Multimodal detection of engagement in groups of children using rank learning

Jaebok Kim[1], Khiet P. Truong[1], Vicky Charisi[1], Cristina Zaga[1], Vanessa Evers[1], and Mohamed Chetouani[2]

[1] Human Media Interaction, University of Twente, Enschede, The Netherlands
{j.kim,k.p.truong,v.charisi,c.zaga,v.evers}@utwente.nl
[2] Sorbonne Universités, UPMC Univ Paris 06, CNRS UMR 7222, Institut des Systémes Intelligents et de Robotique (ISIR), 4 place Jussieu, 75005 Paris, France
mohamed.chetouani@upmc.fr

Abstract. In collaborative play, children exhibit different levels of engagement. Some children are engaged with other children while some play alone. In this study, we investigated multimodal detection of individual levels of engagement using a ranking method and non-verbal features: turn-taking and body movement. Firstly, we automatically extracted turn-taking and body movement features in naturalistic and challenging settings. Secondly, we used an ordinal annotation scheme and employed a ranking method considering the great heterogeneity and temporal dynamics of engagement that exist in interactions. We showed that levels of engagement can be characterised by relative levels between children. In particular, a ranking method, Ranking SVM, outperformed a conventional method, SVM classification. While either turn-taking or body movement features alone did not achieve promising results, combining the two features yielded significant error reduction, showing their complementary power.

Keywords: children, engagement, social signal processing, non-verbal behaviours

1 Introduction

Engagement is often defined as the process of maintaining connections between participants through exchanges of verbal and non-verbal attentional cues to each other [5, 35]. From preschool age onwards, children play with peers in small groups [32, 31] where inter-group dynamics lead to varying engagement behaviours with children [37, 22, 2]. For example, one child does not play with others but plays alone while another child interacts substantially and gets involved with the other children in the group.

Recent advances in the automatic detection of engagement are more and more facilitating the development of robots able to support social interactions among children [34]. Our aim is to endow a social robot with the ability to anticipate children's level of engagement and to interact with children during

playful tasks, and this paper introduces a novel approach to the automatic detection of individual engagement during child-child interaction. Our approach focuses on the analysis of children's non-verbal behaviours which are strong cues of engagement [5, 38].

The automatic analysis and detection of engagement have been studied before using multi-modal cues such as speech activity, gaze, posture, and gestures [29, 6, 19, 21, 4]. These features can be categorised into vocal and visual features, and each category has its own strengths and drawbacks in the wild. For example, vocal features are not useful in situations where there is no speech, which often occur during children's playful tasks. Moreover, visual features such as gaze and gesture have limited accuracy depending on view points and distances. These challenges should be addressed to achieve reliable performances in naturalistic settings where we cannot instruct or restrict the behaviours of the children.

Furthermore, overlooked aspects in previous studies are temporal and group dynamics. Even in group play, an engagement level of each child is modelled by only his own non-verbal behaviours although their engagement and non-verbal behaviours are strongly interrelated with those of the other participants [22, 2]. In other words, a child's level of engagement is shaped by the other participants engagement and it may greatly vary depending on the group composition. Moreover, engagement levels vary over period (i.e. temporal dynamics), which calls for an analysis with fine time resolutions.

In this paper, we present multimodal detection of individual engagement of children in a naturalistic environment. To address the inevitable challenges such as silent moments and noisy viewpoints, we utilised not only vocalic turn-taking features but also body movement features. Moreover, we designed an ordinal annotation scheme and adopted a ranking method considering the great heterogeneity and temporal dynamics of engagement that exist in interactions [22, 2].

This paper is structured as follows. In Section 2, details of the related works will be presented. We will describe our audiovisual corpus and annotation scheme in Section 3. We will explain our method and features in Section 4. In Section 5, the results of our experiments will be presented, and conclusions will be addressed in Section 6.

