Automatic Joint Attention Detection During Interaction with a Humanoid Robot

Dario Cazzato[2], Pier Luigi Mazzeo[1], Paolo Spagnolo[1](✉), and Cosimo Distante[1]

[1] National Research Council of Italy, Lecce, Italy
paolo.spagnolo@cnr.it
[2] Faculty of Engineering, University of Salento, Lecce, Italy

Abstract. Joint attention is an early-developing social-communicative skill in which two people (usually a young child and an adult) share attention with regards to an interesting object or event, by means of gestures and gaze, and its presence is a key element in evaluating the therapy in the case of autism spectrum disorders. In this work, a novel automatic system able to detect joint attention by using completely non-intrusive depth camera installed on the room ceiling is presented. In particular, in a scenario where a humanoid-robot, a therapist (or a parent) and a child are interacting, the system can detect the social interaction between them. Specifically, a depth camera mounted on the top of a room is employed to detect, first of all, the arising event to be monitored (performed by an humanoid robot) and, subsequently, to detect the eventual joint attention mechanism analyzing the orientation of the head. The system operates in real-time, providing to the therapist a completely non-intrusive instrument to help him to evaluate the quality and the precise modalities of this predominant feature during the therapy session.

1 Introduction

Joint attention is an early-developing social-communicative skill in which two people (usually a young child and an adult) share attention with regards to an interesting object or event, by means of gestures and/or gaze. In particular, the work in [16] firstly proposed a preliminary investigation about the extent of the infant's ability to follow changes in adult gaze direction during the first years of life. In particular, on each test trial performed on 34 children between 2 and 14 years old, the enfant first made eye-to-eye contact and then silently turned his or her head, looking at a small concealed signal light for 7 seconds, while the adult head was then turned back to interact with the infant. In the experiment, it was found that the proportion of infants judged as having produced a positive response on one or both trials increased steadily with age. Since its fundamental importance in the fields of cognitive and developmental psychology, a long list of subsequent studies have been proposed in the literature over the years, and it is still an active research topic [12]. For the case of autism, several works

A. Tapus et al. (Eds.): ICSR 2015, LNAI 9388, pp. 124–134, 2015.
DOI: 10.1007/978-3-319-25554-5_13

investigated the meaning and the modalities of joint attention lacks, like in [20], [21] and [7]. In fact, impaired development of joint attention is a predominant feature in children with autism spectrum disorder (ASD), and a set of strategy are used to teach and support joint attention [9]. Although children with ASD can show attention to objects or toys of interest, they have difficulties in sharing attention or interests with the therapist of a relative. For this purpose, the Early Start Denver Model [15] underlines this capacity as a key element of social cognition, working on an improvement of this concerned interaction lack. In [5] a requirement to detect joint attention is that two individuals are attending to the same object, based on one individual using the attention cues of the second individual. Shared Attention is a combination of mutual attention (the attention of two individuals is directed to one another) and joint attention (two individuals are looking at the same object), where at the same time the two individuals have knowledge of the directions of the other individual's attention. In other words, shared attention represents a higher state of the dyadic relationship whereby both individuals are attending the same object, as with joint attention, but both are aware of each other's attentional state. Although they are slightly different, in the literature shared and joint attention are considered as synonyms. Beyond subtle differences, what is important to notice is that for us joint attention means that the child is attending to the same object using the attention cues of the adult, i.e. knowing together that they are attending to the same thing [3].

On the other hand, the usage of Socially Assistive Robotics have spread among recent years, providing a new and useful instrument to elicit interest on the autistic child during the therapy. For a review about the clinical usage of robots in autism research refer to [4] and [17]. In particular, socially assistive robots have been employed also to elicit behavior in children with ASD [6]. Many works in the literature explore this exciting feature, but most of them are based on a simulated robot on a screen, eventually evaluating the participant's movement [10], his visual perspective [22] or his reaction time [11]. In [18] gaze track in a human-robot interaction setting (on a monitor) is analyzed, but the attention is measured by means of eye tracker, that are very expensive, and they needs a user calibration that becomes very difficult in the case of children with ASD. A humanoid robot has been employed to elicit joint attention in [14], but all the acquired data has been manually analyzed a posteriori. In [8] a specific hardware is employed to detect the joint attention in a 1-by-1 human-robot interaction scheme. In particular, the method employs an omnidirectional vision sensor and 16 ultrasonic distance sensors around a movable base in order to localize the person's location and to turn consequently the robot's head. This feature is merged with a speech generation system that can elicit social behaviors. Finally, the work in [1] compares the visual exploration during joint attention elicitation in typical development children and children with ASD by means of a Kinect sensor in order to capture social engagement cues. The system evaluates a possible joint attention event, but the sensor is installed in front of the child (thus visible) and even in this work no sharing with another adult has been taken into account.

