Detecting Conversing Groups Using Social Dynamics from Wearable Acceleration: Group Size Awareness

EKIN GEDIK and HAYLEY HUNG, Delft University of Technology, Netherlands

In this paper, we propose a method for detecting conversing groups. More specifically, we detect pairwise F-formation membership using a single worn accelerometer. We focus on crowded real life scenarios, specifically mingling events, where groups of different sizes naturally occur and evolve over time. Our method uses the dynamics of interaction, derived from people's coordinated social actions and movements. The social actions, speaking, head and hand gesturing, are inferred from wearable acceleration with a transfer learning approach. These automatically labeled actions, together with the raw acceleration, are used to define joint representations of interaction between people through the extraction of pairwise features. We present a new feature set based on the overlap patterns of social actions and utilize some others that were previously proposed in other domains. Our approach considers various interaction patterns of different sized groups by training multiple classifiers with respect to cardinality. The final estimation is then dynamically performed by meta-classifier learning using the local neighborhood of the current test sample. We experimentally show that the proposed method outperforms state of the art approaches. Finally, we show how the accuracy of the social action detection affects group detection performance, analyze the effectiveness of features for different group sizes in detail, discuss how different types of features contribute to the final performance and evaluate the effects of using the local neighborhood for meta-classifier learning.

CCS Concepts: • **Human-centered computing → Empirical studies in ubiquitous and mobile computing**; *Human computer interaction (HCI)*; • **Computing methodologies → Ensemble methods**; *Supervised learning by classification.*

Additional Key Words and Phrases: F-formation detection, wearable acceleration, conversing group detection, dynamic ensemble selection

ACM Reference Format:
Ekin Gedik and Hayley Hung. 2018. Detecting Conversing Groups Using Social Dynamics from Wearable Acceleration: Group Size Awareness. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 163 (December 2018), 24 pages. https://doi.org/10.1145/3287041

1 INTRODUCTION

In most social scenarios, be it a small private gathering or a crowded music festival, people tend to interact with each other, forming groups of varying size. Automatic detection of such conversing groups has a wide variety of possible applications, ranging from surveillance to detailed analysis of socially relevant behavior. For example, a deeper understanding of how people interact throughout an event can provide valuable information regarding the success of the event, such as the mood of the participants, their evaluations, whether they will return and recommend the event to others[26]. Such information is potentially important for organizers. Also, examples in the literature show that when an interacting group is identified, it is possible to estimate social attributes such as dominance [20], leadership[21] and cohesion[18] through behavior of participants, which are especially valuable

Authors' address: Ekin Gedik, e.gedik@tudelft.nl; Hayley Hung, h.hung@tudelft.nl, Delft University of Technology, Delft, Netherlands.

2474-9567/2018/12-ART163 $15.00
https://doi.org/10.1145/3287041

for organizational scenarios. In order to obtain such information, a deeper understanding of social interactions between people is required, which we address in this paper.

This paper focuses on the automatic detection of conversing groups using a single accelerometer in real life crowded social scenarios, more specifically mingling events. Differing from the majority of the existing work that relies solely on the proxemics (spatial distance and relative orientation of the participants), our method focuses on a widely overlooked rich information source: interaction dynamics; the coordinated behavior of people during a conversation. The approach presented in this paper aims to represent interaction dynamics through people's actions and movement patterns which are inferred from a single body worn triaxial accelerometer. Importantly, our proposed approach considers how interaction patterns vary in groups of different sizes. This is achieved by training multiple classifiers with respect to the group cardinality and classifying a new sample using a meta-classifier that is trained with the probability outputs of the group based classifiers. This ensemble fusion technique is known as stacked generalization (stacking in short) in the literature [40]. Unlike the traditional stacking approaches, we use only the data from the local neighborhood of the test sample while training the meta-classifier. Influenced by the findings in social psychology related to the speaker and listener behaviors in groups, we aim to provide a scalable and ubiquitous solution to the detection of conversing groups while preserving the privacy of the participants. Our method works solely on accelerometer data obtained by a custom sensor pack worn around the neck by participants like an ID badge. This makes the approach ubiquitous and viable for dense crowded scenarios, where other modalities such as video and sound may fail due to the crowded characteristics. Results are presented on a real life, in-the-wild mingling dataset, collected during three unique instances, showing the generalization of the proposed method. No recording of raw speech is done; participants' privacy is not invaded. Since only a single sensor is being used, our method is highly power-preserving, making it deployable for large and long scenarios.

We can list the novel contributions of this paper as follows: We (i) propose and utilize a new feature set (overlap statistics) together with some others that were previously used in other domains to capture interaction dynamics through social behavior (ii) propose an approach that considers interaction dynamics of groups related to cardinality and show how it beats the state-of-the-art, (iii) perform various analyses to investigate the nature of this problem further: performance in relation to the group sizes, feature effectiveness, comparison of feature types in terms of performance, and the effects of meta-level classifier on the local neighborhood instead of the whole dataset. Following subsections aim to provide more insight into the problem and the proposed method by presenting a formal problem formulation and discussions related to the use of social dynamics instead of proxemics and the necessity for group size awareness.

1.1 Proxemics vs. Dynamics

Explicitly, we are trying to detect pairwise F-formation membership; if two participants are in the same interacting group or not. We based our problem formulation on the definition of F-formations from the social psychologist Adam Kendon [23]. F-formations are specific types of focused encounter, where participants spatially and orientationally organize themselves into a group which makes it possible to facilitate conversation.

There are many examples of existing work in the literature that aim to F-formations. However, many of these works focus solely on the proxemics, either by working on the proximity information obtained with infrared (IR) sensors [7, 15] or by employing still images and videos for obtaining head and body orientations in addition to the spatial location [1, 33, 38]. This focus on proxemics is understandable and coherent with the definition of F-formation itself. If people were asked to decide from a still image if a participant is part of a group, most probably they will first check the physical proximity and the orientation of the people. However, we argue that even though spatial distance and orientation are extremely strong cues, they overlook one important aspect of the interaction: the dynamic behavior and actions of the participants.

Social scientists have already shown that interacting people tend to coordinate their movements [22] and even start to mimic each other in terms of posture, mannerism and other behaviors [6]. We believe such patterns can be used as informative cues for distinguishing between interacting and non-interacting partners. Thus, in this study we show how the dynamics of the social behavior of participants alone can be exploited to infer the conversational group membership of participants in real life scenarios.

Another reason to use the proposed method in this study is the practical limitations of inferring the spatial location and orientation of the people. In a real life scenario, images or video of the scene might not be always available. A method that relies on video input for conversing group detection will require a similar instrumentation in every event, making the solution less generalizable and non-ubiquitous. Also, characteristics of the event such as the density of the crowd or location of the cameras can affect the performance of the method. Another reason is privacy; not every participant will be comfortable with their videos being taken, since it is more intrusive than recording body acceleration. Another option is to use indoor localization solutions, but they have been shown to perform poorly in a high density scenario, where a commercial indoor GPS system was tested in [17].

A more pervasive solution than video based methods can be obtained with wearable sensors that provide proximity information, but they also have weaknesses. Two technologies are generally used for inferring proximity with wearable sensors; infrared (IR) and Bluetooth. A recent study that evaluated the use of wearable sensors in organizational settings investigated a custom sensor pack, sociometric badges [24], which uses these sensors for proximity detection [5]. Their experiments show that Bluetooth had greater accuracy in detecting the co-location of participants. However, detections were generally overly optimistic, indicating proximity even when there were obstacles between the sensors, such as a separating wall. IR, on the other hand, was more pessimistic in its detections, requiring clean line of sight and strict face-to-face orientation. In the types of crowded scenarios that we are interested in, a Bluetooth based approach will most probably put many participants in the same group whereas an IR based method will tend to miss participants in larger groups, since it requires strict face-to-face orientation, making those less suitable for precise group detection.

1.2 Group Cardinality

We hypothesize that the dynamics of the interaction differ greatly with respect to the cardinality of the group. Previous studies in computer vision have already shown how a cardinality-sensitive approach can improve performance [34]. We believe that the effect of group cardinality on the interpersonal dynamics is even greater. The assumption of one conversational flow per F-formation may not be always true. For large groups, sustaining a single informal conversation is not possible [10]. Even though multiple participants can be in the same F-formation of a larger cardinality, they might still form sub-groups inside, exhibiting behavior of smaller cardinalities. For example, a four person F-formation can have many different interaction characteristics. It could be an egalitarian group where everyone contributes to the interaction equally. However, there can be also two sub-groups exhibiting dyadic interaction characteristics. We hypothesize that with the increasing cardinality, possibilities for different interaction characteristics in the group increase significantly.

