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

Cabrera-Quiros, Laura; Demetriou, Andrew; Gedik, Ekin; van der Meij, Leander; Hung, Hayley

DOI
10.1109/TAFFC.2018.2848914

Publication date
2018

Document Version
Final published version

Published in
IEEE Transactions on Affective Computing

Citation (APA)
Cabrera-Quiros, L., Demetriou, A., Gedik, E., van der Meij, L., & Hung, H. (2018). The MatchNMingle dataset: A novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates. *IEEE Transactions on Affective Computing*, *PP*(99), 1-17. https://doi.org/10.1109/TAFFC.2018.2848914

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

Laura Cabrera-Quiros, Andrew Demetriou, Ekin Gedik, Leander van der Meij, and Hayley Hung, *Member, IEEE*

Abstract—We present *MatchNMingle*, a novel multimodal/multisensor dataset for the analysis of free-standing conversational groups and speed-dates in-the-wild. *MatchNMingle* leverages the use of wearable devices and overhead cameras to record social interactions of 92 people during real-life speed-dates, followed by a cocktail party. To our knowledge, MatchNMingle has 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. It consists of 2 hours of data from wearable acceleration, binary proximity, video, audio, personality surveys, frontal pictures and speed-date responses. Participants' positions and group formations were manually annotated; as were social actions (eg. speaking, hand gesture) for 30 minutes at 20 FPS making it the first dataset to incorporate the annotation of such cues in this context. We present an empirical analysis of the performance of crowdsourcing workers against trained annotators in simple and complex annotation tasks, founding that although efficient for simple tasks, using crowdsourcing workers for more complex tasks like social action annotation led to additional overhead and poor inter-annotator agreement compared to trained annotators (differences up to 0.4 in Fleiss' Kappa coefficients). We also provide example experiments of how *MatchNMingle* can be used.

Index Terms—Multimodal dataset, speed-dates, mingle, f-formation, wearable acceleration, cameras, personality traits

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 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) [65], [66]. 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 [41]) 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 paper 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

- *L. Cabrera-Quiros is with the Department of Intelligent Systems, TU Delft, Delft 2628 CD, The Netherlands and also with the Escuela de Ingeniería Electrónica, Instituto Tecnológico de Costa RicaCartago30101, Costa Rica. E-mail: l.c.cabreraquiros@tudelft.nl.*
- *A. Demetriou, E. Gedik and H. Hung are with the Department of Intelligent Systems, TU Delft, Delft 2628 CD, The Netherlands. E-mail: andrew.m.demetriou@gmail.com, {e.gedik, h.hung}@tudelft.nl.*
- *L. van der Meij is with the Eindhoven University of Technology, Eindhoven 5612 AZ, Netherlands. E-mail: L.v.d.Meij@tue.nl.*

Manuscript received 3 Aug. 2017; revised 26 Mar. 2018; accepted 20 May 2018. Date of publication 25 June 2018; date of current version 1 Mar. 2021. (Corresponding author: Laura Cabrera-Quiros.) Recommended for acceptance by A. A. Salah. Digital Object Identifier no. 10.1109/TAFFC.2018.2848914

TABLE 1
Summary of All the Elements Included in *MatchNMingle*

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 4
Hormone baseline*	Cortisol	Collected using hair samples.
	Testosterone	
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	$30min$ at 20 FPS for the mingle
	Social Actions	(Social actions detailed in Section 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 4.2.2.
***Position and social action annotations were performed by 8 different annotators. More details in Section 5.Unless stated otherwise, all data is publicly available.

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 paper, which introduces the dataset, 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.

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 36,000 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.

1.1 Motivation for *MatchNMingle*

There were 4 main reasons that motivated us to create *MatchNMingle*.

First, we wanted to provide the research community with an open-access resource for the analysis of the nonverbal 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 [41].

Second, 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.

Third, 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.

1.2 What Is Included in MatchNMingle?

A comprehensive summary of all the elements include in *MatchNMingle* is shown in Table 1. Unless specified directly, all the data is publicly available.[1] Also, Fig. 1 presents a visual summary of all the modalities of the dataset. Details about each component or sensor type can be found in Section 3.

Similar to previous efforts ([1], [21], [37]), 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 a total of 3 days) consisted of a speed dating round (3 min date with participants of opposite sex) immediately followed by a mingle party of about an hour where participants could interact freely following

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

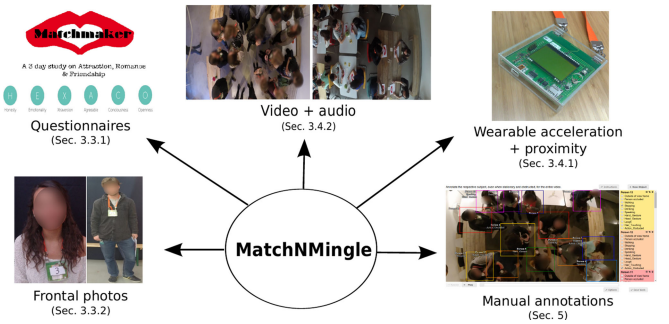


Fig. 1. Visualization of all the modalities included in *MatchNMingle*.

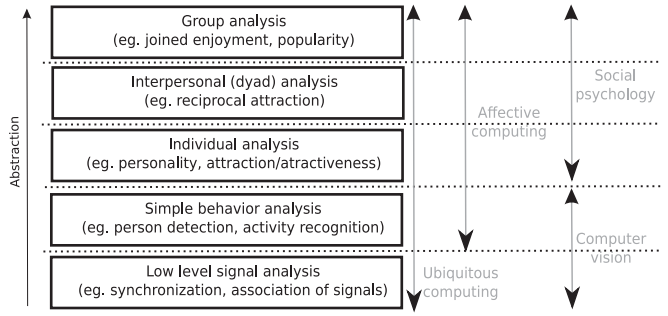


Fig. 2. Levels of abstraction while studying social interactions in which *MatchNMingle* can contribute as a new multimodal resource. A suggestion for the reach of scientific areas (no definitive) is also presented.

their own desires and intentions. Details of data collection procedures can be found in Section 3.5.

