Real-Time Emotion Recognition from Natural Bodily Expressions in Child-Robot Interaction

Weiyi Wang[1](✉), Georgios Athanasopoulos[1], Georgios Patsis[1], Valentin Enescu[1], and Hichem Sahli[1,2]

[1] Department of Electronics and Informatics (ETRO) - AVSP,
Vrije Universiteit Brussel (VUB), Brussels, Belgium
wwang@etro.vub.ac.be

[2] Interuniversity Microelectronics Centre (IMEC), Heverlee, Belgium

Abstract. Emotion perception and interpretation is one of the key desired capabilities of assistive robots, which could largely enhance the quality and naturalness in human-robot interaction. According to psychological studies, bodily communication has an important role in human social behaviours. However, it is very challenging to model such affective bodily expressions, especially in a naturalistic setting, considering the variety of expressive patterns, as well as the difficulty of acquiring reliable data. In this paper, we investigate the spontaneous dimensional emotion prediction problem in a child-robot interaction scenario. The paper presents emotion elicitation, data acquisition, 3D skeletal representation, feature design and machine learning algorithms. Experimental results have shown good predictive performance on the variation trends of emotional dimensions, especially the arousal dimension.

Keywords: Spontaneous emotion recognition · Child-robot interaction · Bodily expressions

1 Introduction

The development of assistive robots aims at designing robots that could help humans in everyday life or on specific tasks. Among which, child companion robot is one of the major applications. Such kind of robots are designed to be able to interact autonomously with children. This requires the robots to correctly interpret the social behaviours of the children, and respond accordingly. Supported by psychological studies, affective phenomena, especially emotions, are the key information conveyed in daily communication among humans [4,26,30,31]. Thus the capability of understanding the emotional states of the children, becomes an asset for child companion robots.

Emotions are multi-component responses that are delivered through various channels such as facial expressions, bodily movements, speech and physiological signals [15]. According to [12], 95% of the emotion recognition study was conducted with facial cues, the majority of the remaining 5% with audio, while

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part III, LNCS 8927, pp. 424–435, 2015.
DOI: 10.1007/978-3-319-16199-0_30

the bodily stimuli were relatively neglected. However, recent empirical study provided the evidence that emotional information could be not only conveyed, but also perceived, via the body as a single channel [2,3,10,35]. Encouraged by those findings, emotion recognition from bodily information attracted increasing interests in recent years, yet most of the work in the literature was focused on acted expressions of adults [6,7,9,19,20,23,32,38]. The main drawback of these studies is the loss of naturalness in the expressions, which makes them not suitable to be utilized in assistive robot applications, especially on child companion robots.

In this paper, we introduce a framework and its preliminary results, for spontaneous emotion recognition from bodily expressions in a child-robot interaction setting. The framework involves natural emotion elicitation, expressive data acquisition, emotion annotation, body feature design and learning models. The remainder of the paper is structured as following: Section 2 reviews different emotion models and gives the explanation of our choice. Section 3 describes our emotion elicitation data acquisition experiments, as well as the annotation scheme. Section 4 gives the details of the features and the recognition model for emotion prediction. We then give some experimental results in Section 5 and the discussion in Section 6.

2 Emotion Modelling

The representation of emotions could be mainly divided into two groups, referred to as *categorical models* and *dimensional models*. The categorical representations are based on selected vocabulary of emotions such as *happy, sad, feared, angry* etc [11]. These discrete labels of emotions have specific social meanings which are, to a large extent, accepted universally by people despite of the regions, cultural backgrounds or genders. Thus, they could be intuitively understood among us. Moreover, categorical models are inherently advantaged to represent simultaneous emotions that occur occasionally in real life [21]. However, the capability to describe the comprehensive emotional states is highly dependent on and constrained by the selected labels. Furthermore, categorical models normally consider the emotions as static temporal segments, which is in conflict with the intrinsic continuity of emotions, and limits the feasibility to describe the variation trends. The dimensional models, on the other hand, were advocated to meet the fact that mental states are much more complicated than the so-called basic emotions [4]. Moreover, they can better cope with the continuous nature of affect. The drawback is that those dimensions are less explicitly interpretable, compared to the categorical labels.

