Tracking posture and head movements of impaired people during interactions with robots

Salvatore Maria Anzalone[1], Elodie Tilmont[2], Sofiane Boucenna[1], Mohamed Chetouani[1], and David Cohen[2]

[1] Institute of Intelligent Systems and Robotics, University Pierre and Marie Curie, 75005 Paris, France, mohamed.chetouani@upmc.fr

[2] Department of Child and Adolescent Psychiatry, Hopital de la Pitie-Salpetriere, University Pierre and Marie Curie, 75013 Paris, France

Abstract. Social robots are starting to be used in assistive scenarios as natural tool to help impaired people in their daily life activities and in rehabilitation activities. A central problem of such kind of systems is the tracking of humans activity in a reliable way. The system presented in this paper tries to address this problem through the use of an RGB-D sensor. State of art algorithms are used to detect and track the body posture the heads pose of each human partner.

Keywords: People tracking, posture estimation, head orientation, human-robot interactino

1 Introduction

Social robotics focus on developing robots able to collaborate with humans as their reliable partners [1] [2]. Particular scenario for such kind of robot is their use as assistive technology. While traditional assistive technologies became more and more robust and reliable [3], researchers started to develop new assistive systems using robots able to help impaired people in a more natural and effective waye [4]. Such kind of applications are very different and each of them uses a particular kind of robot. As instance, due to the rapid ongoing aging of the population, lot of researchers focused on the use of robots to assist elderly people: in this case intelligent environments, domotic tools and robots have been used in conjunction to help elderlies in their daily life activities [5] [6]. Other researches focused on the use of robots for rehabilitation activities. Interesting example of such kind of applications are several studies in which robots have been integrated in therapies with humans, such as improving the attention deficits of autistic children [7]. It is important to point out that researchers focused not only on humanoid robot but also on animal shaped robot or on wheeled robot, accordingly to the constraints of the particular application conceived.

In all these works on assistive robotics, more than in other standard social robotics applications, a key point is the fine perception of the humans. Body behaviours are instinctively used by humans as a natural communication way

with the others [**?**]: a social robot able to interact in a strict and collaborative way with humans, as instance as caregiver of impaired people, should be able not only to recognize the human speech but also to recognize and interpret the body language [8]. Moreover, people behaviours and body movements can contain information that can be useful to the doctors to correctly diagnose and characterize the impairment [9].

The system presented in this paper is a human body tracking system that can be used in conjunction with a robot to organize interactive session of behavioural data acquiring from impaired people. In particular, this system has been conceived as assessment tool of autistic children behaviours during therapeutic sessions involving the use of the robot.

The system proposed uses an RGB-D sensor to detect and to track body features of each human involved in the interactions with a robot caregiver. The state of art about the human behaviour analysis suggests that the body posture, the head movements and the eye gazing are extremely important information on the description of the human behaviours. According to this, the system focuses on the postures, defined as the the tracks of all the joints of the human body, and on the head gesture, that can be seen as a good approximation of the human gaze. Due to technological constraint of such RGB-D sensors used, the eye gazing can not be taken in account: despite its importance, the low resolution of the sensor make the system unable to collect such important information.

Several experiments have been performed to evaluate the performances of the presented system. Results shown in this paper have been collected through the analysis of the behaviour of healthy people, both adult and children.

2 System Overview

The presented system allow robots to gather information about posture and head movements of human partners during direct interactions. In particular, each human that interacts with the robot is modeled through a 3d description that takes in charge about the body posture, in terms of the angles of each limb of the body, and about the head posture, in term of its pitch and yaw. This data is collected by the use of a RGB-D sensor that is conveniently placed in the environment according to its geometry, in order to retrieve a good perspective of the environment.

The system has been implemented through several software modules able to communicate via messages streams or via remote procedure calls, by the use of the ROS platform, Robot Operating System, that offers such kind of capabilities.