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.

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. Fig. 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, Computer Vision-Pattern Recognition and Social Psychology. A suggestion for the reach of these areas (by no means definitive) is also presented in Fig. 2.

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. Second, 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. Third, 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 peoples 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 persons 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 [14], speaker detection from wearable devices [29], [30], personality estimation [13] and perceptions of attractiveness [23]. 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.

The rest of the paper is divided as follows. First, Section 2 presents related efforts about datasets for free-standing conversational groups and speed dates. Section 3 gives a detailed description of the data collection framework, while a description of the general statistics of the dataset is provided in Section 4. The annotation process and a comparison between crowdsourcing and trained annotators is presented in Section 5. Some examples of the use of *MatchNMingle* as multimodal resource are presented in Section 6. Finally, we discuss the limitations of *MatchNMingle* in Section 7 and conclude in Section 8.

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

TABLE 2
Free-standing Convers. Groups and Face-to-Face Interaction

Dataset	Numb. of people	Total time (minutes)	Numb. annotated frames		Max. group size (***)	Scenario (context)	Sensors			
			F-Formations	Social Actions			Video	Audio	Wireless	Accel.
Cocktail [70]	6	30	320	0	6	Mingle in a lab environment	X			
CoffeeBreak [21]	10	-	120	0	-	Outdoor mingle in social event	X			
Big Game [35]	32	30	600	0	4	Indoor quiz game in teams	X		X	
Idiap* [37]	50	360	82	0	5	Indoor poster session	X			
SALSA* [1]	18	60	1200	0	7	Indoor poster session+Mingle	X	X	X	X(4*)
MatchNMingle*	**92**	120	**4200****	**36000**	8	Indoor SpeedDate event+Mingle	X	X	X	X

Numerical and sensor comparison with other datasets.
**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.*
*4*SALSA provides processed accel., instead of raw triaxial.*

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 [69].

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 [28].

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 shows a numerical comparison of datasets oriented specifically to f-formation detection and free-standing conversational groups, and compares these in terms of modalities.

2.1.1 Vision-Only Datasets

The *cocktail party* dataset, published by Zen et al. [70], 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. [21] has been used in several works on the detection of people's position and orientation in images, and in detection of F-Formations [57], [58]. This dataset consists of the free interactions of 10 people. Hung and Kröse proposed the *IDIAP poster* dataset [37], 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.

2.1.2 Wireless Communication Datasets

Along with video, works using wireless communication have had a significant impact in group detection. For example, Cattuto et al. [16] 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 [16] was later used by Isella et al. [39] 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. [44] 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. [7] 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. [46] 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.

2.1.3 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. [35] 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 [36]. Also in [46], Matic et al. collected another set of data in a multimodal approach including accelerometers and proxemics (RSSI values). Although recordings

TABLE 3
Speed Dating Events

Dataset	Numb. dates	Time Date (min)	Sensors			
			Video	Audio	Wireless	Accel.
Madan et al. [43]	57	5		X		
SpeedDate Corpus [40]	991	4		X		
Veenstra and Hung [64]	64	5	X			
MatchNMingle*	674	3	X	X	X	X

Numerical and sensor comparison with other datasets.
**Dataset is (or will be) publicly available.*

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 [18]. In addition, they gathered information about personality traits using the Big-5 [24], 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 4.3.2), 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 [6] for review and Section 3.3.1 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 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 [48]), 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. [21]. 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 section.

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 [42]). 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 ([12], [20]), video ([53], [62]) and audio ([38], [47]). These studies employed ratings of media given by participants or trained raters, with the exceptions of [38], who transcribed interactions and subjected the transcripts to text analysis software, and [47] 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 assess how speed dates unfold. Table 3 compares all these efforts to *MatchNMingle*.

First, Madan et al. [43] and Pentland [52] 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 [52] to levels of attraction and friendship. Jurafsky et al. [40] 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. [40] to detect whether the speaker is awkward, friendly, or flirtatious, and by Ranganath et al. [54], [55] to investigate the difference between intention and perception during speed dates.

To the best of our knowledge, Veenstra and Hung [64] 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 3.3.1).

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]

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 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.

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 3.3.1).

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.

3.3 Offline Data Collection

3.3.1 Questionnaires

In order to test individual differences among participants, the initial online registration survey included 1) the HEXACO personality inventory [3], 2) the brief Self Control Scale (SCS) [60] and 3) the revised Sociosexual Orientation Inventory (SOI) [51].[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 [9], [61]. 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 [6] for review), the six-dimensional HEXACO model has been shown to more optimally capture the data in cross-cultural replications [5], 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 [4]).

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 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 [6] 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).

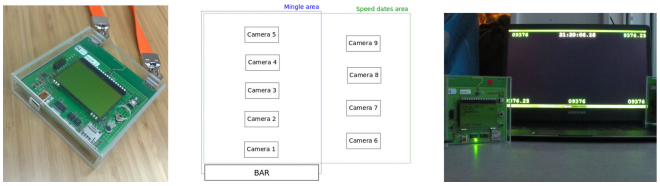
The brief Self Control Scale was designed to assess dispositional self-control and consists of 10 items such as *I am good at resisting temptation*. Each of these items can be rated in a scale of 1 (*not at all like me*) to 5 (*very much like me*).

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:

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.

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.

