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).

1.1.1 Test case: "The library setting"

The School of Computer Science at the University of Birmingham has a library, mostly used by students. An attendant is in charge of the book loans and retrievals, and also to assist the users. The physical scenario is basically a big room. It has some cabinets (where bibliographic material is stored), a reception, some tables with 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, putting things inside a backpack, etc. The objects involved in the scenario is relatively small (books, tables, chairs, laptops, cabinets, etc.).

There are many factors that from which will depend the components and the complexity of an activity recognition system, the targeted activities is an important one. Within the scope of this project, a library offers an ideal stage to test an activity recognition system. It provides a challenging and interesting environment, while still maintain the amount of objects and activities limited. This last one is important, because it maintains *bounded* the amount of domain knowledge that will be required in a library, compared with other more broad environments, in term of activities.

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.

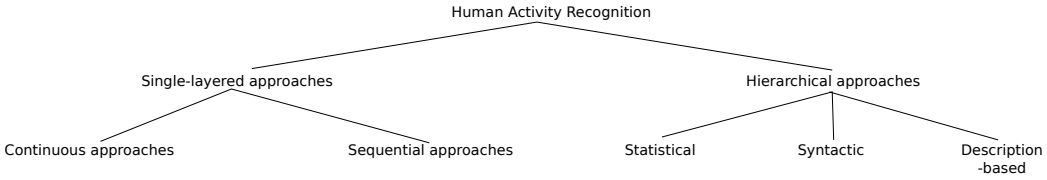

Figure 2.1: The taxonomy of research in activity recognition described in Aggarwal and Ryoo [2011].

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.

2.3.2 Mobile robotics

A robot is an ideal system to perform activity recognition. This is, indeed, a desirable skill for an autonomous robot that will share an environment with people. The robot needs to be *aware* of the surrounding humans. As described in section 2.1, even though AI, robotics and computer vision are relatively recent research fields, the interest in perception and particularly in automatic human analysis have been there since the very beginning.

Activity recognition with a mobile robot offers many potential advantages compared with other systems, however, there are some limitations and challenges to be considered too. The range of action for a robotic system in this context will depend on many factors as sensing and processing capacities, knowledge accessibility, etc. Positively, the mobility of a robot can potentially help to improve its perception capacity by participating in the scene, taking data from different points of view and interacting with the environment, this has been stated as *active perception*. [Bajcsy, 1988].

On the other hand, some challenges arise by using a mobile robot:

sensing Sensing data is usually corrupted because of hardware limitations, presence of statistical noise, discretization by the digitalization process, unstable or moving sensors, unpredictable environment conditions, etc. The collected data is restricted by the location of robot (there is no omni-presence) and as a consequence the robot will only gather information from the visible parts of the environment, loosing the rest of it.

storage and processing Sensory data can grow very easily and its storage and processing becomes a challenge. Knowledge bases may grow very easily too, the massive amount of possible instances, relations and categories forces to restrict the scope of knowledge bases to specific domains. The algorithms' complexity is usually high for the required techniques (e.g. pattern recognition, logic programming, etc.).

time Time constraints are relevant in robotic systems as, for many interesting applications, the data cannot be post-processed, i.e. real-time response is required.

This section presents some previous robotic systems that are relevant to the scope of this project.

Activity recognition from a robotics perspective

During the early years of robotics, much of the effort in the field was put in designing reliable motion planning and control techniques, e.g. [Brooks, 1985, Moravec, 1983]. Meanwhile, new advances were also made in fields as computer vision (e.g. optical flow, visual tracking, etc.), knowledge representation and reasoning (e.g. qualitative reasoning, frame languages), and machine learning (e.g. HMMs, decision trees, etc.). It was until the late 80's and early 90's that robots started facing realistic and non-controlled environments.

One of the first works in activity recognition with a mobile robot can be found in [Bonasso et al., 1996, Kortenkamp et al., 1996]. The authors used a monochromatic stereo-vision system mounted on top of a mobile robot for gesture recognition. They implemented an active approach by dividing the scene in cubic volumes. Volumes with similar motion vectors are merged and they are chained to a known human model to produce a linked representation of the human. The angles of the linked representation are compared to a set of previously defined gestures, to label the execution of the gesture. This work only uses a static representation of gestures within a single layer, and was tested in a scenario where a human points regions of interest to a robot,

and the robot gives a response. The authors point the the consideration of the temporal dimension, group activity recognition and integration with speech recognition as interesting research directions.

