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"

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

The general problem to study in this project is activity recognition with a mobile robot. In this chapter, relevant related work is reviewed.

In humans, activity recognition is a cognitive skill that can be considered mainly into perception. The basis to understand it relies first in Psychology, because it provides the concepts and the evidence of how the mind is constituted (2.1). The next step is to look at the possibilities to mimic a cognitive process into a machine, this problem has been studied widely in Artificial Intelligence. In particular, activity recognition has been studied in Computer Vision. Finally, the problem has to landed to a robotic stage, making emphasis on the advantages and disadvantages of a robotic platform.

2.1 General antecedents - Perception in AI

Perception, as a cognitive process, has been studied widely in Psychology. It refers to the process of organizing and interpreting sensory information so that it has a meaning [King, 2014]. Part of the interest is about how sensory information is processed by the brain, and which parts of it are essential. Also relevant is the domain knowledge that the subject has about a particular context. Together, the sensory input and the domain knowledge are used to interpret a scene.

Sensory input is important for perception, however, not all the data is equally important to interpret a particular scene and conclusions can still be made, even with partial data. In [Heider and Simmel, 1944], an animated film was created using only moving polygons to demonstrate how the motion of abstract entities could be interpreted by human observers in meaningful ways. In [Johansson, 1973], locomotion patterns of living organisms using visual marks were studied. By this mean, the emphasis was put in the

qualitative motion description of the marks rather than in the qualitative motion description of the moving body.

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 was 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 [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.

Since the early 1970s, 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, an initial state A and a desired 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. One important characteristic of the problem is that requires a symbolic representation of the scene. The problem was used as a test case for the robot Shakey [Nilsson, 1984].

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 application will dictate the approach to follow and the kind on sensors that will be required.

First, regarding sensing, two approaches can be followed, environmental and/or pervasive. The first one observes the scene from the distance as it happens with a CCTV camera or a robot. The pervasive approach relies on wearable devices to detect the activity of a person from a first person point of view.

Another possible classification of activity recognition systems focuses on how information is processed. In [Aggarwal and Ryoo, 2011] a taxonomy is proposed as shown in Fig. 2.1.

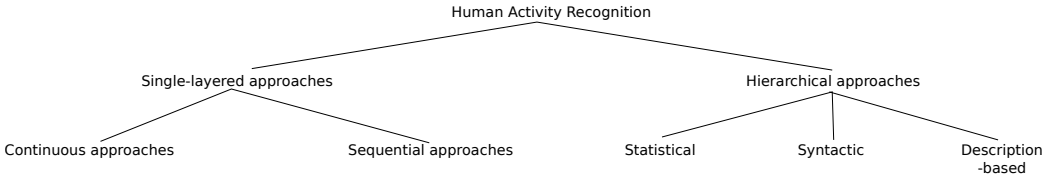

Figure 2.1: The taxonomy of research in activity recognition described in Aggarwal and Ryoo [2011].

2.2.1 Single-layered approaches

They represent activities in terms of raw sensory data[1], because of this, the activity descriptions are trained from datasets.

Single-layered approaches are suitable to recognize short-term and simple activities as gestures, movements of the body or simple interactions with objects. This is mainly because the amount of sensory data grows very easily and long-term activities would require to process larger amounts of data. Also, because activities are not always performed in the same way, even by the same individual; the shorter the activity, the more accuracy that will be attained. An finally, because they are dependant on the sensors and on the environmental conditions (e.g. lighting, point of view).

Continuous approaches[2]

The activities are recognized by analysing continuous sensory data and compare it with an activity pattern.

An activity is represented as a block of data along time where the activity was performed, and it is considered as a whole. A volume (or hyper-volume) is built by concatenating the sensor readings in time. The dimension of the data will depend on the sensing capabilities of the system; for example, a video stream would require 3 dimensions (X, Y, T) and a RGBD camera would be able to use 4 dimensions (X, Y, Z, T), etc. The sensory input is compared with the activity patterns to measure similarity. If a threshold is fulfilled, then the activity is labelled.

