Human action recognition oriented to humanoid robots action reproduction

Stefano Michieletto and Emanuele Menegatti

Intelligent Autonomous Systems Lab (IAS-Lab)
Department of Information Engineering (DEI)
Faculty of Engineering, The University of Padua
Via Ognissanti 72, I-35129 Padova, Italy
{stefano.michieletto,emg}@dei.unipd.it
http://robotics.dei.unipd.it

Abstract. Our research aims at providing a humanoid robot with the ability of observing, learning, and reproducing actions performed by humans in order to acquire new skills. In other words, we want to apply artificial intelligence techniques to automatically recognize a human activity in order to make a humanoid robot able to reproduce it.This system has not only to distinguish between different actions, but also to represent them in a proper manner to allow a robot to reproduce the motion trajectories the demonstrator showed and learn new skills. Since the final system is going to be integrated in an autonomous humanoid robot (specifically model Aldebran Nao), we are working with an RGB-D sensor (Microsoft Kinect) that can be easily applied to it. This objective introduces also strict real-time constrains to the action recognition algorithm: we have opted for a probabilistic approach that offers good modeling and fast recognition performances.

Keywords: Action recognition, humanoid robots, programming by demonstration, imitation learning

1 Introduction

When the Turing Test was proposed, in 1950 [9], the idea was to compare humans and machines by looking at their responses. We would like to extend this criteria and compare humans and machines by looking at their actions. This regards the abilities to perceive the world, to learn from what happen around themselves, and the way they move with respect to the rest of the environment, that is the hallmark of an intelligent system. This project aims to provide a robot the capability to perform a task, recognizing actions from human demonstrators and learning from their movements.

A lot of work has been done from the 1950 in these fields, in particular in computer vision and robotics research. Computer vision goal is to extract information from a scene and structure it, and action recognition is assuming more and more importance in this field for its correlation with wide range of

application such as biometrics, animation and interactive environments, video analysis and surveillance. The identification process can be divided into three stages: feature extraction, feature representation and action classification [10]. Two different approaches characterize features extraction: *model-based* and with *no-model* assumption. The first technique takes advantage of the prior knowledge of the body model to extract the selected features from the scene. The second one, instead, uses a common features extraction method. The features representation is strictly connected to the extraction phase. The *model-based* methods usually keep track of position and orientation of each part that composes the model. *No-model* methods, on the other hand, maintain the features characterize themselves: surface normals, RGB features, time-space information of each point and so on. The action classification can be approached by comparing static images with pre-stored samples (*template-based*[1]), by describing the system with a series of states with a certain probability and allowing switches between different states (*probability-based*[4][6]), or representing the motion with a sequence of grammatical rules (*grammar-based*[2]).

Robotics was defined as the science which studies the intelligent connection between perception and action[8] and mixes together mechanics, controls, computers and electronics, so that it benefits from advances in the different technologies. Recently, robots has been used in more and more areas such as *service robotics*, *field robotics*, or *human augmentation*. By the dawn of the new millennium, the "human" factor pushes robots to rapidly assume anthropomorphic appearance and the more robots interact with people safer, smarter and more independent they need to be, that why other disciplines were involved in the robotics area. The human ability to imitate behaviors is already present in even 12 - 21 day old infants [5]. Despite some hypothesis about the translation between visual and motion features[11], the mechanisms that express a demonstration to a set of internal motor primitives remain largely unknown. Several works in robotic imitation suggest that robots can perform new skills by learning key information from a demonstration [7]. More recently, this idea has been applied on humanoid robots as a natural way to teach how to perform a task in a dynamic environment [3].

The remainder of the paper is organized as follows: Section 2 describes the acquisition system. In Section 3 the features extraction method and their representation are introduced, while the action classification is described in Section 4. Conclusions and future works are contained in Section 5.

2 Acquisition system

The development of affordable RGB-D sensors is giving a rapid boost in computer vision research. These sensors are also reliable: they allow to natively capture RGB and depth information at good resolution and frame rate (the Microsoft Kinect takes 30 frames per second at 640x480 pixels of resolution). On the other hand, the sunlight plays havoc with the infrared pattern projected by the sensor and they cannot be used over eight meters. An indoor environment

is not a real limitation for the humanoid robot we will use, so mounting RGB-D camera can be a very rich source of information for our work. The not invasive impact is another advantage of such sensors with respect to motion sensors and the robot can learn from every person in the scene.

In order to simplify the initial setting, the RGB-D sensor will not be immediately mounted on the robot. At first, the action recognition algorithm will work as a stand-alone system, hence we can assume that the data can be elaborated off-line. This assumption allows us to compare our algorithm with others at the state of the art, without the constrain of a camera moving on a robot. For this purpose, we are thinking of releasing a dataset of actions acquired as RGB-D data. Figure 1 shows some people performing actions for our preliminary tests. Once the algorithm will be efficient and robust enough, the robot movement will be considered and the system will be adjusted to prevent the performances from being affected by the robot movements.

