TUDelft

Delft University of Technology

Detecting F-formations Roles in Crowded Social Scenes with Wearables Combining Proxemics Dynamics using LSTMs

Rosatelli, Alessio; Gedik, Ekin; Hung, Hayley

Citation (APA)
Rosatelli, A., Gedik, E., & Hung, H. (2019). Detecting F-formations Roles in Crowded Social Scenes with Wearables: Combining Proxemics Dynamics using LSTMs. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)* (pp. 147-153). Article 8925179 IEEE. https://doi.org/10.1109/ACIIW.2019.8925179

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

Detecting F-formations & Roles in Crowded Social Scenes with Wearables: Combining Proxemics & Dynamics using LSTMs

Alessio Rosatelli
Ingegneria Informatica e Robotica
University of Perugia
Perugia, Italy
alessio.rosatelli1994@gmail.com

Ekin Gedik
Socially Perceptive Computing Lab
Delft University of Technology
Delft, Netherlands
e.gedik@tudelft.nl

Hayley Hung
Socially Perceptive Computing Lab
Delft University of Technology
Delft, Netherlands
h.hung@tudelft.nl

Abstract—In this paper, we investigate the use of proxemics and dynamics for automatically identifying conversing groups, or so-called F-formations. More formally we aim to automatically identify whether wearable sensor data coming from 2 people is indicative of F-formation membership. We also explore the problem of jointly detecting membership and more descriptive information about the pair relating to the role they take in the conversation (i.e. speaker or listener). We jointly model the concepts of proxemics and dynamics using binary proximity and acceleration obtained through a single wearable sensor per person. We test our approaches on the publicly available MatchNMingle dataset which was collected during real-life mingling events. We find out that fusion of these two modalities performs significantly better than them independently, providing an AUC of 0.975 when data from 30-second windows are used. Furthermore, our investigation into roles detection shows that each role pair requires a different time resolution for accurate detection.

Index Terms—F-formation detection, wearable sensing, conversing groups, recurrent neural networks, role identification

I. INTRODUCTION

As social animals, most people interact and specifically converse with each other on a daily basis. Regardless of the specific type of the event that the interaction takes place, the automatic detection of conversational interaction is the first step to understanding the potential flow of interpersonal influence. For example, being able to accurately identify conversing groups allows us to analyse more complex social phenomena such as attraction [1] or cohesion [2]. The analysis of social behaviour is also shown to be crucial for security and surveillance applications [3], [4]. This information is beneficial for human robot interaction by allowing the robot to identify an ongoing conversation to enter [5]. In this paper, we investigate and provide a novel solution for the detection of conversing groups and participants roles in the interaction, relying solely on the fusion of multimodal sensor data, namely acceleration and proximity and test our solution on a real life crowded mingling event.

We formalise the notion of conversing groups by referring to Kendon's definition F-formations [6]. Kendon defines F-formations as a specific type of focused encounter where participants tend to stay in close proximity and orient themselves to sustain conversation [6]. From the definition, it is clear that proxemics, the spatial distance and bodily orientation of participants, represents valuable information for the identification of F-formations. Indeed, most related work from the wearable and ubiquitous computing community heavily relies on proxemics, extracting proximity and orientation information either from images [7]–[9] and/or wearable sensing such as Infrared trans-receivers [10], [11]. However, there are limitations to such approaches. The estimation of spatial proximity with IR or radio-based sensors are shown to result in various false detections, mainly caused by the surrounding environment and furniture [12]. In crowded scenarios, possible occlusions between people and objects can negatively affect the robust estimation of proximity and orientation. In this paper, we propose an alternate approach fusing the proxemics and dynamics of interaction, based on the findings that people tend to coordinate their movements during conversational interaction [13]. This builds upon the work of Gedik and Hung [14] where they proposed an approach that investigate dynamics alone. Here we argue for the importance of the use of both for better accuracy.