2 Related Work

The automatic detection of engagement using multi-modal cues has been investigated in the field of Human Robot Interaction (HRI) and Social Signal Processing (SSP) [35, 38, 4]. In [29], hand-coded features such as speech, gaze, gesture, and postures of two children were utilised to model individual and group-level engagement. Their F-score based feature ranking showed that gaze-related features were more discriminative than other features (e.g. posture and smiling). Although their feature extraction was based on a fine-temporal resolution (500ms), the authors did not model turn-taking between children. In [6], a correlation between body movement and engagement in playful gaming situations

was investigated. The amount of body movements was quantified by the normalized sum of the angular movements over the total duration of play. They found a positive correlation between the movement and engagement while relying on wearable motion capturing devices which are expensive for practical applications. Moreover, group-level involvement, i.e. the average of individual engagement, was modelled using pitch, hand-coded gaze and blinking [7]. In [19], acoustic features (e.g. energy, pitch, speaking rate) and body movement features (e.g. amount of movements, orientation of head and hands) were automatically extracted to detect group-level engagement.

Although turn-taking features were often neglected in the aforementioned studies, turn-taking features, showed a positive relation to engagement [5, 13]. While individual speaking activity is not often informative to detect engagement, comprehensive behaviours such as speaker-changes, overlaps, and interruptions, demonstrated promising performances in the detection of engagement [28]. However, the performances of turn-taking features still remain doubtful in naturalistic settings where silent situations often occur.

While the HRI studies [4, 29] revealed that gaze-oriented movements and hand gestures were related to engagement, the settings often had regulations on behaviours of subjects or relied on hand-coded features. Without any regulation, the automatic extraction of these features is limited in naturalistic settings. Unlike in these features, body movements are atomic primitives which do not contain any contextual or sequence knowledge of human behaviours such as engagement [8, 1]; however, their statistics (e.g. occurrences) are known to be related with engagement [17, 6]. Moreover, the extraction of body movements is the first step to look into more advanced features (e.g. gaze and gesture). Hence, robust methods to extract and segment movements have been developed (e.g. Motion History Image (MHI) and K-means based segmentation) for identifying individuals in a group [8, 1, 3].

To resolve large variations of human behaviours, a large amount of data is often required, which is challenging for our targeted scenarios where a group of children exhibit social interactions in natural settings. Instead of collecting a large corpus, pairwise based ranking methods, for example, Ranking SVM, can be used to resolve these variations since these methods learn differences between instances in given conditions [23, 26]. For example, Ranking SVM achieved significant improvement compared to conventional methods (e.g. classification) in speech emotion recognition and engagement detection [14, 28]. However, none of these studies revealed limitations in silent situations which are common in child-child interactions.

These studies did not deal with the naturalistic settings where spontaneous interactions without restrictions and inter-group dynamics occur, and these challenges must be addressed to develop practical applications of engagement detection. To resolve the challenges, the annotation scheme, learning methods, and features adopted in this study will be elaborated on in the following sections.

3 Data

We used a corpus containing audiovisual recordings of groups of children [27]. In our corpus, a playful task was used to facilitate children's natural social behaviours. Using 3D cubes, children were asked to build given shapes of animals in collaborative ways as shown in Fig. 1. Dutch children aged 5 - 8 ($6.95 \pm .95$) were recruited from a primary school. We clustered the children by age and then randomly assigned them to a group of three for each session. Eight out of ten sessions were considered in our analyses (two sessions were discarded due to malfunctions of recording), totalling approximately 3 hours. Although we recorded children's behaviours using three different viewpoints focusing on each child as shown in Fig. 1 (a), occlusions caused by children sitting close to each other occurred relatively often which posed a great challenge to the automatic extraction of individual body movement features. Since we did not restrict the movement of the children (except for initial positions), they often moved around and interacted with each other, which led to noisy data.

(a) Overview of recording setup (b) Video still of corpus

Fig. 1. Naturalistic audiovisual corpus used in our study

3.1 Annotation

For our task, we define engagement as verbal and non-verbal exchanges of attention, i.e. attending and responding to each other in a group [28]. During our pilot coding sessions, we provided two coders with the definition of engagement and videos of three sessions. We asked them to label individual levels of engagement in an absolute manner ({low, medium, high}). It turned out that annotators had difficulty labelling these classes, resulting in poor inter-rater agreement (kappa) between the two coders (.57). Hence, we established an annotation scheme by considering relative levels of engagement as follows (from low to high level) [28]:

1 giving relatively less attention to others and receiving relatively less attention from others.

