# Hybrid approaches for context recognition in Ambient Assisted Living systems : application to emotion recognition and human activity recognition and anticipation

Hazem Khaled Mohamed Abdelkawy

Ecole Doctorale

Mathématiques, Sciences de l'Information et de la Communication (MSTIC)

THÈSE

pour obtenir le grade de

Docteur de l'Université Paris-Est Créteil

Spécialité : Informatique

présentée et soutenue publiquement par

**Hazem Khaled Mohamed ABDELKAWY**

le 16 Septembre 2021

# Hybrid approaches for context recognition in Ambient Assisted Living systems: application to emotion recognition and human activity recognition and anticipation.

Directeurs de thèse

**Yacine AMIRAT, Abdelghani CHIBANI**

**Jury**

| | |
|---|---|
| Mounim A. EL YACOUBI | Rapporteur |
| Mounir MOKHTARI | Rapporteur |
| Patrick REIGNIER | Examinateur |
| Marie BABEL | Examinateur |
| Yacine AMIRAT | Directeur |
| Abdelghani CHIBANI | Encadrant |
| Paulo GONÇALVES | Rapporteur |

*To my parents, my wife, my daughters, my sister.*

# Acknowledgments

Foremost, I would like to express my sincere gratitude to my supervisor Prof. Yacine Amirat who has guided and taught me a lot throughout the journey of this Ph.D. thesis, which allowed me to learn and acquire the spirit of scientific research. Because the same thing can also be said about my advisor Dr. Abdelghani Chibani, I would like to thank her too for continuous support.

Besides, I would like to thank all jury members for offering their valuable time to review my thesis and examining my defense. Huge thanks are also due to them for valuable comments, questions, and insightful remarks.

My sincere thanks also go to the staff of the LISSI laboratory represented by the director Prof. Yacine Amirat and also to my colleagues in the department of R&T in the IUT of Créteil/Vitry especially to Dr.Ferhat Attal.

I would also like to thank my lab-mates for the discussions, the inspirational talks, and all the fun we have had together in the last four years. Among them, special thanks go to Roghaeh Mojarad, Elnaz Soliemani, Mohsen Qutbi, Nicolas Khoury, and Arnaud Flori.

Last but not least, I would like to express my gratitude and love to my parents, Eman and Khaled, my lovely wife, Alaa, and my daughters, Hana and Mariem and my sister Heba for their love and support.

# Contents

# List of Figures

# List of Tables

Chapter 1

# General Introduction

Ubiquitous robots and connected objects are intended to become an increasingly important part of our everyday life environments such as homes, hospitals, airports, schools, transports, amongst others. According to the latest estimations of Gartner[1], the US research company, 25 billion objects will be connected in the world by 2021. The ambition of this emerging domain is to provide intelligent assistance services with a high level of performance and acceptability, such as services that adapt to the user's context and naturally interact with them. Designing cognitive capabilities of assistive robots or agents based on the concept of context awareness, is a very challenging topic in the AI community.

The evolution of ubiquitous computing and ambient assisted living (AAL) paradigms, and the recent development of service robotics, have led to the development of a new branch of robotics research, called ubiquitous robotics or networked robotics. The challenges surrounding the ubiquitous robotics are numerous in terms of fields of applications. One of the most important applications is the home care of the elderly and dependent people in the context of

[1] https://www.gartner.com/en/newsroom/press-releases/2018-11-07-gartner-identifies-top-10-strategic-iot-technologies-and-trends

the Silver economy (or senior economy). The integration of service robotics in ambient intelligence environments aims to create cyber-physical spaces that provide, anywhere and anytime, a wide range of services to improve the quality of life, the physical and mental state, and the social well-being of users. The objective is to create a unified ecosystem exploiting all connected objects/entities in the environment (sensors, actuators, smartphones, smart TV, digital tablets, smartwatches, service robots, etc.) to create intelligent services and spaces according to the vision of the Web of Objects (Web of Things). The success of intelligent ambient robotics operating and collaborating with humans in daily living environments depends on their ability to generalise and learn human movements, and obtain a shared understanding of an observed scene. In this context, human daily activity recognition, human emotion recognition, and human intention anticipation are the most challenging cognitive capabilities that should be integrated with any AAL system to guarantee the people's well-being and safety in the ambient intelligence environments. However, to efficiently integrate those capabilities in a highly dynamic environment, multiple modalities sensing systems combined with complex knowledge representation and fusion techniques are required.