In this work, we use the circumplex space [29], specifically, the *arousal* and *valence* dimensions, to model the emotional states of children in their interactions with a robot. *Arousal* values indicate the external expressions between relaxed and aroused, while *valence* values reveal pleasant (positive) or unpleasant (negative) status. This choice was made mainly based on the consideration of natural interaction, by letting the robot continuously adjusting its reactions based on the perceived emotion of the child. More precisely, in our settings it is not necessary for the robot to interpret semantically the child's expressions, while it is essential that it responds quickly according to the changes (even when they are subtle) of the emotional states of the children.

3 Naturalistic Data Acquisition and Annotation

Obtaining expressive data is a vital step for spontaneous emotion modelling. This requires proper emotion elicitation protocol, well arranged recording setup, as well as reliable annotation scheme. In this section, we briefly introduce our spontaneous data acquisition and annotation framework. More details could be found in [37].

3.1 Naturalistic Emotion Elicitation of Children

In general, traditional emotion elicitation approaches employed visual and/or auditory stimuli to induce certain expressions. The most widely applied method is to ask the participants to watch film segments that were pre-selected to deliver strong feelings [14]. The main drawback of this approach lies in its static and passive nature: the participants are hardly expressing externally, especially via the body, in a non-interactive environment. Dyadic interaction tasks also attracted many research work by introducing the communication between participants [28]. A simulated Sensitive Artificial Listener (SAL) with emotional profiles were incorporated in [22] to enhance the affective engagement of the participants. However, it is relatively difficult to design the conversational scope to successfully trigger bodily expressions, specially in the case of children.

In order to deal with the above issues, we designed a child-robot interaction scenario to elicit naturalistic expressions. Each participated child was asked to play the Snakes and Ladders game against the humanoid robot NAO [24]. To better cope with the emotion elicitation purpose, we manually scripted the unfolding of four games (the child and the robot would win two of those respectively, and all the dice throws were predefined). The game steps were designed to be dramatic and therefore produce a clear reaction from the child, either positive or negative.

We hypothesize that a believable interaction should be maintained and hence the robot's verbal communication should be as natural as possible. With this in mind, we opted for a Wizard-of-Oz (WoZ) setting, where the operator's speech was streamed to the robot in real-time, with the voice being modified [34] so that it resembles the robot's voice. Moreover, the robot could display two different affective profiles while playing the game: one *competitive*, where the robot would display self-centred emotions, and one was *supportive*, focusing on the child's performance. The profiles were displayed alternatively in different games. The *competitive* profile made the robot react strongly to positive events and negative events occurring to the robot, making the robot appearing more involved in the game and eager to win it. The behaviours and gestures of the robot appeared

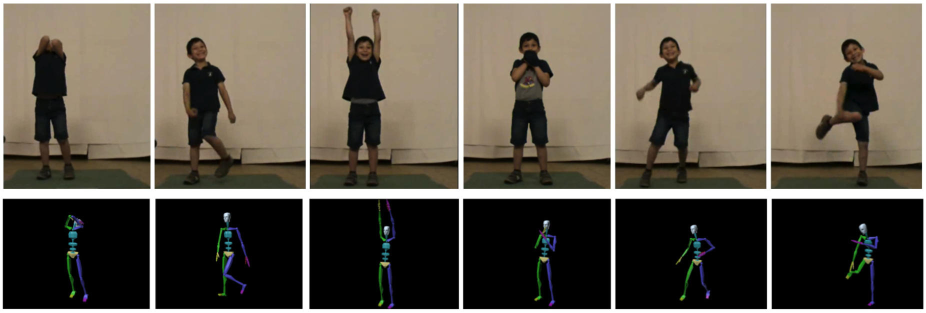

Fig. 1. Examples of the extracted 3D skeletons. Note that in the first and forth columns, the skeletons are well captured even some body parts are occluded.

more aroused and energetic. Following the literature on empathy and sympathy and their importance in peer bonding and fostering trust [27], using the *supportive* profile, the robot displayed and expressed behaviours suggesting a more focused interest on the outcomes for the child's. Additionally, the verbal expressions of the WoZ operator were consistent with the specific affective profile of the robot used at the time of interaction.

