuine responses would either be unavailable or too noisy to be useful.

## 11.3 ECOLOGICAL VALIDITY VS. EXPERIMENTAL CONTROL

In both instantiations of the experiment, we see that there is an inherent trade-off between ecological validity and experimental control; while we want to allow the participants to carry out the task in as realistic a scenario as possible, we need to be able to measure the resulting behavior as accurately as possible too.

This is where social science and computer science diverge. For social science, at least traditionally speaking, to have good experimental control means that it is necessary to have accurate measures of behavior and also

good quality survey responses. This is easier to achieve with laboratory based experiments. In computer science, the whole premise behind 'in the wild' perception is that this trade-off between experimental control and ecological validity can be tuned more precisely to the experimenter's requirements. That is, traditionally speaking, in multimedia tasks we no longer expect uniform lighting, no background noise, and the sample rate or frame rate, bitrate or pixel resolution may all be unknown or may vary. Thus the methods to interpret the data need to be robust to this situation, opening up interesting research challenges for computer science.

## 11.4 ECOLOGICAL VALIDITY VS. ROBUST AUTOMATED PERCEPTION

The notion of 'in the wild' can be taken even one step further using the scenario of mingling that we examine in this chapter by considering whether unconventional sensing modalities can act as a proxy for more traditional sensing modalities. Here, we address this problem for social behavior analysis from wearables. Traditionally, if we want to observe social behavior, extracting turn-taking features has shown to have great discriminative power for a number of tasks related to the analysis of small group behavior [17]. Typically, we would expect to measure this from audio of the speakers recorded with microphones. However, recording audio 'in the wild' can have considerable consequences both from a privacy and from an automated analysis perspective.

Privacy concerns relate to recording unwilling participants accidentally as one person's microphone can easily pick up sounds from the surroundings and other people. Moreover, people may not be willing to have their voices recorded at all, leading to a further sample bias when identifying volunteers (one might expect more sample bias the more experimental control is required). As the scene becomes denser, the background noise can become so great that it hinders robust audio processing of the speech signal. This is where wearables in the wild can address these problems by using accelerometer signals as a proxy for speech and also social behavior. The reason why this unconventional method is still deemed a feasible approach is that we know from social science that when people converse, they gesture and move their bodies [34]. By leveraging these body movements, we hypothesize that we can estimate when someone is performing a social action.

This pushes the boundaries of 'in the wild' processing while trading off ecological validity for the following reasons. First, the wearable sensor we propose to use is an ID badge that is hung around the neck; similar to what one might wear during a conference or festival. Second, we do not record

audio which could make the wearer self-conscious of what they say when interacting with others. In this respect, the sensing is ecologically valid. It is 'in the wild' because the setting does not control for exactly what actions people need to perform when they are socializing. Their social behavior is genuine within the context of the situation.

Finally, a key question one might ask is whether the trade-off in spending so much effort on ensuring an ecologically valid solution is worth it. We can point to a prior study by Ekin and Hung [18] that demonstrated this point clearly. For the task of detecting speaking from body movements, data collected in the lab with acted social behavior yielded easily discriminable features. However, evaluating the same method on data recorded in a more ecologically valid setting led to significantly worse results. Our conclusion here is that laboratory data can lead to an underestimation of the difficulty of a task when transferred to real-life settings.

## 11.5 THIN VS. THICK SLICES OF ANALYSIS

Much work on wearables in the wild has been conducted using smart phones or wearable sensors for analyzing social phenomena on a large scale [37]. Analyzing on a large scale and accumulating observations over weeks or months has the benefit that sensor data such as the frequency of proximity as estimated from blue tooth readings can be used directly as a proxy for the quality of a social relationship. In this chapter however, we aim not to take the sensor data at face value but to investigate how signal processing and machine learning techniques can be used to squeeze out more meaning from noisy sensor data at shorter times scales of minutes or even seconds. The reason why this could even be considered possible is based on the thin slice theory proposed by psychologists Ambady and Rosenthal [3] who discovered that short observations (of typically just a few seconds) of social behavior were often enough to reliably assess some social situations.

## 11.6 COLLECTING DATA OF SOCIAL BEHAVIOR

In pushing to more 'in the wild' sensing, one needs to consider what is appropriate as data to investigate this phenomenon. One can imagine that the act of collecting data exists on a continuum in terms of the research question being addressed. For instance, one might have a specific research question that needs to be answered and so therefore the research is more inductive—the data acts to validate some hypotheses. In other cases, a more deductive approach may be used where the data collection acts as a vehicle to investigate currently unknown patterns of behavior. It is vitally important to consider this when using data to analyze social behavior 'in the wild'.

An individual data sample may exist in an 'in the wild' setting. However, when multiple data samples are aggregated to make an entire corpus, could there be selection bias at this stage? In this chapter, based on the research goals listed in Section 11.1, we focus on data collection using a deductive approach, considering how this further impacts the machine perception process at the end. Note in this case that due to the realistic nature of the data, class labels can have high levels of imbalance.