As mobile robots started to be used in everyday environments, human robot interaction became more relevant. For example, in [Burgard et al., 1999] the authors focused in the problem of motion planning in human environments (a museum). However, they point the relevance of human-robot interaction, a robot is not an isolated agent, and humans provide important information about the environment and to be able to complete the task (e.g. giving a tour). In future work, the same group studied the problem of human tracking from a mobile platform using particle filters [Schulz et al., 2001, 2003] and motion behaviour recognition using the EM algorithm [Bennewitz et al., 2002, 2004]. The authors point interesting research open areas in group tracking (instead of individual tracking), particularly the necessity of a flexible approach to handle individual and group tracks of humans and objects at the same time. Also, as motion patterns have been learnt in a particular environment, an interesting problem is to be able to reuse this knowledge in different scenarios where people show similar behaviours, i.e. a portable gait recognition system.

Human activity recognition also plays a central role in the problem of human robot cooperation. In this context, to achieve cooperation, a robot needs to be aware of its human partner and integrate itself to the common task in a non-obstructive way. With this in mind, a robot needs to be able to observe and also to communicate with his human partner. This requires a sensory approach of activity recognition, but also a high level treatment of the problem, to integrate the context of speech to the task execution. In [Lallee et al., 2010] the authors present an apprentice robotic system that uses speech and gesture recognition to learn new tasks from a human demonstrator. A Spoken Language Programming system (SLP) was developed [Dominey et al., 2007] to map sentences to actions, to allow verbal commands for the robot. The task for the robot is to assist a human to build a table by passing and holding material. SLP enables the system to extract semantic features from a spoken sentence: action, objects, agents. These are mapped to a set of atomic actions for their system to be able to execute the task. While executing the actions or interacting with the demonstrator, the robot visually follow the execution of the action in order to anticipate future actions or to learn new ones. Progressive benchmarking is used by the robot to learn and anticipate actions and interactions, so the robot can eventually gain confidence and take the initiative of the execution of an action. This approach has enabled a robot with defined primitive actions to assist a human demonstrator in the execution of a complex task, to learn new and complex tasks,

and eventually to take the initiative in the execution of subtasks that are necessary to reach the final goal. Human interaction provides robustness for the robot understanding, first by speech recognition, but also by visual scene analysis.

Another work with a similar approach has been described in [Karg and Kirsch, 2012, 2013a,b, Karg et al., 2011], here the goal is to perform a scene diagnosis and to detect abnormal situations on it. An expectations framework has been proposed to create internal representations of *normality* for the environment. In accordance with the authors, expectations should be probabilistic and adaptable; their approach considers to build them by merging information from different sources. The approach relies on motion tracking data and a semantic map of the environment. With this, the authors are able to segment occurring actions and to maintain probabilistic representations of activities, which are used to detect feasible future states and in accordance with a normality metric, to detect current abnormal states. This system has been tested using the TUM kitchen dataset [Tenorth et al., 2009] and simulation data. The authors express their interest in extending the framework to express expectations in a more probabilistic fashion. Also relevant, is to test the system in different scenarios (e.g. a new kitchen), where a semantic map is available and motion tracks of objects can be obtained; and to be able to segment activities properly and to detect abnormal states.

In [Ramirez-Amaro et al., 2014], the aim is to recognize activities by trying to minimize sensory observations (visual) and compensating this with semantic information. The goal is to show that with a simple sensory approach and with enough semantic information, high level activities can be inferred and that this approach is more suitable for a robotics context, mainly because of the requirement for online functionality. The authors track the motion of hands and objects from a visual input. The state of hands and the interaction with the objects is converted to a symbolic representation. They train a decision tree to generate semantic rules, which are used by a reasoning engine to generate a model of human behaviour. They explain the relevance of action segmentation, which is the problem of properly generating and grouping the atomic actions from the sensory data.

2.4 Answer Set Programming

Answer set programming (ASP) is form of declarative programming oriented towards difficult, primary NP-hard, search problems. As an outgrowth of research on the use of non-monotonic reasoning in knowledge representation, it is particularly useful in knowledge intensive applications [Lifschitz, 2008].