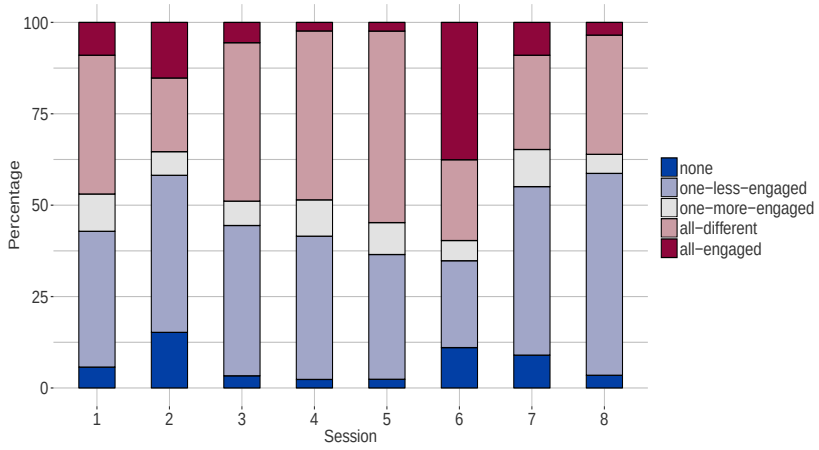


Fig. 2. Distribution of engagement situations

2a giving relatively less attention to others but receiving attention from others.
2b giving attention to others but receiving relatively less attention from others.
3 giving attention to others and receiving attention from others.

In this way, children can be ordered from a low to a high level of engagement. For subsequent analyses, the classes: {2a} and {2b} were equally ranked (in level of engagement) and merged into one class {2}. Moreover, if any differences could not be observed among the three children, ties were allowed (e.g. {1, 1, 1}, {3, 3, 3}). In order to annotate all the recordings, a proper size for an annotation segment needed to be determined. Through pilot coding sessions, we concluded on an empirical basis that 5s-long segments were suitable for the annotators to observe various levels of engagement. The videos and description of relative levels of engagement were given to two annotators to code each child for level every 5s using ELAN [39]. Finally, the average of inter-rater agreement was 0.82 (kappa). In subsequent analyses, we used 1510 segments that both annotators agreed upon which included silent segments ($< 22.5\%$).

If children are equally engaged or disengaged with each other in general, our ordinal annotation is not meaningful. To investigate this issue, we analysed five types of engagement situations: "no-one-engaged (none)", "one-less-engaged", "one-more-engaged", "all-different", and "all-engaged". In "no-one-engaged", none of the children was engaged with any other and just focused on their own task (e.g. {1, 1, 1}). In "one-less-engaged", one child is less engaged and others are more engaged with each other (e.g. {1, 2, 2}). In "one-more-engaged", one child is more engaged than any other (e.g. {3, 2, 2}). "all-different" means all children have different levels of engagement (e.g. {1, 2, 3}). Lastly, "all-engaged" means that all children are equally engaged without observable differences between them (e.g. {3, 3, 3}). Fig. 2 presents each session's proportion of engagement types (the average proportions are $7.1 \pm 4.7\%$, $40.9 \pm 9.1\%$,

Table 1. Feature sets (number of features by functionals) for each child

Category	Features	Functionals
turn-taking (28)	speech (4), pause(4) speaker change (4), speaker change with overlap (4), successful interruption (4), unsuccessful interruption (4), overlap (4)	mean-duration SD-duration total-duration total-count
body (7)	movement (4)	mean-amount SD-amount total-count total-amount
	orientation (1)	mean-orientation
	position x (1) position y (1)	mean-position

$7.9 \pm 2.1\%$, 33.3 ± 11.8, and $10.8 \pm 11.7\%$, respectively). Major portions are "one-less-engaged" and "all-different", which means that the children frequently exhibited different levels of engagement. Moreover, we found variations of engagement situations between the groups, which support findings of previous studies [22, 2].