The advantages of this approach is that it is relatively fast and doesn't require domain knowledge. However, it is very dependant of the sensory

[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, GPS, etc.). However, they can be included too the activities are represented in terms of raw sensory data patterns.

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

input, the continuity of the data, how the activity is performed and of the point of view where the scene is being observed.

There are many examples of this approach. In [Bobick and Davis, 2001] a video stream of aerobics exercises was analysed by attaching to every pixel a vector indicating the presence and recency of motion. Then the stream was compared online with previously described activities to look for matching. In [Ke et al., 2007], volumes were built by attaching similar regions of adjacent frames. Then, the problem was transformed in an object matching problem by comparing the shapes of the volumes (sensory stream and activity patterns).

Sequential approaches

Sequential approaches represent activities as a sequence of states. A state is a vector of features observed in the scene in a specific time. Finally, the sequence is analysed depending on the activity representation. There are two approaches: exemplar-based and model-based.

In exemplar-based approaches, activities, or a class of them, are represented as a sequence of states. Then the sensory input is compared in similarity with the patterns. An example can be found in [Darrell and Pentland, 1993], where states are built from view models. Templates of activities are from sequences of states associated with a physical change (e.g. rotation and scale). The dynamics of articulated objects in scene were recognized using the dynamic time warping algorithm (DTW) to the sequence of states.

In model based approaches, the sequence of states is compared with a set of probabilistic models of activities. The models are built assuming a temporal dependence between the states, so the transitions are modelled probabilistically using hidden Markov models (HMM) or dynamic bayesian networks (DBN).

The first work to use HMM to recognize activities was [Yamato et al., 1992]. They transformed a video stream into a sequence of vectors of image features. Then every vector was transformed to a symbol using vector quantization. Finally, a set of HMMs were created to model the activities, and their parameters were optimized.

2.2.2 Hierarchical approaches

Hierarchical approaches for activity recognition refer to those where complex activities are represented in terms of simpler ones. Multiple layers are defined to represent activities in different levels of complexity. Low level activities can be recognized using single-layered approaches.

Hierarchical approaches are also adequate to represent activities symbolically by using the multi-layered organization to describe semantic relations. By these means, hierarchical approaches are less dependant to training data and they can integrate domain knowledge more easily.

Hierarchical approaches can be categorized, regarding the applied methodology for recognition as statistical, syntactical and description-based.

Statistical

They are based in the hierarchical construction of statistical state-based models, such as HMMs or DBNs.

First the set of activities to work with is defined and organized hierarchically. Complex activities are defined in terms of simpler ones and so on until everything can be synthesized to atomic actions. In this way, many layers are created, from atomic actions to complex activities. In the bottom level, atomic actions are recognized from sensory data using single-layered sequential approaches. As a result, a sequence of feature vectors is transformed into a sequence of atomic actions. This sequence is the input for the next layer, which now will be treated as a new sequence of observations, and the same approach to recognize atomic actions from the first layer will be applied in the second one, and so on.

In [Oliver et al., 2002], the authors present layered hidden Markov models (LHMMs) for online activity recognition using data from video, sound and keyboard data. They divide their system in three layers: the first one is in charge of recognizing features from every source, the second layer trigger short events from the scene, and the last layer is used for longer activities. The hierarchical approach showed an improved performance when compared to single-layered systems. The training data is used more efficiently and it's more easy to add more detail on specific activities.

Some disadvantages of the statistical approaches is their difficulty to model the temporal structure of events (e.g. A occurred 'during'/'before'/'after' B) and also, because of their sequential nature, is hard to handle multiple concurrent tasks.

Syntactic

In the syntactical approach, activities are represented symbolically as a set of production rules generating a string of atomic actions which is later recognized using parsing techniques. Atomic actions are obtained with a single-layered approach, however, in higher layers, recognition is performed symbolically. Context-free grammars (CFGs) and stochastic context-free grammars

(SCFGs) are some of the techniques that have been used to recognize high level activities.