We investigate the fusion of proxemics and dynamics with a publicly available dataset of real-life crowded mingling events [15]. Such events are perfect candidates for analysing group behaviour since groups of different sizes form and break over time naturally, creating rich and varied behaviour to analyse. During the mingling events used in our experiments, participants were fitted with a single sensor pack that was worn over their necks with a lanyard. These sensor packs recorded their bodily acceleration and the IDs of neighbouring sensor packs, acting as binary proximity detections. Our proposed solution uses these information sources as inputs to an Long Short Term Memory (LSTM) network which models the temporal dependencies of the data and learns a joint representation of proxemics and dynamics. Subsequently, we delve deeper into

the nature of the problem by investigating the detection of the roles of the participants in the interaction in terms of whether they are speaking or listening. Based on findings in social science, speakers and listeners behave differently in interaction [13]. We expect these roles to be identifiable and that they could eventually help in distinguishing conversing partners who are more or less involved with each other.

The novel contributions of this paper are as follows: (i) we propose to fuse proxemics and dynamics through LSTM networks and show, with a specific time resolution, it significantly outperforms using separate modalities, (ii) we investigate the automatic detection of roles inside the interaction and find that:

- Group membership and roles in interaction can be jointly detected.
- Each role pair requires a different temporal resolution of analysis for satisfactory detection.

II. RELATED WORK

To better contextualise the nature of this paper, we present a number of studies that focus on the automatic identification of F-formations. We use a rough categorisation based on the sensor types used in these studies. Most of the related literature in this domain focuses on the use of proxemics, the positions and orientations of the participants in a scene, as a cue. Classically favoured modalities for F-formation detection has been static images and videos. Throughout the years, many different methodologies such as Hough voting [16], [17], graph clustering with dominant sets [18], [19] and multi-payoff evolutionary game theory [9] were used to detect F-formations from static images or video.

Another widely used sources of information are wearable and mobile sensing. Such devices provide proximity information through Infrared (IR) and/or radio which is then used to detect interactions between people. In general, the sensor data is taken at face value and used directly as a proxy for conversational interaction which can be used to infer F-formations. However, most works mentioned here do not evaluate the robustness of such sensing systems since most studies aggregate observed face to face detections over sufficiently long period of time with the assumption of relatively low crowd density. Such studies have tended to use these interaction proxies for analysing long term social concepts such as centrality [20], personality traits [21], social patterns in daily life [22], student mental health [23], interest and affiliation [11] and dynamics of interaction networks [24]. These studies might use custom-made sensor packs [11], [20], [24] or employ mobile phones [22], [23]. The temporal resolution of the analysis can range from minutes [11], [24] to days [20] to months [22], [23]. Some studies explicitly infer body orientation in addition to proximity [25], whereas some others infer the proximity from GPS coordinates [26]. In [10], information from wearable sensors is employed together with information from images for F-formation detection.

There exist few studies that consider the dynamics of interaction as cues for F-formation detection. Authors proposed two methods for modelling the dynamics of proxemics from

video data in [5]. In [20], audio is used to infer speaking status which was then used for refining the IR results on interaction detection. Similarly, [27] used acceleration to infer various social actions such as speaking, drinking, etc. Mutual information between different participants' social action streams were then computed and thresholded to obtain F-formation memberships. [14] built on these results and proposed new measures between social action streams and group size based training. To our knowledge, no other studies on F-formation detection focused on fusing the proxemics and dynamics of interaction from wearables.

Determining similar studies for role identification, in the context of this paper, is not trivial. Identification of speakers mainly falls under speaker detection [28] or recognition [29]. They aim to find out who the speaker is from an input stream. These studies mostly employ sensors that registers the flow of speech, such as audio [28], [29] or video that focuses on the movement of the mouth [30]. Only a few studies employ accelerometers for detecting speakers, similar to our proposed approach [31]. However, none considers the roles in the context of an interaction and focuses on the joint estimation as our paper does. The benefit of modelling both jointly is that the correlation between the different behaviours should help us to better estimate conversational involvement and therefore social influence from the perspective of speaker and listener behaviour.