### 11.6.1 Practical concerns when collecting data during social events

In this section, we describe an approach to collect data in what may be considered an extremely uncontrolled and ecologically valid setting where many research challenges lie. Here we focus on mingle scenarios or free-standing conversational gatherings. That is, we address crowded social settings where people come together purely to socialize.

Very few data sets exist to investigate the machine perception of nonverbal social behavior in mingle scenarios with wearables [1,27], although little work analyzing social behavior with wearables does exist [5,18,32,33]. There have been made more efforts in the computer vision community [8,9,22,55], which provide many insights for addressing this problem from a multimodal perspective. Usually researchers who focus on multimedia analysis problems are not likely to have practical experience of wearable sensor system deployment. Therefore we provide a primer on some key issues to consider when collecting data in such settings. Here we focus on lessons learned from capturing our own data set [27] with respect to the use of wearable sensors that capture acceleration and proximity, as well as of cameras that are able to validate the behaviors captured. In moving outside of the lab while still wanting to maintain some experimental control, we encounter important logistic issues that need to be taken into account. While many of these adhere to common sense practices, with so many elements to keep in balance, it can be easy to overlook some aspects. This can have severe negative consequences for the data collection such as lost or unusable data.

#### Requirements of the hardware and software

The selection of adequate sensor or wearable devices is an important aspect of the analysis of social interactions 'in the wild'. And this is strictly related to the event and behavior that is going to be analyzed. Aside from this, the method of analysis will greatly affect the requirements. For example, will the data be analyzed offline or is wireless communication necessary (e.g. for live or realtime processing)?

Moreover, even when each module is independently capable of fulfilling the requirements of the event, all devices (as individual units) have certain restrictions given their hardware (memory, CPU, size of registers) and software (execution time, interruptions and deadlocks, delays). One could, for example, try to use the maximum sample rate of the accelerometer (sensor) but one wonders: can all this information be stored locally in the device or sent wirelessly to a storing unit without critical package loss? And is the software/firmware used capable of handling this rate?

Thus, a balance must be sought between factors such as sensing configurations (e.g. sample rates, sensitivity, mode of sensing, sleep states), storage space, and power consumption. To do so, good practices from the embedded systems community such as efficient programming (e.g. use of idle states and proper scheduling) are advised, so that factors such as power consumption of the final badges are optimal [31].

### Ease of use

For the participants and experimenters, we prefer to have a wearable that is 'grab and go'. That is, no additional registering is required once consent forms have been signed and the sensing device has been fitted. This allows for easier scaling of a data collection event to more and more participants. This has the drawback that associating the correct video data with the sensor requires manual work. However, alternative approaches with automated methods of associating the data together are showing promise [6,47,50].

For the researcher, having a system or firmware that is easily adapted to slight variations in research questions may also be important e.g. adjusting the bit rate, sample rate, wireless sensitivity. This is particularly important if one wants to examine hardware trade-offs e.g. high sample rate vs. long battery life. The sociometer is an example of a custom-made wearable device that could be useful for mingle events. It has several well-selected sensors that comply with the nature of mingle scenarios. However, as a commercial product, the selection and visualization of the raw sensor data are restricted. Reconfiguration of the device in terms of the trade-offs listed above is not possible. Sadly, there are few commercial products that give full access to the device's capabilities for a researcher's needs. For such cases, perhaps an open-source solution such as a platform-based device (e.g. raspberry pi or arduino) can be adapted for use. In our case, custom hardware was used.

### Technical pilot test

Finally, it is vitally important that a technical pilot is carried out, preferably in the venue itself; it can often be the case that even commercial sensors have significant problems in delivering the functionality that they are designed

for. This enables the entire sensor set-up to be tested before participants are involved. Note, however, that in our experience, carrying out a pilot test without participants has one drawback as the participants themselves can cause disruption to sensor signals. For instance, we found in prior data collections that a high density of people led to system performance degradation for an indoor localization system where all people in the same conversing group were detected to be located at exactly the same location in the ground plane, despite having a much better localization resolution when fewer people were present.

It is generally safe to assume that something will break and that all sensors will need to be continuously monitored to ensure that they are recording. This is also a key period to verify that the data across all modalities is correctly synchronized. Even if the primary analysis is carried out using the wearable devices, correctly synchronized video as well as correct data association (knowing where person A wearing sensor X is located in a video) is vital for data labeling.

### Issues during the data collection event

As with any experimental setting, informed consent needs to be sought and ethical approval from an institutional committee needs to be obtained. Since we are dealing with a short term event where multiple people need to attend, overbooking participants is recommended. A financial incentive can be given to those who are turned away to minimize disappointment.

Other practical conditions include pre-defining a clear procedure for the event and if this is particularly intense, rest breaks may need to be planned for participants if there is some level of experimental control necessary. If participants are free to come and go as they please, a mechanism needs to be in place to ensure that they do not leave with the sensor (unless this was planned originally), and it is possible to find out who left (in terms of wearable sensor ID) and when.