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

2 RELATED WORK ON THE DETECTION OF CONVERSING GROUPS

Automatic detection and analysis of interacting groups through computation has been a hot topic for computer science researchers under various names such as F-formation detection, modeling of human networks, detection of free standing conversing groups, etc. In this section we aim to provide a brief overview of such existing studies to show the foundations of our approach. We should also note that the analysis of human behavior in groups and interaction dynamics are extensively studied in other disciplines, especially social psychology [6, 23]. Findings and insights from those studies greatly affected and influenced computer science researchers.

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

2.1 Long-term Studies with Pervasive Devices

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

Using a similar device, a sociometric badge, Olguin et. al. focused on analyzing and measuring organizational behavior in their 2009 work [28]. They employed IR to detect face-to-face interactions and bluetooth for measuring physical proximity. They also made use of accelerometers for detecting physical activity levels and microphones for speech detection. Data collection took 27 days and included 67 participants. Then, interaction characteristics sensed through these multiple modalities were used to classify personality traits of participants into the “Big Five" model.

There are also examples of work focusing on long-term characteristics that do not employ specific wearable sensors. Eagle and Pentland came up with their study “Reality Mining" in 2006, that focused on the utilization of mobile phones for sensing complex social structures [11]. They collected data from 100 mobile phones over the period of 9 months. They aimed to infer various social concepts, such as the detection of social patterns in daily life, identification of significant relations, modeling of organizational rhythms and, most interesting to us, recognition of social interactions. They relied on bluetooth communication of mobile phones to detect people in close proximity. No quantitative evaluation of the proximity networks was provided, but they showed that various social groups, such as friends and daily occurrences can be distinguished and this information can be used for further social understanding.

Similarly, Madan et. al. used mobile phones as social sensors in their 2010 work that aimed to detect behavior changes with respect to illness [25]. For detecting social interactions of participants, they relied mainly on spatial distance, inferred from proximity and cellular-tower identifiers.

Wang et. al. presented their continuous sensing application, StudentLife, in 2014 [39]. A class of 48 students used the application for a 10 week term on their phones. The main aim of the study was to connect the automatic sensor data to the mental health and educational outcomes of the participants. Activity data and indoor and outdoor mobility were inferred from the accelerometer recordings. Audio from microphones was used to extract conversation data. A mix of cues related to light, activity, phone usage and sound were used to detect the sleeping patterns of the participants. Finally, location data was gathered from the GPS and co-location with other students was inferred through bluetooth. A number of significant correlations with various mental well-being surveys were found. Assuming the co-location and conversation-related measures are proxies for interaction, the results indicate that students with frequent social interactions tend to be less depressed and more flourishing.

2.2 Short-term Studies with Pervasive Devices

There are studies in the literature that use pervasive devices, custom sensor packs or mobile phones, for the analysis of short-term social interactions. By short term, we mean studies that focus on a single event that generally spans multiple hours. Gips and Pentland employed a custom sensor pack, UbER-Badge, embedded in a badge worn by conference attendees, to analyze interest and affiliation [14]. Specifically for affiliation, Gips argued for the use of cues related to wearer activity, inferred from accelerometers, in addition to proximity information obtained from IR encounters. In agreement with this study, he proposed to use pairwise measures, more specifically mutual information of the motion energy between pairs (MIME), to detect interacting partners in [15]. However, no qualitative performance evaluation was presented for the detection of interacting partners.

Similarly, Cattuto et. al. analyzed face-to-face interactions in various crowded social settings, including 25 to 575 people, by using custom conference badges equipped with RFID [4]. Exchange of radio packets between badges was treated as a proxy for inferring spatial distance between participants and ultimately used to detect face-to-face interactions. They focused on analyzing the dynamics of interaction networks, showing a super-linear behavior between the number of connections and their durations. However, this study assumed an interaction between two people in close proximity and the actual performance of this assumption was not quantitatively evaluated.

A relatively recent study from Matic et. al. used mobile phones to detect two parameters, interpersonal distance and relative body orientation, which were then used as proxies for inferring social interaction between participants [27]. Authors used a time frame of 10s, aiming to capture dynamic changes in social interactions. Their experiments showed that, using the standard deviation of body orientation throughout 10s windows as a feature in addition to the distance and body orientation, improves the detection of social interactions.

2.3 Static Image Based Methods

More recently, with the increasing success of computer vision methods, researchers started to use images and video as main input modalities. Cristani et. al. focused on unconstrained scenarios and employed solely visual cues to detect social interactions in [8]. Their proposed method took the positions and head orientations of people in the scene as input and employed a voting strategy built on the Hough transform. They presented their results on synthetic data and videos of real life indoor and outdoor scenarios, discussing how automatic detection of positions and head orientations in real life affects the performance. Promising performances on both real life datasets (outdoor and indoor) were presented.

In the same year, Hung and Krose presented their work on detecting F-formations with a graph clustering algorithm that was formulated as the identification of dominant sets [19]. In addition to the proximity between people, body orientation information was used as a cue for detection. They proposed to use socially motivated estimate of focus orientation (SMEFO), which was calculated from location information only, as the body orientation feature and experimentally showed that the addition of this feature to location increased the performance.

In 2013, Setti et. al. compared these two main approaches of detecting F-formations from images and presented their advantages and disadvantages over different scenarios [33]. They concluded that the Hough based method [8] performs better when using position and orientation together, showing good robustness to noise; whereas the dominant set based method [19] is better for scenarios when only position information is available.

In the same year, Setti et. al. published another work that advocates a multi-scale approach for F-formation discovery [34]. It is one of the first papers that takes the cardinality of the interacting groups into consideration in the detection process, as we also advocate. The proposed approach was built on the Hough voting policy of [8] and based on a competition of different voting sessions, specialized for a specific group cardinality, which are then evaluated with an information theoretic criteria to obtain final set of groups. They showed promising results on various datasets, synthetic and real life.

Setti's work in 2015 [35] presented a detailed review of current group detection algorithms for single images, including the ones mentioned here, and proposed a graph-cut based approach that outperformed others. They reported their results in five datasets with various characteristics and presented a deep analysis of methods robustness to noise. This is also one of the few studies (that we are aware of) that includes a performance analysis related to the cardinality of the target groups.

2.4 Video Based Analysis

All the computer vision studies mentioned in the former subsection lack the temporal information. Even though they took video as input, the detection of groups as performed on single frames. Vascon et. al. proposed a game-theoretic approach for detection of F-formations and presented their results on single and multiple frames, integrating temporal information with the multi-payoff evolutionary game theory [38]. They showed that the integration of multiple frames augments the overall group accuracy, especially in cases of strong noise in the positions and orientations.

Another study that uses the temporal information is Alameda-Pineda et.al.' s work[1]. They presented a multimodal approach that combined data from cameras and wearable sensors for estimating head and body pose of participants in the scene. Estimated head and body poses were then used for detection of F-formations and social attractors. Wearable sensors were used to obtain noisy estimates of speaking status and proximity input, which were then used in combination with the visual features in a matrix completion formulation for obtaining head and body poses. Their optimization included a coupling of body and head pose estimates and a temporal constraint, making it certain that detected head and body estimates are jointly estimated and temporally viable.

Depth sensors were also utilized in the literature for the estimation of spatial distance and orientation of the participants in a scene. An example study was presented by Gan et. al. in 2013, that used multiple Kinects for obtaining spatial location and orientation of each participant in the scene [12]. This information was then employed by the heat map based feature representation proposed in the paper. Qualitative evaluation of the proposed feature representation was done on a synthetic dataset only, where the authors found their temporal-encoded Interaction Space (IS) performed slightly worse than the one without temporal information. Authors then argued that this result was mainly caused by the characteristics of their ground truth.

2.5 Moving Beyond Just Group Detection

There are also works in the literature that aim further than the detection of interacting groups. Tran et. al. employed a dominant sets based approach for F-formation detection, which was followed by group activity representation and recognition [37]. For group discovery, authors presented and compared two social cues, personal distance and visual focus of attention, which are basically spatial distance and head orientation. For representing group activity, they used a bag-of-words approach that represented videos as a histogram of codewords and employed Support Vector Machine for activity classification.

Zhang and Hung presented their method to detect differing levels of social involvement, more specifically discovering associates of F-formations; people who are attached to an F-formation but do not have the status of full members [42]. They introduced novel multi-annotator annotations of the associates and compared two methods for detecting them. They also proposed a spatial-context-aware F-formation detector, that focuses on modeling people's frustum of attention. They showed that detecting and cleaning in-group associates improved the performance of F-formation detections.

