A robotic edutainment framework for designing child-robot interaction scenarios

Niki Efthymiou
nefthymiou@central.ntua.gr
School of ECE, National Technical University of Athens
Athens, Greece

Panagiotis P. Filntisis
filby@central.ntua.gr
School of ECE, National Technical University of Athens
Athens, Greece

Gerasimos Potamianos
gpotam@ieee.org
Department of ECE, University of Thessaly
Volos, Greece

Petros Maragos
maragos@cs.ntua.gr
School of ECE, National Technical University of Athens
Athens, Greece

ABSTRACT

This paper presents the development of a child-robot interaction (CRI) system for edutainment scenarios, aiming to provide a framework for their design and to simplify access to social robots by educators with non-specialized technical knowledge in this challenging area. Our framework incorporates powerful robotic perception modules for action and emotion recognition of the interacting child, allowing the robot to exhibit empathy and be informed of the child's activity. Both developed modules are evaluated on respective datasets, outperforming the current state-of-the-art by a significant margin, while retaining low computational cost. The modules are complemented by off-the-shelf automatic speech recognition and synthesis components to further enable and enrich edutainmment-focused CRI. Moreover, the developed framework allows custom CRI scenario-building via a suitable graphical user interface, providing a valuable asset to educators wishing to utilize social robots in the classroom.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; • **Computing methodologies** → **Neural networks**; **Vision for robotics**.

KEYWORDS

Child-Robot Interaction, Robotic Framework for Edutainment, Emotion Recognition, Action Recognition

ACM Reference Format:
Niki Efthymiou, Panagiotis P. Filntisis, Gerasimos Potamianos, and Petros Maragos. 2021. A robotic edutainment framework for designing child-robot interaction scenarios. In *The 14th PErvasive Technologies Related to Assistive Environments Conference (PETRA 2021), June 29-July 2, 2021, Corfu, Greece.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3453892.3458048

PETRA 2021, June 29-July 2, 2021, Corfu, Greece

ACM ISBN 978-1-4503-8792-7/21/06…$15.00
https://doi.org/10.1145/3453892.3458048

1 INTRODUCTION

Child-Robot Interaction (CRI) is an interdisciplinary field that has attracted much scientific interest recently due to the increasing use of robotic agents in everyday life. Many studies in this field focus on how children's mental and cognitive development is affected by CRI [4, 6, 12]. Among the vast number of technologies that have been developed in the last decades for educational and edutainment purposes, social robots stand out due to the wide range of applications they can be used for, e.g., learning handwriting [10], a second language [29], or even social emotional learning [48].

It has been observed that in classrooms where student-centered learning is encouraged, robotic agents create more pleasant learning and motivate children to participate more [27]. When robots encourage students to interact with them, they achieve to arouse their curiosity [26]. Indeed, numerous studies focus on topics related to the conditions under which the integration and use of robotic agents in the classroom can be beneficial to children and, as a result, the interaction between them is designed with specific purposes in mind. Such research goals are the study of children engagement during CRI, the level of the joint attention achieved [17, 30], the analysis of parameters crucial to long-term interaction [12], as well as research for finding appropriate ways to adjust the curriculum on such interaction and the difficulties faced [41].

Most robotic agents used in such studies are semi-autonomous or tele-operated, applying the Wizard-of-Oz technique [3]. More recently though, due to the dramatic advancements in machine learning techniques and neural networks, increasingly more social robots have integrated intelligent perception systems that can interact with humans in a more natural way [9]. Thus, their use by non-experts, e.g., educators or therapists, is expected to increase, and sectors such as education to benefit from this progress.

Motivated by the above, this paper focuses on developing a novel CRI system for edutainment scenarios, aiming to provide a framework for their design and to simplify access to social robots by educators with non-specialized technical knowledge in this challenging area. Specifically, the proposed system:

- Incorporates suitably developed and powerful deep neural-network based architectures for perceiving children actions and decoding their emotional state robustly through visual information, which are evaluated on two children databases where they are shown to outperform the current state-of-the-art at a low computational cost;

