UoB, CS
Report 3
Title: SM of HE with a MR
Student: MABP
Supervisor: NH
Thesis Group: JW PH

Contents

Chapter 1

Introduction

One of the main goals in AI is having robots working autonomously in everyday environments. A robot in this kind of situation is expected to perceive, understand and interact with his environment. However, the environment is dynamic, non-structured and non-deterministic, which makes difficult for a robot to fulfil the assigned tasks. To be able to sort these obstacles, robots need to be provided with cognitive skills.

Everyday environments have many valuable features that a robot needs to understand, among them are human activities. They are a meaningful manifestation of human behaviour. They are important for a robot in order to be able to understand the role of humans in a particular environment, and the occurring interactions with objects and with the environment.

1.1 Human activity analysis with a mobile robot

Activity recognition is the research field that studies the automatic detection and analysis of human activities from the information acquired from sensors [Aggarwal and Xia, 2014]. In the AI context, it is closely related with perception and knowledge processing. The problem of activity recognition has been treated from different perspectives, however, computer vision has been the most popular approach to use.

In principle, robots with appropriate sensing capabilities can perform activity recognition. Moreover, they have some advantages over the use of fixed cameras or wearable devices as they are able to interact with the environment. They are active observers, i.e. they can change their point of view on scene and be selective in the areas of the environment that are more interesting. On the other hand, they have some disadvantages as well. They

don't have omnipresence, so they are not able to sense the environment and will loose information. Also, their sensory data may be noisy or blurry due to movement, erratic hardware, changing environmental conditions, etc. Finally, robots are expected to work in real-time, so online activity recognition would be desirable, however, this puts time constraints in the deliberation process.

The target problem in this project is the study of activity recognition performed with a robot. Particularly in the case where there is not complete information from the environment to have a clear match between the observations and the activity patterns. Here, an interpretation can still be made using previous experience and domain knowledge. Even, if a totally certain interpretation of the scene is not possible, a partial one can still be done with a list of the most probable situations to be happening. This also can be used by a robot to decide to perform new observations of the scene to improve its reasoning conclusions. The chosen technique to do this is Answer Set Programming (ASP).

1.2 Test case: "The library scenario"

To study the proposed problem, a test scenario is presented here.

The School of Computer Science at the University of Birmingham have a library, mostly used by students. An attendant is in charge of the book loans and retrievals, and also to help users using the facility. The physical scenario is basically a big room. It has some cabinets (where bibliographic material is stored), a reception, some tables and chairs and a printing desk. It only has one entrance.

Users mostly use the facility to study, to consult material, to print, to work in team, to do work in PC or simply as a rest area. Because of the rules of the library and nature of the scenario, the amount of activities is restricted by the domain. However, some other activities could appear as using a cellphone, talking, packaging things inside a backpack, etc. The objects involved in the scenario is relatively small (books, tables, chairs, laptops, cabinets, etc.).

Chapter 2

Related Work

2.1 General antecedents - Perception in AI

Perception, as a cognitive process, has been studied widely in Psychology. Some of these studies refer on how information is processed and which parts of it are essential to make sense properly of a sensory input. In [Heider and Simmel, 1944], an animated film was created using only moving polygons to demonstrate how the motion of abstract entities can be interpreted by a human observer. In [Johansson, 1973], locomotion patterns of living organisms using visual marks are studied. By this mean, the emphasis was put in the motion pattern created by the marks rather than in the moving body, whose dimensions and shape were unknown.

In Artificial Intelligence, perception has been treated mostly by the computer vision research community. Earlier works can be traced back to the 1960s, as part of the effort to mimic human-like intelligence using visual perception components. The main difference between computer vision and image processing has been the desire to recover the three-dimensional structure of the world from images, and to use this as a stepping stone towards full scene understanding [Winston and Horn, 1975].

One of the earlier works in 3D reconstruction from a single image is found in [Roberts, 1963]. The developed system as able to reconstruct geometrical bodies with flat surfaces by recognizing the borders of the bodies in the scene and later analysing the shades of their visible surfaces. In the late 1960s, the *block's world* was used as a test scenario for intelligent systems, particularly regarding knowledge representation, reasoning and planning. In the block's world, the actual state A and a desired future state B of the environment are given. The goal is to autonomously generate a plan to transform A into B by the manipulation of the blocks in scene. One important characteristic

of the problem is the requirement of a symbolic representation. The problem was used as a test case for the robot Shakey [Nilsson, 1984]. In [Barrow and Popplestone, 1971] object recognition was studied by decomposing an image into regions and describing the spatial relations between them, in a more qualitative, rather than the traditional quantitative, approach.

2.2 Activity Recognition

Activity recognition is an important research area in the context of automated perception. It has many applications as surveillance, inspection, verification, generation of automated reports, etc. The main goal is to automatically analyse the ongoing activities from a sensory source (a video sequence in most of the cases).