It has its roots in deductive databases, logic programming (with negation), logic-based knowledge representation and (non-monotonic) reasoning, and constraint solving (satisfiability testing).

The basic idea in ASP is to express a problem in a logical format so that the models of its representation provide solutions to the original problem. The resulting models are referred as *answer sets* [Gebser, 2013].

A rule is expressed in ASP as:

$$L_0 \, or \, \ldots \, or \, L_k \longleftarrow \; L_{k+1}, \ldots, L_m, \; not \, L_{m+1}, \ldots, \; not \, L_n,$$

each L_i is a literal in the sense of classical logic. The above rule means that if $L_{k+1}, \ldots, L_m$ are true and if $L_{m+1}, \ldots, L_n$ can be assumed to be false, then at least one $L_0, \ldots, L_k$ must be true [Gelfond and Lifschitz, 1988]. The symbol *not* is called *negation as failure*.

Monotonicity refers to the property of a logic programming system that, when more rules are added, it won't produce a reduction in the set of conclusions of the system. Non-monotonicity allows to a conclusion reduction when more rules are added [Poole, 2010]. This concept is important in systems with incoming knowledge, in dynamic and non deterministic scenarios. Also, allows the assumption of truth states, or belief states and a posterior revision of them when more rules are known. It is clear, that this is a desired property, in a logic system, to handle uncertain and incomplete information.

Negation as a failure symbol *not* L_i it is often read as "it is not believed that L_i is true". However, this does not imply that L_i is believed to be false, *not* L_i is a statement about belief [Gelfond, 2014].

2.4.1 ASP as a declarative problem solving technique

In declarative programming, in stead of coding the method to solve a problem, the idea is to describe the problem and leave the computer to find the solution.

2.4.2 ASP as a knowledge representation language

ASP is well suited for modelling problems in the area of Knowledge Representation and Reasoning involving incomplete, inconsistent and changing information [Schaub, 2013]. Some of its properties, in this context, are [Maximova-Todorova, 2003]:

Restricted monotonicity ASP can behave monotonically which addition of literals about certain predicates.

Language independence The results of a program are not dependant on the ASP solver.

Sort-ignorable The sorts can be ignored through language tolerance.

Knowledge extension Knowledge can be extended by *filtering*, i.e. updating the belief state [Amir and Russell, 2003].

ASP has been particularly applied to reasoning satisfiability problems. However, other problems can be treated too by ASP, as: model enumeration, intersection or unioning, as well as multi-criteria and -objective optimization. Formally, ASP allows for solving all search problems in NP and NP^{NP} in a uniform way.

ASP implementations

ASP implementations work n two steps:

1. A [grounder] builds an intermittent representation of the problem files by generating all possible values of the variables.

2. A **solver** that reads the grounded file and generates the answer sets (solutions).

Since the inception of the concept in the 1980s [Gelfond and Lifschitz, 1988] many implementations of ASP have been created. The majority of them uses the syntax of the language *Lparse*, also called *AnsProlog**. However, *DVL*, one of the most prominent ASP solvers uses a different syntax.

Here is a list of some of the more relevant toolkits available:

potassco Created and maintained by the University of Potsdam [Gebser et al., 2011].

DLV Created and maintained by the Technical University of Vienna and the University of Calabria [Gebser et al., 2011].

2.4.3 ASP for activity recognition

2.4.4 ASP for robotics

Chapter 3

Research Problem

The last chapter presented the state of the art in activity recognition. While the subject has been studied extensively in the recent years, the problem still provide a fertile research field to test different approaches. Activity recognition is particularly relevant for autonomous robots, which provide important benefits but also trigger new challenges. On the other hand, despite the main principles of ASP were stated in the late 1980s, the possibilities of this approach haven't been completely exploited and it still remains as a very active research area in the field of Logic Programming. We're interested in exploring the possibilities of ASP in the context of autonomous robotics. In general, as a tool to handle problems in the knowledge representation and reasoning areas, and in particular, within the problem of activity recognition

This chapter presents the main problem to be studied in this project: activity recognition with a mobile robot. It also introduces the structure of a framework to build up a solution to the problem that puts emphasis the usage of ASP, and that can be integrated with current state of the art hardware and software tools. This is presented in the context of the *Library scenario* (section 1.1.1) to ground the concepts to an example, and to set the path for future work. Finally, the approach is discussed to expose its strengths and weaknesses, and to present expected outcomes.