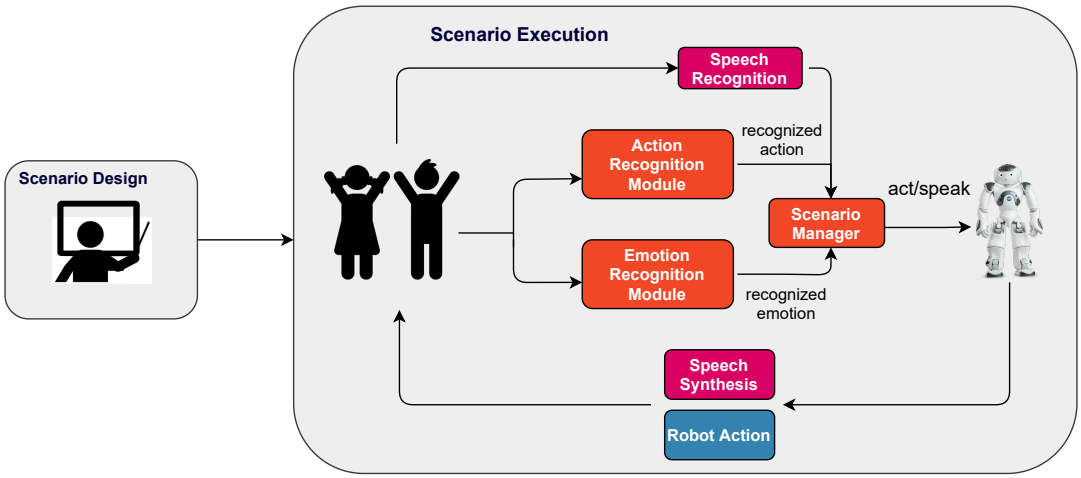

Figure 1: The proposed CRI edutainment framework. Modules shown in orange color are developed in this work, while ones in pink are off-the-shelf (robot action is already available in the NAO robot).

- Integrates off-the-self speech recognition and synthesis modules to further allow pleasant and natural interaction between children and robot through the speech modality as well;
- Provides a novel and user-friendly graphical user interface, designed with educators in mind, allowing them to adjust existing CRI scenarios, or to create custom ones according to the curriculum requirements.

An overview of the proposed framework is depicted in Fig. 1, while in Fig. 2 an example of the developed system is shown set up inside a classroom.

The remainder of this paper is structured as follows: Section 2 reviews related work; Section 3 presents the developed system, detailing its components; Section 4 is dedicated to the evaluation of its perception modules and, finally, Section 5 concludes the work.

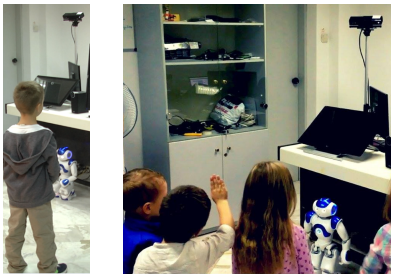

Figure 2: Examples of the developed CRI system set up inside a classroom with children interacting with it.

2 RELATED WORK

Recent works on CRI, from an engineering perspective, have focused on creating complete robotic systems [13, 16, 36], mainly for therapeutic purposes, that can be installed in purpose-built rooms and used by experts or people acquainted with technical issues that might arise. Concerning systems employed in school environments, the work of Shiomi et al. [42] is noteworthy for using a robotic agent, called Robovie, in an elementary school. In that paper, the authors placed the social robot in a science class where the students could interact freely by asking questions about science during their breaks, and Robovie encouraged them to discuss. Even though Robovie was tele-operated and only incorporated basic systems for gesture and speech recognition, it is one of the very few studies that included robots in the educational process in an unconstrained manner. A more recent work by Levinson et al. [33] compared two social robots, NAO [38] and a 3d printed puppet-like robot, during a learning task at a Summer camp. The children, aged five to nine years old, participated in morphology-related activities to groups of up to nine children over the span of a three-week session. The robotic agents didn't include any automatic recognition modules, but only recited scripts and performed movements designed to fit children responses, which the authors concluded to be detrimental to the interaction.

During human-robot interaction, recognizing the human emotional state can play a significant role in developing robotic agents with empathy [11, 25]. Empathy allows a robot to perceive and decode movements and expressions containing information about the emotional state. Consequently, the robotic agents can change and adapt their behavior and actions towards the user appropriately. For example, the Pepper robot is capable of basic emotional analysis [40]. This results in establishing a healthy and long-term interaction and trust relationship [5]. Not surprisingly, the efficacy of using robotic agents with empathy, especially in the education sector, has been studied, concluding that the social robot behavior agreement with the behavioral state of the human has a positive impact on their relationship [32].