2.6 Dynamics Related to Social Behavior

Up until now, all the works we mentioned focused solely on the proxemics of the interaction (aside from [7] and [14] that used speaking status and movement energy, respectively, as cues for detection). Although some studies included temporal information, they mainly focused on modeling the temporal changes in the spatial location only. Perhaps the closest study to our work was published by Hung et. al. in 2014, where the authors used a single accelerometer to classify social actions of participants in a crowded gathering [16]. These classified actions were used to extract pairwise mutual information that aims to capture interaction characteristics between dyads. Interacting partners were then detected by thresholding these values. The authors found that the mutual information computed from the pairwise speaking turns performed the best when using 40 second windows, compared to other social actions, raw acceleration and window sizes. The results were presented on a relatively limited dataset that included 10 minutes data from 26 subjects.

3 DATASET

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

Throughout the event, participants wore a custom made sensor pack around their necks, that records tri-axial acceleration at 20Hz. Top-view videos of each event were also captured, which is only used for the labeling of the ground truth, both participants' social actions and F-formations. Example screenshots are shown in Figure 1.

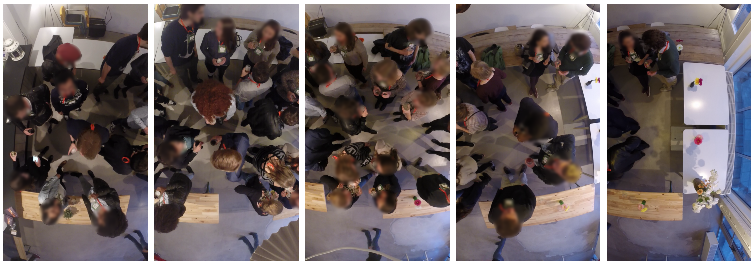

Fig. 1. Example Snapshots from the mingling phase of Day 3. Taken from [3]

3.1 Dataset Statistics

The dataset includes working sensor readings of the mingling session for 26, 22 and 22 participants, respectively for days one, two and three. For each day, a ten minute segment was manually annotated for F-formations, resulting in three different segments from each day. A variety of social actions of participants were also manually annotated, including the speaking status, hand gestures and head gestures that are used in our experiments. We should note that our proposed method is fully automatic, so the ground truth related to these social actions is only used for the training phase of social action classification (see Section 4.1.1 for further details).

Table 1 shows the number of unique interactions and mean and standard deviation of interaction lengths in seconds per group cardinality, for each day.

Table 1. Number of unique interactions, average length and standard deviation of interactions with respect to the group cardinality, for each day. They are extracted from the ground truth labels of F-formations by counting each unique occurrence of a group. All statistics related to length of interactions are in seconds.

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

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

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

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

4 METHODOLOGY

Our method aims to estimate pairwise F-formation membership for each pair in the scene. We define the problem as a binary classification task, where the final aim is to classify whether a pairwise feature representation P_{ij} indicates that person i and person j belong to the same conversing group or not. Thus, data from all participant pairs are used to obtain a joint representation which corresponds to the samples in the classification process. A flow diagram of the proposed method is shown in Figure 2. Here are the basic steps of the method:

(1) Preprocessing
   (a) Social action classification (Section 4.1.1)
   (b) Pairwise feature extraction (Section 4.1.2)
(2) Group-based meta-classifier learning using local neighborhood training (GAMUT)
   (a) Training multiple classifiers with respect to the group cardinality. (Section 4.2.1)
   (b) Prediction of a new test sample with meta-level classifier training using the local neighborhood of the test sample (Section 4.2.2)

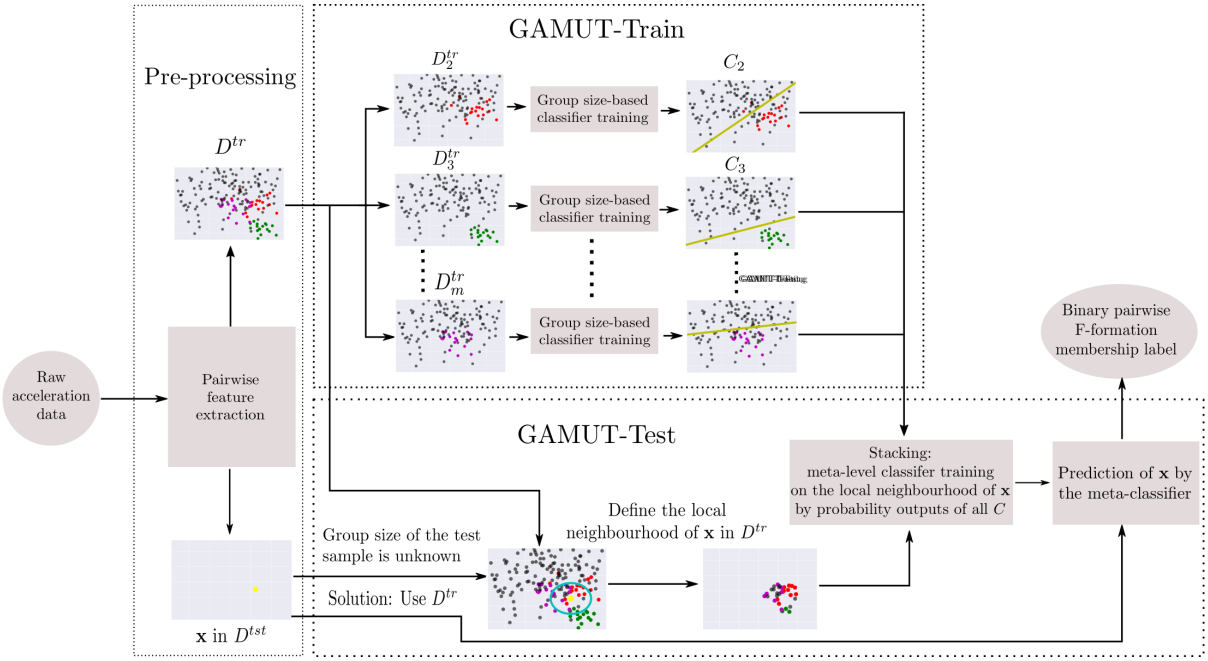
Pre-processing
GAMUT-Train
GAMUT-Test
Raw acceleration data
Pairwise feature extraction
D^{tr}
D_2^{tr}
D_3^{tr}
D_m^{tr}
Group size-based classifier training
Group size-based classifier training
Group size-based classifier training
C_2
C_3
Binary pairwise F-formation membership label
Stacking: meta-level classifer training on the local neighbourhood of **x** by probability outputs of all C
Prediction of **x** by the meta-classifier
Define the local neighbourhood of **x** in D^{tr}
Group size of the test sample is unknown
Solution: Use D^{tr}
x in D^{tst}

Fig. 2. Flow diagram of the proposed method. D^{tr} and D^{tst} correspond to entire training and test sets, respectively. D_n^{tr} is a subset of D^{tr} which is formed by the positive samples coming from groups of cardinality n and all the negative samples. C_n is the group-based classifier that is trained with D_n^{tr}.

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

4.1 Preprocessing

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

4.1.1 Social Action Classification. To provide a generalized solution, our action classification method should be person independent, where data from the test subjects are not used in the training. There are examples of person independent methods for action classification in the literature [2, 30] but they mainly focus on daily activities such as walking and running. On the other hand, manifestations of actions such as speaking and gesturing are highly person specific, making their detections harder tasks for generalization. We employed a transfer learning method, Transductive Parameter Transfer(TPT), which is experimentally shown to outperform traditional person independent approaches for person specific actions [13].

Transductive Parameter Transfer(TPT) is an adaptive transfer learning approach that aims to learn a mapping between the distribution of a dataset and the parameters of the optimal classifier for it. Sangineto et. al. proposed the method in 2014, for personalized facial expression detection from portrait images [31]. A specialized version for social action detection was then proposed by Gedik and Hung [13], which we employ in this study. TPT finds the parameters of the optimal classifier for a target dataset X^t (test set in a traditional setting) by learning a mapping between the marginal distributions of the source datasets (the training set in a traditional setup) and the parameter vectors of their optimal classifiers. A formal definition of the N source datasets (with label information) and the target dataset (without label information) can be made as $D_1^s, ..., D_N^s$, $D_i^s = \left\{x_j^s, y_j^s\right\}_{j=1}^{n_i^s}$ and $X^t = \{x_j^t\}_{j=1}^{n_t}$, respectively . Algorithm 1 presents the main steps of TPT (A more detailed explanation can be found in [13]).