Before the start of the interaction, a short familiarization phase, using animated behaviours, took place so that the children would feel comfortable interacting with the robot, and to familiarize the children with the robot's movements and emotional expressions. Simple gestures (e.g., standing up, waving hello, nodding, etc.) and emotional postures (similar to the ones implemented in [5]) have been used during this phase. These behaviours were triggered by the WoZ operator via a graphical user interface.

3.2 Bodily Data Acquisition

We arranged a dual-Kinect set-up. Two Microsoft Kinect sensors were placed in 90° to record the movement of the child, at the same frame rate. The 3D skeletons were reconstructed offline from the dual recordings, using the iPi Mocap Studio software [25]. Fig. 1 gives some examples of the skeletal representations. Note that even when some body parts were occluded, the skeletons were still well tracked.

For the purpose of our current research on multi-modal emotion recognition [13,17,36], we also recorded high-definition face and frontal body videos, audio from both the robot and the child, and the child-robot interactions. All recordings were synchronized.

3.3 Dimensional Emotion Annotation

A three-view video for each recording session (see Fig.2 as an example), was generated for annotation purposes. Such three-view videos give the raters a better

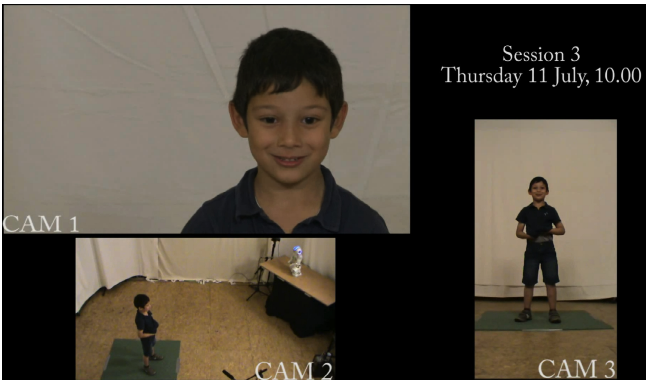

Fig. 2. The synchronised three-view videos (with audio) for annotation purpose

perception of the interaction, hence a more reliable annotation. Moreover, the raters had the possibility to preview the videos as well as repeat the annotation as many times as required. Fig.3 depicts some annotations.

4 Feature Design and Recognition Model

4.1 Feature Extraction

Psychological experiments and statistical analysis conducted in [35] revealed some general relations between the bodily expressive patterns and the emotions. Although this work used the categorical emotion labels, it actually inspired us to design the feature set.

From the 3D skeletal representation, we extracted both low-level postural features and high-level kinematic and geometrical features. Human motions could be thought of as being composed of different physical segments. Each segment can move independently and exhibit an independent degree of activity [1]. These body segments have a hierarchical structure. For instance, the upper body consists of two arms, head, neck and torso. And the left arm is further composed of left hand, left lower arm and left upper arm. [35] has shown that the upper body, especially arms and head, plays the most important role for emotional expressions. Therefore, we calculate the spatial distances among hands, elbows and shoulders in each of the three dimensions, as well as the angles of the two elbows and the angles between the spine and the upper arms. Moreover, we also calculate the distance between the feet, the orientation of the feet, and the orientation of the shoulders. All these lead to 28 postural features in total.

As for the high-level features, they are designed to represent the abstract characteristics of bodily expressions, such as movement power, body spatial extension, head bending etc. in [35]:

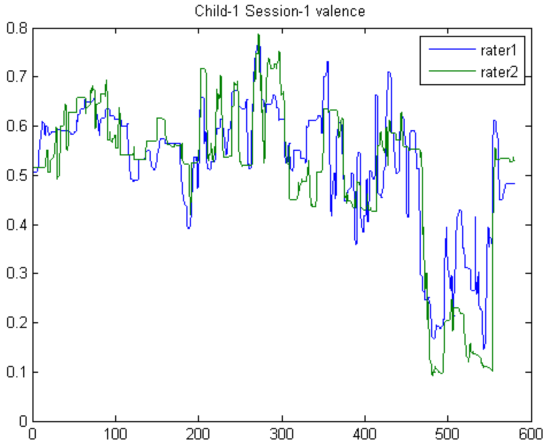

Fig. 3. Example of annotations made by two raters

- *Body movement activity and power:* Using the above described hierarchical structure, along with the mass of each body segment, estimated using the ergonomic definitions of [18], we further calculate the force, kinetic energy and momentum of the movements in a hierarchical manner (from the smallest segment to the whole body):