4. http://hexaco.org/



(a) (b) (c)

Fig. 3. (a) Custom-made wearable device. (b) Distribution of top view cameras on the venue area. (c) Cameras and devices' synchronization.

specifically, unrestricted individuals have been shown to have more short-term sexual encounters, consider uncommitted sexual relationships positively, and engage in more flirtatious behavior [9], [51]. Similarly to the other surveys, this could be answered in a scale of 1 to 5.

3.3.2 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 Fig. 1). We collected also these as prior research has shown that facial attributes, such as facial height-width ratio [63] or closeness of a person's face to the mean face [32] correlate with perceived attractiveness.

3.3.3 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 $\sim$3 mm diameter, with 3 cm lengths from the point closest to the scalp, as in prior research [49]. 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.

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.

3.4.1 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 Fig. 3a) are custom-made wearable devices designed specifically for applications on social interaction and group dynamics analysis [25]. They record triaxial acceleration at 20 Hz with a maximum range of $\pm$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 [25] more for technical details.

3.4.2 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 $\times$ 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). Fig. 3b 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 Fig. 3c. 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. Fig. 4a shows snapshots from 4 of the cameras recording during the speed dates. These snapshots correspond to cameras 6 to 9 on Fig. 3b. In Fig. 4b are shown snapshots from 5

(a) (b)

Fig. 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.

cameras (1 to 5 from Fig. 3b) 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.

3.4.3 Speed Date Responses

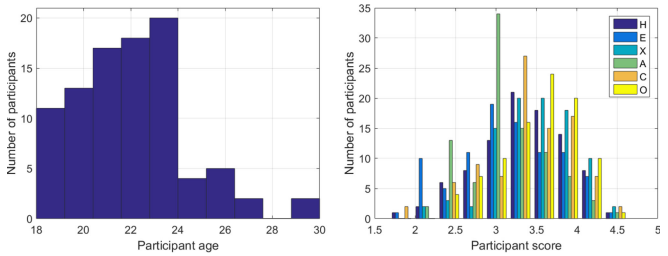
During each date, participants completed a questionnaire in the form of a booklet, designed to resemble materials from a commercial speed-dating event. The booklet format was used so that the participant could hold one end upright, preventing their interaction partner from seeing their responses. After 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.

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.[5] Snacks were also available for purchase.



(a) (b)

Fig. 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.

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.

4.1 Participant Statistics

As introduced in Section 3.2, our participants were mostly students that were not acquainted before the event. Fig. 5a shows the proportion of participants of different ages (mean=22.09, std=2.34). Similarly, Fig. 5b 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 of scales [22]. 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 [27]. 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

5. Bar's staff were not members of our experiment, so a detailed number of alcoholic drinks ingested by participant could not be collected.

TABLE 4
Number of Answers to the Question *Do you want to see this person again?* During the Speed Dates

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

Total of date interactions = 674.

TABLE 5
Summary of Dates and Date Interaction for Participants Carrying a Functional Wearable Device

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%)

Original data: 674 date interactions (Day 1= 224, Day 2=225, Day 3=225)
**Date interactions where the 2 have a functional device.*

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 1,000 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 $t(90) = 0.537$, $p = 0.593$.

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 1,348 match booklet sets of answers, one for each date (6 answers per set/date). From these, half correspond to female responses.

4.2.1 Speed Dates Matches

Table 4 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. Fig. 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.

4.2.2 Wearable Acceleration

Due to hardware malfunctioning, wearable acceleration was not recorded for some of the date interactions, either for both or one participant. Table 5 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. attraction). *MatchNMingle* gives us the possibility to further study these open questions, which we first approach in Section 6.

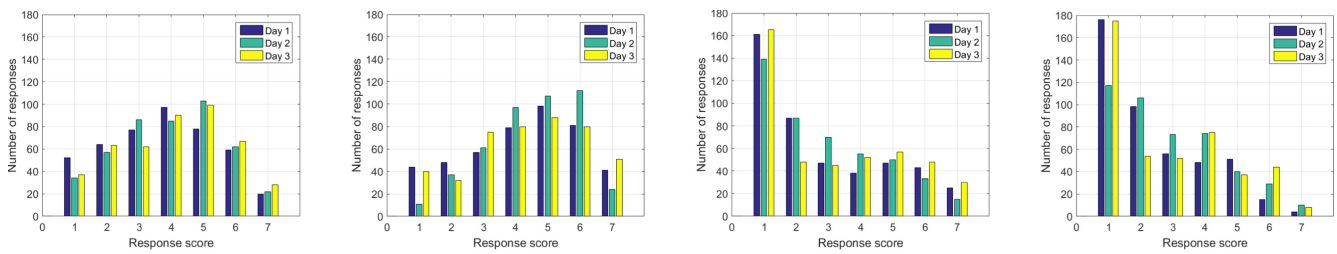


Fig. 6. Distribution of score responses for all dates divided by day. All scores are within a range of 0-low and 7-high. (a) How much would you like to see this person again? (b) How would you rate this person as a potential friend? (c) How would you rate this person as a a short term sexual partner? (d) How would you rate this person as a long term romantic partner?

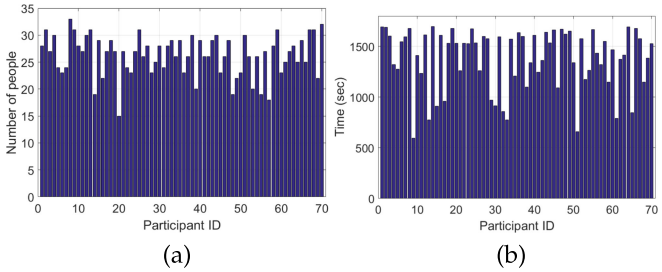


Fig. 7. General statistics from proximity during the mingle. (a) Total number of people interacted with per participant. (b) Longest interaction.

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).

4.3.1 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.

4.3.2 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.

