Service Humanoid Robotics: Review and Design of A Novel Bionic-Companionship Framework

Jiaji Yang[1[0000-0002-1011-9676]] and Esyin Chew [1[0000-0003-2644-9888]]and [2]Pengcheng Liu[0000-0003-0677-4421]

[1]EUREKA Robotics Lab, Cardiff School of Technologies, Cardiff Metropolitan University, Cardiff, CF5 2YB, United Kingdom.

[2] The Department of Computer Science, University of York, York YO10 5GH, United Kingdom.
JYang@cardiffmet.ac.uk

Abstract. At present, industrial robotics focused more on motion control and vision; whereas Humanoid Service Robotics (HSRs) are increasingly being investigated among researchers' and practitioners' field of speech interactions. The problematic and quality of human-robot interaction (HRI) has become one of the hot potatoes concerned in academia. This paper proposes a novel interactive framework suitable for HSRs. The proposed framework is grounded on the novel integration of Trevarthen [23] Companionship Theory and neural image generation algorithm in computer vision. By integrating the image-to-natural interactivities generation, and communicate with the environment to better interact with the stakeholder, thereby changing from interaction to a bionic-companionship. In addition, the article also reviews the research of neural image generation algorithms and summarizes the application cases of the algorithm structure in the field of robotics from a critical perspective. We believe that the new interactive bionic-companionship framework can enable HSRs to further develop towards robot companions.

Keywords: Humanoid Robotics, Human-Robot Interaction, Social Robotics

1 Introduction

Humanoid service robots (HSRs) are a booming reality, and it is reported that HSRs are becoming one of the major technologies that will drive the service industries in the next decade [12]. More and more researchers are committed to using HSRs to help humans complete some simple service tasks and interactive tasks. Delivery robots, concierge robots, chat robots, etc., have been increasingly used by travel and hospitality companies [13]. Although the contribution of these achievements mainly comes from the rapid development of the robotics engineering, Ivanov et al. [14] indicate that the future research focus will gradually shift from robotics engineering to human-robot interaction (HRI), thus opening up new research direction for researchers.

As early as 2003, Fong et al. [30] proposed that in order to make robots perform better, the robot need to be able to use human skills (perception, cognition, etc.) and benefit from human advice and expertise. This means that robots that rely solely on self-determination have limitations in performing tasks. They believe that the collaborative work of humans and robots will be able to break this constraint, and research on human-robot interaction has begun to emerge more and more. Fong et al. [30] believe that to build a collaborative control system and complete human-robot interaction, four key problems must be solved. 1. The robot must be able to detect limitations (what can be done and what humans can do), determine whether to seek help, and identify when it needs to be resolved. 2. The robot must be self-reliant. It must be able to maintain its own security. 3. The system must support dialogue. That is, robots and humans need to be able to communicate with each other effectively. But the dialogue is restricted. Through collaborative control, the dialogue should be two-way and require a richer vocabulary. 4. The system must be adaptive. Although most of the current humanoid service robots already support dialogue and can complete simple interactive tasks, as mentioned in the research, such dialogue seems to be restricted. In the process of interacting with robots, humans always obtain the state of the robot through vision, and then communicate with the robot through the dialogue system. However, humanoid service robots cannot do this, which does not seem to fully satisfy the two-way nature of dialogue. Therefore, this research is different from the current HRI model. This research attempts to introduce vision into the existing dialogue system of humanoid service robots, so as to improve the existing HRI model.