4 Method

In this section, we will present our features: turn-taking and body movement. We will introduce the Ranking SVM algorithm that employs the ordinal annotation scheme. Table 1 summarises all features and their details will be described in the following sections.

4.1 Turn-taking features

Based on previous studies [38, 24, 27], we selected the following turn-taking features: speech, pause, speaker change (change), speaker change with overlap (change.ov), successful interruption (inter), unsuccessful interruption (u.inter), and overlap, as shown in Table 1. More detailed descriptions can be found in [27]. The features were extracted from every 5s long annotation segment using each child's voice stream. First of all, we extracted each child's speech segments using voice activity detection from each voice stream. Then, to correct errors caused by environmental noise and channel-inferences, we employed iterative speaker identification. Similarly to [12], we used Mel-frequency cepstral coefficients (MFCC) features and the Gaussian-Mixture-Model to detect segments of different speakers. In an iterative way, we updated each speaker's model using the previously extracted segments and manually corrected errors until the models became saturated (no more changes of segments were observed). Finally, we extracted each

speaker's speech segments using the saturated models. In real-time applications, on-line speaker segmentation should be applied but we consider this to be future work.

Next, all turn-taking features were extracted from the speech segments detected, and statistical functionals {mean-duration, standard-deviation (SD) of duration, total-duration, and total-count} were applied, as presented in Table 1. Lastly, all values are scaled into the range {0.0 - 1.0} over each session.

4.2 Body movement and segmentation

To extract body movement features, we first performed a foreground segmentation calculating pixel-wise differences between frames, followed by a Gaussian threshold [3]. Next, we identified each child by K-means clustering [3], and extracted movements by using MHI implemented in OpenCV [10, 11]. For its robustness, we did not specify which part of body moves (e.g. legs and hands). Instead, we extracted the number (of changes in pixels between frames), orientation, and position (coordinates of x and y) from each movement. Hence, we applied statistical functionals and obtained mean-amount, SD-amount, total-count, total-amount, mean-orientation, and mean-position (x and y) for each child from every annotation segment. As turn-taking features were normalised for each session, all movement features were also scaled into the same range.

4.3 Ranking SVM

As other ranking methods (e.g. ListNet) are suitable if the number of instances in an order is variant [15], Ranking SVM, categorised as pairwise approach, is more effective in our task where the number of children is invariant. To learn an order of engagement between children, we compare only feature vectors of two children in the same constraint. In our task, the constraint is the period of time. In other words, we do not compare children's feature vectors which have different time periods. Therefore, a value of the constraint, often called *qid*, is each annotation segment's index representing given moments in the range of [0, the total number of segments for each session]. More detailed explanation of Ranking SVM can be found in [23, 26].

5 Analysis and Results

Based on our annotation scheme, we extracted all feature values described in Section 4. In this section, we present the analysis of our features with respect to engagement levels, the detection experiments and their results.

5.1 Feature analysis

Since our annotation schemes are ordinal, we looked into differences of feature values between children depending on their ordinal relations of engagement levels. For example, if one child is more engaged than the other child, is this child

also more active in speaking or moving? Moreover, we do not have prior knowledge of the proper size of a window for feature extraction in our study. In previous work [29, 33, 25], windows of between 0.5s and 5min long were used to predict engagement and dominance. To decide an optimal length of windows for detection experiments, we investigated the effect of different window length varying: {5s, 10s, 15s, 20s, and 25s}. First, we collected all feature values in the pairwise way for each engagement window (grouped by *qid*: segment ID) with different lengths. To decide a new engagement level for each window, we utilised a major voting policy. Next, we grouped feature values by ordinal relations: **higher** and **lower**. All features values extracted from children who had higher ranks are categorised into **higher**. Otherwise, feature values are categorised into **lower**. Next, to validate significance of differences between the feature values of **higher** and **lower**, we conducted a Wilcoxon signed-rank test (alternative: greater) that is a non-parametric paired difference test [36].