Generalizing perception components developed for adults to children is a challenging task, due to the fact that children behavior and natural characteristics, e.g., voice pitch and height, differ from that of adults. This fact necessitates the development of perception components specifically for children [13]. In the case of emotion

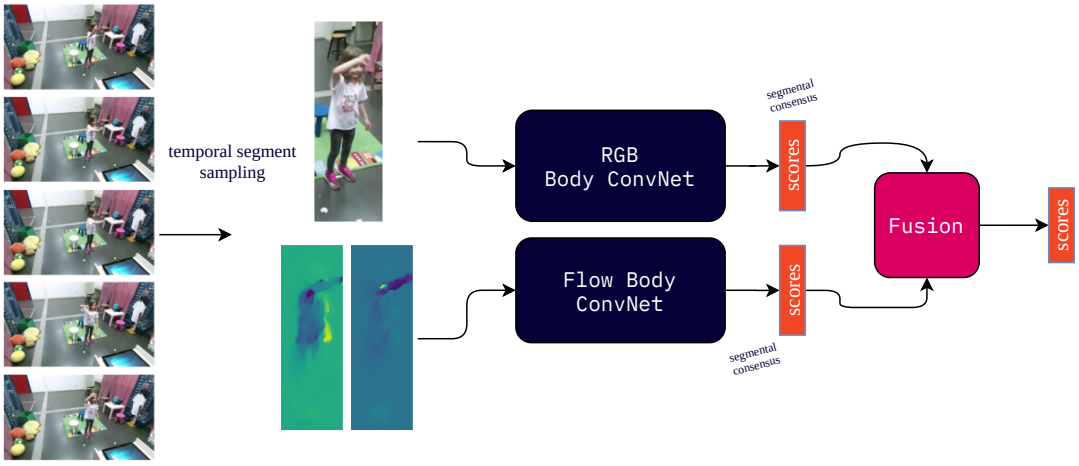

Figure 3: The TSN framework used for action and emotion recognition in the proposed system.

recognition, Goulart et al. propose in [22] a computational system for estimating children emotion during CRI by exploiting visual information from both RGB and infrared thermal cameras. Their system detects facial regions of interest that are relevant to five basic emotions. In other work on this topic, Lopez-Rincon in [34] proposes a Convolutional Neural Network (CNN) combined with a Viola-Jones face detector, trained using the AffectNet database [37] and tuned on the NIMH-ChEF dataset [15] to classify children facial expressions into six basic emotions.

Human action recognition is among the most popular computer vision problems, however action recognition for children has been examined on a small only scale. In our previous work [14], various feature extraction approaches, encoding methods, and fusion techniques were explored, resulting in a robust multi-view fusion action recognition system focused on CRI scenarios. In other work, Marinou et al. [35] proposed an automated approach using 3d skeleton data and a CNN architecture during robot-assisted therapy sessions of children with Autism Spectrum Disorders (ASD). Finally, Zhang et al. [49] also focused on ASD children, proposing an LSTM-based network fed with the extracted children skeleton by the OpenPose algorithm [7] after a denoising filter.

Clearly, various fields of study converge on CRI, broadening the research problems and applications. Some works close to our focus on developing a robotic edutainment system for CRI have been presented above to highlight some of the challenges and state-of-the-art techniques. Undoubtedly, different approaches could also be proposed, for example viewing the action recognition problem as an abnormal event recognition task [31, 44]. However, our focus here lies primarily on developing a robotic system for CRI in edutainment and use by non-expert stakeholders, such as educators.

3 ROBOTIC SYSTEM

The proposed system consists of a NAO robot, a compact camera with microphones such as Kinect in order to be portable and lightweight, and the developed framework. The latter is designed with modularity in mind and contains three primary modules developed in this work and two off-the-self-components that provide necessary additional functionality and are based on existing solutions. The first two primary modules of our framework pertain to the core perception capabilities of the robot: 1) The Action Recognition module, which, as the name suggests, has the responsibility of recognizing the child's actions, and the 2) Emotion Recognition module, which decodes the affective state of the child. The third module is the Scenario Manager that facilitates the educator to design new edutainment scenarios using flowcharts. Alongside these, Speech Recognition and Speech Synthesis modules are also employed, based on existing solutions.