III. DATA

In our experiments, we used the data from the MatchN-Mingle dataset which includes data from three separate speed dating events took place in a local bar and joined by 92 participants in total [15]. In each event, speed dating sessions were followed by a free mingling session of roughly one hour. In the mingling session, participants freely interacted with each other. Each participant wore a single custom badge-like sensor pack on their necks with a lanyard. These sensor packs contain an accelerometer and a radio-based binary proximity sensor, recording at 20Hz and 1Hz, respectively. The recordings of these sensor packs during the mingling sessions are used in our experiments. The video recordings of the scene were only used for obtaining F-formation and speaking status annotations which has been used as ground truth in the roles detection part of this study. 10 minute intervals from the mingling session of each day, totalling in 30 minutes, were annotated for F-formations. Sensor readings and annotations from these 30 minutes form the data used in this paper. During these 30 minutes, groups of various sizes ranging from dyadic interactions to seven people were formed, creating a rich testing environment. For more details regarding the dataset, please refer to [15].

IV. METHODOLOGY

Figure 1 shows an overview of the proposed method and conducted experiments. In this section, the pairwise representation procedure which creates the samples used in our experiments will be first explained. Then, the methodologies and
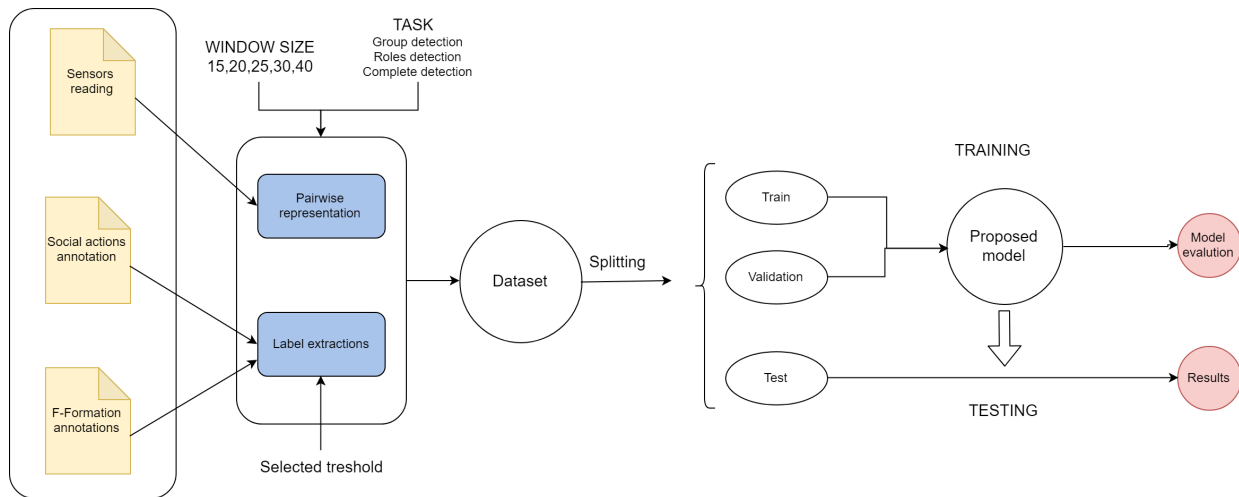
Fig. 1. Flow diagram of the proposed method

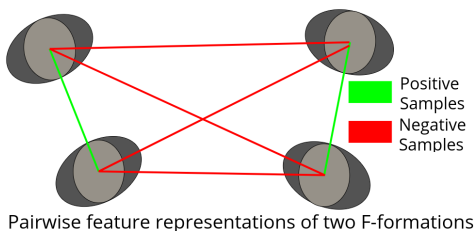Pairwise feature representations of two F-formations

Fig. 2. Synthetic visualisation of pairwise representations (taken from [14])

experiment setups for group membership and roles detection will be presented.

A. Pairwise representation

According to the literature, a group could be represented as a collection of individuals [8] or as a set of dyadic pairs [14], [18], [32]. In this paper, we selected pairwise representations since we utilise, in addition to proxemics, the dynamics of interaction realised through the coordination of the movements of participants. Figure 2 shows a synthetic visualisation of a possible scene with four people and two F-formations. Pairwise representations of such a scene treat each possible connection between people as a sample. These samples are visualised in Figure 2 as green and red lines, representing samples belonging to the positive (pair is in the same F-formation) and the negative (pair is not in the same F-formation) classes, respectively.