This paper introduces a two-level innovation with regards to the state of the art: first of all, it presents a computer vision based system based on non-intrusive and invisible to the patient hardware that can operate in real-time. Moreover, it considers the case of a triadic interaction, involving the child, an adult and a humanoid robot (i.e. the Aldebaran NAO H25) in order to detect the joint attention when the robot performs an ad-hoc movement to elicit joint attention mechanism. The paper is organized as follows. Section 2 introduces the proposed method, while Section 3 shows the experimental setup and the achieved results. Finally, Section 4 concludes the paper.

2 Proposed Method

In the following paragraphs the hardware architecture will be illustrated, as well as the computer vision algorithms implemented for the detection of people, their heads, and correspondent axis for mutual interaction evaluation.

2.1 System Overview

The proposed approach uses images acquired by a Microsoft Kinect device. We arranged a therapy room with a calibrated acquisition device placed on the ceiling, in a non-intrusive position, with the goal of not disturb the children. Our idea is to develop simple and effective algorithms, in order to implement them on low-end hardware. The algorithms run in real time, require no training data (except some seconds of free acquisition for the background modeling, as described below) and can exploit cheap, off-the-shelf sensors. In fig. 1 the flowchart of the whole system is shown. The Kinect device acquires synchronized depth information about the scene in parallel with the RGB video stream. Depth images are segmented in order to detect moving blobs in the scene (calibration data are used to improve segmentation by considering constraints of the system, i.e. the distances between the camera and expected target, as well as the floor and the other planes). After this, Canny operator is applied on the depth images (only on foreground areas) providing an edge map. This map is the input of the head detection algorithm, which processes the edge map with the goal of detect elliptical structures. The outputs of this step are the major and minor axes of the detected ellipses. Finally, the behaviors of children are classified according to mutual positions of the major axes of the educator and the patient.

2.2 Segmentation

The first step of the algorithm is the segmentation of foreground objects. For this purpose, we have chosen to work directly on the depth map: this way, traditional weakness of foreground segmentation algorithms (for example the presence of shadows, reflections on specular surfaces, and so on) are limited or totally avoided. We have implemented the algorithm proposed in [23], which is robust to shadows, reflections, small movements in the background. Even if these

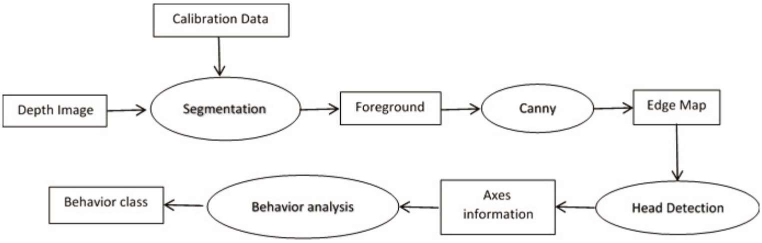


Fig. 1. A schematic diagram of the processing steps

aspects are filtered by using depth images, some artifacts could be present anyhow. To further improve segmentation of desired objects (people and robot), we run the algorithm before patients and educators enter the room. The algorithm is a variation of classic Gaussian Mixture Model approach [19]: each point is represented by a number of Gaussians (with mean and variance), and a variation is considered as foreground if it differs from each gaussian more then the correspondent variance. In fig. 2 we can see a depth image, and the corresponding segmented one.

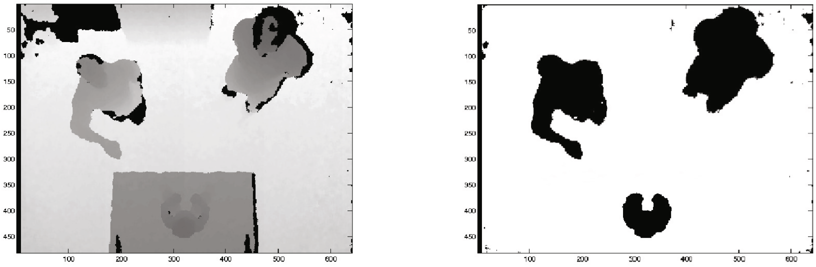


Fig. 2. An example of depth image, and the output of the segmentation procedure.