3.1 Perception Modules

Action Recognition Module. This is based on the Temporal Segment Network (TSN) framework [47], originally introduced for large-scale action recognition. Under this framework, K different segments are randomly sampled from the input video, each consisting of N consecutive frames. Such random sampling helps generalization and reduces the computational cost and redundant information that exists in sequential video frames.

The developed action recognition module architecture can be seen in Fig. 3. Two different streams are used, one spatial that takes as input RGB video frames and one temporal that takes as input the optical flow derived from the video. In order to force the networks to focus on the child and its actions, pose tracking is first performed on the input video, and then the region around the child is cropped during temporal sampling based on the detected skeleton by OpenPose [7]. The same process is applied to the flow stream as well.

After obtaining each segment scores by feeding the data to each network, the segmental consensus function fuses them to obtain each stream predictions. Finally, the resulting predictions are combined again by weighted average fusion to yield the final action classification.

Emotion Recognition Module. For consistency and convenience, the emotion recognition module employs a similar architecture

to that of action recognition. TSNs have already been shown to achieve state-of-the-art results in video emotion recognition [18]. Under this framework, static information is leveraged across different frames to identify a child's expression and combined with the child's movement dynamics by using optical flow as input. As with action recognition, each video is cropped around the child's face after performing face detection on the input [2]. The final result is obtained in the same manner, using segmental consensus and average fusion of the different input modalities.

Off-the-shelf Speech Modules. To enable natural CRI, the robot must also understand the child's speech as well as talk back to the child. Existing cloud-based Text-to-Speech (TTS) [21] and Automatic Speech Recognition (ASR) [20] solutions are employed for this purpose. The integration of the two modules in the main system can be seen in Fig. 4. During the interaction, speech captured by the microphones is continuously streamed to a cloud service, which returns the recognition results. The text is then fed to the scenario manager (described in Section 3.2), which decides if the robot should reply something back. The text to be synthesized is then sent to a cloud TTS service, which synthesizes the speech.

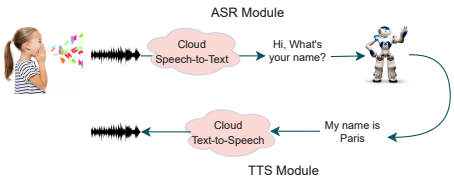

Figure 4: The ASR and TTS modules of the robotic edutainment framework.

3.2 Scenario Manager and Integration

Scenario Modeling. For managing the flow of the dialog, we adopt the "Sense, Think, Act" paradigm [19]. Under this paradigm, the robotic system first uses its perceptual capabilities (*Sense*), then decides on the next course of action (*Think*), and finally executes the chosen action (*Act*). In the proposed framework, we model the paradigm using events [43], similarly to [13]. These are divided into two categories: *Action* events, which command the robot to do something, and *Sense* events, which fire when the perception modules perceive something (i.e., a specific action or emotion). The flow of the interaction is modeled using Harel statecharts [23]. Each state in the chart has hidden parameters that control the flow, along with the received events.

Custom Scenario Design. A major novelty of the developed edutainment framework is that it offers educators the ability to design their own scenarios, using a flexible, user-friendly, and aesthetically appealing drag-and-drop graphical user interface. Via this interface, the teacher is empowered to create the desired complex scenarios by following the aforementioned principles of Scenario Modeling. Then, the graphical scenario gets compiled into a Harel statechart that models the flow and deploys it to the robot to execute it. An example of graphical scenario building can be seen in Fig. 5.

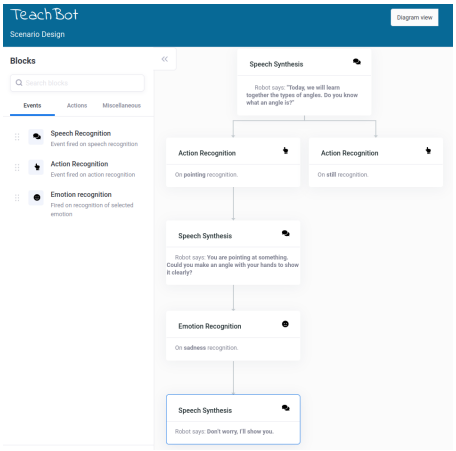

Figure 5: The graphical user interface for building custom edutainment scenarios.

Edutainment Scenario Example. Next, we present an example of an edutainment scenario that can be implemented within the proposed robotic framework, underlining the sequence of the child and robot actions. In this scenario, the robot tries to teach a child the concept of angles in mathematics.