The sensors themselves must be linked to an individual if survey responses are required from the participants. To ensure anonymization of the person's identify from their data, usually a participant is given a number. Their associated sensor then needs to be logged. If sensors and participants are already assigned before the start of the experiment, this can be problematic if there is sensor failure and a replacement needs to be brought in and a new sensor number logged.

For the data collection itself, having at least one person responsible per sensor type or modality ensures that multiple simultaneous failures can be handled relatively quickly.

## 11.7 ANALYZING SOCIAL ACTIONS WITH A SINGLE BODY WORN ACCELEROMETER

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

When analyzing human behavior, past analysis has tended to assume that this is more or less person-independent. Throughout the text, 'person-independent' will be used for settings where data for training a model comes from different sources (people in our case) compared to the test data. Much work has been done on estimating daily activities such as walking or running from accelerometer data, showing promising results with a person-independent setup [4,42]. There is a direct connection between the sensing medium and the physical manifestation of the behavior so that actions such as walking and stepping result in acceleration readings that are easy to discriminate directly from the magnitude of the signal. This makes a person-independent setup for discriminating such behavior quite easy to implement.

However, some of the actions observed in crowded social settings tend to be much more person specific and the connection between the existence of these actions and the accelerometer readings is more ambiguous. In our case, the physical manifestation of speaking comes from vibration of the vocal chords, so unless the subject has a very sensitive accelerometer attached tightly to the body (e.g. the chest [32,33]), there will not be a direct connection between the action and the sensing. However, speaking also has a physical gestural aspect, and it has been shown in previous work that the connection between body movements and speech can still be exploited for detecting if someone is speaking or not [18,23]. Actions like speaking, which are loosely connected with the sensing medium, are expected to be harder to detect and may require specialized approaches.

To examine this, we conducted a number of experiments on a dataset that is collected from a real-life 'in the wild' event. The dataset is comprised of mingling events from three separate evenings where each evening includes data from approximately 32 people. Each participant wore a sensor hung around the neck that records individual tri-axial acceleration at 20 Hz. Note that the sample rate is not high enough to detect vocal chord vibration. However, it is high enough to capture body movements such as

**Table 11.1** AUC scores for various actions

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

gestures. Different social actions are manually labeled by trained annotators for 30 min of the mingling sessions. For more information about the dataset, please refer to [27]. We have focused on the mingling session from the first day for the experiments presented in this section.

### 11.7.1 Feature extraction and classification

We have extracted features for each of the 26 subjects with valid accelerometer data. Statistical and spectral features are extracted from each axis of raw and absolute values of the acceleration and the magnitude of the acceleration, using 3 s windows with 1.5 s overlap. As the statistical features, mean, and variance values are calculated. The spectral features consist of the power spectral density binned into eight components with logarithmic spacing between 0–8 Hz.

We have used the L2 penalized Logistic Regressor as the classifier. Performance evaluation is done with leave-one-subject-out cross-validation. Hyperparameter optimization for regularization is carried out with nested cross-validation. Stepping, speaking, hand and head gestures, and laughter are selected as the target actions. Since the class distributions for each participant are different, we have chosen the AUC (area under the ROC curve) as the performance metric. Performances obtained with the aforementioned setup are presented in Table 11.1. We also present the mean annotator agreement for each action using Fleiss'–Kappa for three annotators. Values higher than 0.4 are considered to be of moderate agreement.

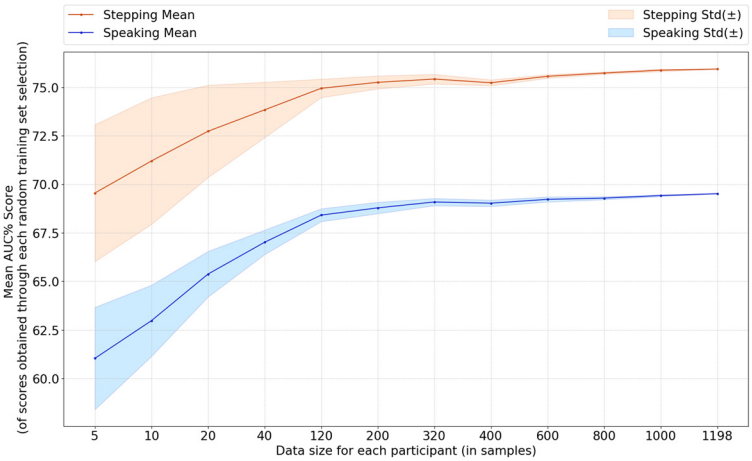
We can see that the results presented in Table 11.1 support the claim that actions that are loosely connected to the physical manifestation of the behavior are harder to detect. Stepping, as expected, has the highest performance of all. We also see that the performance tends to drop as the connection between the physical manifestation of the action itself and the acceleration reduces. For example, head gesture labels in the dataset, the social action with the lowest detection rate, include many subtle nods which are harder to capture via acceleration, compared to a step or hand gesture.