2.3 Head Detection

After the foreground segmentation, we need to detect the heads of subjects present in the scene. To do this, we use the detected foreground as a binary mask on the depth image. This way, we work on a depth image composed by only foreground objects. For the head detection we take advantage of the constraints of the applicative context: autistic children usually do not interact with other people, so it is realistic to consider that each foreground detected region refers to only one person. Firstly, a Canny operator [2] is applied to obtain an edge map. This map is then processed in order to detect ellipses, that correspond to the heads. The algorithm proposed in [13] has been implemented, with some variations: specifically, the detection has been focused on a specific class of ellipses. This because the size and the geometry of the head is known, as well as the position of the camera and the focal lens. So, we can assume that the size of expected ellipses (heads) is known, and can vary in a certain range. In fig. 3 we can see the output of the head detection algorithm.

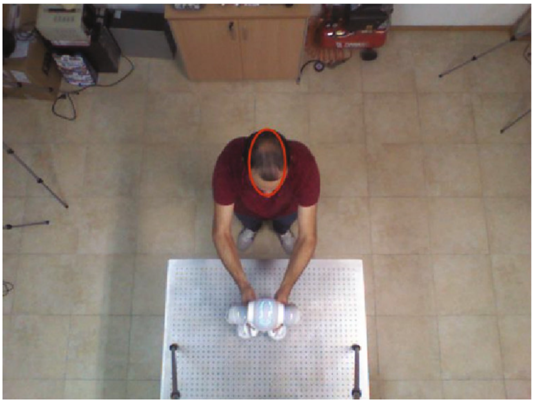


Fig. 3. The output of the head detection algorithm.

2.4 Behavior Understanding

The final step of the whole approach is the behavior detection. The proposed system, as remarked above, has the goal to provide a support to therapists in the analysis of behavior of children with autism spectrum disorder. So, the output of the automatic algorithm needs to be a classification of most common behaviors. According to suggestions provided by therapists, our goal is the detection of three main behaviors:

1. **Joint Attention.** Both adult and child look at the robot **(JA)**;
2. **Child attention.** Educator looks at the child while the child looks at the robot **(A2C2R)**;
3. **Child Adult attention.** Both Adult and child look at each other **(A2C2A)**;

The automatic detection of such behaviors has been done by evaluating the mutual position of the major axes of the detected ellipses (heads), as well as by considering the position (known) of the robot. So, the starting point of this final step is the extraction of the major axes of the ellipses, and the evaluation of their directions. In fig. 4 we can see an example of this approach: the major axes of each head/ellipse are plotted, and the behaviour is evaluated by considering the possible situations (the angle between them, the position of robot, etc).

In fig. 5 we can see some example of images of the desired classes of behaviors, with the correspondent synthetic scheme for the geometric evaluation. Formally, we have assumed these rules for the classification:

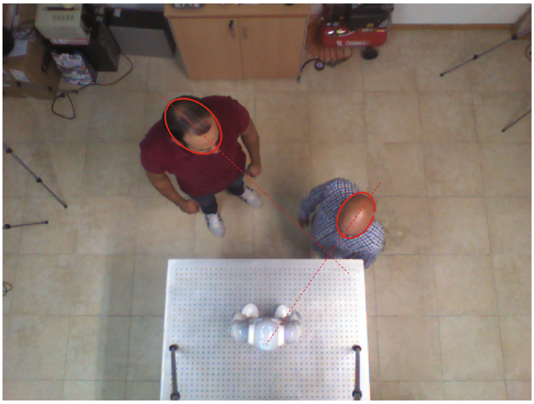


Fig. 4. The major axes and their use in the evaluation of behavior.