$$Force_{segment} = Mass_{segment} \times Acceleration_{segment} \tag{1}$$

$$KineticEnergy_{segment} = 0.5 \times Mass_{segment} \times Velocity^2_{segment} \tag{2}$$

$$Momentum_{segment} = Mass_{segment} \times Velocity_{segment} \tag{3}$$

- *Body spatial extension:* From the positions of the body joints, we calculate the bounding box of the whole body. The spatial extension is calculated by considering the length proportions of the edges, i.e. $\frac{x}{y}$, $\frac{x}{z}$ and $\frac{y}{z}$.

- *Symmetry index*: We calculate symmetry/asymmetry index in x, y, z axis, respectively, based on the positions of two hands. These features highlight the importance of the hands behaviours in emotional expressions.

In total, ten high-level features are extracted. Together with the postural features, we obtain a per-frame-based feature vector of the dimension 38. The features calculation is very computationally efficient and could be done in real-time, provided the skeleton stream.

4.2 Recognition Model

Our prediction task could be abstracted as a time-series regression problem, with the following requirements:

- As demonstrated in [16], each bodily expression consists of different temporal phases, and the states are dependent on the previous ones. Thus the prediction model has to be capable of dealing with the temporal memories.
- Due to our objective to predict effective values continuously in time, sequence segmentation is undesired. This leads to a huge amount of frame-based data to be handled. Therefore, an on-line learning algorithm, instead of batch-based algorithms, is preferable, considering practical issues such as memory and computational capability. Moreover, the learning model has to be able to select the most informative data and discard the redundant ones.
- As both skeleton data and annotations are noisy, the learning model should be tolerant to noise.

Bearing these in mind, we decided to use Gaussian Processes (GP), a kernel-based non-parametric algorithm that achieved great success in time series prediction problems. Specifically, the recursive kernel is used to model the temporal dynamics. In the following, we give a brief description of the GP algorithm, more details could be found in [33].

4.3 Online Recursive Gaussian Processes

A GP is a stochastic process which can be fully determined by its mean function

$$\mu(\mathbf{x}) = \mathbb{E}[Y(\mathbf{x})] \tag{4}$$

and its covariance function

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(Y(\mathbf{x}) - \mu(\mathbf{x}))(Y(\mathbf{x}') - \mu(\mathbf{x}'))] \tag{5}$$

where $\mathbf{x} \in \mathbf{X}$ is the input vector, $Y(\mathbf{x})$ is the random function on $\mathbf{x}$. Normally we assume that $\mu(\mathbf{x}) \equiv 0$, so the GP is only specified by the covariance function $k(\mathbf{x}, \mathbf{x}')$, which has a kernel form. We can then write the GP as:

$$Y(\mathbf{x}) \sim \mathcal{N}(0, k(\mathbf{x}, \mathbf{x}')) \tag{6}$$

Given the training samples $(\mathbf{x}_i, y_i) \in \mathcal{D}$, where y_i is the target value at data point i, the matrix of covariances between the training points $\mathbf{K} = [k(\mathbf{x}_i, \mathbf{x}_j)]$ is called *Gram Matrix*. We also define $\mathbf{k}(\mathbf{x}') = [k(\mathbf{x}_i, \mathbf{x}')]_{i=1}^{N}$, N being the number of training samples. Then, for a new input data point $\mathbf{x}^\star$, the distribution of the prediction is:

$$p(Y^\star | \mathbf{x}^\star, \mathcal{D}) \sim \mathcal{N}(Y^\star | \mu^\star, \sigma^{\star 2}) \tag{7}$$

where

$$\mu^\star = \mathbf{k}(\mathbf{x}^\star)^T (\mathbf{K} + \sigma^2 \mathbf{I}_N)^{-1} \mathbf{y} \tag{8}$$

$$\sigma^{\star 2} = k(\mathbf{x}^\star, \mathbf{x}^\star) - \mathbf{k}(\mathbf{x}^\star)^T (\mathbf{K} + \sigma^2 \mathbf{I}_N)^{-1} \mathbf{k}(\mathbf{x}^\star) \tag{9}$$

The variance of the prediction $\sigma^{\star 2}$ could be used as the uncertainty measure.