ALGORITHM 1: Transductive Parameter Transfer approach (Taken from [13])

```
Input: Source sets $D_1^s, ..., D_N^s$ with labels and the target set $X^t$
Output: $\mathbf{w_t}, c_t$
Compute $\{\boldsymbol{\theta_i} = (\boldsymbol{w_i}, c_i)\}_{i=1}^N$ using L2 penalized Logistic Regression.
Create training set $\tau = \{X_i^s, \boldsymbol{\theta_i}\}_{i=1}^N$.
Compute the Earth Mover's Distance (EMD) kernel matrix $\boldsymbol{K}$ that defines
  distances between distributions where $\boldsymbol{K_{ij}} = \kappa(X_i^s, X_j^s)$.
Given $\boldsymbol{K}$ and $\tau$, compute $\hat{f}(.)$ by Kernel Ridge Regression.
Compute $(\boldsymbol{w_t}, c_t) = \hat{f}(X^t)$ using the mapping obtained by step 4.
```

Table 2. Performances of social action detection with TPT

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

We used the same feature extraction and classification setup of [13] for obtaining the social action labels for all the participants in the experiment. Statistical (mean and variance) and spectral (power spectral density with 8 logarithmically spaced bins between 0-8 Hz) features were extracted from each axis of raw and absolute values of acceleration and the magnitude of the acceleration. 3s windows with 1.5s overlap, experimentally shown to perform well in [13], were used.

Classification was done in a Leave-one-subject-out fashion. So, in each fold, we treated one participant as the target set and all others (including participants from other days) as the source sets. This procedure was replicated for each corresponding social behavior type; speaking, hand gestures and head gestures. Since the labels were imbalanced for many participants, we chose to evaluate using Area Under Curve(AUC). The performances obtained with TPT are shown in Table 2. For each participant, we obtained classified labels corresponding to a 3s window for each action, with 1.5s overlap. In terms of the 1.5s overlap of labels, we favored positive ones. Specifically, if a positive label was followed by a negative one, or vice versa, the overlapping 1.5s was considered to be positive.

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

Figure 3 is a synthetic visualization of a simple possible scene, where four people are interacting in two groups. We assume that every participant is connected to all other participants in the scene with hypothetical connections and these hypothetical connections represent samples in the classification process. The aim of the classification is to decide if these hypothetical connections connect two participants that are in the same conversing group or not. Here, green lines indicate positive samples or true connections and red lines are negative samples or false connections.

These hypothetical connections correspond to our pairwise features that provide a joint representation of the data of the participants connected by the line. We have used various features that are already employed in the literature and proposed a new one, which we named overlap statistics. Table 3 shows all the features that are used in our experiments. It also presents the dimensionality of the aforementioned feature, from which pairwise

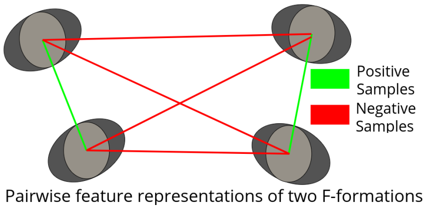

Fig. 3. Synthetic visualization of two F-formations with cardinality of two and their pairwise representations (lines). All possible pairwise representations form a sample.

Table 3. Pairwise features for joint representation. Acronyms used: mag: Magnitude, Y: Y axis, Z: Z axis, S: Binary speaking status, Ha: Binary hand gesture status, He: Binary head gesture status, (N): Shows that the normalized mutual information is also calculated in addition to non-normalized one, Numbers: Shows which participant in the pair that the stream comes from.

Feature	Dim.	Computed on	ID
Correlation	1	Acceleration[mag1-mag2,Y1-Y2,Z1-Z2]	0-2
(N)Mutual information	1	Acceleration[mag1-mag2, Y1-Y2, Z1-Z2, Y1-Z2], S1-S2 S1-Ha2, S1-He2, S2-Ha1, S2-He1	3-21
Boolean turn activity[9]	6	S1-S2, S1-Ha2, S1-He2, S2-Ha1, S2-He1	22-52
Overlap statistics	4	S1-S2, S1-Ha2, S1-He2, S2-Ha1, S2-He1	53-73
Event synchrony[29]	2	S1-S2, S1-Ha2, S1-He2, S2-Ha1, S2-He1	74-84

data streams they are extracted and their assigned IDs. The table is followed by the detailed explanation of each feature.

Correlation: Pearson correlation coefficient of two input streams. Calculated as:

$$\rho = \frac{cov(X,Y)}{\sigma X, \sigma Y} \tag{1}$$

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

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{2}$$

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

$$NI(X;Y) = \frac{I(X;Y)}{\sqrt{H(X)H(Y)}} \tag{3}$$

Both normalized and non-normalized mutual information values are used in the experiments.

Boolean turns activity(BTA): Selection of measures presented in [9], that aims to provide statistics related to turn taking in dyadic interaction. This is applied to two binary social action streams where a 1 indicates the presence of the action. For the sake of easier representation, we will call it being 'active'. Here are the measures defined for this feature set:

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

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

(1) Number of unique times that participant 2 was active while participant 1 was active.

(2) Mean length of the active intervals of participant 2 where participant 1 is active.
(3) Median length of the active intervals of participant 2 where participant 1 is active.
(4) Standard deviation of the length of the active intervals of participant 2 where participant 1 is active.

Event synchrony: A method to measure synchronicity and time delays between two univariate signals was presented in [29]. The synchronicity and time delay patterns of two signals are represented by symmetrical (Q_τ) and anti-symmetrical combinations (q_τ) of events happening in the signals. Events correspond to unique continuous active regions of the signals. The formulation for two univariate signals x and y is as follows:

$$c^\tau(x|y) = \sum_{i=1}^{m_x} \sum_{j=1}^{m_y} J_{ij}^\tau \tag{4}$$

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

$$J_{ij}^\tau = \begin{cases} 1 & \text{if } 0 < t_i^x - t_j^y \leq \tau \\ 1/2 & \text{if } t_i^x = t_j^y \\ 0 & \text{else} \end{cases} \tag{5}$$

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

$$Q_\tau = \frac{c^\tau(y|x) + c^\tau(x|y)}{\sqrt{m_x m_y}}, q_\tau = \frac{c^\tau(y|x) - c^\tau(x|y)}{\sqrt{m_x m_y}} \tag{6}$$

With this feature extraction setup, we end up with samples with a dimension of 85. As it can be seen, each feature tries to represent some type of pairwise measure between data streams of participants, may it be correlation, synchrony, lag, commonly occurring events, etc. The IDs correspond to the order of the streams given in Table 3 and their explanations. For example, IDs 53 to 57 map to overlap statistics, as in given order in the explanation, of the speaking streams.

4.2 GAMUT: Group-based Meta-classifier Learning Using Local Neighborhood Training

As mentioned earlier, the scenarios we are interested in can include a variety of groups. Differing interaction dynamics are expected to arise in groups of different cardinalities. For example, we will expect two groups in Figure 3 to have relatively similar characteristics with identifiable turn-taking patterns. However, the pairwise interactions in a group of three with participants A, B and C might differ compared to a group of two. On the other hand, participants A and B can be in a dyadic interaction where both are active in the conversation, still resembling the dynamics of the two person groups mentioned. However, in such a case, the pairwise interaction dynamics of participant C with the others is expected to be different, since C will be in a listener role. Variations in interaction dynamics increase with the increasing cardinality.

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

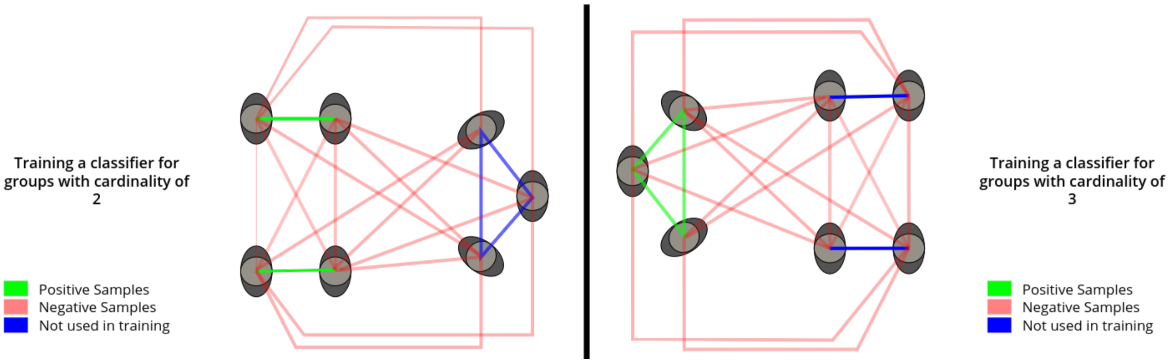

Fig. 4. Visual explanation of the training of group cardinality based classifiers, namely C_2 (left) and C_3 (right). The same convention as Figure 4 is used where all possible pairwise representations are shown with connecting lines between participants. The scene includes three conversing groups, two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

a meta-level classifier is trained only with the probability outputs of the group based classifiers based on the training samples from this local neighborhood. This meta-level classifier is then used to classify the test sample.