We found that 20s long windows produced the largest number (7) of features that have significant differences of both mean-count and mean-amount (or length) between ranks ($p < .0001$). 5s long windows produced the smallest number (4) of features, which means that 5s long windows are not sufficiently long to capture turn-taking and movement features. Hence, we decided to choose 20s long windows for subsequent analyses and detection experiments. Table 2 summarises our findings of 20s long windows. Note that we list only significant results (with $p < .0001$), and present the normalised values of mean-count. Moreover, we analysed feature values of all segments to look at overall characteristics of feature values while cross-validation was employed for evaluation in Section 5.2.

Table 2. Average of feature values (mean-count) with respect to ordinal relations of engagement levels (**higher**: if the engagement level is higher, **lower**: the engagement level is lower)

Engagement	speech	pause	change.ov	inter	u.inter	overlap	movement
higher	.242	.233	.198	.080	.052	.139	.187
lower	.178	.164	.164	.068	.042	.115	.168

Except for speaker change, all turn-taking features showed significant differences. From these findings, we concluded that as some children are more engaged than others at given moments, they tend to show more active turn-taking in conversations. For movement features, amount and count of movement were the most significantly discriminative between higher and lower ranks. Possibly, orientation and positions might not be related with ranks in a linear way.

5.2 Detection experiments

In this section, we present detection experiments using Ranking SVM. As a baseline, SVM classification (**SVM**) was compared to Ranking SVM (**SVMRANK**).

As an exploratory study, we did not select our features using selection methods for ranking [20]. Rather, we compared performances of feature groups: turn-taking and movement. Furthermore, we combined these features at feature level to see if they would complement each other and increase performances in two different situations: **all** and **speech**. **all** includes silent samples while **speech** excludes silent samples. We investigated how much movement features complement turn-taking features in these challenging situations.

For purposes of reproduction, we utilised the implementation of LIBSVM and its extensions [16, 30]. Parameters of each model were optimised by a simple grid search. For evaluation, we used the normalised Kendall tau distance, which is a widely used evaluation method for rank learning [18]. To calculate it for two lists (e.g. X_1 and X_2), it is defined as follows:

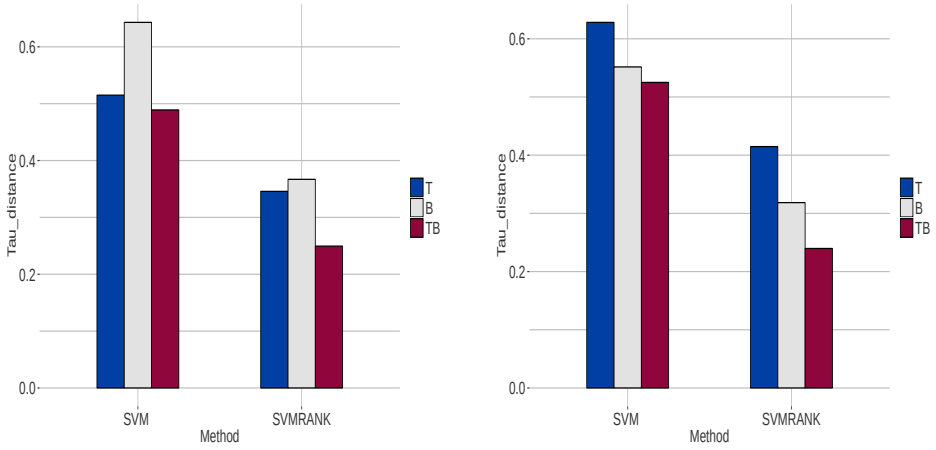
$$K(X_1, X_2) = \frac{D}{N(N-1)/2} \tag{1}$$

where D is the total number of swapped pairs and N is the total number of elements in a list. If all orders are incorrect, then it becomes 1.0 while indicating 0.0 for completely correct orders, which can be regarded as an error rate. To test the statistical significance of differences between the methods, we employed a paired corrected t-test [9] (p-values are separately provided with the results).