The aim of this research is focused on how to improve the interaction between HSRs and humans. Inspired by the recent advance in the field of neural image caption generation that is currently receiving much attention in computer vision, this article proposes a novel humanoid service robot and human interaction framework centered on the neural image caption generation algorithm. The framework is anticipated to enhance HRI to reach a new state, making it possible for HSRs to become bionic companions of humans. The concept of the bionic companionship comes from the Trevarthen Companionship Theory [23], which describes that the companionship should have the ability and interest to interact with the dynamic thoughts and enthusiasm of the partner's relationship, and can recognize what others think is meaningful and the emotions of these things express sympathy. The earlier concept of the robot companion is mentioned in the research of Dautenhahn et al [5]. Their study suggested that autonomous robot companions may be regarded as a special kind of service robot. Robot companions can communicate with non-experts in a natural and intuitive manner, and HSRs need to have a high degree of awareness and sensitivity to social environment. Based on the recent advancement in neural image caption generation, related technologies that automatically generate text descriptions for pictures and even media streams are becoming more mature and accurate [15][26], we propose to apply this technology to HSRs, so that the robots can automatically convert pictures or data information that caught by cameras or sensors into texts or sentences. Hence, HSRs can communicate more naturally with humans by using these texts or sentences. This suggested novel interactive framework not only satisfies the concept of robot companionship, which requires the robot to have a high degree of bionic awareness and sensitivity to the environment, but also enhance the HRSs active communication with humans in a natural humane feature with 7 senses. The study will contribute to the further development from HRI to HRC (Human-Robot Companionship). This research will mainly adopt the method of convolutional neural network combined with recurrent neural network to realize the processing of visual data and text generation by the robot. The conversion from image to text is realized by inputting the image features extracted by the convolutional neural network into the recurrent neural network. This is also one of the main methods in the field of image captioning. Since this algorithm is adapt from the machine translation algorithm, most research use BLEU scores as the evaluation criteria of the algorithm. However, the BLEU score is not accurate due to the introduction of image features. Research believes that a more accurate model evaluation method should be proposed in the future.

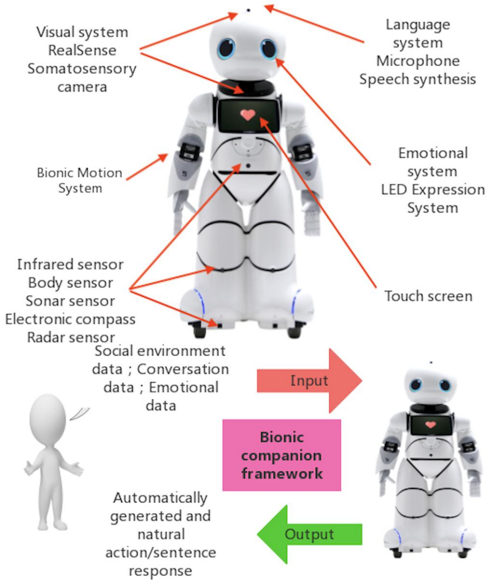


Fig. 1. Conceptual Model of Bionic Humanoid Robot with 7 senses: Upper: Humanoid service robot [1] with various sensors support that will be used in research; Lower: Input and output through the interactive framework.

1.1 Human Robot Companionship

The implementation of this HRC framework is on the Canbot UU humanoid robot (Fig. 1) [1]. The robot's 22-degree-of-freedom motion joints can perform a variety of simulated movements, such as raising head, turning head, raising arm, shaking crank, shaking hands, leaning back, walking or turning, and producing excellent interaction with humans. Thanks to the robot drive neural network motion control algorithm and MCU microchips unit function, the robot arm can move flexibly in various directions.In addition, Canbot UU's advanced vision system and various sensors can collect more complete data for our proposed framework and make our models in the framework more robust. The robot's design based on imitating the human's seven senses also provides a strong foundation for the concept of the bionic partner designed in this study.