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

ALGORITHM 2: Training and testing phases for the GAMUT

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

4.2.1 Training Multiple Classifiers With Respect to Group Cardinality. Figure 4, which includes two conversing groups of size two and one of size three, is provided to visualize the training of group level classifiers. It uses a similar representation to Figure 3, where lines correspond to pairwise features; a sample in the classification process. As it can be seen, while training a classifier for detection of groups of size two, pairwise samples from the two person groups are treated as positives and all samples that are extracted from a pair that is not in the same group as negative. Positive samples from the three person group are not included in the training. Similarly, the right side of the image visualizes the case for the classifier of size three, where positive samples come from the three person group and positive samples from two person groups are not used.

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

$$\underset{(w,c)}{min}\frac{1}{2}w^Tw + C\sum_{i=1}^{n} log(exp(-y_i(\mathbf{x}_i^T w + c)) + 1), \tag{7}$$

where x_i is the pairwise feature vector, y_i is the binary pairwise F-formation membership labels, c and C are the bias and regularization terms, respectively. We used stochastic average gradient descent as the optimizer in our experiments[32]. The pairwise formulation causes the classes to be extremely imbalanced. Participants are not in the same conversing group with the majority of the others at the events, resulting in many negative samples. This phenomena can be easily seen from Figures 3 and 4, where the number of negative samples is much higher than the positive ones, even in these simple scenarios. To account for the imbalance, the weights are adjusted to be inversely proportional to the class frequencies.

4.2.2 Class Prediction of a New Sample by Meta-classifier Training (Stacking) Using Its Local Neighborhood. For a test sample, we first define its local neighborhood in the training set, similar to a transductive setting. We obtain probability outputs for training samples in this local neighborhood with each of the (already trained) group size based classifiers. Then, we train a meta-level classifier using these probability outputs and the labels of these samples in the local neighborhood. This process, stacking, is known to reduce the generalization error rate of multiple classifiers by reducing their biases [40]. Using only the samples from the local neighborhood makes it possible to consider samples with similar characteristics to the current test sample, for further tuning the weights learned for the meta-level classifier. This process is repeated for every test sample. In other words, different meta-level classifiers are trained for each test sample. Similar to group size based classifiers, we chose a L2 Regularized Logistic Regressor as the meta-level classifier.

5 RESULTS

5.1 Experiments Setup

We tested GAMUT on the dataset of Section 3. We randomly kept 10% of the dataset as the test set while the remaining samples were used in the training. In order to test the generalization capabilities of the proposed method, this random selection process and the following evaluation were repeated 500 times, each producing a performance score of its own. While forming the training and test sets, we made sure that there is no data from the same pair of people in both training and test sets to avoid contamination.

Our proposed approach is compared to various baselines. Firstly, we implemented the method proposed in [16], which is closest to our setting in terms of approach and modality; the state-of-the-art in our problem. This approach finds the optimal threshold value using the mean mutual information values calculated over speaking and gesturing streams of pairs in the training set. This threshold value is then used to classify the samples in the test set. Secondly, we considered an approach where group cardinalities were not considered in the training. In this approach, all the features we propose are used, but training is performed with Logistic Regression on all the dataset without distinction of group sizes.

For selecting the size of the local neighborhood (K value), we used an empirical approach. Since we have samples coming from six different group sizes, we expected $n^{tr}/6$ samples should be representative as a local neighborhood. In an optimal case where the number of samples are equally divided between cardinalities and the samples from same group sizes have similar representations in the feature space, this neighborhood will be formed by the training samples of the same corresponding group cardinality as the test sample. However, this is not always the case and there might be regions where distinguishing between the characteristics of different sized

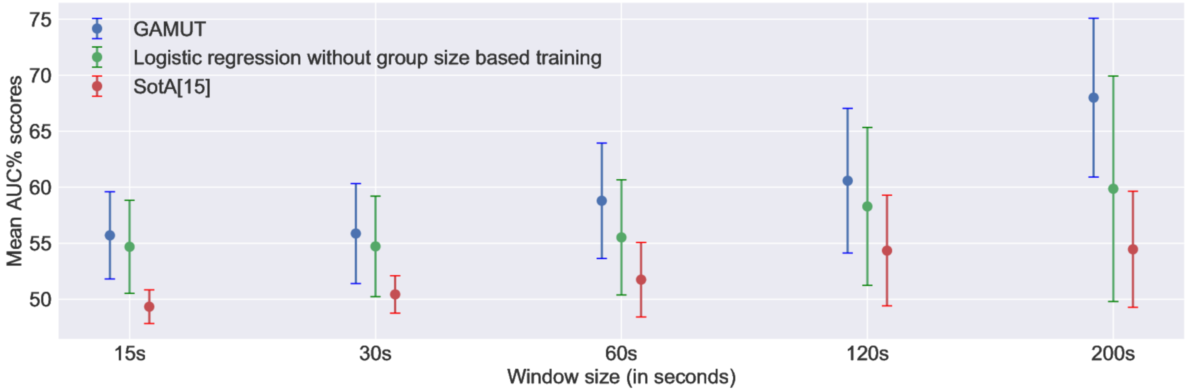

Fig. 5. Performance scores (mean AUC±STD (%)) of various approaches on pairwise F-formation membership detection. Mean and standard deviation of the AUC scores of 500 runs are visualized with the points and error bars, respectively.

groups are harder, so we experimented with various neighborhood sizes, ranging from n^{tr} (no local information, all samples are used in the stacking process) to $n^{tr}/6$.

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

5.2 Performance Scores

The method presented in [16] performed worst, even providing AUC scores lower than the random baseline (AUC of 50%). The reason for this becomes clear when the statistics of the dataset presented in [16] are investigated. The authors reported a performance value better than random on a subset of their dataset that includes nine participants with only dyadic interactions. The low performance obtained with this approach is extremely important since up until this point, existing work solely relied on pairwise mutual information of various streams for investigating speaking turns and conversing groups detection through dynamics [16, 41]. Our empirical results show that when a realistic scenario with groups of various sizes is considered, such approaches fail to provide satisfying results.

The contribution of our newly proposed features is already demonstrated by the performance of the logistic regressor without the group size based training. Even with this setup, AUC scores that are better than random and outperforming state-of-the-art are always obtained. A more detailed analysis of the effectiveness of features will be presented in Section 6.

Our proposed approach, GAMUT, performs significantly better than all other approaches regardless of the window size ($p < 0.01$, with a paired t-test computed on the performance values of 500 runs). The contribution of our method is more clearly visible for larger window sizes. Also, we generally see a pattern for all methods: Increasing window size results in better performance. This is expected, since various interaction dynamics can arise in longer intervals and it becomes easier to capture such dynamics in longer window sizes. Even though the performance was increasing, we stopped our experiments at 200 seconds, since the number of usable samples

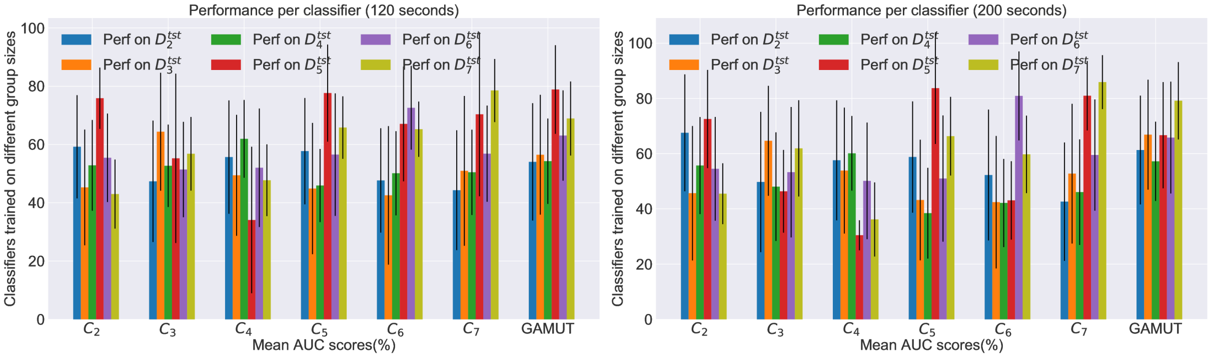

Fig. 6. Mean and standard deviation of AUC% scores of C_k and GAMUT on D_k^{tst} for all k with window sizes of 120 (left) and 200 (right) seconds. Group of bars visualizes one classifiers performance on various group cardinality based subsets.

from some group sizes would reduced drastically, making training and testing impossible. These results clearly show that in order to capture variations in real life, a method that is group size aware is definitely required.

6 FURTHER ANALYSIS

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

6.1 Performances of Group Size Based Classifiers and GAMUT on Datasets of Different Group Cardinalities

In this subsection, we provide an analysis of how group size based classifiers and GAMUT perform on datasets that include positive samples only from a specific group cardinality. This way, we empirically show that our hypothesis in Section 4.2 (that a classifier trained specifically for capturing the dynamics of a single cardinality should perform optimally for the samples of the same group size) holds.