Activities play a relevant role in the interpretation of a scene, not only in physical terms (space and time), but also symbolically as they can usually be associated with a meaning and a domain.

Human activities are difficult to classify because they cover a broad range of situations in different contexts, and they depend on many parameters. Regarding their complexity, activities can be treated as hierarchical entities because high-level activities are usually composed of simpler actions. In [Turaga et al., 2008], two non-exclusive categories are used: actions and activities. The first one is used for simple actions performed preferably by a individual, and activities are treated as a complex sequence of actions performed by several individuals. In [Aggarwal and Ryoo, 2011], a four layers categorization is proposed:

Gestures Elementary movements of a person's body part, and are the atomic components describing a meaningful motion of a person. E.g. 'stretching an arm', 'raising a leg'.

Actions Single person activities that may be composed of multiple gestures organized temporally. E.g. 'walking', 'waving'.

Interacions Human activities that involve two or more persons and/or objects. E.g. 'Two persons fighting', 'a person eating an apple'.

Group activities The activities performed by conceptual groups of multiple persons and/or objects. E.g. 'a football team playing a match', 'a group of students making an exam'.

The research in activity recognition goes in different directions depending on many factors as specific domains (e.g. robotics, gaming), features of interest in the scene (e.g. abnormality detection, verification), sensory capabilities

of the system (e.g. computer vision, pervasive). It usually reflects interest in specific parts of the problem as sensing, scene reconstruction, representation of activities, pattern recognition, reasoning, etc.

In [Aggarwal and Ryoo, 2011] a taxonomy is proposed to organize the research in the area, Fig. 2.1.

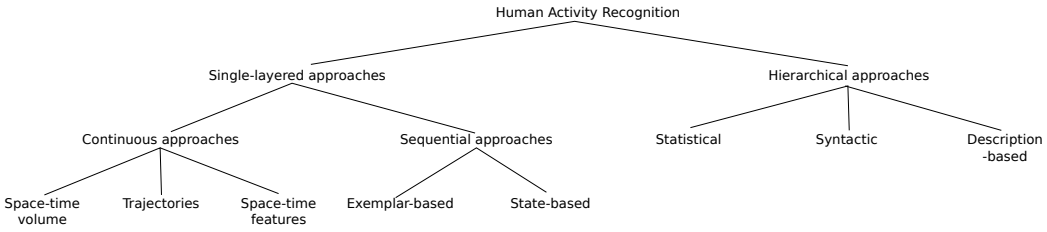

Figure 2.1: The taxonomy of research in activity recognition described in Aggarwal and Ryoo [2011].

2.2.1 Single-layered approaches

The representation and recognition of activities is performed using raw sensory data[1]. Sensory data is processed to obtain particular descriptive features of the scene which are compared with known activity patterns. These patterns can be obtained in a supervised or unsupervised fashion (e.g. common occurrences of a specific action). Most of them are based in computer vision and machine learning techniques.

Single-layered approaches are suitable for recognition of short-term and simple activities as gestures, movements of the body or simple interaction with an object. This is because the amount of collected data, which is going to be processed, grows very easily and long-term activities would require the processing of larger amounts of data. Also, because activities are not always performed in the same way, even by the same person; the shorter the activity, the more accuracy that will be attained. An finally, because they are dependant on the conditions of the sensors and from the environment (e.g. different light conditions, a particular point of view) and noise should be considered.

[1]The original survey [Aggarwal and Ryoo, 2011] describes single layered approaches as image-based approaches, but it leaves out the systems with other sensing capabilities (e.g. 3D sensors, sonars). However, they can be included too if it is clear that the representation and recognition of activities is based in the processing of raw sensory data.

Continuous approaches[2]

The activities are recognized by analysing continuously sensory data projected in time. A volume (or hyper-volume) is built from the sensory data and particular features are extracted and compared with known patterns. The dimension of the data will depend on the sensing capabilities of the system; for example, video analysis would require 3 dimensions (X, Y, T) and a RGBD camera would be able to use 4 dimensions (X, Y, Z, T), etc. Continuous approaches can also be classified depending on the features that are used to describe activities (volumes, trajectories, point descriptors, etc.).

In [Bobick and Davis, 2001] a video signal of aerobics exercises is analysed by attaching to every pixel a vector indicating the precence of motion and the recency of motion in a sequence. The vector sequences are compared in time with known pattern of exercises. The system was able to recognize the activities in real time, and with a linear temporal temporal variance.

In [Ke et al., 2007] activity recognition is performed by extracting from a video signal sequences that are similar to the known activity pattern, using a shape-based representation. Then a volume is built by concatenating the video frames in time. Similar neighbour regions are then clustered to create a volume. Finally the shapes of the volumes are compared to known patterns of activities.