Fig. 7a shows the total of number of people with whom each participant interacted during the event. In addition, Fig. 7b shows the longest interaction of each person during the event. These figures show the dynamic nature of the event. 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 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 percent, 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 7.

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.

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 ([11], [19], [56]). 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 [29].

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. [65] 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).

5.1.1 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.

5.1.2 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 [34], we treated gesture (hand or head) as any intentional or unintentional movement of the hand or head. These annotations can be later subcategorized according to each researcher's needs.

5.1.3 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 ([8], [17], [59]). 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.

5.1.4 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.

5.1.5 Space and Environment

This category describes 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).

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 section) and 10 minutes for the F-Formations.[6] 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 [48]. Also, the action for *walking* represents a long spatial displacement, whereas *stepping* is used for changes of posture in a limited space.

6. A detailed description of times is provided with the dataset.

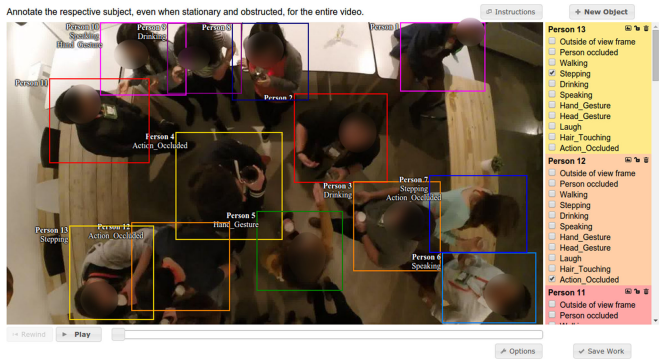


Fig. 8. Modify vatic tool [67] for our manual annotation process.

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*.

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.[7] These were enough to ensure that all participants had annotations for at least 75 percent 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 Fig. 4b).

The Vatic tool proposed by Vondrick et al. [67] 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 Fig. 8). 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

7. 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.

bounding box included check box es for each of the 8 social actions, as can be seen in on the right in Fig. 8. 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.

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,[8] 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 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.

5.4.1 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):

$$r(t) = \frac{Area(BB_1(t)) \cap Area(BB_2(t))}{min(Area(BB_1(t)), Area(BB_2(t)))}. \quad (1)$$

We used the minimum area as denominator in Eq. (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.

5.4.2 Complex HIT (Social Actions)

For the pilot tests for this task, the 8 social actions presented in Section 5.1 were annotated for 12 participants for an interval of 2 minutes. This was done using 3 different trained 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.[9] All on-site annotators passed the this run at the first try.

Fig. 9a shows the distributions of the Fleiss'-Kappa coefficient for all actions over participants (eg. one different k

8. Widely used to assess multiple inter-annotator agreement.

9. The practice run is passed if the worker accomplishes 70 percent overlap with the gold standard for a 1-minute HIT.

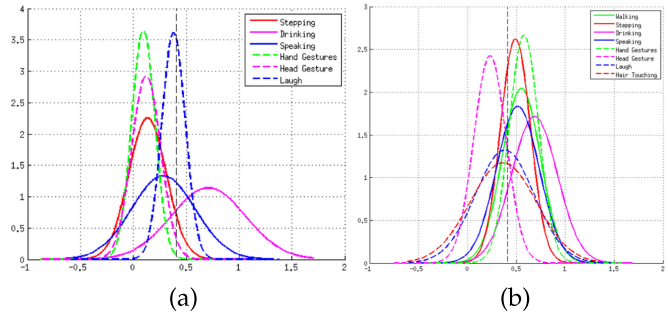


Fig. 9. Distribution of inter-agreement Fleiss'-Kappa coefficient (k) over participants annotated for 3 annotators. (a) MTurk workers. (b) On-site trained annotators. Dotted vertical line represents the inter-agreement threshold for moderate agreement. *Walking* and *Hair Touching* are excluded for the MTurk workers as they ignored these classes (no labels).

value per participant annotated) in the MTurk pilot test, while Fig. 9b shows the same for the on-site annotators. As can be seen on Fig. 9a, 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 Fig. 9b. 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 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 crowd-sourcing 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.

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 Fig. 9b, 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 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.

Overall,[10] 51 of the 92 participants are visible in one of the annotated cameras for the entire 30 minutes, 69 for 90 percent of this time, 82 for 80 percent, 88 for 70 percent, 88 for 60 percent, and 89 for 50 percent. Only one participant was missing for more than 80 percent of the time, as he left the event early.

At a resolution of 20 FPS, the 30 minutes annotated correspond to 36,000 samples. We calculated the percentage of this time interval in which each social was annotated as occurring. 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 observerd 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.

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. Fig. 10 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 (as seen also in Fig. 7) and variations through time for the same person when they leave, form or merge groups.

6 Experiments using *MatchNMingle*

In the following section, we provide examples of some of the different research questions that can be answered using *MatchNMingle*. To illustrate the richness of the

10. These statistics correspond to the *version 2* of our dataset.

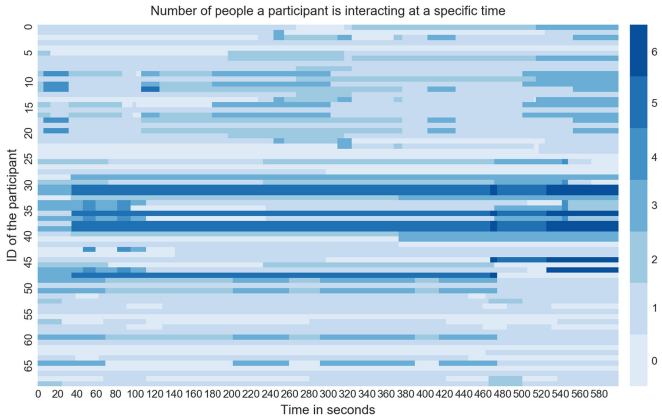


Fig. 10. Number of people each participant is interacting with at a specific time.

dataset, our examples address research questions at differing levels of abstraction; from cue/action detection and prediction to behaviors. Note that these tasks still have several open questions. Moreover, our examples demonstrate the range of possibilities of *MatchNMingle* as a multimodal resource.