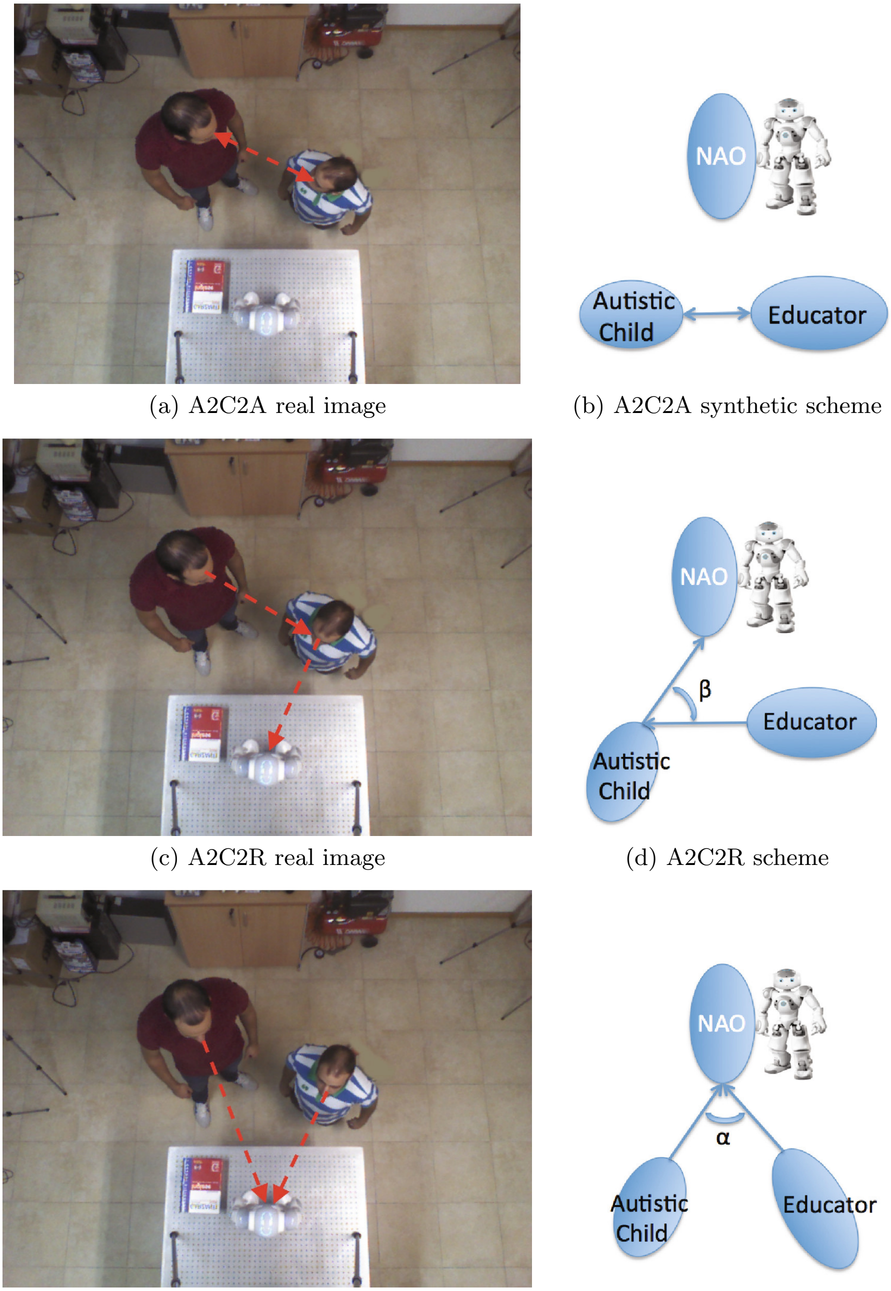
1. **JA.** The intersection of the major axes of the subjects (child and adult) corresponds to the position of the robot;
2. **A2C2R.** The intersection of the major axes corresponds to the head of the child AND the major axis of the child corresponds to the position of the robot (i.e. the gaze of the child is on the robot);
3. **A2C2A.** The major axes of both child and adult are about congruent (i.e. they are congruent, OR they intersect with a very small angle, less then 10 degrees).

3 Experiments

In the next subsections, firstly the setup and the main characteristics of the test sequences will be illustrated; then, the results obtained by using the proposed algorithms will be presented.