This thesis proposes three novel hybrid approaches that enable the AAL system to detect the emotional states, actions, and intentions of users, taking advantage of their context and the benefits of combining data-driven and knowledge-based techniques. Firstly, a hybrid approach for contextual emotion recognition for cognitive assistance services in ubiquitous environments is proposed. The proposed approach is based on integrating deep-learning models with possibilistic logic and expressive emotional knowledge representation and reasoning model. Secondly, a hybrid approach for human activity-aware AAL system is proposed. The proposed approach is based on the integration

of spatial and temporal deep-learning models for recognizing human activities from IMUs and skeleton key-points with higher knowledge representation and reasoning engine. Finally, a hybrid approach is proposed to recognize and anticipate the human contextual daily activities from the current partially observed activities and events in an ambient environment. The proposed approach combines both of human and robot egocentric vision perspectives is proposed to recognize and proactively anticipate human daily activities. The manuscript is organized as follows :

- **In chapter 2**, we first present the fundamental principles of ambient intelligence and ubiquitous systems, and their exploitation in the development of a new variant of service robotics commonly called ubiquitous or ambient robotics. In the remainder of the chapter, we review the existing technologies that enable the implementation of ambient intelligence systems. Then, we discuss the benefits of robotic systems for personal assistance with a focus on the challenges of ubiquitous robotics. Finally, in the last part, we present the objectives of the thesis in terms of the ambient environment perception, knowledge management, and reasoning based on deep learning and n-ary ontologies for better management of human-robot interaction in Ambient Assisted Living (AAL).

- **In chapter 3**: the objective of this chapter is to present and also analyze the state of the art of human activity recognition, human intention anticipation, and human emotion estimation in an ambient environment. Firstly, we we reviewed and analyzed the literature of the human action recognition approaches. Secondly, a detailed analysis of the state of the art human intention anticipation approaches is presented. Finally, a comparative study of different emotion estimation approaches is presented and analyzed.

- **In chapter 4**: we present the foundations of the Narrative Knowledge Representation Language (NKRL) used for context representation in AAL environments. First, we present the definitions used throughout this chapter. Then, we describe, on the one hand, the foundations of the modelling of narrative knowledge by exploiting the HClass and HTemp ontologies, and on the other hand, the NKRL reasoning mechanisms.

- **In chapter 5**: a hybrid approach for contextual emotion recognition for cognitive assistance services in ubiquitous environments is proposed. The proposed approach is able to recognize accurate explicit discrete emotions using a multilayer perceptron neural network fusion model combined with possibilistic logic. Besides, the proposed approach is able to recognize non-directly observable emotions using expressive emotional knowledge representation and reasoning model.

- **In chapter 6**: a novel hybrid approach for human activity-aware AAL system is proposed. A combaniation of Spatio-Temporal Joint based Convolutional Neural Network (STJ-CNN) and Hierarchical Multichannel deep Residual Network (HMResNet) is proposed to recognize human activities from both skeleton keypoints and multichannel IMUs's raw data. Besides, the NKRL representation and inference is exploited to represent and combine the detected human activities with the ambient events, and infer the semantic context of the detected activity.

- **In chapter 7**: a novel hybrid approach is proposed to recognize and anticipate the human contextual daily activities from the partially observed activities and events in an ambient environment.

- **In chapter 8**: we summarized the contributions of the proposed thesis, and the different future directions are highlighted.