6.1 Attraction Detection

Speed-dating events offer an ecologically-valid context to study initial interactions, while collecting relevant questionnaire data [26]. Although the use of automated techniques has shown promise [31], behavioral assessments are often gathered by human raters in the social sciences when analyzing nonverbal behavior.

Here, we instead attempt to automatically classify attraction levels using movement-based features. Thus, we recreate as close as possible the features presented by Veenstra and Hung [64], using information from the wearables instead of video, leveraging movement 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.

As seen in Table 5, we have wearable acceleration and booklet responses from 10,052 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 1s sliding window with a shift of 0.5s. Similar to [45], we found this preprocessing step provided a good indication for movements in the acceleration signal. 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. Similar to those in [64], these features based on movement are an indicative of arousal.

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, [64] 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 [40], [52]. 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 $p < 0.001$). The remaining classes cannot be compared directly to [64] as the questions in that study differ from ours.

This experiment shows that using movement-based features from the accelerometers to measure arousal, emulating what was done by Veenstra and Hung [64] in a simple classification setup, provides rather acceptable F-scores for the two classes related to attraction (Romantic and Sexual interest). Furthermore, one could use *MatchNMingle* to further investigate the relation between non-verbal behavior and attraction and increase this performance.

6.2 Speaker Detection

Another possible research direction with *MatchNMingle* is the analysis and automatic detection of socially relevant cues. There are studies that have investigated social behavior in various time resolutions (long-term [50], [68] versus short term [2], [16]), employing different modalities (multimodal sensors [50], video [2], accelerometers [35], etc.). One specific direction is the automatic classification of social actions, such as speaking or gesturing, in a short-term dense crowded scenario [35], [36]. Here, we present a simple baseline experiment for such a scenario where we use acceleration data and speaking annotations 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 [29].

As explained in Section 5.5, due to the dynamic nature of the event, all participants are not always visible under the camera and consequently have different amounts of annotation labels. 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 acclimitization, we refer to the warming up period of the mingle where people are entering and introducing or re-introducing themselves to one another. Unlike [29] 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

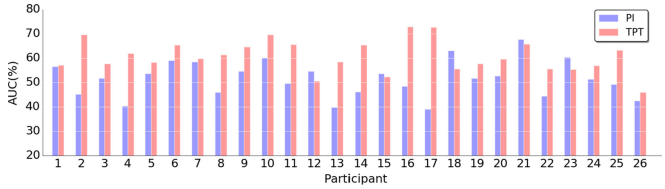


Fig. 11. Performance in terms of AUC for speaking status detection. PI: Traditional person independent setup. TPT: Transductive Parameter Transfer.

from the second and third days. We have selected the 10 minutes from each day that had the maximum number of participants with consecutive acceleration data and speaking status labels without any interruption.[11] Thus, we obtained a diverse subset, including different types of interactions and varying total speaking times per participant. For the training set, the percentages of the speaking intervals for participants differed between 4 to 59 percent, with the mean and standard deviation of 32 and 15 percent, respectively. Similarly, the percentages for the test set was between 6 to 66 percent, with mean and standard deviation of 27 to 16 percent. 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 [29] 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 [29]. 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 Fig. 11.

As can be seen from Fig. 11, 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 percent), 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 [29].

6.3 Personality Estimation

Thanks to the self-assessed personality traits provided by the HEXACO inventory in *MatchNMingle*, it is possible for researchers to analyze and develop new methodologies for the automatic classification of such traits in a dynamic (mingle) and/or sitting (speed dates) scenario. One such study, which uses *MatchNMingle*, is presented by Cabrera-Quiros et al. [13]. Their main results are summarized below.

In this study, the authors used wearable acceleration and proximity data for automatic personality estimation during the mingle scenario. They leveraged three different behavioral modalities (speaking turns, body movement energy and proximity), originating from the two aforementioned digital modalities. Instead of employing audio, this study used the Transductive Parameter Transfer method mentioned in the former section to obtain speaking turns from the acceleration. The personality detection was treated as a binary classification problem, where each item of the HEXACO inventory yielded one label for each participant, as positive or negative. This labelling was obtained by finding the median values for each item and placing participants with higher values into the positive class and vice a versa. A logistic regressor was selected as the classifier where the regularization parameter was set with 10-fold cross validation. Performances are then obtained with the single and whole possible combinations of the features originating from different behavioural modalities.

Apart from eXtraversion, authors obtained performances better than random for all traits. For different traits, different feature combinations yielded the best performance, but generally, multi-modality increased the performance. For the Honesty, Conscientiousness and Openness to Experience, the authors obtained an accuracy close to 70 percent. More interestingly, using two behavioral modalities originating from the same digital modality (movement and speaking turns) increased the performance for most of the items. Also, adding proximity, an additional digital modality, provided an increase in the accuracy of almost all traits, further showing the benefits of a multi-modal approach. For a more details please refer to [13].

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 percent) for the speed dates and 22 (24 percent) 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 5).

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 percent 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.[12]

This is the reason why the recall reported in Section 4.3.2 (90 percent) is high while the precision is rather low

11. Participant numbers and times used in this study (subset) will be provided with the dataset.

12. The human body has proved to be a natural damping for radio communications [33], limiting back to bakc detections.

(33 percent). 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 Fig. 4b), 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 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.[13] The limitations of this dataset for the analysis of facial cues was also discussed in Section 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. 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.