It should be noted that there might be a second factor at play here. In real-life events, it is generally harder to obtain annotations. Thus, the annotations must be made later manually. This of course introduces some differences in annotator agreement with respect to the type of action. Table 11.1 shows the annotator agreements as reported on a subset of the data taken from [27]. It can be seen that the lowest annotator agreement values are for the head gestures, followed by laughter. Variation in agreement (due to behavioral ambiguity or visual occlusion of the person being annotated) in the labels might have also contributed to the low performance of these actions, in addition to the nature of the connection between the action and the sensing medium. Thus, noisy labels, at least for some actions, are a reality of data collection in the wild, which needs to be taken into account when evaluating the perception performance. A further discussion of the trade-offs between using crowd sourced annotations compared to onsite annotators are also discussed in [27].

### 11.7.2 Performance vs. sample size

In the former experiment, 30 min of data from each participant was used. The results obtained showed that 30 min was enough to capture a variety of actions with various different situational contexts (i.e. differing conversing partners with different levels of conversational involvement), obtaining acceptable performance even for more subtle actions. But what is the minimum required amount of data for acceptable performance? Will the patterns be similar if we had less data? Since it is not guaranteed that we would have a continuous stream of 30 min of data, we conducted another experiment, where we used the earlier setup but with gradually increasing amounts of data for each participant, starting from five samples to a total of 1198 that covers the entire 30 min period.

As mentioned in the former section, each sample is extracted with a sliding window of 3 s with 1.5 s shift. Thus, we can say that five samples roughly corresponds to 7.5 s of data, 40 samples correspond to 1 min, and so on. We still used a leave-one-subject setup where for each fold, all the data from one participant corresponded to the test set. However, the training set is formed randomly by selecting $n$ samples from each of the other participant's data. Since the selection is random, the process is repeated $m$ times, which was also dependent on the number of samples selected. For computational reasons, we gradually reduced the number of repetitions from 150 to 15, resulting in 5 to 1000 samples for each repetition. We have selected two relatively well performing actions, stepping and speaking. These actions have different characteristics as described earlier with stepping being




■ **FIGURE 11.2** AUC scores of stepping and speaking with respect to data size.

more closely connected to the physical manifestation of the behavior compared to speaking, which relies on detecting bodily gestures that are related to speech. In addition, this selection is based on former studies that showed that the connection between speech and acceleration is highly person specific, compared to stepping–walking [18]. The mean of the AUC scores of all repetitions, with increasing data size, are shown in Fig. 11.2 with standard deviation.

First, from Fig. 11.2 we observe the higher standard deviation for smaller sample sizes. This is related to the decreasing number of repetitions but we argue that is not the only factor. We believe there are parts of the event that are less informative than others and if the selected samples are coming from such intervals the performance tends to be low, and therefore fails to generalize over the whole event. This issue will be discussed further later in this section where we will present results of an experiment where the samples are not randomly sampled but selected chronologically. We also observe that the standard deviation for both actions converges to small values with increasing sample size.

We can see from Fig. 11.2 that the pattern for the two actions are quite similar. Performances for the actions increase with a steep curve in the beginning and after 120 samples the increase gets smaller. This suggests that 3 min of data from each person is enough to cover the variations in each type of action in such an event. The question then becomes if it is possible to provide a specialized solution which can guarantee better results even if the number of samples is relatively low.

### 11.7.3 Transductive parameter transfer (TPT) for personalized models

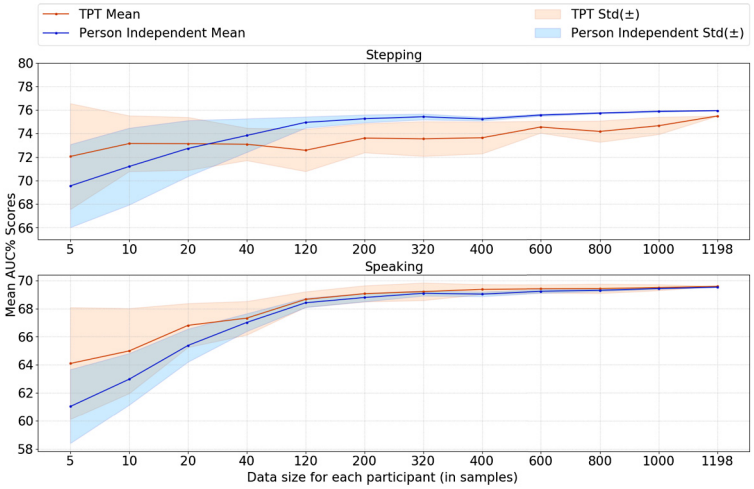
Following on from the results of [18], where it was shown that a transfer learning approach that guarantees personalized models in a person-independent setup tends to perform better for person specific actions, we repeated the former experiment with a personalized model. The method is named Transductive Parameter Transfer (TPT) and was first proposed for personalized facial expression recognition [45] and then modified for social action detection from a body worn accelerometer in [18].