Chapter 2

# Overview of research context

## 2.1 Introduction

In this chapter, we first present the fundamental principles of ambient intelligence and ubiquitous systems, and their exploitation in the development of a new variant of service robotics commonly called ubiquitous or ambient robotics. In the remainder of the chapter, we review the existing technologies that enable the implementation of ambient intelligence systems. Then, we discuss the benefits of robotic systems for personal assistance with a focus on the challenges of ubiquitous robotics. Finally, in the last part, we present the objectives of the thesis in terms of the ambient environment perception, knowledge management, and reasoning based on deep learning and n-ary ontologies for better management of human-robot interaction in Ambient Assisted Living (AAL).

## 2.2 Ambient intelligence: principles and definitions

In recent years, mobile computing has become the core of many daily objects such as smartphones, smart TVs, smart glasses, digital tablets, smartwatches,

etc. Currently, the connected objects are beginning to become noticeable in the market, which are new communicating and intelligent devices intended to be used as monitoring, assistance, communication, and information tools. These objects are starting to be more and more present in our everyday places, such as houses and workplaces. For example, Samsung Vapor Cook smart oven, General Electric ChillHub smart refrigerator, Google Home, and iRobot's Roomba 980 smart vacuum cleaner. The variety of the connected objects used in our day-to-day activities, their number, and level of sophistication will continue to grow in the future. The American research agency Gartner estimates that 21 billion number of objects that will be connected by 2021.

Ubiquitous robots and connected objects are intended to become an increasingly important part of our everyday life environments such as homes, hospitals, airports, schools, transports, amongst others. According to the latest estimations of Gartner[1], the US research company, 25 billion objects will be connected in the world by 2021. The ambition of this emerging domain is to provide intelligent assistance services with a high level of performance and acceptability, such as services that adapt to the user's context [1] and naturally interact with them. Designing cognitive capabilities of assistive robots or agents based on the concept of context awareness, is a very challenging topic in the AI community.

Given the development of these new technologies, in the near future, it is expected that the systems will provide countless functions that will be accessible at all times and in any place and several modes of interactions. Such systems, described as ubiquitous or omnipresent, can continuously and transparently adapt the same function or service (e.g., medication reminder, appointment reminder, etc.), to the context of use (e.g., displaying a text message on

[1] https://www.gartner.com/en/newsroom/press-releases/2018-11-07-gartner-identifies-top-10-strategic-iot-technologies-and-trends

the user's PC if the user is at work, or display a text message on the user's PC if the user is at work, or displaying the same message on the dashboard of the user's vehicle if the user is driving). Besides, these systems can dynamically discover the objects available in the ambient environment; these objects may be intangible, such as information services, or physical, such as robots, sensors/actuators, multimedia equipment, etc. The services provided by these objects are then exploited to provide the user with contextual value-added services. According to A. K. Dey [2], context refers to any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and the applications themselves. By linking the physical world to the digital world, the objective is to transform the devices and equipment that we use in our daily lives into communicating objects endowed with perception, interpretation, decision, and action capabilities, allowing to proactively offer value-added assistance services to the user. The development of radio identification, wireless communication networks, and Internet protocols, allows the creation of large scale networks of connected objects according to the paradigm of the Internet of Things; an object can be a sensor, an actuator, or any physical object. The thematic expression of ubiquitous computing, called Ambient Intelligence (AmI), has been introduced in 1998 by the Philips company as part of a prospective study about the evolution of household electronics by 2020. Concurrently, the Information Societies Technology Advisory Group (ISTAG) published in 2001 a document of four scenarios illustrating the notion of "Ambient Intelligence" by 2010 [3]. The work carried out by the ISTAG aimed to provide a vision to orient the European program of the IST (Information Societies Technology) towards emerging fields of research, such as ambient intelligence.

