Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 50 (2018) 16–39

www.elsevier.com/locate/csl

Automatic temporal ranking of children's engagement levels using multi-modal cues☆

Jaebok Kim*, Khiet P. Truong, Vanessa Evers

Department of Electrical Engineering, Mathematics and Computer Science, University of Twente, 5 Enschede, Drienerlolaan, The Netherlands

Received 1 May 2016; received in revised form 11 June 2017; accepted 13 December 2017
Available online 19 December 2017

Abstract

As children of ages 5–8 often play with each other in small groups, their differences in social development and personality traits usually cause various levels of engagement among others. For example, one child may just observe without engaging at all with others while another child may be interested in both the other children as well as the activity. To develop child-friendly interaction technology such as social robots that can adapt robot behaviours to the social situation of a group of children and facilitate harmonious engagement, we aim to study how we can automatically detect these children's engagement levels. In this paper, we present a novel automatic method that ranks children in a group according to their engagement level in a temporal way based on non-verbal cues that are robust in naturalistic group settings. Our method combines the omission probability of each rank transformed from discriminative outputs from an SVM ranking method and the transition probability between ranks in time. In comparing our proposed method to other existing methods (such as rule-based ranking, basic SVM, SVM ordinal regression, SVM ranking, and SVMHMM), we found that our novel method yields promising results.

Keywords: Engagement; Play; Children conversation; Non-verbal; Multi-modal

1. Introduction

As said by Aristotle: "Man is by nature a social animal; an individual who is unsocial naturally" – people's social behaviours should be understood in consideration of their group (Vinciarelli, 2009; Gatica-Perez, 2009). In a similar sense, young children in the age range of 6–9 years are developing their social behaviours and skills most often in small groups (Stangor, 2004), and hence, it is necessary to understand how they behave and interact in small groups. Among several contexts, social or group-play is one of the most important and experienced environments where children have dynamic and comprehensive interactions with their peers and learn new skills while having fun.

Engagement is a broadly used term characterised by interactions between people (social engagement), involvement in tasks (task engagement), or both of them (social-task engagement) (Corrigan et al., 2013b; 2013a). Engagement between people, social engagement (Corrigan et al., 2013b), is the process of maintaining connections through

☆ This paper has been recommended for acceptance by R. K. Moore.

* Corresponding author.
E-mail address: j.kim@utwente.nl (J. Kim).

http://dx.doi.org/10.1016/j.csl.2017.12.005
0885-2308/

exchanges of verbal and non-verbal attention between participants (Argyle, 1973; Sidner et al., 2005; Kim et al., 2016). When children play in a small group, they often exhibit different levels of engagement with each other and/or the task because of their variations in personality, social skills, and inter-group dynamics (Stangor, 2004; Piaget, 1972; Al Moubayed and Lehman, 2015). For example, some children pay attention to only his/her own play while others interact substantially and get involved with the other children in the group. Within the context of child-robot interaction, understanding distinctive levels among a group of children might be greatly helpful (Robins et al., 2005) for the development of social robots that can facilitate the flow of engagement (Matsuyama et al., 2015) or maintain attentions in educational setting (Szafir and Mutlu, 2012).

How can we sense and interpret children's engagement in group-play? The automatic detection of engagement has been studied using various non-verbal cues such as vocal activity, gaze, gestures, and postures (Leite et al., 2015; Bianchi-Berthouze et al., 2007; Gatica-Perez et al., 2004; Gupta et al., 2012; 2013; Anzalone et al., 2015; Gatica-Perez, 2009; Nakano and Ishii, 2010). While their features showed promising contribution to engagement detection, they have limitations in naturalistic settings for children's group-play. For example, unless we pose strict regulations on children's movement and positions, it is challenging to extract vocal bursts (Tahon et al., 2012) or back-channels because of noise in the multi-party condition. Moreover, semantic visual features such as gaze, gesture, and posture become unreliable when children freely move around but viewpoints are limited, which is natural in group-play. These issues should be addressed to achieve reliable performances in naturalistic settings for children's group-play where we cannot restrict the behaviours of the children.

More importantly, we should consider the temporal dynamics of children's social behaviours and the large dependency of the children's behaviours on the groups they are in. Firstly, the temporal dynamics of social behaviours and its vocal cues have been rarely studied (Al Moubayed and Lehman, 2015). Since levels of engagement and dominance of children change dynamically over time, a finer resolution for the detection of these levels is necessary. However, coarser resolutions, e.g. whole sessions, have often been used in previous works (Sidner et al., 2005; Leite et al., 2015). Secondly, the non-verbal expression of social behaviours heavily depend on the participants in the groups (Hall et al., 2005). Since individuals show distinctive trends of behaviours and cues depending on whom they are interacting with at given moments, it is not desirable to define absolute levels of engagement such as "high" or "low" for individuals in the group. Rather, labels that express relative orders and that take into account the behaviours of others are more suitable for modelling social behaviours such as engagement in small groups.

In the current study, we present a method towards automatic temporal ranking of individual engagement levels in small groups of children using non-verbal cues in a naturalistic setting. We define engagement as "verbal and non-verbal exchanges of attention, i.e. attending and responding to each other in a group" (Argyle, 1973; Kim et al., 2016). To deal with the inevitable challenges such as noisy multi-party and viewpoints, we utilised not only vocal turn-taking features but also body movement features. Moreover, we propose to use a ranking algorithm, more specifically, pairwise ranking which might be effective in modelling speaker or group dependencies as shown in Cao et al. (2015). As far as we know, no previous work has demonstrated how effective ranking methods are for the current task, and integration of temporal domains has remained unexplored. The main contributions of this paper can be summarized as follows: (1) we present a novel algorithm that integrates ranking characteristics and temporal dynamics in a seamless way which yielded the best performance in our experiments, (2) we present a novel annotation scheme (based on Argyle, 1973; Sidner et al., 2005) representing relative levels of children's engagement with a fine time resolution (each 5 s), and (3) we present a robust feature extraction of non-verbal cues such as vocalic turn-taking and primitive body movement in order to deal with noisy of the naturalistic setting.

This paper is structured as follows. We first give a short overview of related previous work in Section 2, and present our spontaneous corpus of children behaviours in a social play setting in Section 3. In addition to, the annotation scheme that defines categories of engagement levels identifying distinctive patterns in a relative manner instead of an absolute manner is also presented in Section 3. We define and analyse non-verbal features in Section 4. Our novel algorithm based on ranking and temporal decoding is presented in Section 5. The results will be reported with cross-session validation in Section 6 and discussed in Section 7.

2. Related work

We present related previous work on small group analysis in the context of adults and children, as well as related work on ranking algorithms and temporal recognition models.