In order to update the GP with sequentially arriving data points, [8] proposed a sparse on-line GP algorithm. The main idea is to keep the size of the model by controlling the number of data points that are used for prediction. Those remained data points are called "*Basic Vectors*". Each sample is scored by a "novelty" measure:

$$\gamma(\mathbf{x}^\star) = k(\mathbf{x}^\star, \mathbf{x}^\star) - \mathbf{k}_{\mathcal{B}}^{\star T}\mathbf{K}_{\mathcal{B}}^{-1}\mathbf{k}_{\mathcal{B}}^{\star} \tag{10}$$

where $\mathbf{k}_{\mathcal{B}}^{\star T} = [k(\mathbf{b}_i, \mathbf{x}^\star)]$ and $\mathbf{K}_{\mathcal{B}} = [k(\mathbf{b}_i, \mathbf{b}_j)]$, with $\mathbf{b}_i, \mathbf{b}_j \in \mathcal{B}$, the basic vectors. The highly scored new sample will be absorbed in the set of basic vectors, while the lowest scored one will be discarded from the set. The number of basic vectors, as a global hyperparameter, plays the role to balance the prediction strength and the computational efficiency, which is generally determined by the calculation capacity. Refer to [8] for more details.

To cope with the temporal dynamics in a systematic way, a recursive kernel is applied on the GP, to form a recursive GP [33]. For the widely used squared exponential kernel:

$$k_{SE}(\mathbf{x}, \mathbf{x}') = exp(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2l^2}) \tag{11}$$

the corresponding recursive version is:

$$\kappa^{(t)}(\mathbf{x}, \mathbf{x}') = exp(-\frac{||\mathbf{x}^{(t)} - \mathbf{x}'^{(t)}||^2}{\sigma^2})exp(\frac{(\kappa^{(t-1)}(\mathbf{x}, \mathbf{x}') - 1)}{\rho^2}) \tag{12}$$

5 Experimental Results

The 3D skeletal frames, annotations (arousal and valence separately), as well as the video recordings of face and frontal body, were synchronised and temporally aligned to have the same frame rate of $30Hz$. The feature vectors, as described in section 4.1 were generated from the skeletal frames, on a frame basis, and all values of the features were normalised to have the same order of magnitude.

Firstly, we evaluated the emotion prediction model with same-subject sequence. The model was trained with a full recording sequence, using the online updating. Then the same sequence was tested on the model. The size of the GP *basic vector* was set to 300 (maximally 300 samples in the training set were kept in the model for prediction calculation). The results are shown in Fig. 4.

As it can be seen, both the arousal and valence dimensions were well predicted, with the 300 training samples stored in the GP. This result proved the effectiveness of the proposed features to describe bodily expressions.

To evaluate the generalization ability of our approach, we applied the trained model on a sequence performed by another child. Fig. 5 illustrates the obtained results. For the arousal dimension, as shown in the first sub-figure, the trends were well followed. The vertical shifts could be explained by the slightly different scales used for the annotations of the two sequences. As for the valence dimension, the result was less good compared to the arousal predictions. There were several opposite predictions. For instance, around point D in the figure, the annotated negative expression was predicted as positive. However, if we review the