As shown in figure 1, data coming from the RGB-D sensor is used by the Skeleton Tracker module to detect the presence of humans in the environment. In this case the module is able to localize and track in three dimensions the articulations of each person. These identified positions are used to select inside the RGB images the areas in which the corresponding heads of each person should appear.A Face Tracking module is then able to detect the faces from these selections, to estimate their pose in terms of pitch and yaw. Two face trackers

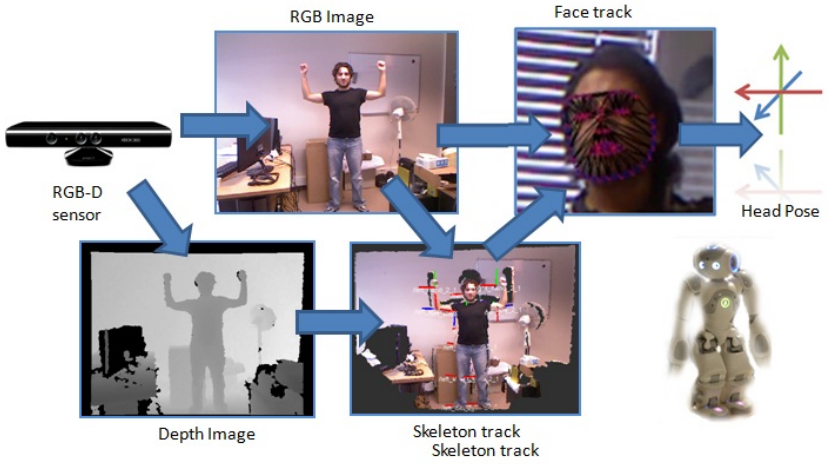

Fig. 1. The gaze recognition pipeline: skeleton traking; head detection; gaze estimation.

based on Active Appearance Models [10] have been used: one based on the Constrained Local Models algorithm [11]; another one based on the Generalized Adaptive View algorithm [12]. chosen.

Data collected by this system can be used in real time to obtain a feedback about the human activity: the robot can use the collected information to act in a coherent way. Moreover, the same data can be treated offline in order to extract information that can be relevant for the doctors in the assessment of the impairment.

3 People Detection

Human activity is detected by the use of the 3d information perceived by the RGB-D data through the use of a multiple skeleton tracking system provided by OpenNi [13]. In particular, each person will be tracked in the space in terms of their joints information. This data will permit the retrieving of posture, limb and arm movements and gestures information.

Depth information acquired by the sensor is elaborated to distinguish the body of each person from the environment. This is achieved through a background subtraction technique applied to the depth image.

The depth image left is then segmented and classified according a per-point approach, labeling each depth point of the body as a particular body part. As shown in figure 2, a total of 31 body patches distributed among the body has been considered. The body patch labeling uses depth invariants and 3D translation invariants features that try to describe to which part of the body each depth pixel belongs to. The classification process is based on a randomized decision forest of such features that results in a dense probabilistic skeleton with body parts labeled accordingly. The training process of such classifier employed

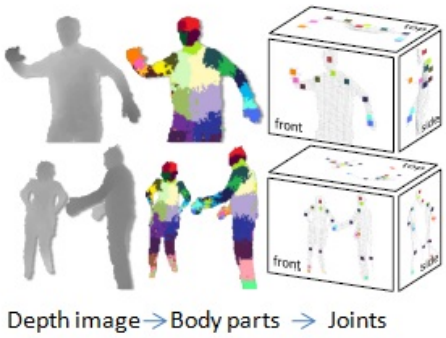

Fig. 2. The process of extraction and traking of people joints (Images from [13]).

a database of 500k labeled frames captured through motion capture in hundred of different scenarios, such as dancing, kicking, running. Each estimated patch is finally employed, according to its density, to extract the position of each joint of the body, correspondent to each patch.

4 Head Tracking

Once located an estimation of the 3d position of the head of each human in the environment, the system reproject it to the RGB image, in order to select on it a a section in which the head of each subject should appear. A face tracker algorithm is then applied to the crops in order to estimate a model in terms of pitch and yaw of each head found.

Two algorithms have been used: the Constrained Local Models algorithm and the Generalized Adaptive View-based Appearance Model. Both of them can be used with three dimensional, depth map based data, but, due to the huge noise of the sensor used, the simple use of RGB images has been chosen.