3.1 Setup and Dataset

The acquisition setup is composed by a calibrated Kinect Camera installed on the ceiling (about 3 meters from the floor). We created a simulated operative scene containing a work table (1.5 meters from the floor), NAO humanoid robot located on the table, an autistic child and an adult. We acquired several sequences in which we have simulated the three different kinds of attention interactions among the adult, the child and the humanoid robot described in section 2.4. The proposed setup has to be considered stationary, because our goal, for the future, is to arrange a therapy room with this camouflaged hardware. We use the term 'simulation' because the acquisition sessions, even if performed following instructions of therapists, have been performed in absence of a therapist. Moreover, the acquisitions have been made by using children without ASD. In the future, the



(a) A2C2A real image

(b) A2C2A synthetic scheme

(c) A2C2R real image

(d) A2C2R scheme

(e) JA real image

(f) JA scheme

Fig. 5. Some images acquired during experimental phase. Red dotted lines highlight the gaze orientation.

acquisition sessions will be made in a clinical context, in presence of a therapist, and actions will be performed by real ASD patients. In this step of our work, our goal is the test of computer vision algorithms, so the presence of real ASD patients or non ASD patients is irrelevant.

Another observation is necessary: due to severe laws about acquisition and publications of images with minors, we present images in which actors are adults (we hope in the future to obtain the necessary permissions to use minor images).

It is important to note that we are testing the computer vision algorithms, instead of clinical ones: so, we are interested in results in terms of correct detection of heads, angles between them, and so on, while an analysis of medical implications of these results will be examined in depth in the future.

The presence of the robot during acquisitions is strategic for the correct evaluation of child attention: during the different simulations, the robot performed generic behaviours (it stood up, or it sat down), with the goal of attract the gaze of children; this way, we can test if the children react to external impulses (where external has to be intended as 'not from the therapist').

3.2 Results

We implemented the proposed method and tested it on three video sequences of 1600 frames each one. We have manually labelled each frame in order to create a ground truth of the three different behavior classes described in section 2.4. We have introduced a new class **Non Classified Behavior (NCB)** for labelling of all frames that do not belong to the three main behavior classes. This Ground Truth creation process generated four labelled classes and the population of each is given by table 1.

Table 1. Behavior classes population

	JA	**A2C2R**	**A2C2A**	**NCB**
#	1480	957	1043	1320

Table 2 shows the obtained results; it contains a confusion matrix with the percentage of correct detection in the diagonal (bolded). As it can be highlighted the proposed algorithm gives very good preliminary results; the worst detection percentage (75%) is obtained for the A2C2R class. This is because it is more difficult to detect the angle orientation of the adult ellipse towards the child one when he gazes the humanoid robot. Other mis-classified errors are given by the wrong detection of the ellipse location. It should be noted that the system works in real time: it consumes around $25ms$ to process each frame and give the estimated behavior class (on a standard PC equipped with Intel I7 processor and 8 GB RAM).

Table 2. Results of the proposed approach

	JA	**A2C2R**	**A2C2A**	**NCB**
JA	**81%**	10%	4%	5%
A2C2R	11%	**75%**	8%	6%
A2C2A	7%	4%	**87%**	2%
NCB	1%	11%	1%	**87%**

4 Conclusions and Future Improvements

This work presented a novel automatic system able to detect joint attention by using completely non-intrusive depth camera installed on the room ceiling. Preliminary experiments conducted on three different video sequences showed that the proposed methodology is able to detect not only the joint attention, but also different interaction behaviors between adult, child and robot, analyzing the head orientation. This real-time system provides an useful non-intrusive tool to report the dominant behavior during the therapy session. Even if the obtained results still does not perform an hit rate of 100%, we point out the fact that, in the context of automatic or semi-automatic evaluation tools, each additional tool that can provide help in supporting therapists in this difficult theme is fundamental: autism, as known, is a generic term to indicate a disorder whose level can vary in a range (called *spectrum*), and each kind of therapy can be useful to produce an improvement in this range, with the (very difficult) goal of producing an exit of the patient from this spectrum.

The future works will be addressed to better estimate the angle among the head and to improve the ellipse detection algorithm. We will also acquire additional sequences during therapy sessions in order to evaluate the algorithm performances in real context. Furthermore, future works will investigate the possibility to create a common dataset in order to provide a comparison measure for this new innovative and non-invasive approach to automatically detect such events, as well as to test the system in clinical settings in real triadic interactions, thus involving therapists and children with ASD.