2 literature reviews

2.1 Neural image caption generation development and reviews

The challenge of generating natural language descriptions from visual data has been extensively investigated in the field of computer vision. The early research mainly focused on generating natural language descriptions from video-type visual data [10][17]. These systems convert complex visual data into natural language through rule-based systems. However, because these rules are artificially designed, these systems are not considered to be more robust and have been proven to be used only in limited applications such as traffic scenarios [26]. Among the research achievements in the past five years, many researchers have been inspired by the successful use of sequence to sequence training with neural networks for machine translation and consequently put forward a method for generating image description based on recurrent neural network [4][21]. In facts, this way of replacing the encoder in the encoder-decoder framework in machine translation with image features makes the original complex task of generating image data caption into a simple process of 'translating' the image into a sentence [4]. In the same period, Donahue et al. [8] used long short-term memory (LSTM) for its model, which is suitable for end-to-end large-scale visual learning process. In addition to images, Donahue et al [8] also applied LSTM to video, allowing their models to generate video descriptions. Vinyals et al [26] and Kiros et al [20] inventively explored the structure of the currently popular neural image generation algorithm. This algorithm is based on the combination of Convolutional Neural Network (CNN) image recognition model and natural language processing (NLP) structured model. At the same time, the neural image captioning algorithm based on the attention mechanism has also attracted extensive attention in the field of computer vision, Denil et al. [6] recommended a real-time target tracking and attention recognition model driven by computer visual data. Tang et al. [27] proposed an attention generation model based on deep learning. The model is inspired by visual neuroscience and collects data with the object as the center in model generation. After these, Mnih et al. [25] investigated a new recurrent neural network model, which has the ability to automatically select specific regions from the images and videos for feature extraction.As the algorithm becomes more and more mature, the application of the algorithm in some fields has also been broken through recently, such as the caption generation of car images [3], the description generation of facial expressions [19], humanoid robots driven with the image caption generation for children education [15]. Recent research on Image caption generation also shows that the accuracy and reliability of the technology is getting higher and higher [7]. Even research on the use of reinforcement learning to automatically correct image caption generation networks has emerged [9]. These studies on the generation of neural image captions have undoubtedly laid a solid foundation for their application for HRSs . This makes it possible for humanoid robots to interact while recognizing the social environment, thereby improving the interactive service quality of the HSRs.

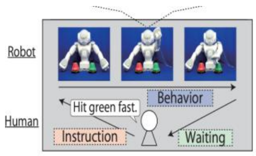
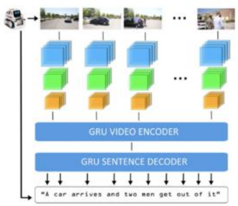
2.2 Neural image caption generation algorithm 'crash into' robot

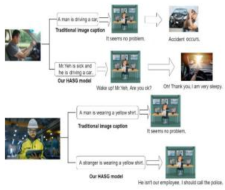
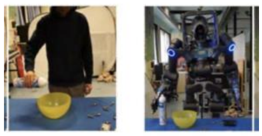
Recently, increasingly more studies have been conducted on the HRI combining with the algorithm structure of image caption generation (as shown in Tab. I). Kim et al. [15] used the structure of CNN combined with recurrent neural network (RNN) + Deep concept hierarchies (DCH) to design and develop an educational intelligent humanoid robot system. The system is used to play video games with children. In the research, CNN was used to extract and pre-process some cartoons with educational features, and RNN and DCH were used to convert the collected video features into Q&A about cartoons. During the game, the child and the robot ask and answer questions each other based on the content of the cartoon after watching the same cartoon. The study results show that such a system can interact effectively with children. However, for HRI, the simple and limited question-and-answer conditions cannot satisfy all the interaction scenarios required. Cascianelli et al. [2] used Full-gated recurrent unit (GRU) encoder-decoder architecture to develop a human-robot interface that provides interactive services for service robots. This research solves a problem called natural language video description (NLVD). Meanwhile, they also compared the performance when using LSTM and GRU two different algorithms to solve this problems. They prove that the GRU algorithm runs faster and consumes less memory. We think this kind of model may more suitable for HSRs. Although the research model is competitive on public data sets, the experimental results on the designed data sets show that the model has serious overfitting. This proves that in the actual model training process, a specific training data set for HSRs interaction should be established and other methods such as transfer learning should be considered to improve the generalization ability of the model on interactive tasks. Luo et al [16] created a description template to add various image features collected by the robot, such as face recognition and expression, to the generated description. Compared with the previous models, their interaction is more natural and close to human description. But the purpose of them is to use the model to provide services to industry managers and not to use it to conduct an entire HRI framework.