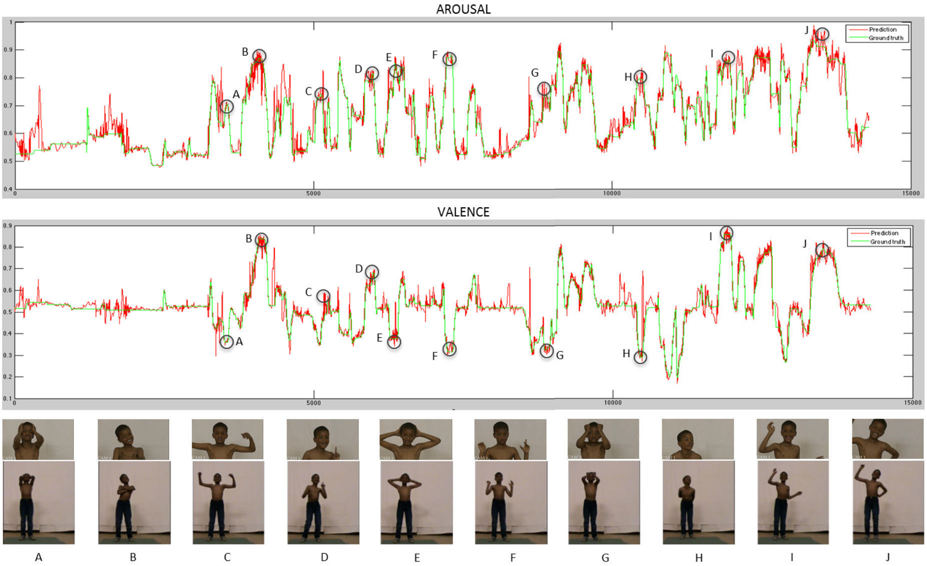

Fig. 4. Same-subject prediction. The solid and dashed curves are the prediction and the annotation, respectively. The first and second sub-figures are arousal and valence dimension, respectively. Ten frames (A-J) are marked, the corresponding facial and bodily expressions are given in the third sub-figure.

recorded videos, the negative emotion was actually delivered via the facial and vocal expressions, while the "jump and turn-around" body motion alone could be interpreted as a positive expression, that happened occasionally in other recordings. Another interesting pattern in both the arousal and valence prediction is that we can see a clear lag between the ground-truth and the prediction. This is due to the delay of the annotation (normally less than one second) reported by the raters. This delay had been compensated by the model during the prediction, which is a merit in real-time applications, as child-robot interaction.

6 Discussion

In this work, we present our initial attempt to recognize children's affective states from their spontaneous bodily expressions, in child-robot interactions. The predicted child's arousal and valence values, are used by the robot's behavioural control system, so as to achieve a more natural and comfortable interaction. We designed an emotion elicitation scenario. The preliminary experiments have shown encouraging results for both arousal and valence predictions from the stand-alone bodily cues, which has been widely recognised as a very challenging problem, especially under naturalistic settings. Moreover, our experiments further demonstrated the importance of bodily information in emotion modelling tasks. For example, position E in Fig. 5, where the face shows a positive

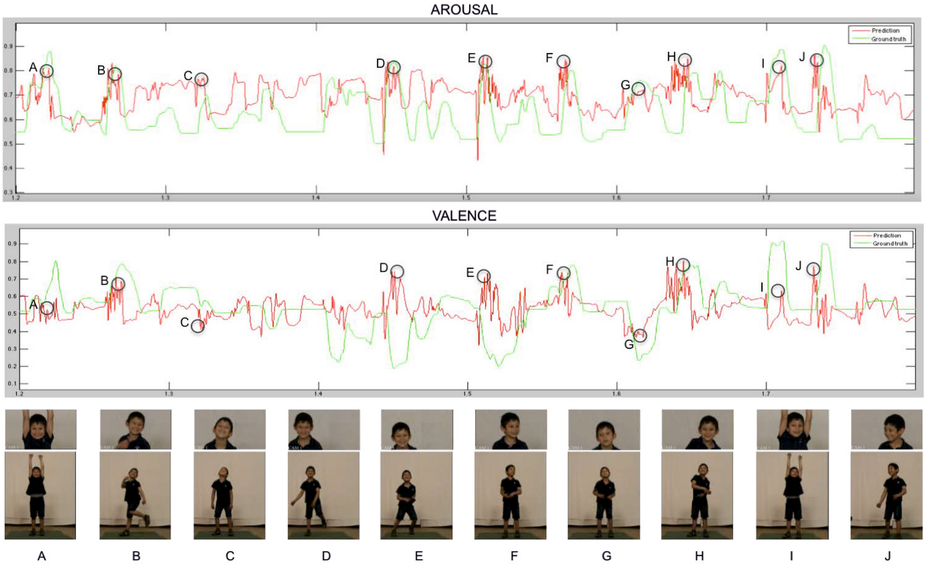

Fig. 5. Cross-subject prediction. The layout is the same as in Fig. 4.

expression that is similar to the one at position H, while the body displays a strongly negative state, which is consistent with the interaction circumstance at that moment.