8 Conclusions and Future Work

In this paper we introduced *MatchNMingle*, a multimodal resource for the analysis of human behavior in social interactions during standing conversations in groups (mingle), and in seated dyads (speed date). *MatchNMingle* is, to the best of our knowledge, the dataset with the largest number of participants (92), longest recording time (2 hours), and largest set of manual annotations for social actions (36,000 samples) that is publicly available in the context of free-standing conversations and speed dates.

Multimodal data were collected according to a framework that adheres to the standards of computer sciences and social psychology, allowing for future research in each field as well as interdisciplinary collaboration.

We detailed the framework used to collect the dataset and the main statistics of the data collected. In addition, we described the process of manually annotation for the social actions (or social cues) and provided a comparison of the performance of trained annotators with workers hired using crowd-sourcing tools. For the latter, we separate this comparison by task complexity and demonstrate that crowdsourcing is a good option for simple tasks (such as manually tracking a person in a video), while more complex tasks such as social actions should preferably be relegated to more qualified annotators.

In addition, we provided example experiments of some of the different research questions that can be answered using *MatchNMingle*, this addressing research questions at differing levels of abstraction.

We strongly believe that there is a wide 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.

13. The OpenFace [10] and OpenPose [15] toolbox were applied to this data, with discouraging results. Please refer to our website for more about this topic and the results mentioned.

Acknowledgments

Thanks to Sebastian Deuten, Veronica de Groene, Shari Molawi, Julia Emmer, Phuong Khanh Vu, Marina Tulin, Clara De Inocencio, Maria I. Rinderu, Jordy Jouby and Catherine Mohlo for their help collecting this dataset; Maarten Jonker and Leon Harmsen, managers of *Il Cafe*; and also our anonymous participants. This paper was partially funded by the Dutch national program COMMIT, the Instituto Tecnológico de Costa Rica and the Netherlands Organization for Scientific Research (NWO) under project number 639.022.606. Laura Cabrera-Quiros and Andrew Demetriou are contributed equally.

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 Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1707–1720, Aug. 2016.

[2] X. Alameda-Pineda, Y. Yan, E. Ricci, O. Lanz, and N. Sebe, "Analyzing free-standing conversational groups: A multimodal approach," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 5–14.

[3] M. C. Ashton and K. Lee, "Empirical, theoretical, and practical advantages of the HEXACO model of personality structure," *Personality Social Psychology Rev.: An Official J. Society Personality Social Psychology*, vol. 11, no. 2, pp. 150–166, 2007.

[4] M. C. Ashton and K. Lee, "The prediction of Honesty–Humility-related criteria by the HEXACO and Five-Factor Models of personality," *J. Res. Personality*, vol. 42, pp. 1216–1228, 2008.

[5] M. C. Ashton, K. Lee, M. Perugini, P. Szarota, R. E. de Vries, L. DiBlas, K. Boies, and B. De Raad, "A six-factor structure of personality-descriptive adjectives: Solutions from psycholexical studies in seven languages," *J. Personality Social Psychology*, vol. 86, no. 2, pp. 356–366, 2004.

[6] 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 Social Psychology Rev.*, vol. 18, no. 2, pp. 139–152, 2014.

[7] M. Atzmueller, T. Thiele, G. Stumme, and S. Kauffeld, "Analyzing group interaction and dynamics on socio-behavioral networks of face-to-face proximity," in *Proc. Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2016, pp. 1231–1238.

[8] S. O. Ba and J.-M. Odobez, "Recognizing visual focus of attention from head pose in natural meetings," *IEEE Trans. Syst. Man Cybern., Part B (Cybern.)*, vol. 39, no. 1, pp. 16–33, Feb. 2009.

[9] 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," *Eur. J. Personality*, vol. 25, pp. 120–132, 2011.

[10] T. Baltrušaitis, P. Robinson, and L. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–10.