ROBOT: Today, we will learn together the types of angles. Do you know what an angle is?

CHILD: [*Points to a corner of the room while saying*] There!

ROBOT: [*Recognizes the gesture and speech. Then responds*] You are pointing at something. Could you make an angle with your hands to show it clearly?

CHILD: I can't! [*says while expressing sadness*].

ROBOT: [*Recognizes the emotion and speech. Then says*] Don't worry, I'll show you. [*Then performs an action. The robot makes an angle with its hands.*]

Integration. The lightweight system modules communicate via a message broker, implemented using the TCP/IP protocol. More specifically, during the interaction, the perception modules send *Sense* events to the broker, which in turn transmits the events to the scenario manager. In addition, when the scenario manager demands that the robot does something, it sends an *Act* event to the corresponding module: the Speech Synthesis module or the robot itself in the case of an action. All perception modules and scenario management are deployed in a Linux machine with an RTX 2080 graphics processor unit.

4 EXPERIMENTAL RESULTS

In this section, we proceed to evaluate the proposed architectures for children action and emotion recognition. Due to the ongoing CoViD-19 pandemic, we have not yet been able to conduct a system-wide user evaluation by multiple children and teachers. As a result, we only evaluate the perception modules separately on available children data. Examples of such data are depicted in Fig. 6.

Figure 6: Example images from the BabyRobot action database (upper row) and the EmoReact dataset (lower row).

4.1 Action Recognition

The action recognition module is evaluated on the BabyRobot action database [14], that has been created during the BabyRobot project [1] and contains various interactions among children and multiple robots. The dataset includes 25 children performing actions belonging to 13 classes, while interacting in a pantomime game with a NAO robot. We present the experimental results of various explored methods compared to the state-of-the-art for single-view action recognition on this database.

We use a BNInception [45] architecture for the RGB and Flow streams and consider two different pretraining schemes: pretraining on the Kinetics [8] action recognition dataset and pretraining on the ImageNet database. We train each network for 60 epochs with stochastic gradient descent and cross-entropy loss, sampling 5 segments with length 1 for RGB and length 5 frames for the Flow stream.

We choose leave-one-out cross-validation in order to allow comparisons with our earlier work [14], and we present the results in Table 1. We can observe that pretraining both the RGB and Flow models on the Kinetics database achieves significantly higher accuracy than pretraining on ImageNet. Besides, both the Flow module and the weighted average fusion with RGB outperform the previous state-of-the-art method of Dense Trajectory Ensemble features (which includes Histogram of Oriented Gradients - HOG, Histogram of Optical Flow - HOF, and Motion Boundary Histogram - MBH features), as well as the C3D convolutional network of [14].

Regarding the computational costs for training the network, we note that one epoch of training (with a batch size of 8) using the TSN framework takes ~ 11 seconds to complete for the RGB modality and ~ 62 seconds for the Flow modality. During inference, using the RGB modality takes ~ 0.8 seconds to classify all 13 classes, while using the Flow modality takes ~ 1.5 seconds. Based on the above, for the action recognition module, we select the best model: weighted average fusion of the RGB-TSN and Flow-TSN, both pretrained on the Kinetics dataset.

4.2 Emotion Recognition

In order to evaluate the emotion recognition module on children data, we employ the EmoReact [39] dataset. The dataset contains videos of 63 children (32F, 31M) reactions to various subjects, collected from YouTube. The number of all videos across the training, validation, and test set is 1102. Each video is annotated with one or more emotions from a total of 8 emotion labels.

Table 1: Results of the action recognition module on the BabyRobot action dataset.

model	Accuracy (%)
RGB-Kinetics	47.14
RGB-ImageNet	42.75
Flow-Kinetics	74.75
Flow-ImageNet	63.49
RGB-Kinetics + Flow-Kinetics	**76.23**
RGB-ImageNet + Flow-ImageNet	64.10
Dense Traj. Ensemble [14]	74.15
C3D [14]	59.38

We train each TSN for 60 epochs, similarly to the action recognition module, with stochastic gradient descent and binary cross-entropy loss, using a batch size of 16, sampling 5 segments from each video, and selecting the epoch with the best validation area under the curve (ROC AUC). For the RGB modality, the segment length is set to 1, while for the Flow stream to 5. The CNN architecture we use is a residual network with 50 layers (ResNet50) [24]. We also consider two different pretraining methods: 1) using the standard pretrained ImageNet weights or 2) the weights of a ResNet50 trained on the most extensive facial expression dataset, AffectNet [37]. We have trained a ResNet50 on AffectNet, achieving 58.60% accuracy on the validation set (the test set is not available). Because the label distribution of AffectNet is highly skewed, we employ balanced sampling so that the network sees the underrepresented classes more often.