The CLM algorithm follows the Active Appearance Model approach: both try to model faces via a statistical description based on a set of landmarks. Shapes of faces are deformed iteratively according to the landmarks positions, to find a best fit with the actual face image. In particular, standard AAM algorithm tries to compute a model of faces according to their shape and their appearance. The shape model is calculated from a set of keypoints spread over a face, as its contour, the border of the lips, the nose and the eyes. In particular, a data set of labeled faces has been chosen as training set of the algorithm. The shape model is obtained from the mean and the variance of the PCA transformation of the facial key points of all the faces in the training set. It will be described as the mean shape parametrized by the variance. The appearance model will be built by normalizing the grayscale image of the face, wrapping it over its mean shape. Also in this case, the model will be obtained from the PCA transformation of the wrapped faces in the training set, described by its mean and parametrized

according to the variance. The shape model and the appearance model will be fused to obtain a full face model using another PCA transformation: this will result in a parametrized model able to take in account both shape and appearance of the faces. During its normal usage, the AAM algorithm will calculate the error between the current model and the its actual appearance. The error will control the change of the model parameters iteratively, to better approximate the actual appearance, minimizing it. In particular, the error will encode how the parameters of the model should be changed: this relation is learnt in the training step and is used iteratively, providing also high speed performances to the whole system.

The CLM algorithm can be seen as a slightly variation of the AAM algorithm: also in this case the face model is composed by a conjunction of space model and appearance model. However, in this case the appearance model is built employing local features: a patch of pixels around each keypoint is considered, instead of using the whole wrapped face as in the AAM algorithm. Moreover, also the model fitting is different: to adapt the parameters to the actual face the Nelder-Meade simplex algorithm is employed.

The GAVAM algorithm tracker follows a different path: it tries to integrate several state of art approaches in order to obtain a reliable head pose estimation. In particular, a static pose estimator has been fused with a derivative tracker and with a keypoint-based pose estimator. The static pose estimator is able to recognize faces in a single frame, but neglects all the useful temporal information. This issue is coped by a derivative approach: the head position and orientation is tracked through the frames sequence. However, this last system gives high precision in short time scales, but its accuracy becomes very weak over the time. Then, this is integrated with a local keypoint approach that uses templates, in a similar way of AAM and CLM, to track the head over the time.

5 Experimental Results

The system has been evaluated in direct interaction scenarios, involving a Nao robot. A population of both adult and children has been chosen to estimate the performances of the system.

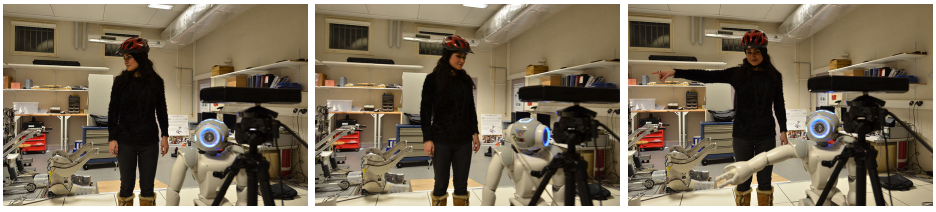

Fig. 3. Data capture session: motion capture makers are placed on the helmet and on the arm.

A motion capture system have been used to capture the pose of the head of three adult subjects. Markers have been placed over the ears, on the left and on the right, and above the forehead. Such information has been used as ground truth of the system.

As shown in figure 3, each human partner was asked to stand in front of the robot, at 1.5mt from the RGB-D sensor, and to imitate the robot movements, that turned his head on the left and on the right, for three times.

Head Pose	Gavam	CLM
Pitch	61%	49%
Yaw	79%	93%
Overall	70%	71%

Table 1. Head's pose estimation performances using Gavam and CLM approaches.

Data obtained by presented system has been evaluated, correlating it with the ground truth perceived by the motion capture: the whole performances of the two algorithms are similar, but, as shown in table 1, a fine analysis of the the pitch recognition and the yaw recognition results shows that the best performances are obtained by the usage of both algorithm together: the Gavam approach to recognize the heads pitch, while the CLM approach to recognize the heads yaw.