[11] L. Bao and S. Intille, "Activity recognition from user-annotated acceleration data," in *Proc. Int. Conf. Pervasive Comput.*, 2004, pp. 1–17.
[12] R. Bowers, S. Place, P. M. Todd, L. Penke, and J. B. Asendorpf, "Generalization in mate-choice copying in humans," *Behavioral Ecology*, vol. 23, pp. 112–124, 2012.
[13] L. Cabrera-Quiros, E. Gedik, and H. Hung, "Estimating self-assessed personality from body movements and proximity in crowded mingling scenarios," in *Proc. Int. Conf. Multimodal Interaction*, 2016, pp. 238–242.
[14] L. Cabrera-Quiros and H. Hung, "Who is where?: Matching people in video to wearable acceleration during crowded mingling events," in *Proc. Conf. Multimedia*, 2016, pp. 267–271.
[15] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1302–1310.
[16] 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*, vol. 5, 2010, Art. no. e11596.
[17] I. Chamveha, Y. Sugano, Y. Sato, and A. Sugimoto, "Social group discovery from surveillance videos: A data-driven approach with attention-based cues," *British Mach. Vis. Conf. (BMVC)*, 2013.
[18] T. Choudhury and A. Pentland, "Sensing and modeling human networks using the sociometer," *IEEE Int. Symp. Wearable Comput.*, 2003, pp. 216–222.
[19] D. Cook, K. D. Feuz, and N. C. Krishnan, "Transfer learning for activity recognition: A survey," *Knowl. Inf. Syst.*, vol. 36, no. 3, pp. 537–556, 2013.
[20] J. C. Cooper, S. Dunne, T. Furey, and J. P. O. Doherty, "Dorsomedial prefrontal cortex mediates rapid evaluations predicting the outcome of romantic interactions," *J. Neurosci.*, vol. 32, pp. 15647–15656, 2012.
[21] 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," in *Proc. Brit. Mach. Vis. Conf.*, 2011, pp. 23.1–23.12.
[22] L. J. Cronbach, "Coefficient alpha and the internal structure of tests," *Psychometrika*, vol. 16, no. 3, pp. 297–334, 1951.
[23] A. Demetriou, "Rose colored lenses: The role of testosterone and cortisol in mate assessment," Master's thesis, Social Psychology Department, VU University of Amsterdam, Amsterdam, 2015.
[24] J. Digman, "Personality structure: Emergence of the five-factor model," *Annu. Rev. Psychology*, vol. 41, pp. 417–440, 1990.
[25] M. Dobson, "Low-power epidemic communication in wireless ad hoc networks," PhD thesis, Faculty of exact sciences, Vrije Universiteit Amsterdam, 2013.
[26] E. Finkel, P. Eastwick, and J. Matthews, "Speed-dating as an invaluable tool for studying romantic attraction: A methodological primer," *Pers. Relationships*, vol. 14, pp. 149–166, 2007.
[27] R. Furr and R. Bacharach, *Psychometrics An Introduction*, 2nd ed. Thousand Oaks, CA, USA: SAGE Publications Ltd, 2014.
[28] D. Gatica-Perez, O. Aran, and D. Jayagopi, "Small group analysis," in *Social Signal Processing*. Cambridge, U.K.: Cambridge Univ. Press, 2017.
[29] E. Gedik and H. Hung, "Personalised models for speech detection from body movements using transductive parameter transfer," *Pers. Ubiquitous Comput.*, vol. 21, pp. 723–737, 2017.
[30] E. Gedik and H. Hung, "Speaking status detection from body movements using transductive parameter transfer," in *Proc. Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2016, pp. 69–72.
[31] K. Grammer, M. Honda, A. Juette, and A. Schmitt, "Fuzziness of nonverbal courtship communication unblurred by motion energy detection," *J. Personality Social Psychology*, vol. 77, pp. 487–508, 1999.
[32] K. Grammer and R. Thornhill, "Human (Homo-sapiens) facial attractiveness and sexual selection–The role of symmetry and averageness," *J. Comparative Psychology*, vol. 108, no. 3, pp. 233–242, 1994.
[33] P. Hall, Y. Hao, V. Nechayev, A. Alomain, C. Constantinou, C. Parini, M. Kamarudin, T. Salim, D. T. M. Heel, R. Dubrovka, A. Owadall, W. Song, A. Serra, P. Nepa, M. GalIo, and M. Bozzetti, "Antennas and propagation for on-body communication systems," *IEEE Antennas Propag. Mag.*, vol. 49, no. 3, pp. 41–58, Jun. 2007.
[34] R. A. Hinde, *Non-Verbal Communication*. Cambridge, U.K.: Cambridge Univ. Press, 1971.
[35] H. Hung, G. Englebienne, and J. Kools, "Classifying social actions with a single accelerometer," in *Proc. Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2013, pp. 207–210.
[36] H. Hung, G. Englebienne, and L. Cabrera-Quiros, "Detecting conversing groups with a single worn accelerometer," in *Proc. ACM Int. Conf. Multimodal Interaction*, 2014, pp. 84–91.
[37] H. Hung and B. Kröse, "Detecting F-formations as dominant sets," in *Proc. Int. Conf. Multimodal Interfaces*, 2011, pp. 231–238.
[38] 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 Sci.*, vol. 22, pp. 39–44, 2011.
[39] 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," *J. Theoretical Biol.*, vol. 271, pp. 166–180, 2011.
[40] D. Jurafsky, R. Ranganath, and D. McFarland, "Extracting social meaning: Identifying interactional style in spoken conversation," in *Proc. Human Lang. Technol.: The Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2009, pp. 638–646.
[41] A. Kendon, "Conducting interaction: Patterns of behavior in focused encounters," Cambridge University Press, 1990.
[42] D. A. Kenny, "Interpersonal perception: A social relations analysis," *J. Social Pers. Relationships*, vol. 5, no. 2, 1988.
[43] A. Madan, R. Caneel, and A. Pentland, "Voices of attraction," 2004.
[44] C. Martella, M. Dobson, A. van Halteren, and M. van Steen, "From proximity sensing to spatio-temporal social graphs," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun.*, 2014, pp. 78–87.
[45] C. Martella, E. Gedik, L. Cabrera-Quiros, G. Englebienne, and H. Hung, "How was it?: Exploiting smartphone sensing to measure implicit audience responses to live performances," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 201–210.
[46] A. Matic, V. Osmani, and A. Maxhuni, "Multi-modal mobile sensing of social interactions," in *Proc. Int. Conf. Pervasive Comput. Technol. Healthcare*, 2012, pp. 105–114.
[47] D. A. McFarland, D. Jurafsky, and C. Rawlings, "Making the connection: Social bonding in courtship situations," *Amer. J. Sociology*, vol. 118, pp. 1596–1649, 2013.
[48] M. M. Moore, "Nonverbal courtship patterns in women: Context and consequences," *Ethology Sociobiology*, vol. 6, no. 4, pp. 237–247, 1985.
[49] G. Noppe, Y. B. De Rijke, K. Dorst, E. L. Van DenAkker, and E. F. Van Rossum, "LC-MS/MS-based method for long-term steroid profiling in human scalp hair," *Clinical Endocrinology*, vol. 83, no. 2, pp. 162–166, 2015.
[50] D. Olguín, B. N. Waber, T. Kim, A. Mohan, K. Ara, and A. Pentland, "Sensible organizations: Technology and methodology for automatically measuring organizational behavior," *IEEE Trans. Syst. Man Cybern. Part B: Cybern.*, vol. 39, no. 1, pp. 43–55, Feb. 2009.
[51] L. Penke and J. B. Asendorpf, "Beyond global sociosexual orientations: A more differentiated look at sociosexuality and its effects on courtship and romantic relationships," *J. Personality Social Psychology*, vol. 95, no. 5, pp. 1113–1135, 2008.
[52] A. Pentland, "Social dynamics: Signals and behavior," in *Proc. Int. Conf. Digital Libraries*, 2014, pp. 271–275.
[53] 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 Human Behavior*, vol. 33, pp. 547–550, 2012.
[54] R. Ranganath, D. Jurafsky, and D. McFarland, "It's not you, it's me: Detecting flirting and its misperception in speed-dates," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2009, pp. 334–342.
[55] R. Ranganath, D. Jurafsky, and D. McFarland, "Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates," *Comput. Speech Lang.*, vol. 27, pp. 89–115, 2013.
[56] N. Ravi, N. Dandekar, P. Mysore, and M. Littman, "Activity recognition from accelerometer data," in *Proc. 17th Conf. Innovative Appl. Artif. Intell.*, 2005, pp. 1541–1546.
[57] F. Setti, H. Hung, and M. Cristani, "Group detection in still images by F-formation modeling: A comparative study," in *Proc. Int. Workshop Image Anal. Multimedia Interactive Serv.*, 2013, pp. 1–4.
[58] F. Setti, O. Lanz, R. Ferrario, V. Murino, and M. Cristani, "Multi-scale f-formation discovery for group detection," in *Proc. Int. Conf. Image Process.*, 2013, pp. 3547–3551.