We look into performances of each feature set using Leave-One-Session-Out-Cross-Validation (LOSOCV). In each fold, one session is used for validation and all other sessions are used for training. Since we have 3 children's samples per segment, a total of 4530 samples (including silent samples 1011) were used and the average number of test samples is 566 and that of training is 3735. Fig. 3 summarises two results, all samples (**all**) and samples without silent segments (**speech**).

5.3 Results

First of all, **SVMRANK** outperformed **SVM** with significant differences ($p < .0001$) in both **speech** and **all**. In other words, **SVMRANK** was effective in modelling relative levels of engagement using not only vocal interactions (i.e turn-taking), but also body movement in the multimodal detection. In Fig. 3 (a) showing cases of **speech**, performances of turn-taking features (T) were slightly superior to those of body movement features (B). However, the differences between T and B are not significant ($p = .6$). Second, in cases of **all** (see Fig. 3 (b)), B outperformed T with significant differences ($p = .0290$). Although turn-taking features showed discriminative power between higher and lower ranks in the previous section, they did not show promising results. Since we conducted neither non-linear correlation analyses nor error analyses of **speech** and **all**, separately, our findings are not conclusive yet. However, combined features showed the best performances and reduced error rates of both turn-taking and body movement features with significant levels ($p = .0014$ and $p = 0.0043$, respectively) for **all**. In other words, turn-taking and body movement features complemented each



(a) Performances in **speech** (b) Performances in **all**

Fig. 3. Summary of performances (normalised Kendall tau distance): T (turn-taking), B (body movement), TB (combined)

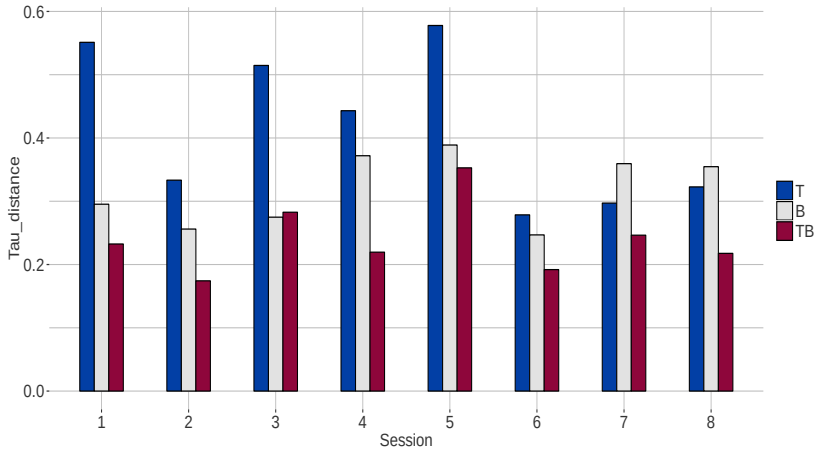


Fig. 4. Intra-session performances: T (turn taking), B (body movement), TB (combined features)

other, leading to the best results. In addition, Fig. 4 presents results of inter-session performances of **all**. As displayed, depending on the groups of children, the performances fluctuated. While gains of TB with respect to T and B vary, TB reduced errors for most sessions. In particular, for session 1, TB reduced errors of T by nearly .33. Moreover, TB showed the smallest variation over sessions (.003). In other words, the combined feature set was robust against inter-group dynamics.

5.4 Limitations and future work

While turn-taking features outperformed movement features in **speech**, their performances were degraded in **all** where "silent" moments often occurred. Since the combined features achieved the best performances in both **speech** and **all**, we concluded that turn-taking and movement features complemented each other, which is promising for applications in the wild. However, non-linear relations between features and engagement levels still remain unexplored. Thus, levels of feature or decision fusion should be investigated. For example, we could build separated classifiers for different situations. Furthermore, we might be able to utilise more advanced visual features (e.g. gaze, blinking, gesture) which have semantic information for engagement [1]. While these features were widely used in controlled or laboratory settings [29, 4, 6], we should investigate rigorous methods to extract these features in naturalistic settings where we cannot regulate children's behaviours.