Since we are interested in the dynamics of interaction which unfolds over time, the temporal resolution of the data gains importance [9], [19]. An instantaneous representation of the scene is not sufficient for exploiting the coordination patterns between interacting partners. Hence, we used a sliding window approach, with 50% overlap, over the sensor data of participants. Samples in our experiments are formed by continuous sequences of sensor data obtained from two participants. Following sections will provide more information about the dimensions of these input samples with respect to experiment setups and chosen sliding window sizes.

B. Group membership detection

Since our samples are represented as time-series data, we used recurrent neural networks. The proposed architecture is visualised in Figure 3. Each sample is fed into a three-layer LSTM [33] with 16 hidden units per cells. The cell number varies according to the selected window size and equals to the input sequence length. The number of layers and hidden units were empirically found. The outputs of the last LSTM layer cells are discarded except the last one which should ideally contain the information extracted from the previous elements of the sequences. This representation vector is then fed into a feed-forward network which is composed of two linear layers with Relu activation functions. This network reduces the vector dimension from 16 to 8 and then maps it into an output vector with dimensions identical to the number of classes for the current experiment.

The model is trained end to end with a weighted cross entropy loss on the output vector, defined by the equation below:

$$\text{loss(x, class)} = \text{weight[class]} \left(-x[class] + \log \left(\sum_j \exp(x[j]) \right) \right)$$

The pairwise representation creates high class imbalance in the data, favouring the negative samples. Hence, we used a weighted loss where the weights were set to be inversely proportional to the label distribution. The network was optimised with Adam. [34].

As shown in Figure 2, there are two classes for group detection experiments; pairs that are in the same F-formation and pairs that are not. A sample is treated as positive if the two participants forming the sample are annotated as being in the same group in the ground truth at least 66% of the sliding window that the sample is extracted from.

We experiment with three different input combinations: acceleration only, proximity only and fusion of the two.
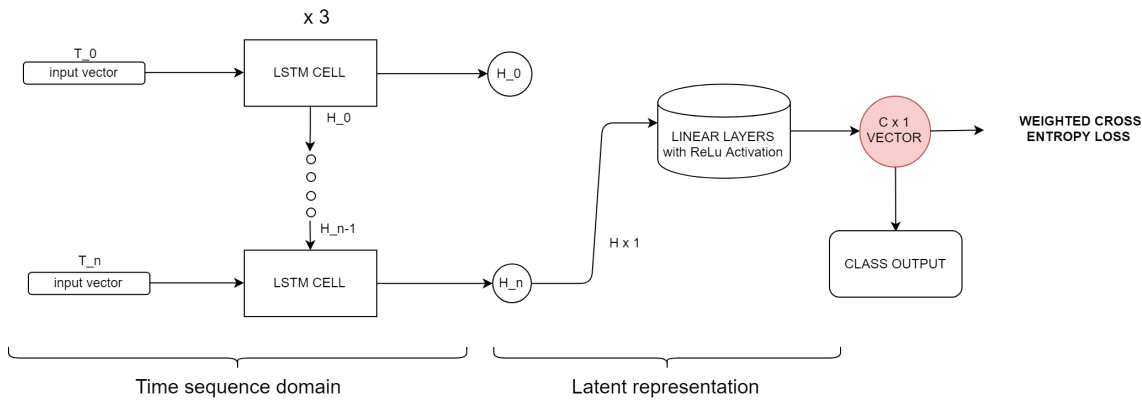
Fig. 3. Proposed model structure and information propagation through the network.

Input ID	Definition
0 : 2	X,Y,Z accelerometer readings for P1
3 : 5	X,Y,Z accelerometer readings for P2
6	Binary proximity sequence between P1 and P2

Figure I shows the inputs used in these combinations. For the first setup, tri-axial accelerations from two participants are concatenated to form the samples (IDs 0 to 5). The second setup uses the binary proximity stream only (ID 6). Samples for the fusion setup is formed by the concatenation of the acceleration and the proximity streams (IDs 0 to 6). In order to have the same sampling rate, binary proximity readings are up-sampled to 20Hz. As an example, with a window size of 15 seconds and an input combination of fusion, each sample will be 300x7. The final size of the resulting dataset varies according to the window size, both in the number of samples, which is conversely proportional to the window size, and the sequence length of each sample which is directly proportional to it.