Similar to the conclusion in [23], the valence dimension is much more difficult to model, as shown in our preliminary results. Therefore, in future work we planed to fuse different modalities including bodily, facial and vocal/verbal signals. Additionally, considering the fact that spontaneous emotion annotation is a very subjective task that has strong dependency on individual's perception, modelling the emotional changes instead of the absolute values might be more practical. Last but not least, a more sophisticated sample selection algorithm, could benefit the predictive performance, by keeping the most informative and contributing data frames in the model.

Acknowledgments. This work has been made under the EU-FP7 ALIZ-E Project (No. 248116), also supported by the CSC grant 2011688012. We also would like to thank the members of Beeldenstorm and KIK project in Brussels, Belgium, who provided us the opportunity to implement our data acquisition scenario. We also thank the participated children and their parents.

References

1. Aggarwal, J., Cai, Q.: Human motion analysis: a review. In: Proc. of Nonrigid and Articulated Motion Workshop, pp. 90–102 (1997)
2. Alaerts, K., Nackaerts, E., Meyns, P., Swinnen, S.P., Wenderoth, N.: Action and emotion recognition from point light displays: an investigation of gender differences. PLoS ONE **6**(6), e20989 (2011)
3. Atkinson, A., Dittrich, W., Gemmell, A., Young, A.: Emotion perception from dynamic and static body expressions in point-light and full-light displays. Perception **33**(6), 717–746 (2004)
4. Baron-Cohen, S., Tead, T.: Mind Reading: the Interactive Guide to Emotions. Jessica Kingsley Publishers Ltd. (2003)
5. Beck, A., Stevens, B., Bard, K., Canamero, L.: Emotional Body Language Displayed by Artificial Agents. ACM Transactions on Interactive Intelligent Systems **2**(1), 1–29 (2012). Special issue on Affective Interaction in Natural Environments
6. Bernhardt, D.: Emotion inference from human body motion. Tech. Rep. 787, Computer Laboratory, University of Cambridge, Cambridge (2010)
7. Bianchi-Berthouze, N., Kleinsmith, A.: A categorical approach to affective gesture recognition. Connection Science **15**(4), 259–269 (2003)
8. Csato, L., Opper, M.: Sparse On-line Gaussian Processes. Neural Computation **14**(3), 641–668 (2002)
9. De Silva, P., Osano, M., Marasinghe, A., Madurapperuma, A.: Towards recognizing emotion with affective dimensions through body gestures. In: Proceedings of 7th International Conference on Automatic Face and Gesture Recognition (FG 2006), pp. 269–274. IEEE (2006)
10. Dittrich, W., Troscianko, T., Lea, S., Morgan, D.: Perception of emotion from dynamic point-light displays represented in dance. Perception **25**(6), 727–738 (1996)
11. Ekman, P.: Basic emotions. In: Handbook of Cognition and Emotion, chap. 3. No. 1992 (1999)
12. de Gelder, B.: Why bodies? Twelve reasons for including bodily expressions in affective neuroscience. Philosophical Tran. of the Royal Society B: Biological Sciences **364**, 3475–3484 (2009)
13. Gonzalez, I., Sahli, H., Enescu, V., Verhelst, W.: Context-independent facial action unit recognition using shape and gabor phase information. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011, Part I. LNCS, vol. 6974, pp. 548–557. Springer, Heidelberg (2011)
14. Gross, J.J., Levenson, R.W.: Emotion Elicitation Using Films. Cognition and Emotion **9**(1), 87–108 (1995)
15. Gross, M.M., Crane, E.A., Fredrickson, B.L.: Methodology for Assessing Bodily Expression of Emotion. Journal of Nonverbal Behavior **34**(4), 223–248 (2010)
16. Gunes, H., Piccardi, M.: Automatic temporal segment detection and affect recognition from face and body display. IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics: A Publication of the IEEE Systems, Man, and Cybernetics Society **39**(1), 64–84 (2009)
17. Jiang, D., Cui, Y., Zhang, X., Fan, P., Ganzalez, I., Sahli, H.: Audio visual emotion recognition based on triple-stream dynamic bayesian network models. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011, Part I. LNCS, vol. 6974, pp. 609–618. Springer, Heidelberg (2011)
18. Kahol, K., Tripathi, P., Panchanathan, S.: Gesture segmentation in complex motion sequences. In: Proc. of International Conference on Image Processing (ICIP 2003) (2003)