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

Figure 6 presents the results for two window sizes, 120 and 200 seconds, for the sake of space. In the plots, each collection of bars corresponds to the mean AUC score of one classifier (C_k or GAMUT) on six different group cardinality based subsets of the data, calculated over 500 runs with leaving 10% of the data out for testing. The error bars correspond to the standard deviation of the AUC scores.

Figure 6 supports our hypothesis and shows the power of our proposed approach. For both window sizes, the best performing group size based classifier for a subset is the one with the matching cardinality. We also see that some group based classifiers performed relatively well on subsets with different cardinalities. This supports our second hypothesis that for some cases, a classifier might capture dynamics of the groups of other sizes. More importantly, nearly for each subset regardless of the window size, GAMUT guarantees to be the second best performing classifier after the matching group size based classifier. Note that in practice, GAMUT is therefore the best performing classifier as the group size from which a test sample is drawn is not known.

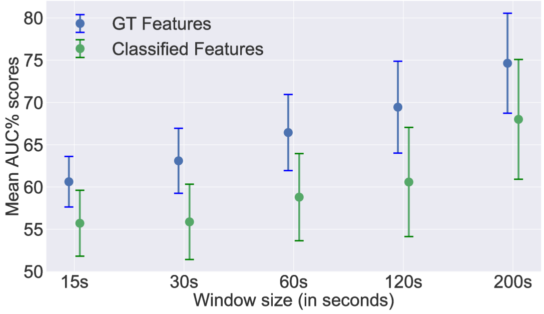

Fig. 7. Performances of GAMUT with the ground truth or classified social action labels.

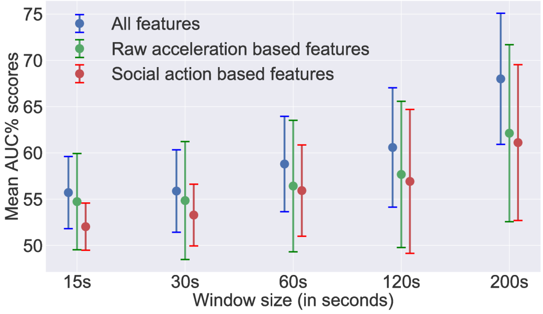

Fig. 8. Performances of GAMUT with raw acceleration or social action based features

6.2 Effects of Social Action Classification Performance on Pairwise F-formation Membership Detection

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

In order to see the effects of the performance of social action detection on the final goal, a follow up experiment was performed where we used the human annotated labels (ground truth) for speaking, hand gesturing, and head gesturing to extract pairwise features, instead of the automatically generated labels with TPT. Figure 7 shows the pairwise F-formation membership detection scores with ground truth and classified social action labels.

As expected, GAMUT with ground truth social action labels always performs better. This result shows that our features tend to perform better in capturing interaction dynamics if the social action labels are entirely correct. Fortunately, the difference in performance is marginal, showing that our social action detection approach still provides valuable information that can be used to infer pairwise F-formation membership relatively satisfactorily.

6.3 Contributions of Raw Acceleration and Social Action Based Features

Pairwise features used in GAMUT can be grouped into two categories with respect to which type of streams are used in their extraction: Raw acceleration (IDs 0-11) or social action labels (IDs 12-84). In this section, we present an ablation study where we use these feature groups separately in the training, to further understand their contribution to the final performance. Figure 8 shows the performances obtained when these features are used separately and together.

Using both sets results in higher performance compared to using either separately. This is an interesting outcome showing how additional higher level information extracted from the same source can act in a complementary manner. We also see that the raw acceleration based features tend to perform better than the social action based ones. This is especially true for small window sizes where raw acceleration based features clearly outperform the social action based ones. However, with the increasing window size, the gap between the performances of two feature sets seems to close, showing that the social action based features require more time to be informative. This is expected since many social concepts require time to unfold and it is harder to capture them in shorter time resolutions. Also, the social action labels used for extracting the features are results of a classification process (Section 4.1.1) and by no means perfect. As discussed in Section 6.2, if the ground truth labels are used for the extraction of social action based features, the overall performance increases significantly. This is another factor that can explain the gap between the performances of the two feature sets.

6.4 Correlation Analysis of Features and Pairwise F-formation Labels

To have a deeper understanding about the contributions of each feature, we conducted a correlation analysis between the vectors of single features and the ground truth for pairwise F-formation membership. In order to investigate how the correlations of features change with respect to the group cardinality, subsets of the whole dataset that had positive samples coming from a single group cardinality were used. We used the whole dataset for computing the Pearson correlation coefficients, without any distinction between training and test sets. $D_k \cup D_0$, a subset where positive samples are coming from the groups of cardinality k, will be denoted as D_k for simplicity. The correlation coefficients are calculated for all window sizes per subset, but to preserve space, only the results for two window sizes are presented in Figure 9.

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

The most striking observation from Figure 9 is how the features with highest correlation coefficients vary with respect to the group cardinalities. This supports our claim that groups with differing sizes have different interaction characteristics and might be more discriminative with different features. Another interesting aspect is how correlations of some features increase (or decrease) with the increasing window size, supporting that some dynamics of interactions are only captured in specific temporal resolutions. In the following paragraphs, we will analyze informative features (according to the correlation values) per group size in detail.

- $\mathbf{D_2}$: Over the range of all window sizes, the correlation of the Z-axis of the accelerometer readings (ID 2) seems to have the highest correlation coefficients. Z-axis captures the forward-backward acceleration of the body. The high correlation value is not surprising, since in two person groups, people are expected to move synchronously. Couple of features that are easily noticeable in 200 second windows are the synchrony ratio of speaking and hand gestures (ID 32) and the median length and standard deviation of the hand gestures co-occurring between the participants (IDs 68 and 69). With the increasing window size, correlation coefficients of the features related to the co-occurrence of speaking and hand gesturing increase, pointing to a more involved interaction.
- $\mathbf{D_3}$: It can be directly seen that there are four features with high coefficients that are consistent over different windows, IDs 46, 47, 49 and 71. The first three correspond to the Boolean Turn Activity features between streams of speaking and hand and head gesturing, more specifically the synchrony ratio and co-occurrence of these actions. Feature 71 is the mean length of head gestures occurring while the other person is speaking. Features with high correlations seem to be more representative of the listening behavior, such as head nods occurring while the other participant is speaking. Features based on mutual information of the raw acceleration improve in correlation with the increasing window size, such as IDs 4, 5, 6, 9 and 10. This might be connected to the increasing variance in interaction characteristics, pointing to mimicry and synchrony emerging between the pairs within longer intervals.
- $\mathbf{D_4}$: Features with relatively high correlations are sparse for D_4. The correlations of features 5 and 9, non-normalized and normalized mutual information between the Z-axes of the acceleration, seem to be comparatively higher than the rest for the window size of 60 seconds. Correlations for features 31 (co-occurrence of not speaking and not hand gesturing) and 63 (standard deviation of the length of the head gestures during speaking) marginally improve with the window size of 200 seconds. This collection of comparatively highly correlated features covers concepts both from D_2 (measures related to the Z axis of raw acceleration) and D_3 (measures between social actions of speaking and gesturing), and represents the dyadic interactions where both participants are active in conversation and the listener behavior in a three person group, that might both occur in a group of four.