One limitation of this approach is the difficulty to handle concurrent activities, and also to consider unexpected events that are not integrated in the grammar.

An example can be found in [Ivanov and Bobick, 2000]. The authors aim to recognize complex activities in sequences of video. Two layers are defined; in the lower level, atomic actions are recognized using HMMs, and in the upper one uses SCFGs. The approach showed to be able to handle longer time activity constraints and more robust to uncertain detections in the lower level.

Description-based

This approach represent activities as a hierarchy of events, making emphasis in their spatial, temporal and logical structures.

A complex activity is modelled from de occurrence of its sub-events that satisfies certain relations. The temporal relation between sub-events is also considered in the representation, Allen's calculus is frequently used for this [Allen, 1983]. Atomic actions are obtained from sensory data and summarized.

Now to recognize activities the problem becomes a *constraint satisfaction problem*, which is NP-hard. This approach allows a good integration of additional knowledge sources. Particularly, the

There are many possibilities to treat the problem. In [Nevatia et al., 2004, Ryoo and Aggarwal, 2006], CFGs are used to represent activities hierarchically, defining temporal relations between sub-events. In [Sridhar et al., 2010], relevant features from the scene are extracted and their behaviour is represented using qualitative spatio-temporal relations (QSTR), then patterns of activities are learnt using Markov chains.

2.3 Description-based activity recognition and mobile robotics

In this project, the selected approach to follow is a description-based one. This section presents relevant related work in this line, and in the Answer Set Programming (ASP) paradigm for Also, here are presented the precedents of activity recognition with mobile robots.

2.3.1 Description-based activity recognition

As mentioned in section 2.2.2, description-based approaches represent activities hierarchically by decomposing complex activities in sub-events. The representation should also make emphasis in the spatial, temporal and logical structures. The recognition is performed by obtaining features from scene (spatial, temporal, logical) and creating a scene description as a *list* of facts, then the problem becomes a constraint satisfaction problem, to find the best activity match for these set of observations.

Representation

An activity is represented as a set of *facts* that needs to be fulfilled with the observations. This is important, because the facts can be used as logical predicates. These facts act as constraints between the activity patterns and help to discriminate between them, some of them may be more relevant than the others.

The execution of an activity depends on the subject, and even a particular subject doesn't execute the same activity in the same way. This is the reason why activities are usually defined in qualitative terms, i.e. in a symbolic more human-like manner. Quantitative descriptions of activities are still interesting, however, they are restricted to specific domains as rehabilitation or sports.

Regarding the temporal dimension, time can be represented as an instant t or as an interval (t_1, t_2). For the instants, simple temporal logic can be used to represent these kind of statements. Intervals have been typically treated with two aproaches [Fisher, 2008]: Interval temporal logics [Moszkowski, 1983] and Allen's interval algebra [Allen, 1983].

Allen's interval algebra was introduced as a calculus for temporal reasoning. It defines 13 possible relations between intervals, and provides a composition table that can be used for reasoning about temporal descriptions of events.

Qualitative spatial representations are the focus of study of Qualitative Spatial Representations (QSR). These are a set of calculi which allow a machine to represent and reason about spatial entities [Cohn and Renz, 2007], e.g. lines, dots, regions, etc. They are usually combined with a temporal representation (e.g. Allen's interval algebra) to represent behaviour dynamics.

In [Sridhar, 2010], activities are learnt, in an unsupervised fashion, and recognized from video sequences by reasoning under qualitative spatio-temporal representations (QSTR). Objects positions and their trajectories are ex-

tracted from scenes and represented in QSTR. Activities are learnt using a Markov Chain Monte Carlo (MCMC) procedure to find the maximum a posteriori probability (MAP) of candidate interpretations. This work shows an example of unsupervised learning of activities, using simulated and real examples. The qualitative approach (QSTR) is robust to changes in the execution of actions and to sensory errors. Finally, the categorization of activities showed to be reliable in learning functional object categories which provides semantic information from the scene. On the other hand, some limitations of this work are a fixed point of view and a posterior analysis. The analysis is performed only in short video sequences, and in short activities. The search space grows very easily as the scene becomes more complex.