3.1 Problem Description

The subject of study in this project is ASP-based activity recognition with a mobile robot.

As mentioned at the beginning of this chapter, the problem of activity recognition is relevant in the context of robotics. At the same time, ASP offers an interesting and novel approach to the problem. These three parts,

Activity Recognition, Robotics and ASP, have been studied separately in extension, however, their joint possibilities are still uncovered. The focus of this project is to study the problem of activity recognition with an autonomous robot, and particularly how to exploit an ASP-based approach.

Human activities can be classified in different ways and with different grade of detail. In [Turaga et al., 2008], two non-exclusive categories are used: actions, performed by a person, and activities, performed by many persons. One more descriptive categorization has been given in [Aggarwal and Ryoo, 2011], separating activities in four classes:

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'.

Interactions 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'.

All of these categories require to be able to sense humans with different level of detail. For example, gestures require specific algorithms (e.g. hand detection, face detection, skeleton tracing, etc.) while long distant pedestrian tracking algorithms may consider persons as *moving dots*. With this in mind, target activities should be defined within the sensory constraints of a mobile robot.

The goal application is to be able to build a system on top of a robot that can be able to observe and follow the ongoing activities in a location (e.g. a library), from different points of view, but from a single one. This is the way in which humans perceive, however, this also drives to a situation of incomplete information as the robot won't have enough sensory information from all the environment. It is a thesis in this project that the lack of sensory information can be complemented with a stronger cognitive approach, in this case knowledge representation and reasoning capabilities.

This activity information can be used in different ways, but particularly as semantic knowledge, which has a meaning and a context and that can be used for reasoning. This is useful for a robot to have a qualitative description of the environment, to augment its navigation capabilities and task planning, and

also to bridge the gap in human-robot interaction [Kostavelis and Gasteratos, 2015].

Going back to the library setting described in section 1.1.1.

3.2 Methodology

This section presents the proposed approach to tackle the problem.

First, the target platform is an autonomous mobile robot. In a general fashion, a control system for a robot can be simplified as a perception-action loop. An activity recognition system fits in by providing interesting features from the environment (activities) that a robot can use to improve its performance.

The system requires some input, particularly observations from the environment, which are compulsory. Additionally to this, symbolic information is also going to be required in the form of a semantic map and domain knowledge. Finally, the representations of activities are also considered as an input. The output of the system is a set of activities and features found on the environment, within a degree of confidence.

Now let's look the inside of the proposed system.

3.2.1 Sensing

The end target for this project are autonomous mobile robots, with this in mind, all the sensing in charge of the robot. No building sensing (e.g. CCTV) or portable devices (e.g. cellphone, laptop) are allowed, this is the goal. We are looking forward to use state-of-the-art sensing techniques, however, as completely reliable sensing is not assured and improving sensing algorithms is beyond the scope of this project, the use of simulations and environment marks is considered for experimental purposes.

The features to be sensed will depend on the type of activities to be recognized. By using the categorization of activities mentioned in section 3.1, we are interested in the mid-level activities. This considers single human activities and interactions with other humans and/or objects, and excludes gestures and group activities. So, human and object sensing are needed. In the same fashion, the aim is to provide a robot with some understanding of activities within a spatio-temporal conceptualization by building a semantic map, so location and mapping are also a requirement, i.e. environment sensing. All these give us a path to decompose the scene in three different categories within a 3D space.

humans Detection, recognition, pose, motion, trajectory, skeleton, etc.

objects Detection, recognition, pose, motion, trajectory.

environment Location.

3.3 Evaluation

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`.

E. Amir and S. J. Russell. Logical filtering. In G. Gottlob and T. Walsh, editors, *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 75–82, San Francisco, 2003. Morgan Kaufmann.