In addition to the research on robot vision-language, the research on robot vision-action is endless. Yamada et al. [28] use recurrent neural networks to enable humanoid robots to online learn commands from humans and generate corresponding behaviors as the response. This article provides a reference and pavement for humanoid robots to use deep learning to obtain online learning capabilities for human commands. Inspired by them, the idea proposed in this paper is that the description generated by the neural image captions can drive HSRs to perform appropriate behaviors, and HSRs can even obtain online learning capabilities of interacting with surrounding people through social environments. Tremblay et al. [22] and Nguyen et al [18] believe that non-experts often lack the rationality of task description when issuing instructions to robots. They use deep learning to allow robots to automatically generate human-readable instructions' description according to the surrounding social environment. Furthermore, Nguyen et al. [18] also use visual data to make humanoid robots imitate and learn human actions under corresponding commands, so that the robot can learn how to complete the corresponding tasks only through visual data. Although they pointed out that humanoid robots cannot complete precise control of movements when they imitate movements of visual data.

Through the discussion of these studies, it is plausible and feasible to establish an interactive framework for HSRs using deep learning and neural image caption generation. The existing technology is sufficient to help HSRs cope with the simple interactive tasks of the service industry an pave the way of the evolution from HRI to robot companion.

Table 1. Neuro image caption Structural algorithm drive robots

Reference	Algorithm	Robot Model	Studies or Results
Kim et al(2015,). Pororobot: A deep learning robot that plays video Q&A games.	Convolutional neural networks (CNNs) ; Recurrent neural net- works (RNNs) ;deep concept hierarchies (DCH)	Humanoid robot Pororobot (NAO V5) 	The robot can online interact with people on a given task, and according to the instructions issued by humans to judge and generate the task content to be completed.
Yamada et al(2016). Dynamical integration of language and behavior in a recurrent neural network for human–robot interaction.	Recurrent neural network (RNN)	Humanoid robot NAO V5 	The robot can online interact with people on a given task. And automatically judge and generate motions completion instructions based on the instructions issued by humans.
Cascianelli et al(2018). Full-GRU natural language video description for service robotics applications.	Full- GRU encoder-decoder architecture;	Service robot 	This research focuses on the natural language video description (NLVD) task, and proposes a complete GRU encoder-decoder architecture for solving this problem. They show that compared with other latest algorithms, the proposed method is faster to train and consumes less memory.
Tremblay et al (2018, May). Synthetically trained neural networks for learning human-readable plans from real-world demonstrations.	Neural networks; Convolutional pose machines	Baxter robot Non-humanoid robot 	This research developed a system composed of multiple neural networks, with the help of which the robot can automatically generate human-readable command descriptions.

Luo et al (2019, June). Multi-Modal Human-Aware Image Caption System for Intelligent Service Robotics Applications.	CNN-LSTM structure Human-Aware Context Generator (HACG)	Non-humanoid robot 	This paper proposes the HACG model, which is a multi-modal fusion of image captions, facial expression recognition and facial recognition. The study uses a template-based method to replace some keywords in the generated description with replacements, so as to achieve the purpose of making the description more natural.
Gui et al (2018, October). Teaching robots to predict human motion.	Generative adversarial networks.	Humanoid robot Pepper 	This paper have developed a deep learning based system that enables robots to predict and demonstrate human motion. To this end, we propose a novel motion GAN model to improve the prediction plausibility from a global perspective.
Nguyen et al(2018). Translating videos to commands for robotic manipulation with deep recurrent neural networks.	LSTM and GRU	Humanoid robot WALK-MAN 	This paper proposed a new method to translate human demonstration videos to commands using deep re- current neural networks.