C. Joint prediction of membership and roles

We define two basic roles in an interaction: speaker and listener. Since we use a pairwise representation, this results in three different classes for each sample from pairs that were already in an interaction: speaker&speaker, speaker&listener and listener&listener. We use the speaking status ground truth present in the MatchNMingle [15] to label someone as a speaker or a listener. If a person is annotated as speaking at least 30% of the window that the sample is extracted from, this person is labelled as a speaker and listener otherwise. This threshold was selected to allow a continuous turn-taking behaviour to properly represent speaker&speaker pairs.

We formulate the experiments on role detection as four-class classification problems that focus on jointly detecting the existence of an interaction and the roles participants take in it. These joint prediction experiments include all samples used in the group detection experiments but the positive samples are now relabelled with respect to the role pairs. This formulation results in the following classes: pairs that are not in the same F-formation, speaker&speaker pairs, speaker&listener pairs and listener&listener pairs.

V. RESULTS AND DISCUSSION

Following subsections present the results of the group and roles detection experiments. For both experiment types, we used 80-10-10 training-validation-test split. While splitting, we made sure there are no samples belonging to the same pair of people in the train, validation and test sets to avoid contamination.

To test the generalisation ability of our methods, we repeated the random splitting step 20 times, resulting in 20 different networks and 20 performances for each. Results for group membership detection are then presented as the means of these 20 repetitions. For the joint role and membership detection experiments, confusion matrices for each run are accumulated and normalised at the end, for a more detailed evaluation in the multiclass case. Each model is trained for 50 epochs on the training set and the model with the lowest loss on the validation set is used for evaluation on the test set.

A. Group membership detection

Due to highly imbalanced data, we selected Area Under Receiver Operating Characteristic Curve (AUC for short) as the evaluation metric. Figure 4 visualises the mean AUC scores obtained for group detection with different input combinations explained in the former section, with window sizes ranging from 15 to 40 seconds. Firstly, acceleration only results seem to be poor compared to the others, ranging from 0.52 to 0.57 (standard deviations are close to 0.06). Still, a one-sided paired t-test showed that the results are significantly better than a random classifier, regardless of the window size ($p < 0.05$). We could also see that increasing the window size causes a marginal increase in the performance. This is expected since the dynamics of interaction, which we aim to capture through acceleration, are expected to unfold in larger time resolutions.

We can see that proximity provides satisfactory performance by itself, providing AUC scores ranging from 0.87 to 0.94.
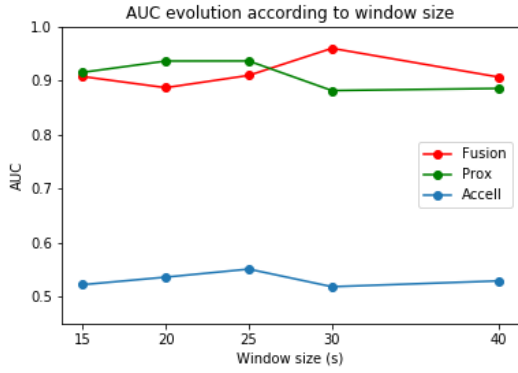
Fig. 4. Mean AUC scores for group membership detection with different window lengths and inputs

Fig. 5. Joint prediction of group membership and roles confusion matrix for 10-second windows (accumulated over 20 repetitions and normalised)

Contrary to the pattern we saw for the acceleration only performances, performances with the proximity information tend to drop with the increasing window size. This is might suggest that proximity itself is not enough for detecting reasonably longer interactions.

Performances obtained with the fusion of acceleration and proximity supports our claims regarding the nature of interactions including both proxemics and dynamics. We see that the performances obtained with fusion tend to be slightly lower than the proximity ones for window sizes of 20 and 25. However, a paired one-tailed t-test showed no significant difference. In other words, fusion always guarantees performances at least as good as proximity. More interestingly, we can see that when 30-second windows are used, fusion provides a noticeable increase over the proximity with an AUC score of 0.975. This result is shown to be significantly better than the proximity one ($p < 0.01$). When analysed together with the former results, we can conclude that dynamics captured through body acceleration indeed provides more information regarding the existence of an interaction between partners. However, it should be noted that this effect heavily depends on the size of the window, since the coordination of people can only be robustly sensed in specific time resolutions.