TPT aims to find the parameters of the classifier for the target dataset $X^t$, without using any label information of $X^t$, by learning a mapping between the marginal distributions of the source datasets and the parameter vectors of their classifiers. $N$ source datasets with label information and the unlabeled target dataset are defined as $D_1^s, ..., D_N^s$, $D_i^s = \{x_j^s, y_j^s\}_{j=1}^{n_i^s}$ and $X^t = \{x_j^t\}_{j=1}^{n_t}$, respectively. The main steps of the TPT are shown below (for a detailed explanation, please refer to [18]).

1. Compute $\{\boldsymbol{\theta}_i = (\boldsymbol{w}_i, c_i)\}_{i=1}^N$ using L2 penalized logistic regression.
2. Create the training set $\tau = \{X_i^s, \boldsymbol{\theta}_i\}_{i=1}^N$.
3. Compute the kernel matrix $\boldsymbol{K}$ that defines the distances between distributions where $\boldsymbol{K}_{ij} = \kappa(X_i^s, X_j^s)$.
4. Given $\boldsymbol{K}$ and $\tau$, compute $\hat{f}(.)$ with kernel ridge regression.
5. Compute $(\boldsymbol{w}_t, c_t) = \hat{f}(X^t)$ using the mapping obtained in former step.

We conducted the performance vs. sample size experiment explained in the former section, with the addition of TPT. TPT is also used in a person-independent setup, where data from other participants are treated as source datasets with label information whereas the data to be classified is the target dataset. Although [18] suggests the use of an Earth Mover's Distance (EMD) kernel for computing the distance between distributions, we employed a density estimate kernel [45], since it is computationally less complex and more suitable for many random repetitions. The resulting AUC scores are plotted in Fig. 11.3.

According to Fig. 11.3, TPT outperforms a traditional person-independent setup when using small sample sizes for both actions. It seems to generalize better even with a small amount of data. For speaking, with the increasing data size, the gap between the two methodologies starts to close, showing that the single logistic regressor in the person-independent setup has seen enough diverse cases to generalize better. A one tailed paired t-test between AUC scores showed that, up until 320 samples, TPT provides significantly better performance ($p < 0.05$ for 40 samples and $p < 0.01$ for the rest).

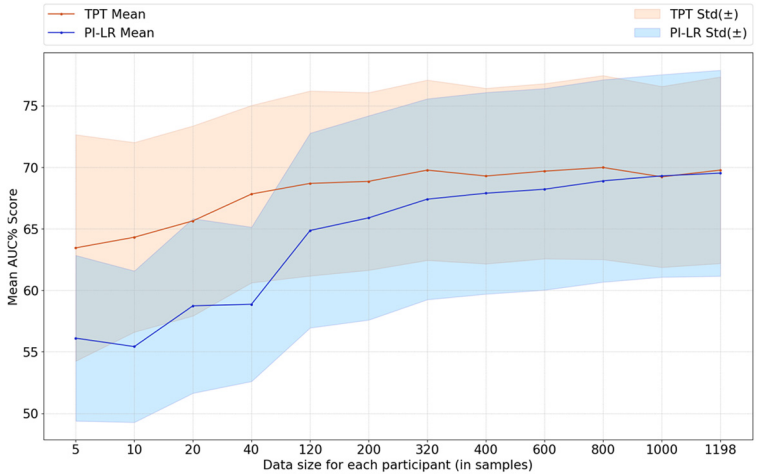



■ **FIGURE 11.3** AUC scores of stepping and speaking with respect to data size.

After that point, the mean scores provided by TPT seemed to be still higher than the person-independent setup but the significance is not guaranteed (some results such as those at 400 and 600 samples are still significant, though). We can say that with the increasing data size, the two methods converge to similar performances. However, especially for smaller sample sizes, we can still conclude that for estimating an action in a person specific manner, TPT is more robust.

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

Now, we want to go back to our claim that some parts of the event are more informative than others. The first parts of the dataset correspond to the be-




■ **FIGURE 11.4** AUC scores of speaking with temporally increasing data size.

ginning of the event, when groups are just starting to be formed. We might expect people to be less involved in the conversation as the discussions are not yet in full flow. This might result in samples that are not representative of all variations of actions that can occur in a real-life event, throughout time. So, we did a follow-up experiment where we compared the performances of TPT and the traditional person-independent setup for speaking detection. However, this time for each participant in the training set, we increased the number of samples in chronological order. Thus, $n$ samples for a participant correspond to the first $n$ samples in time. Since there are no repetitions, the means and the standard deviations are computed on the individual performances of all participants. The results of this experiment are shown in Fig. 11.4.