We also built a traditional baseline scheme by extracting for each video HOG, HOF, and MBH features, aggregating them using Fisher Vectors [28], and employing a linear SVM for classification.

The results are presented in Table 2. We observe that both ImageNet and AffectNet RGB pretrained models achieve similar performance, while the Flow network achieves a lower ROC AUC. Average fusion increases the final ROC AUC. In the same Table, we also list the current state-of-the-art result on the EmoReact database [39] and the non-deep learning baseline, showing that our method achieves better emotion recognition performance.

The computational burden of the emotion recognition module using RGB is ~ 40 seconds per epoch of training and ~ 26 seconds for inference, while using the Flow modality ~ 166 seconds per epoch of training and ~ 115 for inference. Considering that the system needs to be lightweight, for the deployment of the emotion recognition module, we have selected only the RGB-TSN pretrained on the AffectNet dataset, since the Flow-TSN not only demands significantly larger computational resources but also has a minuscule effect on performance.

4.3 Speech Recognition

As a final experiment, we investigated the performance of the Google Cloud Speech-To-Text Engine as the ASR module of our

Table 2: Results of the emotion recognition module on the multi-label EmoReact dataset.

model	ROC AUC
RGB-AffectNet	0.648
RGB-ImageNet	0.636
Flow-ImageNet	0.588
RGB-ImageNet + Flow-ImageNet	0.650
RGB-AffectNet + Flow-ImageNet	**0.652**
OpenFace with SVM [39]	0.620
HOG, HOF, MBH Ensemble with SVM	0.600

system. For this purpose, we used the BabyRobot Distant Speech Recognition dataset [46], which includes speech from 25 children while interacting with multiple robots. We randomly selected 5 utterances from each child (resulting in a total of 125 utterances) from the dataset and used the Speech-to-Text Engine to recognize the sentences, achieving a very promising Word Error Rate of 25%.

5 CONCLUSION

In this paper, we proposed a novel lightweight edutainment robotic framework for CRI. The developed framework aims to become a valuable educational tool by easily allowing teachers to integrate robots into the education process without requiring specialized technical knowledge. This is accomplished through a novel and flexible graphical user interface that offers easy creation of the desired CRI scenarios.

In addition, the system employs two deep neural network based perception modules to recognize actions by children, as well as their emotions. We have evaluated the proposed modules on databases of children performing actions and emotions, respectively, and showed that they achieve high performance, surpassing the current state-of-the-art. Based on our experimental analysis, and taking into account the need to balance computational cost and performance, we have selected the RGB-TSN pretrained on the AffectNet dataset as the architecture of the emotion recognition module, and the fusion of Flow-TSN and RGB-TSN, both pretrained on the Kinetics dataset, for the action recognition module.

In the future, we aim to improve the scenario design module further, as well as to conduct extensive system-wide evaluations with numerous children and multiple use-cases customized by educators. Such evaluations are unfortunately currently not possible due to the CoViD-19 pandemic.

ACKNOWLEDGMENTS

This research is carried out/funded in the context of the project “Intelligent Child-Robot Interaction System for designing and implementing edutainment scenarios with emphasis on visual information" (MIS 5049533) under the call for proposals “Researchers’ support with an emphasis on young researchers- 2nd Cycle”. The project is co-financed by Greece and the European Union (European Social Fund- ESF) by the Operational Programme Human Resources Development, Education and Lifelong Learning 2014-2020.