References

1. Anzalone, S.M., Tilmont, E., Boucenna, S., Xavier, J., Jouen, A.L., Bodeau, N., Maharatna, K., Chetouani, M., Cohen, D., Group, M.S., et al.: How children with autism spectrum disorder behave and explore the 4-dimensional (spatial 3d+ time) environment during a joint attention induction task with a robot. Research in Autism Spectrum Disorders **8**(7), 814–826 (2014)
2. Canny, J.: A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. **8**(6), 679–698 (1986)
3. Carpenter, M., Liebal, K.: Joint attention, communication, and knowing together in infancy (2011)

4. Diehl, J.J., Schmitt, L.M., Villano, M., Crowell, C.R.: The clinical use of robots for individuals with autism spectrum disorders: A critical review. Research in Autism Spectrum Disorders **6**(1), 249–262 (2012)
5. Emery, N.: The eyes have it: the neuroethology, function and evolution of social gaze. Neuroscience & Biobehavioral Reviews **24**(6), 581–604 (2000)
6. Feil-Seifer, D., Matarić, M.J.: Toward socially assistive robotics for augmenting interventions for children with autism spectrum disorders. In: Khatib, O., Kumar, V., Pappas, G.J. (eds.) Experimental Robotics: The Eleventh International Symposium. Springer Tracts in Advanced Robotics, vol. 54, pp. 201–210. Springer, Heidelberg (2009)
7. Gulsrud, A.C., Hellemann, G.S., Freeman, S.F., Kasari, C.: Two to ten years: Developmental trajectories of joint attention in children with asd who received targeted social communication interventions. Autism Research **7**(2), 207–215 (2014)
8. Imai, M., Ono, T., Ishiguro, H.: Physical relation and expression: Joint attention for human-robot interaction. IEEE Transactions on Industrial Electronics **50**(4), 636–643 (2003)
9. Jones, E.A., Carr, E.G.: Joint attention in children with autism theory and intervention. Focus on Autism and Other Developmental Disabilities **19**(1), 13–26 (2004)
10. Khoramshahi, M., Shukla, A., Billard, A.: From joint-attention to joint-action: effects of gaze on human following motion. In: 6th Joint Action Meeting (2015)
11. Li, A.X., Florendo, M., Miller, L.E., Ishiguro, H., Saygin, A.P.: Robot form and motion influences social attention. In: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, pp. 43–50. ACM (2015)
12. Moore, C., Dunham, P.: Joint attention: Its origins and role in development. Psychology Press (2014)
13. Prasad, D.K., Leung, M.K.: Methods for ellipse detection from edge maps of real images. In: Machine Vision - Applications and Systems, pp. 135–162. InTech (2012)
14. Robins, B., Dickerson, P., Stribling, P., Dautenhahn, K.: Robot-mediated joint attention in children with autism: A case study in robot-human interaction. Interaction Studies **5**(2), 161–198 (2004)
15. Rogers, S.J., Dawson, G.: Early Start Denver Model curriculum checklist for young children with Autism. Guilford Press (2009)
16. Scaife, M., Bruner, J.S.: The capacity for joint visual attention in the infant. Nature (1975)
17. Scassellati, B., Admoni, H., Mataric, M.: Robots for use in autism research. Annual Review of Biomedical Engineering **14**, 275–294 (2012)
18. Staudte, M., Crocker, M.W.: Investigating joint attention mechanisms through spoken human-robot interaction. Cognition **120**(2), 268–291 (2011)
19. Stauffer, C., Grimson, W.: Adaptive background mixture models for realtimetracking. In: IEEE Int. Conf. on. Comp. Vision and Patt. Recognition. vol. 2, p. 252 (1999)
20. Warreyn, P., Paelt, S., Roeyers, H.: Social-communicative abilities as treatment goals for preschool children with autism spectrum disorder: the importance of imitation, joint attention, and play. Developmental Medicine & Child Neurology **56**(8), 712–716 (2014)

21. Warreyn, P., Roeyers, H.: See what i see, do as i do: Promoting joint attention and imitation in preschoolers with autism spectrum disorder. Autism **18**(6), 658–671 (2014)
22. Zhao, X., Cusimano, C., Malle, B.F.: Do people spontaneously take a robot?s visual perspective? In: Proc. of the ACM/IEEE Intern. Conf. on Human-Robot Interaction Extended Abstracts, pp. 133–134. ACM (2015)
23. Zivkovic, Z.: Improved adaptive gaussian mixture model for background subtraction. In: Proc. of the IEEE Intern. Conf. on Patt. Recogn. vol. 2, pp. 28–31 (2004)