B. Joint prediction of group membership and roles

Since we are dealing with a 4-class problem in joint prediction, we present normalised confusion matrices which make a more detailed evaluation possible. Following the group detection results, we stick to the fusion of proximity and acceleration for the experiments of the current section. Figures 5, 6, 7 show the performances obtained with window sizes 10, 15 and 25, respectively. These three window sizes are empirically found out to be the best selections for the three distinct role pairs. For the 4-class classification, a random classifier is expected to obtain a normalised confusion matrix with scores of 0.25 in the diagonal, which acts as the baseline.

In all of the figures, we can see that our model easily identifies the no interaction class, providing normalised scores
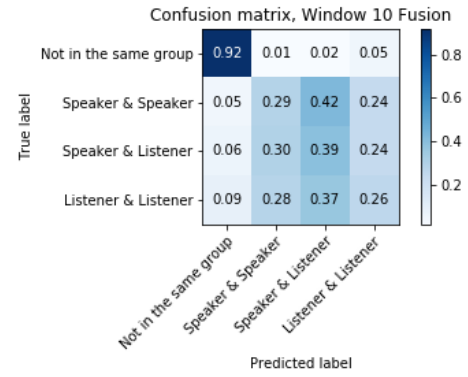
close to 0.9. This is compatible with the results of the former subsection where the proposed method with the fusion of both modalities successfully managed to distinguish between the presence and absence of an interaction between people.

The proposed method is relatively successful in distinguishing between role pairs and the performances have some interesting implications. The highest performances for each role pair are obtained with different window sizes. For the speaker&speaker class, the highest performance is obtained when 25 second windows are used. It is an expected result, since we already discussed that in order to detect coordination between interacting partners, longer time intervals are generally needed. This is especially true for speaker&speaker pairs where the emergence of turn-taking patterns are necessary for robust identification. We also see that, regardless of the window size, many speaker&speaker pairs are falsely labelled as speaker&listener pairs. One possible cause for this is the speakers that were not expressive during interactions. Another one is the limited amount of samples for speaker&speaker pairs in the training set where they were at least three times less than the others, resulting in them not being fully represented. On the other hand, results show that both other role pairs are equally misclassified as speaker$speaker pairs. We can assume that listeners that were quite active in terms of their movement might be the reason for this.

Highest performance for speaker&listener pairs, 0.66, is obtained with the window size of 15 seconds. However, we also see that nearly two-thirds of other role pairs are also misclassified as speaker&listeners. With this window size, no other role pair can be detected with a performance better than random. This suggests that the model frequently favoured the speaker&listener pairs over others. This is not directly related to the distribution of the role pairs in the training since the percentages of speaker&listener pairs were close to listener&listener pairs. We can speculate that speaker&listener pairs cover the largest variance of interaction patterns since they include two different roles, which might be the reason for these results.

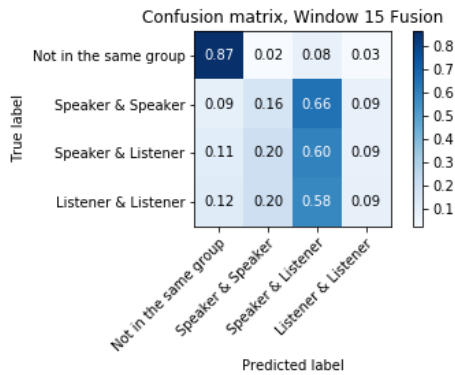For listener&listener pairs, the only performance which is

Fig. 6. Group membership and roles confusion matrix for 15-second windows (accumulated over 20 repetitions and normalised)

Fig. 7. Group membership and roles confusion matrix for 25-second windows (accumulated over 20 repetitions and normalised)

better than random is obtained with a window size of 10 seconds. The performance, 0.26, is only marginally higher than a random baseline. This is not surprising. There is no specific pattern for listener&listener pairs as the turn-taking for speaker&speaker pairs. Speaker&listener pairs can be still detected through the coordination of the listeners' backchannels as they react to the speakers. Lack of such behavioural patterns for listener&listener pairs make them relatively harder to model. These results show that more research into listener behaviour is needed if listener&listener pairs need to be robustly detected.