[59] 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, in *Proc. 15th ACM Int. Conf. Multimodal Interaction*, 2013, pp. 3–10.
[60] J. P. Tangney, R. F. Baumeister, and A. L. Boone, "High self-control predicts good adjustment, less pathology, better grades, and interpersonal success," *J. Personality*, vol. 72, no. 2, pp. 271–324, 2004.
[61] 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," *Pers. Relationships*, vol. 20, pp. 199–215, 2013.
[62] 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. Nat. Acad. Sci. United States America*, vol. 113, pp. 4009–4014, 2016.
[63] 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 Sci.*, vol. 25, pp. 806–811, 2014.
[64] A. Veenstra and H. Hung, "Do they like me? using video cues to predict desires during speed-dates," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2011, pp. 838–845.
[65] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image Vis. Comput.*, vol. 27, pp. 1743–1759, 2009.
[66] 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 Trans. Affective Comput.*, vol. 3, no. 1, pp. 69–87, Jan.–Mar. 2012.
[67] C. Vondrick, D. Patterson, and D. Ramanan, "Efficiently scaling up crowdsourced video annotation," *Int. J. Comput. Vis.*, vol. 101, pp. 184–204, 2013.
[68] D. Wyatt, "Collective modeling of human social behavior," in *Proc. AAAI Spring Symp.: Human Behavior Model.*, 2009.
[69] K. W. Z., Y. Xiang, M. Aalsalem, and Q. Arshad, "Mobile phone sensing systems: A survey," *IEEE Commun. Surv. Tut.*, vol. 15, no. 1, pp. 402–427, Jan.–Mar. 2013.
[70] G. Zen, B. Lepri, E. Ricci, and O. Lanz, "Space speaks: Towards socially and personality aware visual surveillance," in *Proc. Int. Workshop Multimodal Pervasive Video Anal.*, 2010, pp. 37–42.

Laura Cabrera-Quiros received the 'licenciatura' and master degrees from the Instituto Tecnólogico de Costa Rica, in 2012, 2014, respectively. She is working toward the PhD degree in the Pattern Recognition and BioInformatics Group, Delft University of Technology, working on automatic social behavior analysis using multimodal streams. In 2014, she received a full scholarship by the Costa Rican goverment to pursue her postgraduate studies abroad. Her main interest is the use and fusion of wearable sensing and computer vision for applications oriented to analysis of social behavior.

Andrew Demetriou received the research master's degree in social psychology from VU Amsterdam, in 2015, with a focus on biological data collection methods, and mate choice/romantic attraction. He is working toward the PhD degree in the Multimedia Computing group, TU Delft and the Psychology Department, University of Northumbria at Newcastle. His research has been published in Letters on Evolutionary Behavioral Science, the *Journal of Crime and Delinquency*, Proceedings of ISMIR 2016, and Proceedings of 10th ACM Conference on Recommender Systems. His research interests include social and romantic bonding, optimal mental/physiological states (e.g., flow, mindfulness), and how music, along with biological and sensor data, can be used to study these phenomena.

Ekin Gedik received the bachelor's and master's degrees from the Middle East Technical University, Turkey, in, 2010, 2013, respectively. He is currently working toward the PhD degree in the Pattern Recognition and Bioinformatics Group of Delft University of Technology. His research interests include but are not limited to social behaviour analysis, wearable sensing, affective computing and pattern recognition. He is currently focused on analysis and detection of social behaviours, interaction and their connection to various social phenomena.

Leander van der Meij received the PhD degree in neuroscience with the University of Valencia, Spain and University of Groningen, The Netherlands, in 2012. He is an assistant professor at Eindhoven University of Technology since 2016. Between 2013-2016 he was an assistant professor with the Vrije Universiteit Amsterdam. He studies how hormones may stimulate affiliation or aggression and how it depends on social context.

Hayley Hung received the PhD degree in Computer Vision from Queen Mary University of London, in 2007 and her first degree from Imperial College, United Kingdom in Electrical and Electronic Engineering. She is an assistant professor with the Pattern Recognition and Bioinformatics group, TU Delft, The Netherlands, since 2013. Between 2010-2013, she held a Marie Curie Fellowship with the Intelligent Systems Lab, University of Amsterdam. Between 2007-2010, she was a post-doctoral researcher with Idiap Research Institute in Switzerland. Her research interests are social computing, social signal processing, computer vision, and machine learning. She is a member of the IEEE.