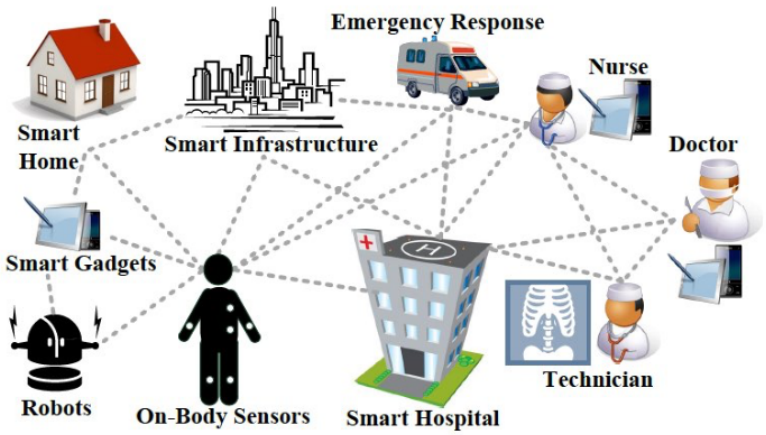



Figure 2.1 – Example of the wide range of services that an ambient intelligence system may provide for a dependent person

The exploitation of the ambient intelligence paradigm aims to design and implement intelligent environments or ecosystems that provide a variety of reactive or proactive services to improve the quality of life, physical and mental well-being, comfort, and safety of users. In the context of dependent people's home care, as shown in Fig.2.1, it is possible to design an ambient intelligence system (AIS) to provide a wide range of services such as reminders to drink, take medication or carry out tasks; alarms to hospital services or caregivers in the event of an accident (fall, injury, burn, intoxication, asphyxiation, medical complication, etc.), etc. Other types of assistance services can be developed to support a balanced lifestyle and reduce the risk factors of potential chronic diseases such as diabetes, Alzheimer's, and Parkinson's (physical activity coaching, cognitive stimulation, etc.). Finally, it is also possible to implement environmental services to match the user's preferences and needs, for example, adjusting the opening of shutters, switching lights on/off, adjusting the ambient temperature and light, turning on an oven, etc.

The paradigm of ambient intelligence has been defined in several ways,

but they are generally quite similar. According to the ISTAG [3], ambient intelligence is the creation of environments capable of considering the personal characteristics of different users, adapting and intelligently responding to their specific requests, acting in a non-intrusive and often invisible way, and allowing them to access the offered services most naturally and intuitively, using voice recognition, gesture recognition or the manipulation of tangible objects. Reignier [4] defines ambient intelligence as a paradigm resulting from the intersection of the ubiquitous computing and the artificial intelligence where the objective is to exploit the perceptual capabilities offered by all sensors to analyze the environment, the users and their activities, and enable the system to react according to the context. The key properties of this paradigm are the ability to analyze the context and to adapt dynamically to the changes that occur in it.

## 2.3 Identification and localization technologies

### 2.3.1 Radio-Identification

Radio Frequency Identification (RFID) is a technology for identifying and tracking a person or an object. RFID radio tags are capable of transmitting and receiving data transmitted remotely by a reader, such as the identifier of an object or a person, see Figure.2.2. There are two types of tags: Passive tags and active tags. The passive tags do not use any external power supply such as battery; they are powered by the electromagnetic field produced by the RFID reader when sending a request. Compared to passive tags, active tags require an external power supply. Active tags can be operated as range-finders; the distance can be estimated from the received power measurements of the incoming radio signal. In terms of standards, four main frequency bands are re-