Sequential approaches

Sequential approaches look to recognize activities by analysing a sequence of extracted features. First the sensory data is processed to extract particular features, which will be concatenated in time. Then sequential methods are applied to to search for sequences that coul eventually match with the pattern of a known activities. In case of that the sequence is corrupted, a probabilistic approach can be applied to decide the occurrence of the activity.

Sequential approaches can be classified depending on the used recognition methodology in exemplar-based and model-based.

In exemplar-based sequential approaches a sequence in time is created by extracting particular features from the incoming sensory data. In [Veeraraghavan et al., 2006] high-level actions are treated as a sequence in time of atomic actions. The same activity performed in two different occasions can create two different sequences because of an execution in a different speed. The authors develop a method to learn the the pattern time variances in the activity sequence to be able to recognize activities performed at different speeds, or with eventual pauses.

[2] 'Space-time approaches' in [Aggarwal and Ryoo, 2011]

The second one, creates a sequence in time of states, which might be separated in time, and creates a statistical model to test the belonging of that pattern to a known class of activity. Hidden Markov models (HMMs) and Dynamic Bayesian networks (DBNs) are widely used in this approach. The activity is modelled in terms of hidden states and then transition probabilities are trained. The model will reflect the similarity of a sequence of states with a probability value. These methods can be robust in realistic cases where the sequence of states is corrupted or incomplete.

The first work to use probabilistic graphical models to recognize activities is [Yamato et al., 1992]. They transformed a sequence of images into an image feature vector sequence, and then into a symbol sequence by vector quantization. They used a set of HMMs to model the activities to be recognized and dataset to optimize the parameters of the model. Their results reflect a good and reliable performance of HMMs to model human activities.

2.2.2 Hierarchical approaches

They describe high-level activities in terms of simpler ones, building multiple layers that are suitable to represent complex activities.

Hierarchical approaches can be classified regarding the used recognition methodology as statistical, syntactical and description-based.

2.2.3 Statistical

They are based in the construction of statistical state-based models concatenated hierarchically (e.g. layered hidden Markov models) to represent and recognize high-level human activities.

Syntactic

A grammar syntax is used to model sequential activities (e.g. stochastic context-free grammar). By this mean, a high level activity is represented as a string of atomic level activities that takes part.

Description-based

Activities are represented by the description of sub-events and their spatial, temporal and logical structures.

Bibliography

J. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, 43(3):16, 2011.

J. Aggarwal and L. Xia. Human activity recognition from 3d data: A review. *Pattern Recognition Letters*, 48(0):70 – 80, 2014. ISSN 0167-8655. doi: http://dx.doi.org/10.1016/j.patrec.2014.04.011. URL `http://www.sciencedirect.com/science/article/pii/S0167865514001299`. Celebrating the life and work of Maria Petrou.

H. Barrow and R. Popplestone. Relational descriptions in picture processing. In *Machine Intelligence 6*, page 377, 1971.

A. Bobick and J. Davis. The recognition of human movement using temporal templates. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(3):257–267, Mar 2001. ISSN 0162-8828. doi: 10.1109/34.910878.

F. Heider and M. Simmel. An experimental study of apparent behavior. *The American Journal of Psychology*, pages 243–259, 1944.

G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2):201–211, 1973. ISSN 0031-5117. doi: 10.3758/BF03212378. URL `http://dx.doi.org/10.3758/BF03212378`.

Y. Ke, R. Sukthankar, and M. Hebert. Spatio-temporal shape and flow correlation for action recognition. In *CVPR*. IEEE Computer Society, 2007. URL `http://dblp.uni-trier.de/db/conf/cvpr/cvpr2007.htmlKeSH07`.

N. Nilsson. Shakey the robot. Tech Note 323, AI Center, SRI International, 1984.

L. G. Roberts. *Machine Perception of Three-Dimensional Solids*. PhD thesis, Massachusetts Institute of Technology, June 1963.

P. K. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Trans. Circuits and Systems for Video Technology*, 18(11):1473–1488, Nov. 2008. URL `http://dx.doi.org/10.1109/TCSVT.2008.2005594`.

A. Veeraraghavan, R. Chellappa, and A. Roy-Chowdhury. The function space of an activity. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 959–968, 2006. doi: 10.1109/CVPR.2006.304.

P. H. Winston and B. Horn. *The psychology of computer vision*. McGraw-Hill computer science series. McGraw-Hill, New York, 1975. ISBN 0-07-071048-1. URL `http://opac.inria.fr/record=b1083572`. Includes index.

J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR '92., 1992 IEEE Computer Society Conference on*, pages 379–385, Jun 1992. doi: 10.1109/CVPR.1992.223161.