(a) (b) (c) (d)

Fig. 1. Actions collected in our preliminary tests: pointing (a), sitting down(b), walking (c), and waving (d)

3 Features extraction and representation

A *model-based* approach will be adopted for features extraction. At the beginning of the project the OpenNI[1] freeware implementation of skeleton tracker will give us information about 25 skeletal joints covering the entire body. For each joint we will record position and orientation with respect to the camera frame. The `ROS::tf`[2] standard message will be used in order to exchange information between the acquisition process and the other algorithms we will implement. This choice is due to ROS[3] as common framework in this work. An example of the 3D model obtained from the tracker is showed in Figure 2. The OpenNI skeleton tracker is not open source and it will be probably replaced with a different one, in order to allow us to modify some internal parameters and improve the system.

[1] http://openni.org/
[2] http://www.ros.org/wiki/tf/
[3] http://www.ros.org/

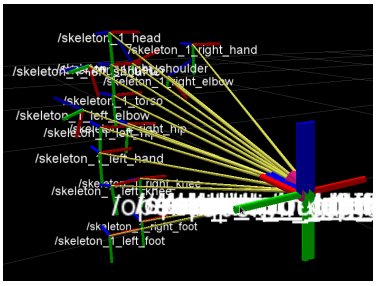
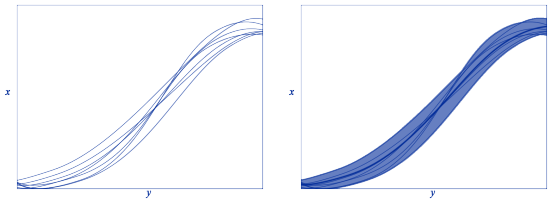


Fig. 2. 3D skeleton model (joints data obtained using OpenNI tracker). Each coordinates system (the RBG axis) represents position and rotation of a skeleton joint with respect to (yellow arrows) the camera frame.

4 Action classification

The action classification will lean on a *probability-based* model. Having a probabilistic model allows us to classify actions in a compact and efficient representation extracting the principal action constraints. This peculiarity is fundamental in working with actions involving a large number of degree of freedom, like in the human body. In fact, the demonstrations can vary one from each other and some movements could be not essential to reach the goal. A probabilistic approach makes also possible to perform incremental construction of a model, thus making easier to refine the representation with new demonstrations. The purpose of this work is make the robot able to reproduce the skills learned using the action recognition system. A regression process will be developed to archive an effective robot control with smooth trajectories in a fast and analytic way. The idea is to extract a "mean" trajectory that would change in the "variance" field of the acquired trajectories depending on the situation. Figure 3 shows the idea for an example trajectory in 2D space.

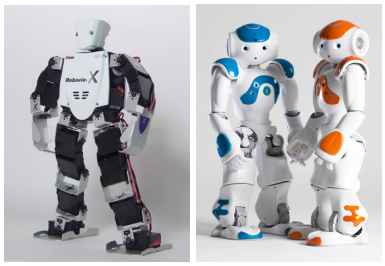


(a) (b)

Fig. 3. Example of 2D trajectories extraction (a) and their regression (b).

5 Conclusion

In this paper we presented an human action recognition system designed to reproduce the acquired skills on a humanoid robot. We plan to test our system first in a simulated environment, and then with two different humanoids: VStone Robovie-X and Aldebaran Nao (Figure 4). Robovie-X is an affordable small robot perfect to test robustness and flexibility. Nao is bigger than Robovie-X and it can mount on it the RGB-D sensor to perform real-time tests on a moving robot.



(a) (b)

Fig. 4. Robot proposed for our tests: Robovie-X (a) and Nao (b).

References

1. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. Pattern Analysis and Machine Intelligence, IEEE Transactions on 23(3), 257–267 (Aug 2002)
2. Brand, M.: Understanding manipulation in video. In: Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition (FG '96). pp. 94–. FG '96, IEEE Computer Society, Washington, DC, USA (1996)
3. Calinon, S.: Robot Programming by Demonstration: A Probabilistic Approach. EPFL/CRC Press (2009)
4. Gong, S., Xiang, T.: Recognition of group activities using dynamic probabilistic networks. In: Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2. pp. 742–. ICCV '03, IEEE Computer Society, Washington, DC, USA (2003)
5. Meltzoff, A.N., Moore, M.K.: Imitation of Facial and Manual Gestures by Human Neonates. Science 198(4312), 75–78 (Oct 1977)
6. Moënne-Loccoz, N., Brémond, F., Thonnat, M.: Recurrent bayesian network for the recognition of human behaviors from video. In: Proceedings of the 3rd international conference on Computer vision systems. pp. 68–77. ICVS'03, Springer-Verlag, Berlin, Heidelberg (2003)
7. Schaal, S.: Learning from demonstration. In: Advances in Neural Information Processing Systems 9 (1997)

8. Siciliano, B., Khatib, O. (eds.): Springer Handbook of Robotics. Springer (2008)
9. Turing, A.M.: Computing machinery and intelligence (1950)
10. Wei, W., Yunxiao, A.: Vision-based human motion recognition: A survey. In: Proceedings of the 2009 Second International Conference on Intelligent Networks and Intelligent Systems. pp. 386–389. ICINIS '09, IEEE Computer Society, Washington, DC, USA (2009)
11. Wolpert, D.M., Ghahramani, Z., Jordan, M.I.: An internal model for sensorimotor integration. Science 269, 1880–1882 (1995)