R. Bajcsy. Active perception. *IEEE Journal on Computer Vision*, 76(8): 996–1006, Aug. 1988.

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

M. Bennewitz, W. Burgard, and S. Thrun. Learning motion patterns of persons for mobile service robots. In *Proceedings of the 2002 IEEE International Conference on Robotics and Automation, ICRA 2002, May 11-15, 2002, Washington, DC, USA*, pages 3601–3606, 2002. doi: 10.1109/ROBOT.2002.1014268. URL `http://dx.doi.org/10.1109/ROBOT.2002.1014268`.

M. Bennewitz, J. Pastrana, and W. Burgard. Active localization of people with a mobile robot based on learned motion behaviors. In *Workshop on Selforganization of Adaptive Behavior (SOAVE)*, 2004.

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.

R. P. Bonasso, E. Huber, and D. Kortenkamp. Recognizing and interpreting gestures within the context of an intelligent robot control architecture. In *Technical Report, Metrica Inc. Robotics and Automation Group, NASA Johnson Space*, 1996.

R. A. Brooks. A robust layered control system for a mobile robot. Technical Report AIM-864, AI Lab, MIT, Sept. 1985.

W. Burgard, A. B. Cremers, D. Fox, D. Hahnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun. Experiences with an interactive museum tour-guide robot. *Artificial Intelligence*, 114(1-2):3–55, Oct. 1999. URL `http://www.sciencedirect.com/science/article/B6TYF-3Y0JBMM-2/2/313003341edf00f`

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.

P. Dominey, A. Mallet, and E. Yoshida. Real-time cooperative behavior acquisition by a humanoid apprentice. In *Humanoid Robots, 2007 7th IEEE-RAS International Conference on*, pages 270–275, Nov 2007. doi: 10.1109/ICHR.2007.4813879.

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.

M. Gebser. *Answer set solving in practice*. Morgan & Claypool Publishers, San Rafael, 2013. ISBN 1608459713.

M. Gebser, R. Kaminski, B. Kaufmann, M. Ostrowski, T. Schaub, and M. Schneider. Potassco: The Potsdam answer set solving collection. *AI Communications*, 24(2):107–124, 2011.

M. Gelfond. *Knowledge representation, reasoning, and the design of intelligent agents : the answer-set programming approach*. Cambridge University Press, New York, NY, 2014. ISBN 1107029562.

M. Gelfond and V. Lifschitz. The stable model semantics for logic programming. In *5th Conference on Logic Programming*, pages 1070–1080. Seattle, 1988.

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`.

M. Karg and A. Kirsch. Acquisition and Use of Transferable, Spatio-Temporal Plan Representations for Human-Robot Interaction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012.

M. Karg and A. Kirsch. An Expectations Framework for Domestic Robot Assistants. In *Conference on Advances in Cognitive Systems*, 2013a.

M. Karg and A. Kirsch. Simultaneous Plan Recognition and Monitoring (SPRAM) for Robot Assistants. In *Proceedings of Human Robot Collaboration Workshop at Robotics Science and Systems Conference (RSS) 2013*, 2013b.

M. Karg, M. Sachenbacher, and A. Kirsch. Towards expectation-based failure recognition for human robot interaction. In *22nd International Workshop on Principles of Diagnosis, Special Track on Open Problem Descriptions*, 2011.

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.

D. Kortenkamp, E. Huber, and R. P. Bonasso. Recognizing and interpreting gestures on a mobile robot. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 2*, AAAI'96, pages 915–921. AAAI Press, 1996. ISBN 0-262-51091-X. URL http://dl.acm.org/citation.cfm?id=1864519.1864523.

I. Kostavelis and A. Gasteratos. Semantic mapping for mobile robotics tasks: A survey. *Robotics and Autonomous Systems*, 66(0):86 – 103, 2015. ISSN 0921-8890. doi: http://dx.doi.org/10.1016/j.robot.2014.12.006. URL http://www.sciencedirect.com/science/article/pii/S0921889014003030.

S. Lallee, E. Yoshida, A. Mallet, F. Nori, L. Natale, G. Metta, F. Warneken, and P. Dominey. Human-robot cooperation based on interaction learning. In O. Sigaud and J. Peters, editors, *From Motor Learning to Interaction Learning in Robots*, volume 264 of *Studies in Computational Intelligence*, pages 491–536. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-05180-7. doi: 10.1007/978-3-642-05181-$4_2$1.URL