19. Kapur, A., Kapur, A., Virji-Babul, N., Tzanetakis, G., Driessen, P.F.: Gesture-based affective computing on motion capture data. In: Tao, J., Tan, T., Picard, R.W. (eds.) ACII 2005. LNCS, vol. 3784, pp. 1–7. Springer, Heidelberg (2005)
20. Kleinsmith, A., Bianchi-Berthouze, N.: Affective Body Expression Perception and Recognition: A Survey. IEEE Transactions on Affective Computing **4**(1), 15–33 (2013)
21. Larsen, J.T., McGraw, A.P.: Further evidence for mixed emotions. Journal of Personality and Social Psychology **100**(6), 1095–1110 (2011)
22. Mckeown, G., Valstar, M., Cowie, R., Pantic, M., Member, S., Schr, M.: The SEMAINE database: annotated multimodal records of emotionally coloured conversations between a person and a limited agent. IEEE Transactions on Affective Computing **3**(1), 5–17 (2012)
23. Metallinou, A., Katsamanis, A., Narayanan, S.: Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. Image and Vision Computing, September 2012
24. N/A: Aldebaran Robotics. http://www.aldebaran.com
25. N/A: Ipi Mocap Studio. http://ipisoft.com/
26. Picard, R.: Affective computing. Tech. Rep. 321, MIT (1995)
27. Preston, S., de Waal, F.: Empathy: Its Ultimate and Proximate. Behavioral and Brian Sciences **252**, 1–72 (2002)
28. Roberts, N.A., Tsai, J.L., Coan, J.A.: Emotion elicitation using dyadic interaction tasks. In: Handbook of Emotion Elicitation and Assessment, pp. 106–123 (2007)
29. Russell, J.A.: A Circumplex Model of Affect. Journal of Personality & Social Psychology **39**, 1161–1178 (1980)
30. Russell, J.A.: Core affect and the psychological construction of emotion. Psychological Review **110**(1), 145–172 (2003)
31. Scherer, K.R.: What Are Emotions? And How Can They Be Measured. Social Science Information **44**(4), 695–729 (2005)
32. Schindler, K., Van Gool, L., de Gelder, B.: Recognizing emotions expressed by body pose: a biologically inspired neural model. Neural Networks: The Official Journal of the International Neural Network Society **21**(9), 1238–1246 (2008)
33. Soh, H.: Online spatio-temporal gaussian process experts with application to tactile classification. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (2012)
34. Verhelst, W., Roelands, M.: An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech. In: ICASSP 1993, vol. 2, pp. 554–557 (1993)
35. Wallbott, H.G.: Bodily Expression of Emotion. European Journal of Social Psychology **28**(6), 879–896 (1998)
36. Wang, F., Verhelst, W., Sahli, H.: Relevance vector machine based speech emotion recognition. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011, Part II. LNCS, vol. 6975, pp. 111–120. Springer, Heidelberg (2011)
37. Wang, W., Athanasopoulos, G., Yilmazyildiz, S., Patsis, G., Enescu, V., Sahli, H., Verhelst, W., Hiolle, A., Lewis, M., Cañamero, L.: Natural emotion elicitation for emotion modeling in child-robot interactions. In: Proc. of Workshop on Child Computer Interaction (WOCCI 2014) (2014, to appear)
38. Wang, W., Enescu, V., Sahli, H.: Towards real-time continuous emotion recognition from body movements. In: Salah, A.A., Hung, H., Aran, O., Gunes, H. (eds.) HBU 2013. LNCS, vol. 8212, pp. 235–245. Springer, Heidelberg (2013)