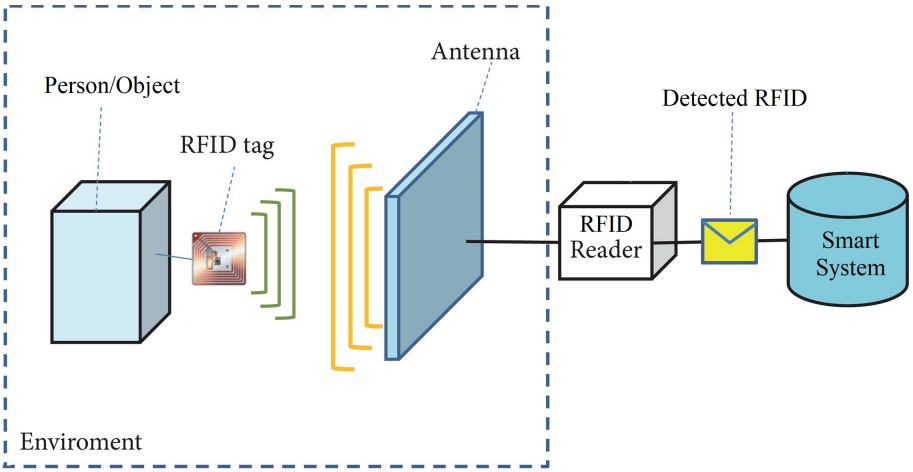



Figure 2.2 – Radio Frequency Identification (RFID) system architecture

served for the RFID applications: the low-frequency band, around 125 kHz, the high-frequency band at 13, 56 MHz, the UHF band, between 800 and 900 MHz, and finally the microwave band with frequencies at 2, 45 GHz and 5, 8 GHz. In terms of RFID applications, Georgia Tech developed the EL-E robot [5], which is a home assistance RFID-based robot to support dependent people in daily activities. On one hand, long-range antennas are used to enable the robot to detect the presence of an object or a person in a given space (bedroom, kitchen, etc.). On the other hand, Short-range antennas are used to detect the presence of objects near the robot's hand. The RFID enables the robot to perceive and semantically understand its environment. For example, the robot can deliver an object to a person wearing an RFID tag, or handling the ambient objects, such as: delivering a phone to a dependent person, as shown in Fig.2.3. Besides, as shown in Figure.2.4, a living lab apartment was developed by the University of Orebro-Sweden [6] as part of the European project Robot-ERA, which aims to develop and evaluate robotic technologies for elderly people. The ambient environment consists of a grid of 1900 passive RFID tags that were deployed under the floor to localize and guide the navigation system of different robots, where

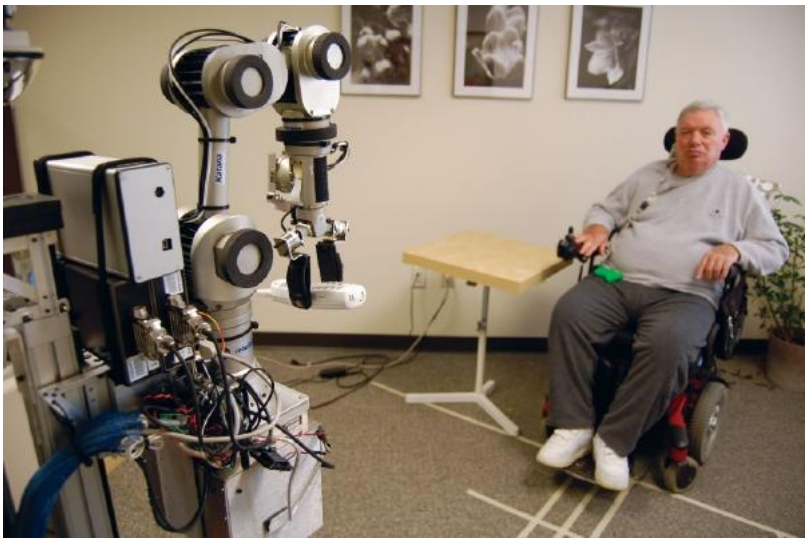
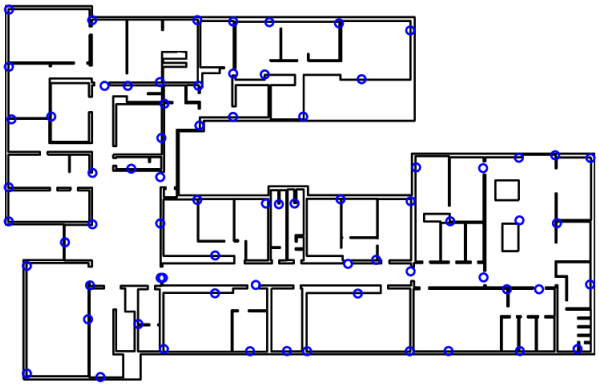