REFERENCES

[1] BabyRobot project 2019. http://babyrobot.eu.
[2] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *Proc. FG*.
[3] T. Belpaeme. 2020. The Wizard is Dead, Long live Data: towards Autonomous Social Behaviour using Data-driven Methods. In *Companion Publication of ICMI*.
[4] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka. 2018. Social robots for education: A review. *Science Robotics* 3 (2018).
[5] T. W. Bickmore and R. W. Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions On Computer-Human Interaction* 12 (2005), 293–327.
[6] L. Boccanfuso, E. Barney, C. Foster, Y. A. Ahn, K. Chawarska, B. Scassellati, and F. Shic. 2016. Emotional robot to examine different play patterns and affective responses of children with and without ASD. In *Proc. HRI*.
[7] Z. Cao, T. Simon, S. Wei, and Y. Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proc. CVPR*.
[8] J. Carreira and A. Zisserman. 2017. Quo vadis, action recognition? a new model and the Kinetics dataset. In *Proc. CVPR*.
[9] G. Chalvatzaki, P. Koutras, A. Tsiami, C. S. Tzafestas, and P. Maragos. 2020. i-Walk intelligent assessment system: activity, mobility, intention, communication. In *Proc. ECCVW ACVR-2020*.
[10] S. Chandra, P. Dillenbourg, and A. Paiva. 2019. Children teach handwriting to a social robot with different learning competencies. *International Journal of Social Robotics* (2019), 1–28.
[11] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. 2001. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine* 18 (2001), 32–80.
[12] D. P. Davison, F. M. Wijnen, V. Charisi, J. van der Meij, V. Evers, and D. Reidsma. 2020. Working with a social robot in school: a long-term real-world unsupervised deployment. In *Proc. HRI*.
[13] N. Efthymiou, P. P. Filntisis, P. Koutras, A. Tsiami, J. Hadfield, G. Potamianos, and P. Maragos. 2020. ChildBot: multi-robot perception and interaction with children. *arXiv preprint arXiv:2008.12818* (2020).
[14] N. Efthymiou, P. Koutras, P. P. Filntisis, G. Potamianos, and P. Maragos. 2018. Multi-view fusion for action recognition in child-robot interaction. In *Proc. ICIP*.
[15] H. L. Egger, D. S. Pine, E. Nelson, E. Leibenluft, M. Ernst, K. E. Towbin, and A. Angold. 2011. The NIMH Child Emotional Faces Picture Set (NIMH-ChEFS): a new set of children’s facial emotion stimuli. *International Journal of Methods in Psychiatric Research* 20 (2011), 145–156.
[16] P. G. Esteban, P. Baxter, T. Belpaeme, E. Billing, H. Cai, H. L. Cao, M. Coeckelbergh, C. Costescu, D. David, A. De Beir, et al. 2017. How to build a supervised autonomous system for robot-enhanced therapy for children with autism spectrum disorder. *Paladyn, Journal of Behavioral Robotics* 8 (2017), 18–38.
[17] C. Filippini, E. Spadolini, D. Cardone, D. Bianchi, M. Preziuso, C. Sciarretta, V. del Cimmuto, D. Lisciani, and A. Merla. 2020. Facilitating the child–robot interaction by endowing the robot with the capability of understanding the child engagement: the case of Mio Amico robot. *International Journal of Social Robotics* (2020), 1–13.
[18] P. P. Filntisis, N. Efthymiou, G. Potamianos, and P. Maragos. 2020. Emotion understanding in videos through body, context, and visual-semantic embedding loss. In *Proc. ECCV*.
[19] E. Gat. 1998. On three-layer architectures. *Artificial Intelligence and Mobile Robots* (1998), 195–210.
[20] Google Speech-to-Text 2021. https://cloud.google.com/speech-to-text.
[21] Google Text-to-Speech 2021. https://cloud.google.com/text-to-speech.
[22] C. Goulart, C. Valadão, D. Delisle-Rodriguez, D. Funayama, A. Favarato, G. Baldo, V. Binotte, E. Caldeira, and T. Bastos-Filho. 2019. Visual and thermal image processing for facial specific landmark detection to infer emotions in a child-robot interaction. *Sensors* 19 (2019), 2844.
[23] D. Harel. 1987. Statecharts: A visual formalism for complex systems. *Science of Computer Programming* 8 (1987), 231–274.
[24] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *Proc. CVPR*.
[25] K. Hone. 2006. Empathic agents to reduce user frustration: The effects of varying agent characteristics. *Interacting with Computers* 18 (2006), 227–245.
[26] T. Kanda, R. Sato, N. Saiwaki, and H. Ishiguro. 2007. A two-month field trial in an elementary school for long-term human–robot interaction. *IEEE Transactions on Robotics* 23 (2007), 962–971.
[27] T. Kanda, M. Shimada, and S. Koizumi. 2012. Children learning with a social robot. In *Proc. HRI*.
[28] V. Kantorov and I. Laptev. 2014. Efficient feature extraction, encoding and classification for action recognition. In *Proc. CVPR*.
[29] J. Kennedy, P. Baxter, E. Senft, and T. Belpaeme. 2016. Social robot tutoring for child second language learning. In *Proc. HRI*.
[30] T. Komatsubara, M. Shiomi, T. Kaczmarek, T. Kanda, and H. Ishiguro. 2019. Estimating children’s social status through their interaction activities in classrooms with a social robot. *International Journal of Social Robotics* 11 (2019), 35–48.