In [Young and Hawes, 2013, 2014], QSTR are applied to the analysis of multi-agent behaviour in the RoboCup simulation league, particularly in estimating future behaviours. Positions, trajectories and orientation of the agents in scene are represented using region connection calculus (RCC), qualitative trajectory calculus (QTC) and the Star calculus respectively. Other agent's behaviours are learnt by using a HMM, which is fed with the current observations and a window of previous ones. As not all the data is relevant, this is first filtered. This work presents a study of activity prediction. The results show that qualitative representations are more easy to treat general cases of activities and require less training data. Some drawbacks are that the system posses global information from the environment, and the domain is restricted.

Reasoning

2.3.2 robotics

2.3.3 Activity recognition with mobile robotics

2.4 Answer Set Programming

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.

J. F. Allen. Maintaining knowledge about temporal intervals. *Commun. ACM*, 26(11):832–843, Nov. 1983. ISSN 0001-0782. doi: 10.1145/182.358434. URL `http://doi.acm.org/10.1145/182.358434`.

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.

A. G. Cohn and J. Renz. Qualitative Spatial Representation and Reasoning. In F. van Harmelen, V. Lifschitz, and B. Porter, editors, *Handbook of Knowledge Representation*, pages 551–596. Elsevier, Oxford, 2007.

T. Darrell and A. Pentland. Space-time gestures. In *Computer Vision and Pattern Recognition, 1993. Proceedings CVPR '93., 1993 IEEE Computer Society Conference on*, pages 335–340, Jun 1993. doi: 10.1109/CVPR.1993.341109.

M. Fisher. Temporal representation and reasoning. In F. van Harmelen, V. Lifschitz, and B. Porter, editors, *Handbook of Knowledge Representation*, pages 513–550. Elsevier, Amsterdam, 2008.

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

Y. Ivanov and A. Bobick. Recognition of visual activities and interactions by stochastic parsing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):852–872, Aug 2000. ISSN 0162-8828. doi: 10.1109/34.868686.

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`.

L. King. *The science of psychology : an appreciative view*. McGraw-Hill Education, New York, NY, 2014. ISBN 0078035406.

B. C. Moszkowski. *Reasoning About Digital Circuits*. PhD thesis, Stanford, CA, USA, 1983. AAI8329756.

R. Nevatia, J. Hobbs, and B. Bolles. An ontology for video event representation. In *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW '04. Conference on*, pages 119–119, June 2004. doi: 10.1109/CVPR.2004.27.

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

N. Oliver, E. Horvitz, and A. Garg. Layered representations for human activity recognition. In *Multimodal Interfaces, 2002. Proceedings. Fourth IEEE International Conference on*, pages 3–8, 2002. doi: 10.1109/ICMI.2002.1166960.

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

M. Ryoo and J. Aggarwal. Recognition of composite human activities through context-free grammar based representation. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1709–1718, 2006. doi: 10.1109/CVPR.2006.242.

M. Sridhar. *Unsupervised Learning of Event Classes from Video*. PhD thesis, University of Leeds, 2010.

M. Sridhar, A. G. Cohn, and D. C. Hogg. Unsupervised learning of event classes from video. In *Proc. AAAI*, pages 1631–1638. AAAI Press, 2010.

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.

J. Young and N. Hawes. Predicting situated behaviour using sequences of abstract spatial relations. In *Proceedings of the AAAI 2013 Fall Symposium How Should Intelligence be Abstracted in AI Research: MDPs, Symbolic Representations, Artificial Neural Networks, or _____?*, 2013.

J. Young and N. Hawes. Effects of training data variation and temporal representation in a qsr-based action prediction system. In *AAAI Spring Symposium 2014 on Qualitative Representations for Robots*, Stanford University in Palo Alto, California, US, March 2014.