6 Conclusions

We explored the multimodal engagement detection of individuals using non-verbal features, turn-taking and body movement, in the context of children's collaborative play. To observe spontaneous engagement in groups of three children, we did not impose any restriction on children's conversations and their movements. As a consequence, there were silent situations and limited viewpoints that hindered the automatic extraction of non-verbal features. Moreover, groups of three children exhibited large variations of interactions with temporal dynamics. To address the large variations, we showed that levels of engagement can be characterised by relative levels between children. Moreover, we conducted detection experiments of individual engagement levels using turn-taking and movement features. The Ranking SVM outperformed the SVM classification, which means that the ranking method could be better suited for the multimodal detection of engagement in groups of children. Furthermore, while each feature set alone did not achieve promising results, the combined feature set showed significant error reduction, which means that turn-taking and body movement features complemented each other. As future work, we will conduct more detailed feature analysis including non-linear correlation analysis and investigate methods of integrating our multimodal features.

Acknowledgements

The research leading to these results was supported by the European Community's 7th Framework Programme under Grant agreement 610532 (SQUIRREL - Clearing Clutter Bit by Bit). This work was also partially performed within the Labex SMART (ANR-11-LABX-65) supported by French state funds managed by the ANR within the Investissements dAvenir programme under reference ANR-11-IDEX-0004-02. We would like to thank S.M. Anzalone, S.V. Waveren and F.V. Dixhoorn.

References

1. Aggarwal, J.K., Park, S.: Human motion: Modeling and recognition of actions and interactions. In: 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004. Proceedings. 2nd International Symposium on. pp. 640–647. IEEE (2004)
2. Al Moubayed, S., Lehman, J.: Toward better understanding of engagement in multiparty spoken interaction with children. In: Proceedings of the International Conference on Multimodal Interaction. pp. 211–218. ACM (2015)
3. Antić, B., Letić, D., Crnojević, V., et al.: K-means based segmentation for real-time zenithal people counting. In: Proceedings of International Conference on Image Processing (ICIP). pp. 2565–2568. IEEE (2009)
4. Anzalone, S.M., Boucenna, S., Ivaldi, S., Chetouani, M.: Evaluating the engagement with social robots. International Journal of Social Robotics 7(4), 465–478 (2015)
5. Argyle, M.: Social interaction, vol. 103. Transaction Publishers (1973)
6. Bianchi-Berthouze, N., Kim, W.W., Patel, D.: Does body movement engage you more in digital game play? and why? In: International Conference on Affective Computing and Intelligent Interaction. pp. 102–113. Springer (2007)
7. Oertel gen bierbach, C.: On the use of multimodal cues for the prediction of involvement in spontaneous conversation. In: Proceedings of the INTERSPEECH. pp. 1541–1544 (2011)
8. Bobick, A.F.: Movement, activity and action: the role of knowledge in the perception of motion. Philosophical Transactions of the Royal Society of London B: Biological Sciences 352(1358), 1257–1265 (1997)
9. Bouckaert, R.R., Frank, E.: Evaluating the replicability of significance tests for comparing learning algorithms. In: Advances in knowledge discovery and data mining, pp. 3–12. Springer (2004)
10. Bradski, G., Kaehler, A.: Learning OpenCV: Computer vision with the OpenCV library. " O'Reilly Media, Inc." (2008)
11. Bradski, G.R., Davis, J.W.: Motion segmentation and pose recognition with motion history gradients. Machine Vision and Applications 13(3), 174–184 (2002)
12. Busso, C., G. Georgiou, P., Narayanan, S.S.: Real-time monitoring of participant's interaction in a meeting using audio-visual sensors. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE (2007)
13. Campbell, N., Scherer, S.: Comparing measures of synchrony and alignment in dialogue speech timing with respect to turn-taking activity. In: Proceedings of the INTERSPEECH. pp. 2546–2549 (2010)
14. Cao, H., Verma, R., Nenkova, A.: Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech. Computer speech & language 29(1), 186–202 (2015)
15. Cao, Z., Qin, T., Liu, T.Y., Tsai, M.F., Li, H.: Learning to rank: from pairwise approach to listwise approach. In: Proceedings of international conference on Machine learning. pp. 129–136. ACM (2007)
16. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2(3), 27 (2011)
17. Dittmann, A.T., Llewellyn, L.G.: Body movement and speech rhythm in social conversation. Journal of Personality and Social Psychology 11(2), 98 (1969)