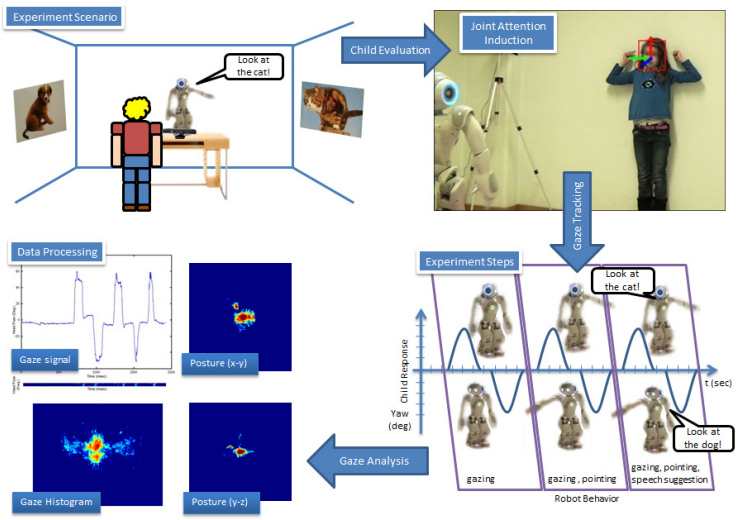

Fig. 4. A children during the joint attention induction experiment.

Other experiments have been conducted with 15 healthy children in an age between 6 and 10 years old. The goal has been to collect data of their behaviour in

a simple joint attention experiment [14] [15]. As show in figure 4, the robot Nao tries to induce in them joint attention by looking towards two animal figures placed on the two opposite sides of the room, alternatively. The induction is repeated three times: the first time the robot just looks towards each figure; the second time it tries to add more informative content by looking and pointing each figure; the third time it will add to these behaviours a vocalization of the object, look at the cat, look at the dog.

Head pose and body posture information of each child can be recorded and analyzed off-line. In particular, the head pose estimation has been collected using the Gavam approach to retrieve the heads pitch and using the CLM approach to retrieve the heads yaw, according to the results here discussed. Figures 5 show the heads yaw of three typical healthy children: each of them respond to the joint attention induction pushed by the Nao robot, moving the head towards the left and the right sides.

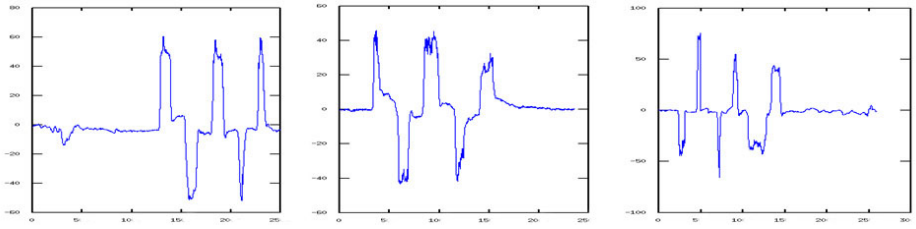

Fig. 5. The head's yaw variation among the time (deg/sec) during the joint attention experiment of three different healthy children.

A different way to gather important information about the head pose behavior of the children is by considering the histogram of their head movements, on the yaw-pitch plane. The histogram in figure 6 shows how the attention of the child is captured on the left, on the right and on the center, characterized to the three spots corresponding to the two focus of attention, the figures of animals on the two sides, and to the robot, placed just above the RGB-D sensor, in front of the child.

The same histogram can be assessed for the children pose: in particular, figure 7 shows the histogram of the displacements from the average positions of the body.

6 Conclusions and Future Works

In this paper a system able to recognize and track human body activities has been presented. The system is conceived for human robot interaction contexts, in particular where the robot is conceived as an assistive tool for impaired people. Using a RGB-D sensor, the system is able to capture human postures and head

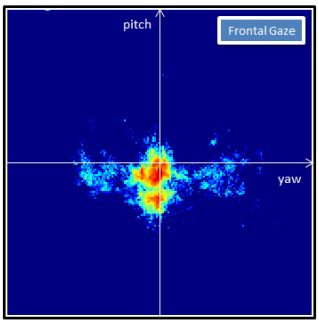

Fig. 6. The average histogram of the head movements (yaw-pitch) of healthy children.