The first thing we observe from Fig. 11.4 is how the performances of the person-independent method is lower compared to those from Fig. 11.3. Using random selection of the samples throughout the event, the person-independent method was providing an AUC of roughly 61% for 5 samples. However, in the temporally increasing setup, the performance for the same number of samples is roughly 56%. The pattern is similar for the following sample sizes and the performance of temporally increasing selection is only able to reach the level of random selection if at least 320 samples are used for training. TPT on the other hand still provides similar results to the random selection method and provides relatively satisfactory results even with samples that were less informative for a traditional person-independent approach.

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

### 11.7.4 Discussion

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

Specifically, for the case study presented in this chapter, when focusing on the detection of actions through wearables, there are some important points to consider. First, one should understand the connection between the physical manifestation of the action, and the sensing medium they are using. This is required for the valid selection of features and models that will be used for classification. In real-life scenarios, it is not guaranteed to have each action perfectly represented in all its possible variations for each participant. This is particularly true because natural 'in the wild' behavior samples only come into being as the result of the dynamics of a conversation as it unfolds over time. That is, a monologue in a group would yield more positive examples of speaking for the speaker of the group but no speaking samples for the members of the group who are just listening. So, the experimental setup and methodology chosen should encapsulate this together with the physical nature of the action. The experiments presented in this section are good examples of this, where two approaches for the detection of two actions tend to perform differently, because of the physical nature of the actions in relation to the sample sizes.

## 11.8 CHAPTER SUMMARY

This chapter has introduced some basic concepts of how to perform social behavior analysis 'in the wild' and specifically in the case of analyzing

complex conversational scenes. In conducting research in this area, we have discussed two conceptual concerns: how to consider the relationship between ecological validity and 'in the wild' automated perception. Next, we provided concrete guidelines on how to collect data in such settings, differentiating between inductive versus deductive research practices and how this influences the data collection process. Finally, we provide some experiments on doing social action detection during complex conversational scenes using just accelerometer data recorded from a body worn sensor pack. Within this setting, we address challenging questions related to recording data in a deductive setting; When is there enough data? Does the learning model change depending on the amount of data available? How does the amount of data and the learning model vary with respect to the level of physical connection between the social behavior being detected? All these investigations provide an initial glimpse of what could be further investigated when considering automated social behavior analysis 'in the wild'. We have presented behavior analysis from the perspective of just a single modality (acceleration). However, further sensing modalities such as proximity or other more traditional modalities such as video and audio could also be combined opportunistically to provide richer representations for social behavior understanding.

## REFERENCES

[1] X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrinca, E. Ricci, B. Lepri, O. Lanz, N. Sebe, Salsa: a novel dataset for multimodal group behavior analysis, IEEE Trans. Pattern Anal. Mach. Intell. 38 (8) (2016) 1707–1720.

[2] X. Alameda-Pineda, Y. Yan, E. Ricci, O. Lanz, N. Sebe, Analyzing free-standing conversational groups: a multimodal approach, in: ACM International Conference on Multimedia, 2015.

[3] N. Ambady, R. Rosenthal, Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis, Psychol. Bull. 111 (2) (1992) 256.

[4] Ling Bao, Stephen Intille, Activity recognition from user-annotated acceleration data, Pervasive Comput. (2004) 1–17.

[5] Laura Cabrera-Quiros, Ekin Gedik, Hayley Hung, Estimating self-assessed personality from body movements and proximity in crowded mingling scenarios, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction, ICMI 2016, New York, NY, USA, ACM, 2016, pp. 238–242.

[6] Laura Cabrera-Quiros, Hayley Hung, Who is where?: Matching people in video to wearable acceleration during crowded mingling events, in: Proceedings of the 2016 ACM on Multimedia Conference, MM '16, New York, NY, USA, ACM, 2016, pp. 267–271.

[7] G. Chittaranjan, H. Hung, Are you a werewolf? Detecting deceptive roles and outcomes in a conversational role-playing game, in: International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2010.

[8] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, A. Del Bue, D. Tosato, G. Menegaz, V. Murino, Social interaction discovery by statistical analysis of F-formations, in: British Machine Vision Conference, August 2011.