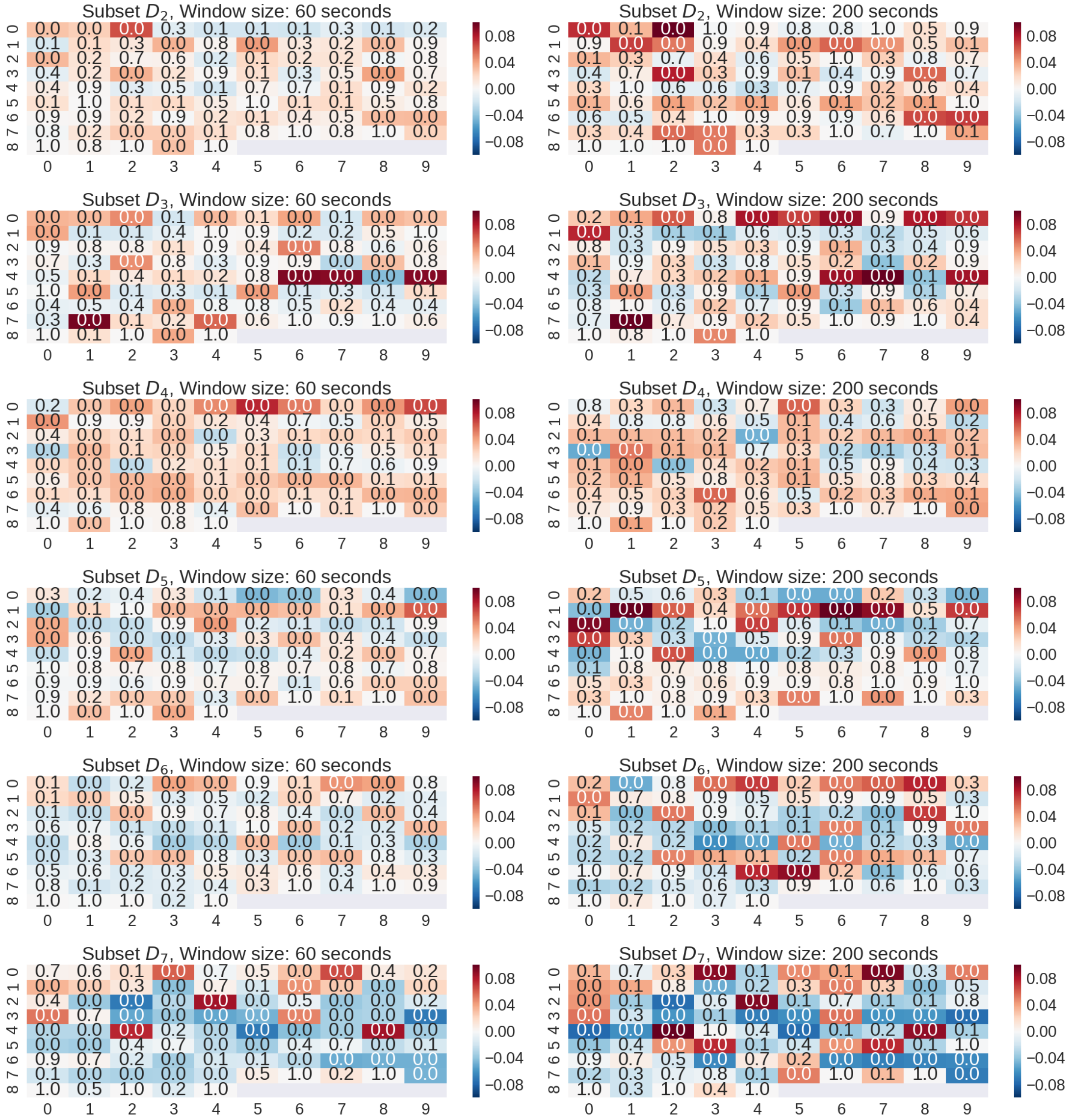

Fig. 9. Pearson correlation coefficients (r) and significance (p) values calculated between feature vectors and the F-formation membership ground truth on group cardinality based subsets (D_k) of the data. Correlation values (r) are presented as the color of the cells where as the values inside the cells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

- **$\mathbf{D_5}$** : Similar to D_4, there are not many features with high correlations for the window size of 60 seconds. However, correlation coefficients of various features improve with the increasing window size. The most notable features are 11, 16, 17 and 20, which are all (normalized) mutual information measures between two speaking status streams, speaking-hand gesturing streams and speaking-head gesturing streams. As the group size increases, we see that many social action pairs are more strongly correlated.
- **$\mathbf{D_6}$** : Comparatively high correlations are only present in the window size of 200 seconds, most notably features with IDs of 4, 8, 64 and 65. The first two are (normalized) mutual information values between the Y-axis, the right to left acceleration of the participant. The second two are the overlap statistics calculated from the streams of speaking and gesturing axes. This is the only group cardinality where the movement in the Y-axis has a higher correlation value than any other acceleration based measure. This can be related to the more frequent occurrences of listener-listener pairs in such large groups, where synchronized posture shifts are captured through side to side movements of the body.
- **$\mathbf{D_7}$** : Unlike other large sized groups, we a see a few features with comparatively higher correlation coefficients at the window size of 60 seconds, such as 3, 7, 24, 42 and 48. This collection of features covers different concepts, such as the mutual information between raw acceleration streams and boolean turn activity features between two speaking statuses, speaking status and head and hand gestures. These features also either retain or increase their correlation values with increasing window size. There are also features with relatively higher negative correlation coefficients in both window sizes. Two examples are features 39 and 45, synchrony ratios of speaking with hand and head gestures. This result suggests that in larger groups, there might be multiple parallel interactions happening at the same time, resulting in pairs with non-synchronized social actions.

In summary, our proposed feature sets, perhaps apart from event synchrony, intrinsically carry information about the different aspects of the problem. In particular, for dyadic interactions measures between the raw acceleration readings are seen as desirable cues. Compatible with the observations of Section 6.3, even with the second level of information encoded in the social actions, raw acceleration still holds much information. Especially for groups with high cardinalities, measures that include gestures, especially head gestures, seem to gain importance, probably representing active listening behavior.

6.5 Comparison of Ensemble Learning Methods

GAMUT combines the prediction probabilities of the group based classifiers by training a meta-classifier (stacking). There are other options for combining predictions of multiple classifiers in the literature [36]. In this section, we compare the performance of GAMUT to two other ensemble learning techniques; maximum and mean fusion. In these methods, final prediction for a test sample is obtained by computing the maximum or mean of the probabilities provided by the multiple classifiers. Figure 10 shows the performances of maximum and mean fusion in addition to GAMUT.

GAMUT outperforms both ensemble learning methods regardless of the window size. Since the performance differences between the methods are relatively low, we applied a paired t-test (computed on the results of 500 repetitions for each) to see if these differences are statistically significant. Apart from the mean fusion for 120s, all results are found to be statistically significant with $p < 0.01$. In other words, GAMUT guarantees better performance for almost all cases in comparison to other ensemble learning methods. When compared to Figure 5, we can see that even the mean and maximum fusion methods outperform logistic regression without group size based training and the method of [15]. Thus, we can conclude that, regardless of the combination technique, employing group based classifiers will always provide better performance than methods that ignore this.

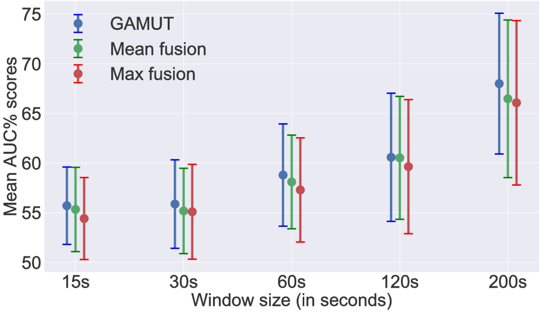

Fig. 10. Performances of GAMUT, mean and max fusion methods

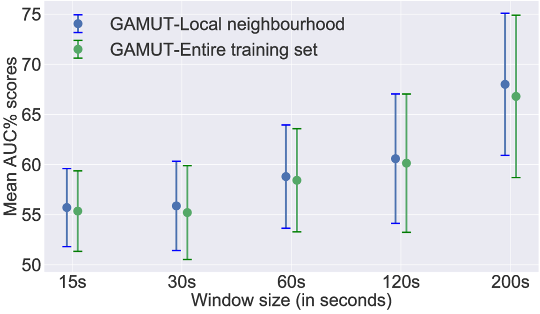

Fig. 11. Performances of GAMUT with stacking on local neighborhood and the entire training data

6.6 Effects of Using the Local Neighborhood in Meta-classifier Training

GAMUT uses data only from the local neighborhood of a test sample while training the meta-classifier, expecting the local neighborhood to be more informative than the entire training set. This subsection investigates the effects of this setup by comparing the performances of GAMUT where the meta-classifier training is either performed on the local neighborhood or the entire training set. Results are shown in Figure 11.

GAMUT with the local neighborhood meta-classifier training always outperforms the cases where the entire training set is used. Paired t-tests for each window size showed that the differences are significant with $p < 0.01$. Still, the differences are marginal in terms of percentage. We believe this is caused by the dataset statistics and the random selection process used while forming the training and testing splits. Investigation of the distributions of group sizes in different runs have shown that in some runs, the number of samples in the training set coming from some group sizes are too low to be representative. In such cases, the probability of having representative samples in the local neighborhood reduces and using the entire training set might prove to be superior. However, with more data, it will be possible to have a training set that includes enough samples from all group sizes. Then, the use of the local neighborhood should be more optimal, providing a bigger increase in the performance.

7 CONCLUSION AND FUTURE WORK

7.1 Conclusion

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

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

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

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

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

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

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

7.2 Future Work

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

The main aim of the method was to provide pairwise F-formation membership. This information can be used as a starting point for creating a connectivity graph, that includes all the participants in the scene. While doing so, incorrect estimations of our method can be refined by introducing constraints related to the group membership and temporal consistency. A possible option is to use the posterior probability estimates that our method provides as edge strengths in the connectivity graph. There are already successful methods in the literature mainly used for optimizing connectivity graphs, that can be modified to be suitable for our problem formulation [19].