These results present the first attempts on investigating a formerly understudied phenomenon. We can conclude that proxemics and dynamics provide valuable information on the roles of people in interaction. However, relatively low performances, especially for listener&listener pairs, show that much work needs to be done for obtaining robust solutions. Next section discusses possible directions for further research.

VI. Conclusion and Future work

In this study, we investigated the automatic detection of pairwise F-formation memberships and roles using wearable sensors. We built our solution on proxemics and dynamics of interaction which were sensed through proximity and acceleration. With an LSTM network, we were able to learn joint pairwise representations of these modalities. These representations are then used to identify the presence of interaction and the roles participant took in it. Our solution was fully automatic and trained end-to-end, converting raw input data to output labels. We tested our method on the publicly available MatchNMingle dataset which was collected during real-life mingling events. We also investigated different setups in terms of input modalities and time resolutions. We tested our method on 20 different splits of our dataset to evaluate its generalisation capabilities.

Our experiments showed that using acceleration and proximity together guarantees performance at least as good as using separate modalities. With a window size of 30 seconds, which was suffic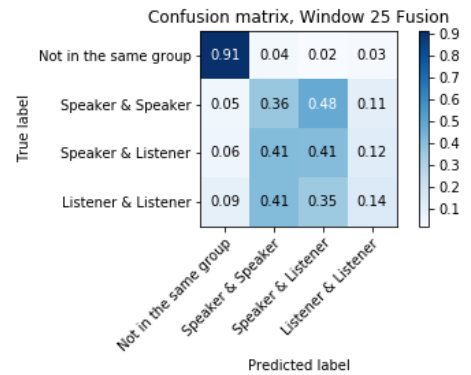ient for interaction dynamics to arise, the fusion of these modalities provided a mean AUC score of 0.975, significantly outperforming others. Our experiments on the detection of role pairs showed that different role pairs are better represented in different window sizes. Even though we were able to detect each role pair better than a random classifier, relatively low performances of these detections, especially for the listener&listener pairs, showed there is still room for improvement.

We believe there are many possibilities for extending this work, both for group membership and role detection parts. First of all, our formulation of the problem was based on a pairwise representation of interactions. As a refinement step to group membership detection, pairwise predictions of our method can be used to create a complete scene representation, similar to a proximity graph. There exist studies in the literature which provide a solution for this step [18]. By employing one, we can refine the predictions of our method. Secondly, our model treats the proximity and acceleration streams as equals while learning joint representations. More sophisticated formulations are also possible, such as using proximity streams as pooling layers for acceleration, which might result in better representation.

The fixed threshold used to identify speakers for training might result in mislabelling some instances. For example, in a 5 person group, if everyone speaks for an equal amount of time in a given interval, their speaking lengths will correspond to one-fifth of the window. In our current setup, none of these people will be identified as speakers. A proper labelling approach which considers the group size should result in better identification. Moreover, previous work has already identified the importance of using the group size information in the training [14]. Using the group size as additional information in the training phase might cover more varied behaviours of different roles. Finally, in this work, we only focused on two basic roles: speakers and listeners. However, social science literature has vast knowledge of group behaviour and the roles people take in interaction. Identifying more representative classes and creating sophisticated models for representing them should result in more satisfactory results.

REFERENCES

[1] C. R. Madan A. and P. A., "Voices of Attraction," in *proceedings of Augmented Cognition, (AugCog) HCI*, 2005.

[2] M. C. Nanninga, Y. Zhang, N. Lehmann-Willenbrock, Z. Szlávik, and H. Hung, "Estimating verbal expressions of task and social cohesion in meetings by quantifying paralinguistic mimicry," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ser. ICMI '17. New York, NY, USA: ACM, 2017, pp. 206–215. [Online]. Available: http://doi.acm.org/10.1145/3136755.3136811

[3] C. Gárate, S. Zaidenberg, J. Badie, and F. Bremond, "Group tracking and behavior recognition in long video surveillance sequences," in *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, vol. 2. IEEE, 2014, pp. 396–402.

[4] M. Cristani, R. Raghavendra, A. D. Bue, and V. Murino, "Human behavior analysis in video surveillance: A social signal processing perspective," *Neurocomputing*, vol. 100, pp. 86–97, 2013.