[9] Marco Cristani, Anna Pesarin, Alessandro Vinciarelli, Marco Crocco, Vittorio Murino, Look at who's talking: voice activity detection by automated gesture analysis, in: Constructing Ambient Intelligence, Springer, 2012, pp. 72–80.
[10] T.M.T. Do, D. Gatica-Perez, GroupUs: smartphone proximity data and human interaction type mining, in: 2011 15th Annual International Symposium on Wearable Computers, June 2011, pp. 21–28.
[11] Trinh Minh Do, Daniel Gatica-Perez, Human interaction discovery in smartphone proximity networks, Pers. Ubiquitous Comput. 17 (3) (March 2013) 413–431.
[12] Wen Dong, Bruno Lepri, Fabio Pianesi, Alex Pentland, Modeling functional roles dynamics in small group interactions, IEEE Trans. Multimed. 15 (1) (2013) 83–95.
[13] Nathan Eagle, Alex Pentland, Social serendipity: mobilizing social software, IEEE Pervasive Comput. 4 (2) (April 2005) 28–34.
[14] Nathan Eagle, Alex Sandy Pentland, Reality mining: sensing complex social systems, Pers. Ubiquitous Comput. 10 (4) (2006) 255–268.
[15] Sergio Escalera, Oriol Pujol, Petia Radeva, Jordi Vitrià, María Teresa Anguera, Automatic detection of dominance and expected interest, EURASIP J. Adv. Signal Process. 2010 (2010).
[16] D. Gatica-Perez, I. McCowan, D. Zhang, S. Bengio, Detecting group interest-level in meetings, in: IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2005.
[17] Daniel Gatica-Perez, Automatic nonverbal analysis of social interaction in small groups: a review, Image Vis. Comput. 27 (12) (November 2009) 1775–1787.
[18] Ekin Gedik, Hayley Hung, Personalised models for speech detection from body movements using transductive parameter transfer, Pers. Ubiquitous Comput. 21 (4) (Aug. 2017) 723–737.
[19] Juan Lorenzo Hagad, Roberto Legaspi, Masayuki Numao, Merlin Suarez, Predicting levels of rapport in dyadic interactions through automatic detection of posture and posture congruence, in: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), IEEE, 2011, pp. 613–616.
[20] H. Hung, Y. Huang, C. Yeo, D. Gatica-Perez, Associating audio-visual activity cues in a dominance estimation framework, in: Computer Vision and Pattern Recognition Workshop on Human Communicative Behaviour, 2008.
[21] H. Hung, D. Jayagopi, S.O. Ba, J.-M. Odobez, D. Gatica-Perez, Investigating automatic dominance estimation in groups from visual attention and speaking activity, in: International Conference on Multi-modal Interfaces, 2008.
[22] H. Hung, B.J.A. Krose, Detecting F-formations as dominant sets, in: International Conference on Multimodal Interfaces (ICMI), November 2011.
[23] Hayley Hung, Gwenn Englebienne, Jeroen Kools, Classifying social actions with a single accelerometer, in: ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, 2013, pp. 207–210 (Oral presentation).
[24] Hayley Hung, Daniel Gatica-Perez, Estimating cohesion in small groups using audio-visual nonverbal behavior, IEEE Trans. Multimed. 12 (6) (October 2010) 563–575.
[25] Hayley Hung, Yan Huang, Gerald Friedland, Daniel Gatica-Perez, Estimating dominance in multi-party meetings using speaker diarization, IEEE Trans. Audio Speech Lang. Process. 19 (4) (May 2011) 847–860.
[26] D. Jayagopi, H. Hung, C. Yeo, D. GaticaPerez, Modeling dominance in group conversations from non-verbal activity cues, IEEE Trans. Audio Speech Lang. Process. (2008).
[27] L. Cabrera-Quiros, A. Demetriou, E. Gedik, L.v.d. Meij, H. Hung, The MatchNMingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates, IEEE Trans. Affect. Comput. (2018), https://doi.org/10.1109/TAFFC.2018.2848914.