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

REFERENCES

[1] Xavier Alameda-Pineda, Yan Yan, Elisa Ricci, Oswald Lanz, and Nicu Sebe. 2015. Analyzing free-standing conversational groups: a multimodal approach. In *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 5–14.

[2] Ling Bao and SS Intille. 2004. Activity recognition from user-annotated acceleration data. *Pervasive Computing* (2004), 1–17. http://link.springer.com/chapter/10.1007/978-3-540-24646-6_1

[3] Laura Cabrera-Quiros, Andrew Demetriou, Ekin Gedik, Leander van der Meij, and Hayley Hung. 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* (2018).

[4] Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani. 2010. Dynamics of person-to-person interactions from distributed RFID sensor networks. *PloS one* 5, 7 (2010), e11596.

[5] Daniel Chaffin, Ralph Heidl, John R Hollenbeck, Michael Howe, Andrew Yu, Clay Voorhees, and Roger Calantone. 2017. The promise and perils of wearable sensors in organizational research. *Organizational Research Methods* 20, 1 (2017), 3–31.

[6] Tanya L Chartrand and John A Bargh. 1999. The chameleon effect: the perception–behavior link and social interaction. *Journal of personality and social psychology* 76, 6 (1999), 893.

[7] Tanzeem Choudhury and Alex Pentland. 2003. Sensing and Modeling Human Networks using the Sociometer. In *Proceedings of the Seventh IEEE International Symposium on Wearable Computers (ISWCâĂŹ03)*, Vol. 1530. 17–00.

[8] Marco Cristani, Loris Bazzani, Giulia Paggetti, Andrea Fossati, Diego Tosato, Alessio Del Bue, Gloria Menegaz, and Vittorio Murino. 2011. Social interaction discovery by statistical analysis of F-formations.. In *BMVC*, Vol. 2. 4.

[9] Emilie Delaherche, Mohamed Chetouani, Fabienne Bigouret, Jean Xavier, Monique Plaza, and David Cohen. 2013. Assessment of the communicative and coordination skills of children with autism spectrum disorders and typically developing children using social signal processing. *Research in Autism Spectrum Disorders* 7, 6 (2013), 741–756.

[10] Robin IM Dunbar, NDC Duncan, and Daniel Nettle. 1995. Size and structure of freely forming conversational groups. *Human nature* 6, 1 (1995), 67–78.

[11] Nathan Eagle and Alex Sandy Pentland. 2006. Reality mining: sensing complex social systems. *Personal and ubiquitous computing* 10, 4 (2006), 255–268.

[12] Tian Gan, Yongkang Wong, Daqing Zhang, and Mohan S Kankanhalli. 2013. Temporal encoded F-formation system for social interaction detection. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 937–946.

[13] Ekin Gedik and Hayley Hung. 2017. Personalised models for speech detection from body movements using transductive parameter transfer. *Personal and Ubiquitous Computing* (2017), 1–15.

[14] Jonathan Gips and Alex Pentland. 2006. Mapping human networks. In *Pervasive Computing and Communications, 2006. PerCom 2006. Fourth Annual IEEE International Conference on*. IEEE, 10–pp.

[15] Jonathan Peter Gips. 2006. *Social motion: Mobile networking through sensing human behavior.* Ph.D. Dissertation. Massachusetts Institute of Technology.

[16] Hayley Hung, Gwenn Englebienne, and Laura Cabrera Quiros. 2014. Detecting conversing groups with a single worn accelerometer. In *Proceedings of the 16th international conference on multimodal interaction*. ACM, 84–91.

[17] Hayley Hung, Gwenn Englebienne, and Jeroen Kools. 2013. Classifying social actions with a single accelerometer. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. ACM, 207–210.

[18] Hayley Hung and Daniel Gatica-Perez. 2010. Estimating cohesion in small groups using audio-visual nonverbal behavior. *IEEE Transactions on Multimedia* 12, 6 (2010), 563–575.

[19] Hayley Hung and Ben Kröse. 2011. Detecting f-formations as dominant sets. In *Proceedings of the 13th international conference on multimodal interfaces*. ACM, 231–238.

[20] DB Jayagopi, H Hung, C Yeo, and D Gatica-Perez. 2009. Modeling Dominance in Group Conversations from Nonverbal Activity Cues In Trans. on Audio, Speech, and Language Processing, Special Issue on Multimodal Processing for Speech-based Interactions.

[21] Dinesh Babu Jayagopi and Daniel Gatica-Perez. 2010. Mining group nonverbal conversational patterns using probabilistic topic models. *IEEE Transactions on Multimedia* 12, 8 (2010), 790–802.

[22] Adam Kendon. 1970. Movement coordination in social interaction: Some examples described. *Acta psychologica* 32 (1970), 101–125.

[23] Adam Kendon. 1990. *Conducting interaction: Patterns of behavior in focused encounters*. Vol. 7. CUP Archive.

[24] Taemie Kim, Erin McFee, Daniel Olguin Olguin, Ben Waber, Alex Pentland, et al. 2012. Sociometric badges: Using sensor technology to capture new forms of collaboration. *Journal of Organizational Behavior* 33, 3 (2012), 412–427.

[25] Anmol Madan, Manuel Cebrian, David Lazer, and Alex Pentland. 2010. Social sensing for epidemiological behavior change. *Proceedings of the 12th ...* (2010). http://dl.acm.org/citation.cfm?id=1864394

[26] Claudio Martella, Ekin Gedik, Laura Cabrera-Quiros, Gwenn Englebienne, and Hayley Hung. 2015. How was it?: exploiting smartphone sensing to measure implicit audience responses to live performances. In *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 201–210.

[27] Aleksandar Matic, Venet Osmani, and Oscar Mayora-Ibarra. 2012. Analysis of social interactions through mobile phones. *Mobile Networks and Applications* 17, 6 (2012), 808–819.

[28] Daniel Olguín Olguín, Benjamin N Waber, Taemie Kim, Akshay Mohan, Koji Ara, and Alex Pentland. 2009. Sensible organizations: Technology and methodology for automatically measuring organizational behavior. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39, 1 (2009), 43–55.

[29] R Quian Quiroga, Thomas Kreuz, and Peter Grassberger. 2002. Event synchronization: a simple and fast method to measure synchronicity and time delay patterns. *Physical review E* 66, 4 (2002), 041904.

[30] Nishkam Ravi, Nikhil Dandekar, Preetham Mysore, and ML Littman. 2005. Activity recognition from accelerometer data. *AAAI* (2005), 1541–1546. http://www.aaai.org/Papers/IAAI/2005/IAAI05-013

[31] Enver Sangineto, Gloria Zen, Elisa Ricci, and Nicu Sebe. 2014. We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer. In *Proceedings of the ACM international conference on multimedia*. ACM, 357–366.

[32] Mark Schmidt, Nicolas Le Roux, and Francis Bach. 2013. Minimizing finite sums with the stochastic average gradient. *arXiv preprint arXiv:1309.2388* (2013).

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

[34] Francesco Setti, Oswald Lanz, Roberta Ferrario, Vittorio Murino, and Marco Cristani. 2013. Multi-scale F-formation discovery for group detection. In *Image Processing (ICIP), 2013 20th IEEE International Conference on*. IEEE, 3547–3551.

[35] Francesco Setti, Chris Russell, Chiara Bassetti, and Marco Cristani. 2015. F-formation detection: Individuating free-standing conversational groups in images. *PloS one* 10, 5 (2015), e0123783.

[36] David MJ Tax, Martijn Van Breukelen, Robert PW Duin, and Josef Kittler. 2000. Combining multiple classifiers by averaging or by multiplying? *Pattern recognition* 33, 9 (2000), 1475–1485.

[37] Khai N Tran, Apurva Gala, Ioannis A Kakadiaris, and Shishir K Shah. 2014. Activity analysis in crowded environments using social cues for group discovery and human interaction modeling. *Pattern Recognition Letters* 44 (2014), 49–57.

[38] Sebastiano Vascon, Eyasu Z Mequanint, Marco Cristani, Hayley Hung, Marcello Pelillo, and Vittorio Murino. 2016. Detecting conversational groups in images and sequences: A robust game-theoretic approach. *Computer Vision and Image Understanding* 143 (2016), 11–24.

[39] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 3–14.

[40] David H Wolpert. 1992. Stacked generalization. *Neural networks* 5, 2 (1992), 241–259.

[41] Danny Wyatt, Tanzeem Choudhury, and Jeff Bilmes. 2007. Conversation detection and speaker segmentation in privacy-sensitive situated speech data. In *Eighth Annual Conference of the International Speech Communication Association*.

[42] Lu Zhang and Hayley Hung. 2016. Beyond F-formations: Determining Social Involvement in Free Standing Conversing Groups from Static Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1086–1095.

Received February 2018; revised May 2018; accepted October 2018