[31] S. Lee, H. G. Kim, and Y. M. Ro. 2019. BMAN: Bidirectional multi-scale aggregation networks for abnormal event detection. *IEEE Transactions on Image Processing* 29 (2019), 2395–2408.
[32] I. Leite, G. Castellano, A. Pereira, C. Martinho, and A. Paiva. 2014. Empathic robots for long-term interaction. *International Journal of Social Robotics* 6 (2014), 329–341.
[33] L. Levinson, O. Gvirsman, I. M. Gorodesky, E. Perez, A.and Gonen, and G. Gordon. 2020. Learning in summer camp with social robots: a morphological study. *International Journal of Social Robotics* (2020), 1–14.
[34] A. Lopez-Rincon. 2019. Emotion recognition using facial expressions in children using the NAO Robot. In *Proc. CONIELECOMP*.
[35] E. Marinoiu, M. Zanfir, V. Olaru, and C. Sminchisescu. 2018. 3D human sensing, action and emotion recognition in robot assisted therapy of children with autism. In *Proc. CVPR*.
[36] F. S. Melo, A. Sardinha, D. Belo, M. Couto, M. Faria, A. Farias, H. Gambôa, C. Jesus, M. Kinarullathil, P. Lima, L. Luz, A. Mateus, I. Melo, P. Moreno, D. Osório, A. Paiva, J. Pimentel, J. Rodrigues, P. Sequeira, R. Solera-Ureña, M. Vasco, M. Veloso, and R. Ventura. 2019. Project INSIDE: towards autonomous semi-unstructured human–robot social interaction in autism therapy. *Artificial Intelligence in Medicine* 96 (2019), 198–216.
[37] A. Mollahosseini, B. Hasani, and M. H. Mahoor. 2017. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* 10 (2017), 18–31.
[38] NAO, Softbank Robotics [n.d.]. https://www.softbankrobotics.com/.
[39] B. Nojavanasghari, T. Baltrušaitis, C. E. Hughes, and L.P. Morency. 2016. EmoReact: a multimodal approach and dataset for recognizing emotional responses in children. In *Proc. ICMI*.
[40] A. K. Pandey and R. Gelin. 2018. A mass-produced sociable humanoid robot: Pepper: The first machine of its kind. *Robotics & Automation Magazine* 25 (2018), 40–48.
[41] E. Senft, S. Lemaignan, M. Bartlett, P. Baxter, and T. Belpaeme. 2018. Robots in the classroom: learning to be a good tutor. In *Proc. HRI*.
[42] M. Shiomi, T. Kanda, I. Howley, K. Hayashi, and N. Hagita. 2015. Can a social robot stimulate science curiosity in classrooms? *International Journal of Social Robotics* 7 (2015), 641–652.
[43] G. Skantze and S. Al Moubayed. 2012. IrisTK: a statechart-based toolkit for multi-party face-to-face interaction. In *Proc. ICMI*.
[44] C. Sun, Y. Jia, H. Song, and Y. Wu. 2020. Adversarial 3D convolutional auto-encoder for abnormal event detection in videos. *IEEE Transactions on Multimedia* (2020).
[45] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proc. CVPR*.
[46] A. Tsiami, P. Koutras, N. Efthymiou, P. P. Filntisis, G. Potamianos, and P. Maragos. 2018. Multi3: Multi-sensory perception system for multi-modal child interaction with multiple robots. In *Proc. ICRA*.
[47] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *Proc. ECCV*.
[48] E. Wolfe, J. Weinberg, and S. Hupp. 2018. Deploying a social robot to co-teach social emotional learning in the early childhood classroom. In *Proc. HRI*.
[49] Y. Zhang, Y. Tian, P. Wu, and D. Chen. 2021. Application of skeleton data and long short-term memory in action recognition of children with autism spectrum disorder. *Sensors* 21 (2021), 411.