[5] O. A. I. Ramírez, G. Varni, M. Andries, M. Chetouani, and R. Chatila, "Modeling the dynamics of individual behaviors for group detection in crowds using low-level features," in *25th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2016, New York, NY, USA, August 26-31, 2016*, 2016, pp. 1104–1111. [Online]. Available: https://doi.org/10.1109/ROMAN.2016.7745246

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

[7] F. Setti, C. Russell, C. Bassetti, and M. Cristani, "F-formation detection: Individuating free-standing conversational groups in images," *PloS one*, vol. 10, no. 5, p. e0123783, 2015.

[8] L. Bazzani, M. Cristani, D. Tosato, M. Farenzena, G. Paggetti, G. Menegaz, and V. Murino, "Social interactions by visual focus of attention in a three-dimensional environment," *Expert Systems*, vol. 30, no. 2, pp. 115–127, 2013.

[9] S. Vascon, E. Z. Mequanint, M. Cristani, H. Hung, M. Pelillo, and V. Murino, "Detecting conversational groups in images and sequences: A robust game-theoretic approach," *Computer Vision and Image Understanding*, vol. 143, pp. 11–24, 2016.

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

[11] J. Gips, "Alex (sandy) pentland, mapping human networks," in *Proceedings of the Fourth Annual IEEE International Conference on Pervasive Computing and Communications*, 2006, pp. 159–168.

[12] D. Chaffin, R. Heidl, J. R. Hollenbeck, M. Howe, A. Yu, C. Voorhees, and R. Calantone, "The promise and perils of wearable sensors in organizational research," *Organizational Research Methods*, vol. 20, no. 1, pp. 3–31, 2017.

[13] A. Kendon, "Movement coordination in social interaction: Some examples described," *Acta psychologica*, vol. 32, pp. 101–125, 1970.

[14] E. Gedik and H. Hung, "Detecting conversing groups using social dynamics from wearable acceleration: Group size awareness," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 4, p. 163, 2018.

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

[16] G. P. A. F. D. T. A. D. B. G. M. V. M. Marco Cristani, Loris Bazzani, "Social interaction discovery by statistical analysis of f-formations," in *Proceedings of the British Machine Vision Conference*. BMVA Press, 2011, pp. 23.1–23.12, http://dx.doi.org/10.5244/C.25.23.

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

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

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

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

[21] D. O. Olguín, B. N. Waber, T. Kim, A. Mohan, K. Ara, and A. Pentland, "Sensible organizations: Technology and methodology for automatically measuring organizational behavior," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 1, pp. 43–55, 2009.

[22] N. Eagle and A. S. Pentland, "Reality mining: sensing complex social systems," *Personal and ubiquitous computing*, vol. 10, no. 4, pp. 255–268, 2006.

[23] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell, "Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones," in *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*. ACM, 2014, pp. 3–14.

[24] C. Cattuto, W. Van den Broeck, A. Barrat, V. Colizza, J.-F. Pinton, and A. Vespignani, "Dynamics of person-to-person interactions from distributed rfid sensor networks," *PloS one*, vol. 5, no. 7, p. e11596, 2010.

[25] A. Matic, V. Osmani, and O. Mayora-Ibarra, "Analysis of social interactions through mobile phones," *Mobile Networks and Applications*, vol. 17, no. 6, pp. 808–819, 2012.

[26] P. Casagranda, M. L. Sapino, and K. S. Candan, "Audio assisted group detection using smartphones," in *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2015, pp. 1–6.

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

[28] N. Brümmer and J. Du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

[29] J. P. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[30] R. Cutler and L. Davis, "Look who's talking: Speaker detection using video and audio correlation," in *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532)*, vol. 3. IEEE, 2000, pp. 1589–1592.

[31] E. Gedik and H. Hung, "Personalised models for speech detection from body movements using transductive parameter transfer," *Personal and Ubiquitous Computing*, vol. 21, no. 4, pp. 723–737, 2017.

[32] S. Inaba and Y. Aoki, "Conversational group detection based on social context using graph clustering algorithm," in *2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE, 2016, pp. 526–531.

[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.