3 Bionic-Companionship Framework

In the previous two sections, this paper reviewed the neural caption generation approach and its application to robots. This section will summarize the previous points and depict a navel humanoid service robot and human interaction framework with neural image subtitles as the core (as shown in Fig. 2). The framework uses the structure of the NIC algorithm to better realize the interaction of HSRs from HRI to the direction of bionic companionship. According to the description of robot companions as given in [24] and [15],, the proposed framework should provide HSRs with a more natural interaction and a more sensitive understanding of the environment, so the system is divided into three subsystems.

3.1 Image / video description generation system

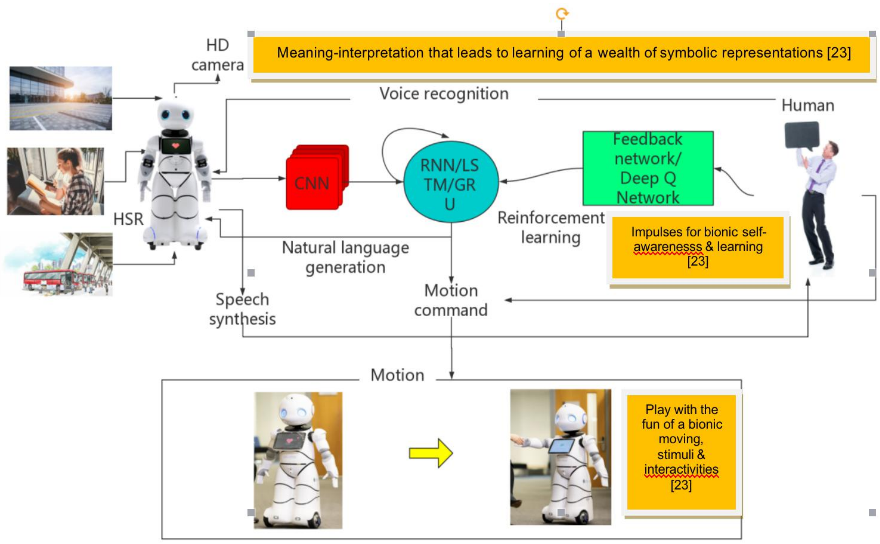
This subsystem is the core system of the entire interactive framework. HSRs collect visual data of the surrounding environment through the equipped visual sensors (such as HD cameras). The type of visual data collected depends on the complexity of the interactive task to be completed by the HSRs. It is generally considered that more complex interactive tasks require the use of continuous pictures or and real-time videos. The system uses the latest neural image generation algorithm structure, uses CNN to perform feature extraction on the pictures and video data of the surrounding social environment, and converts the data into feature vectors' sequence that can be used by RNN. Finally, RNN completes the process of generating interactive description from visual data. HSRs use a speech synthesis system that converts those descriptions into voices to communicate with humans. This process is different from the past mode of using HSRs human sensing sensors and setting fixed interactive feedback, The innovation of this system is that HSRs can automatically and naturally generate interactive feedback. This means that the change of the scene during the interaction will cause the continuous change of the interaction feedback and this change is not preset by humans. In addition, in further conversation interactions, human's voice response, and social environment data will be coordinated by HSRs and produce continuous conversation interaction behavior.

3.2 Command-robot behavior system

For HSRs, simple conversation interaction is not enough. HSRs should generate corresponding motion based on visual data and human behavior data. For example, when humans wave to the robot, the robot should also actively wave to respond. The hypothesis of this study is to classify or cluster description text generated from visual data, and use these classified description texts to control the motions of HSRs in response to complex interactive tasks. For example, when the description generated by neural image captions is 'hello', then HSRs will automatically determine whether 'hello' matches a category that requires interactive motion and perform corresponding motions such as waving.

3.3 Human-supervised feedback reinforcement learning system

Since the core of the framework is an image / video description generation system, Therefore, the interactive performance of the framework is affected by the accuracy of the generated description. In the absence of a large amount of high-quality data, it may even lead to overfitting or underfitting of the model. Inspired by Gui et al [11], we believe that human-supervised feedback reinforcement learning subsystem should be added to the interaction framework. The human-supervised feedback reinforcement learning system will allow humans to correct the generated descriptions, thereby maximizing optimization and supporting the entire interactive framework.