[28] Bruno Lepri, Ankur Mani, Alex Pentland, Fabio Pianesi, Honest signals in the recognition of functional relational roles in meetings, in: AAAI Spring Symposium: Human Behavior Modeling, 2009, pp. 31–36.
[29] Anmol Madan, Ron Caneel, Alex Pentland, Voices of Attraction, 2004.
[30] R. Caneel, A. Madan, A. Pentland, Voices of attraction, in: Proceedings of Augmented Cognition (AugCog), HCI, 2005.
[31] Peter Marwedel, Embedded System Design: Embedded Systems Foundations of Cyber-Physical Systems, and the Internet of Things, Springer, 2011.
[32] Aleksandar Matic, Venet Osmani, Alban Maxhuni, Oscar Mayora, Multi-modal mobile sensing of social interactions, in: Pervasive Health, IEEE, 2012, pp. 105–114.
[33] Aleksandar Matic, Venet Osmani, Oscar Mayora-Ibarra, Mobile monitoring of formal and informal social interactions at workplace, in: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, UbiComp '14 Adjunct, ACM, 2014, pp. 1035–1044.
[34] D. McNeill, Language and Gesture, Cambridge University Press, New York, 2000.
[35] Marjolein C. Nanninga, Yanxia Zhang, Nale Lehmann-Willenbrock, Zoltán Szlávik, Hayley Hung, Estimating verbal expressions of task and social cohesion in meetings by quantifying paralinguistic mimicry, in: Edward Lank, Alessandro Vinciarelli, Eve E. Hoggan, Sriram Subramanian, Stephen A. Brewster (Eds.), Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI 2017, Glasgow, United Kingdom, November 13–17, 2017, ACM, 2017, pp. 206–215.
[36] Catharine Oertel, Céline De Looze, Stefan Scherer, Andreas Windmann, Petra Wagner, Nick Campbell, Towards the automatic detection of involvement in conversation, in: Proceedings of the 2010 International Conference on Analysis of Verbal and Nonverbal Communication and Enactment, COST'10, Springer-Verlag, Berlin, Heidelberg, 2011, pp. 163–170.
[37] Daniel Olguin Olguin, Benjamin N. Waber, Taemie Kim, Akshay Mohan, Koji Ara, Alex Pentland, Sensible organizations: technology and methodology for automatically measuring organizational behavior, IEEE Trans. Syst. Man Cybern., Part B 39 (1) (2009) 43–55.
[38] K. Otsuka, J. Yamato, Y. Takemae, H. Murase, Quantifying interpersonal influence in face-to-face conversations based on visual attention patterns, in: Proc. ACM CHI Extended Abstract, Montreal, Apr. 2006.
[39] Emanuele Principi, Rudy Rotili, Martin Wöllmer, Stefano Squartini, Björn Schuller, Dominance detection in a reverberated acoustic scenario, in: Jun Wang, Gary G. Yen, Marios M. Polycarpou (Eds.), International Symposium on Neural Networks, in: Lect. Notes Comput. Sci., vol. 7367, Springer, 2012, pp. 394–402.
[40] N. Raiman, H. Hung, G. Engliebienne, Move, and I will tell you who you are: detecting deceptive roles in low-quality data, in: International Conference on Multimodal Interfaces (ICMI), November 2011.
[41] Rajesh Ranganath, Dan Jurafsky, Dan McFarland, It's not you, it's me: detecting flirting and its misperception in speed-dates, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, vol. 1, Association for Computational Linguistics, 2009, pp. 334–342.
[42] Nishkam Ravi, Nikhil Dandekar, Preetham Mysore, Michael L. Littman, Activity recognition from accelerometer data, in: AAAI, vol. 5, 2005, pp. 1541–1546.
[43] Rutger Rienks, Dirk Heylen, Dominance detection in meetings using easily obtainable features, in: Machine Learning for Multimodal Interaction, Springer, 2006, pp. 76–86.
[44] Rutger Rienks, Anton Nijholt, Dirk Heylen, Verbal behavior of the more and the less influential meeting participant, in: Proceedings of the 2007 Workshop on Tagging, Mining and Retrieval of Human Related Activity Information, TMR '07, New York, NY, USA, ACM, 2007, pp. 1–8.

[45] Enver Sangineto, Gloria Zen, Elisa Ricci, Nicu Sebe, We are not all equal: personalizing models for facial expression analysis with transductive parameter transfer, in: Proceedings of the 22nd ACM International Conference on Multimedia, ACM, 2014, pp. 357–366.
[46] Xiaofan Sun, Anton Nijholt, Khiet P. Truong, Maja Pantic, Automatic understanding of affective and social signals by multimodal mimicry recognition, in: Affective Computing and Intelligent Interaction, Springer, 2011, pp. 289–296.
[47] T. Teixeira, D. Jung, A. Savvides, Tasking networked CCTV cameras and mobile phones to identify and localise multiple persons, in: ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp), 2010.
[48] Ryoko Tokuhisa, Ryuta Terashima, Relationship between utterances and "enthusiasm" in non-task-oriented conversational dialogue, in: Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue, SigDIAL '06, Stroudsburg, PA, USA, Association for Computational Linguistics, 2006, pp. 161–167.
[49] A. Veenstra, H. Hung, Do they like me? Using video cues to predict desires during speed-dates, in: International Conference on Computer Vision Workshop on Socially Intelligent Surveillance Monitoring, November 2011.
[50] A. Wilson, H. Benko, CrossMotion: fusing device and image motion for user identification, tracking and device association, in: International Conference on Multimodal Interaction (ICMI), 2014.
[51] Hans-Georg Wolff, Klaus Moser, Effects of networking on career success: a longitudinal study, J. Appl. Psychol. 94 (1) (2009) 196.
[52] Britta Wrede, Elizabeth Shriberg, Spotting "hot spots" in meetings: human judgments and prosodic cues, in: INTERSPEECH, 2003.
[53] Zhou Yu, David Gerritsen, Amy Ogan, Alan Black, Justine Cassell, Automatic prediction of friendship via multi-model dyadic features, in: Proceedings of the SIGDIAL 2013 Conference, Metz, France, Association for Computational Linguistics, August 2013.
[54] Massimo Zancanaro, Bruno Lepri, Fabio Pianesi, Automatic detection of group functional roles in face to face interactions, in: Proceedings of the 8th International Conference on Multimodal Interfaces, ACM, 2006, pp. 28–34.
[55] Lu Zhang, Hayley Hung, Beyond F-formations: determining social involvement in free standing conversing groups from static images, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016, pp. 1086–1095.