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

# Contents

# Chapter 1

# Related Work

## 1.1 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.

## 1.2 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. 1.1.

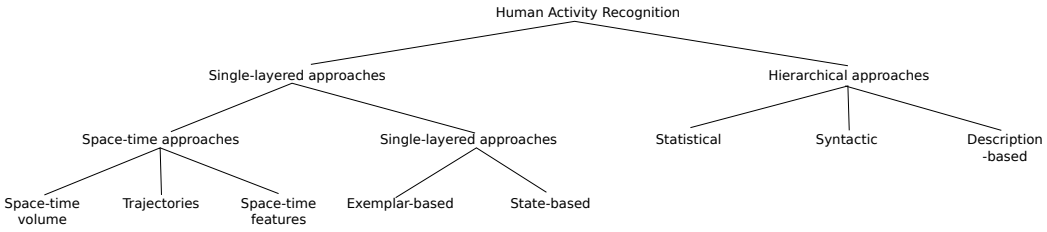


Figure 1.1: The taxonomy of research in activity recognition described in Aggarwal and Ryoo [2011].

### 1.2.1 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.

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

Continuous approaches can also be classified depending on the features that are used to describe activities:

- Volumes:
- Trajectories:
- Point descriptors:

**Sequential approaches**

In these approaches the goal is to interpret a sequence of observations.

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

### 1.2.2 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.

### 1.2.3 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.

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

### 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.

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`.

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`.

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.