4 Conclusion

This study presents a review of neural image generation algorithms and application cases in the field of robotics, then proposes a novel humanoid service robot and human interaction framework based on the bionic companionship theory. The three subsystems of the bionic companionship framework are designed and introduced in details. The proposed novel interactive framework not only satisfies the concept of robot companion, which requires the robot to have impulses for bionic self-awareness and learning to the environment, but also enhances HRSs' meaning-interpretation capabilities that lead to the learning of a wealth of symbolic representations with 7 senses [23] . The proposed framework will contribute to the further development from HRI to HRC (Human-Robot Companionship). The future work will focus on implementing each of the subsystems in the framework and applying the framework to HSRs to verify its performance. It is also expected to establish an interactive model training data set dedicated to HSRs which will report in another journal article in due course for the reference of humanoid and social robotics researches and practitioners.

References

1. CANBOT Homepage, https://www.canbotrobots.com/html/yy-detail.html, last accessed 2020/06/04
2. Cascianelli, S., Costante, G., Ciarfuglia, T. A., Valigi, P., & Fravolini, M. L.: Full-GRU natural language video description for service robotics applications. IEEE Robotics and Automation Letters, 3(2), 841-848 (2018).

3. Chen, L., He, Y., & Fan, L.: Let the robot tell: describe car image with natural language via LSTM. Pattern Recognition Letters, 98, 75-82 (2017).
4. Cho, Kyunghyun, van Merrienboer, Bart, Gulcehre, Caglar, Bougares, Fethi, Schwenk, Holger, and Bengio, Yoshua.:Learning phrase representations using RNN encoder-decoder for statistical machine translation. In EMNLP (2014).
5. Dautenhahn, K ; Woods, S ; Kaouri, C ; Walters, M.L ; Kheng Lee Koay & Werry, I.,: What is a robot companion - friend, assistant or butler? IEEE/RSJ International Conference on Intelligent Robots and Systems, pp.1192–1197 (2005).
6. Denil, M., Bazzani, L., Larochelle, H., & de Freitas, N.: Learning where to attend with deep architectures for image tracking. Neural computation, 24(8), 2151-2184 (2012).
7. Ding, S., Qu, S., Xi, Y., Sangaiah, A.K. and Wan, S.: Image caption generation with high-level image features. Pattern Recognition Letters, 123, pp.89-95 (2019).
8. Donahue, Jeff, Hendrikcs, Lisa, A., Guadarrama, Segio, Rohrbach, Marcus, Venugopalan, Subhashini, Saenko, Kate, and Darrell, Trevor.: Long-term recurrent convolutional networks for visual recognition and description. arXiv:1411.4389v2 (2014).
9. Fidler, S. : Teaching machines to describe images with natural language feedback. In Advances in Neural Information Processing Systems pp. 5068-5078, (2017).
10. Gerber, R. & Nagel, N.-H.,: Knowledge representation for the generation of quantified natural language descriptions of vehicle traffic in image sequences. Proceedings of 3rd IEEE International Conference on Image Processing, 2, pp.805–808 vol.2 (1996).
11. Gui, L. Y., Zhang, K., Wang, Y. X., Liang, X., Moura, J. M., & Veloso, M.: Teaching robots to predict human motion. In 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 562-567). IEEE (2018)|.
12. Harris, K., Austin ,K., and Andrew, S.,: "Why the Automation Boom Could Be Followed by a Bust," Harvard Business Review (March 13), https://hbr.org/2018/03/why-the-auto mation-boom-could-be-followed-by-a-bust (2018).
13. Ivanov, S. Ultimate transformation: How will automation technologies disrupt the travel, tourism and hospitality industries? Zeitschrift für Tourismuswissenschaft, 11 (1), 25–43 (2019).
14. Ivanov, S., Gretzel, U., Berezina, K., Sigala, M., & Webster, C.. :Progress on robotics in hospitality and tourism: A review of the literature. Journal of Hospitality and Tourism Technology. https://doi.org/10.1108/JHTT-08-2018-0087 (2019).
15. Kim, K. M., Nan, C. J., Ha, J. W., Heo, Y. J., & Zhang, B. T.: Pororobot: A deep learning robot that plays video Q&A games. In 2015 AAAI Fall Symposium Series (2015).
16. Luo, R. C., Hsu, Y. T., & Ye, H. J.: Multi-Modal Human-Aware Image Caption System for Intelligent Service Robotics Applications. In 2019 IEEE 28th International Symposium on Industrial Electronics (ISIE) (pp. 1180-1185). IEEE (2019).
17. Mitchell, M., Han, X., Dodge, J., Mensch, A., Goyal, A., Berg, A., Yamaguchi, K., Berg, T., Stratos, K., & Daumé III, H. Midge: Generating image descriptions from computer vision detections. In Proceedings of the 13th conference of the European chapter of the association for computational linguistics(pp. 747–756). Association for Computational Linguistics (2012).
18. Nguyen, A., Kanoulas, D., Muratore, L., Caldwell, D. G., & Tsagarakis, N. G. :Translating videos to commands for robotic manipulation with deep recurrent neural networks. In 2018 IEEE International Conference on Robotics and Automation (ICRA) (pp. 1-9). IEEE (2018).
19. P. Kuznetsova, V. Ordonez, T.L. Berg, Y. Choi, TREETALK: Composition and compression of trees for image descriptions, Trans. Assoc. Comput.Ling. 2 (1) 351–362 (2014).