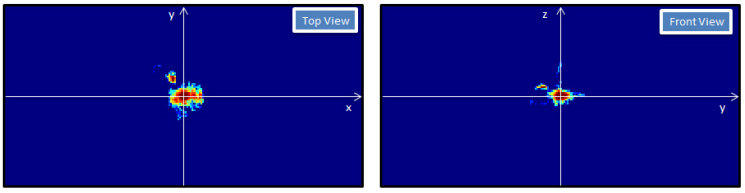

Fig. 7. The average histogram of the position (top-view and front-view) of healthy children.

poses. The information collected can used as real-time input of a robotic system, as well as, to help doctors in the correct assessment of an impairment.

Several experiments have been conducted to evaluate the performances of the system in real contexts. Results shown the potentialities and the lacks of this approach. Results also encourage on the use of such system as an assessment tool of able to retrieve social engagement cues. In particular, the system will be used to help therapists on stimulating autistic children in joint attention activities.

Acknowledgements

Authors would like to thank to Dr. A. Carbone and T. Luiz for their kind collaboration. The current study was supported by a grant from the European Commission (FP7: Michelangelo under grant agreement n. 288241), and the fund Entreprendre pour aider.

References

1. C.L. Breazeal. *Designing sociable robots.* The MIT Press, 2004.
2. C. Breazeal. Toward sociable robots. *Robotics and Autonomous Systems*, 42(3-4), 2003.

3. Dorit Maor, Jan Currie, and Rachel Drewry. The effectiveness of assistive technologies for children with special needs: a review of research-based studies. *European Journal of Special Needs Education*, 26(3):283–298, 2011.
4. Adriana Tapus, Maja J Mataric, and Brian Scassellati. Socially assistive robotics. *IEEE Robotics and Automation Magazine*, 14(1):35, 2007.
5. Amedeo Cesta, Gabriella Cortellessa, M Vittoria Giuliani, Federico Pecora, Massimiliano Scopelliti, and Lorenza Tiberio. Psychological implications of domestic assistive technology for the elderly. *PsychNology Journal*, 5(3):229–252, 2007.
6. Salvatore M Anzalone, Stefano Ghidoni, Emanuele Menegatti, and Enrico Pagello. A multimodal distributed intelligent environment for a safer home. In *Intelligent Autonomous Systems 12*, pages 775–785. Springer Berlin Heidelberg, 2013.
7. Giovanni Pioggia, Roberta Igliozzi, Marcello Ferro, Arti Ahluwalia, Filippo Muratori, and Danilo De Rossi. An android for enhancing social skills and emotion recognition in people with autism. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, 13(4):507–515, 2005.
8. M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino, and C. Yoshida. Cognitive developmental robotics: a survey. *Autonomous Mental Development, IEEE Transactions on*, 1(1):12–34, 2009.
9. Brian Scassellati, Henny Admoni, and Maja Mataric. Robots for use in autism research. *Annual Review of Biomedical Engineering*, 14:275–294, 2012.
10. Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(6):681–685, 2001.
11. David Cristinacce and Tim Cootes. Feature detection and tracking with constrained local models. In *Proc. British Machine Vision Conference*, volume 3, pages 929–938, 2006.
12. L-P Morency, Jacob Whitehill, and Javier Movellan. Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–8. IEEE, 2008.
13. J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 1297–1304, Washington, DC, USA, 2011. IEEE Computer Society.
14. H. Sumioka, K. Hosoda, Y. Yoshikawa, and M. Asada. Acquisition of joint attention through natural interaction utilizing motion cues. *Advanced Robotics*, 21(9):983–999, 2007.
15. Y. Yoshikawa, T. Nakano, M. Asada, and H. Ishiguro. Multimodal joint attention through cross facilitative learning based on μx principle. In *Development and Learning, 2008. ICDL 2008. 7th IEEE International Conference on*, pages 226–231. IEEE, 2008.