18. Fagin, R., Kumar, R., Sivakumar, D.: Comparing top k lists. SIAM Journal on Discrete Mathematics 17(1), 134–160 (2003)
19. Gatica-Perez, D., McCowan, I.A., Zhang, D., Bengio, S.: Detecting group interest-level in meetings. Tech. rep., IDIAP (2004)
20. Geng, X., Liu, T.Y., Qin, T., Li, H.: Feature selection for ranking. In: Proceedings of the international conference on Research and development in information retrieval. pp. 407–414. ACM (2007)
21. Gupta, R., Lee, C.c., Lee, S., Narayanan, S.: Assessment of a child's engagement using sequence model based features. In: Workshop on Affective Social Speech Signals (2013)
22. Hall, J.A., Coats, E.J., LeBeau, L.S.: Nonverbal behavior and the vertical dimension of social relations: a meta-analysis. Psychological bulletin 131(6), 898 (2005)
23. Hang, L.: A short introduction to learning to rank. IEICE TRANSACTIONS on Information and Systems 94(10), 1854–1862 (2011)
24. Heldner, M., Edlund, J.: Pauses, gaps and overlaps in conversations. Journal of Phonetics 38(4), 555–568 (2010)
25. Jayagopi, D.B., Ba, S., Odobez, J.M., Gatica-Perez, D.: Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues. In: Proceedings of international conference on Multimodal interfaces. pp. 45–52. ACM (2008)
26. Joachims, T.: Optimizing search engines using clickthrough data. In: Proceedings of the international conference on Knowledge discovery and data mining. pp. 133–142. ACM (2002)
27. Kim, J., Truong, K.P., Charisi, V., Zaga, C., Lohse, M., Heylen, D., Evers, V.: Vocal turn-taking patterns in groups of children performing collaborative tasks: an exploratory study. In: Proceedings of the INTERSPEECH. pp. 1645–1649 (2015)
28. Kim, J., Truong, K.P., Evers, V.: Automatic detection of children's engagement using non-verbal features and ordinal learning. In: Workshop on Child Computer Interaction (2016)
29. Leite, I., McCoy, M., Ullman, D., Salomons, N., Scassellati, B.: Comparing models of disengagement in individual and group interactions. In: Proceedings of Annual ACM/IEEE International Conference on Human-Robot Interaction. pp. 99–105. ACM (2015)
30. Li, L., Lin, H.T.: Ordinal regression by extended binary classification. In: Advances in neural information processing systems. pp. 865–872 (2006)
31. Parten, M.B.: Social participation among pre-school children. The Journal of Abnormal and Social Psychology 27(3), 243 (1932)
32. Piaget, J.: The psychology of the child. Basic Books, New York (1972)
33. Pianesi, F., Zancanaro, M., Lepri, B., Cappelletti, A.: A multimodal annotated corpus of consensus decision making meetings. Language Resources and Evaluation 41(3-4), 409–429 (2007)
34. Robins, B., Dautenhahn, K., Te Boekhorst, R., Billard, A.: Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills? Universal Access in the Information Society 4(2), 105–120 (2005)
35. Sidner, C.L., Lee, C., Kidd, C.D., Lesh, N., Rich, C.: Explorations in engagement for humans and robots. Artificial Intelligence 166(1), 140–164 (2005)
36. Siegel, S.: Nonparametric statistics for the behavioral sciences. McGraw-hill (1956)
37. Stangor, C.: Social groups in action and interaction. Psychology Press (2004)

38. Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., Schröder, M.: Bridging the gap between social animal and unsocial machine: A survey of social signal processing. Affective Computing, IEEE Transactions on 3(1), 69–87 (2012)
39. Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H.: Elan: a professional framework for multimodality research. In: Proceedings of LREC. pp. 5–8 (2006)