20. R. Kiros, R. Salahutdinov, R. Zemel,:Multimodal neural language models, in: International Conference on Machine Learning, pp. 595–603 (2014).
21. Sutskever, Ilya, Vinyals, Oriol, and Le, Quoc VV.:Sequence to sequence learning with neural networks. In NIPS, pp. 3104– 3112 (2014).
22. Tremblay, J., To, T., Molchanov, A., Tyree, S., Kautz, J., & Birchfield, S.: Synthetically trained neural networks for learning human-readable plans from real-world demonstrations. In 2018 IEEE International Conference on Robotics and Automation (ICRA) (pp. 1-5). IEEE (2018).
23. Trevarthen, C.: Intrinsic motives for companionship in understanding: Their origin, development, and significance for infant mental health. Infant Mental Health Journal: Official Publication of The World Association for Infant Mental Health, 22(1 - 2), 95-131 (2001).
24. Turkle, S. A nascent robotics culture: New complicities for companionship. American Association for Artificial Intelligence Technical Report Series AAAI (2006).
25. V. Mnih, N. Hees, A. Graves, K. Kavukcuoglu, Recurrent models of visual atten- tion, NIPS (2014).
26. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D.: Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164) (2015).
27. Y. Tang, N. Srivastava, R.R. Salakhutdinov, Learning generative models with vi- sual attention, in: NIPS , pp. 1808–1816 (2014).
28. Yamada, T., Murata, S., Arie, H., & Ogata, T. Dynamical integration of language and behavior in a recurrent neural network for human–robot interaction. Frontiers in neurorobotics, 10, 5 (2016).
29. C. Trevarthen. Play with infants: The impulse for human story-telling,In, Tina Bruce, Pentti Hakkarainen and Milda Bredikyte (Eds.)The Routledge International Handbook of Play in Early Childhood.Abingdon: Taylor & Francis/Routledge, Chapter 15. http://www.becera.org.uk/BECERA%202017/CT%20ON%20PLAYRoutledge%20Handbook%202017.pdf (2016)
30. Fong, T., Thorpe, C., & Baur, C. Collaboration, dialogue, human-robot interaction. In Robotics Research (pp. 255-266). Springer, Berlin, Heidelberg (2003).