Figure 2.3 – EL-E robot delivering the phone equipped with an RFID tag to dependent person

each one is equipped with an RFID tag reader. As an extension of RFID technology, Field Communication (NFC) is a short-range, high-frequency wireless communication technology that allows the exchange of information between devices up to a distance of approximately 10 cm. Currently, This technology is integrated into smartphones, tablets, connected audio-video players, smart-payment terminals, etc. Finally, the iBeacon technology was developed by Apple based on Bluetooth Low Energy (BLE) technology. The beacon device will transmit a signal carrying a unique universal identifier to interact with a smartphone present in a given area. The identifier will be detected by a smartphone and will be converted into a physical location or can be used to trigger a specific action.

### 2.3.2 Localization systems

In this section, we present the main indoor localization systems which can be divided into two categories [7] :



(a) Distribution map of the RFID tags inside the living lab apartment (in red)

(b) Snap-shot of the apartment

Figure 2.4 – Living Lab at the University of Orebro

- Wifi Positioning Systems (WPS): These systems exploit the distributed WiFi access terminals (hotspots) to localize the ambient objects and humans in different environments such as urban environments, buildings, and apartments. This localization system can be implemented using different techniques; on one hand, the simplest one is the triangulation between different WiFi hotspots. On the other hand, the most precise techniques are based on the radio fingerprint mapping such as Ekahau's RTLS [2] (Real-Time Location System), which has a high localization precision up to few centimeters. The RSS (Received Signal Strength) is the most commonly used parameter in WPS (triangulation, mapping) localization systems to estimate the working area boundaries, which is defined as the distance between the transmitter node and the receiver node. To achieve good performance, this system requires a well-defined attenuation model for different places such as obstacles, walls, etc.

- Localization systems based on short-range wireless technologies such as Bluetooth, Infrared, Zigbee, Ultra-Wideband (UWB), etc. The localization techniques used in WPS systems can also be used for this type of system. In [8], Active Badge is proposed as a localization system based on infrared (IR) technology. The IR sensors are distributed at specific places in the ambient environment to capture the IR signals emitted by the badge worn by a person or an object to be tracked. Therefore, The badge is localized using the triangulation technique. The drawbacks of the Active Badge system can be summarized as follow: (i) the system accuracy is relatively low up to several meters; (ii) the IR sensor should be on the line of sight with the badge to be located without any obstacles; (iii) the system is so sensitive for the fluorescent light and sunlight. In [9],

[2] https://www.ekahau.com/products/ekahau-connect/pro/

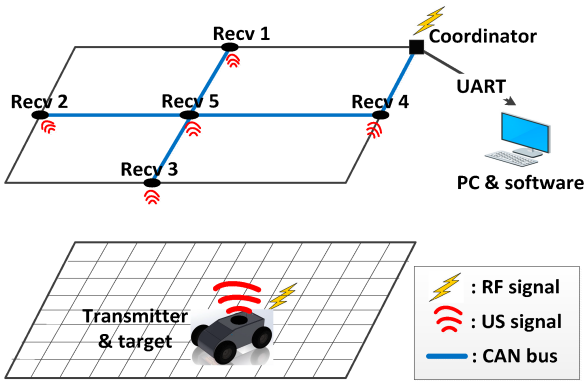



Figure 2.5 – Cricket localization system

the Active Bat localization system is proposed; the system is based on the time-of-flight measurement of the ultrasonic signals. Compared to Active Badge, the Active Bat localization system is more accurate, and less sensitive to the obstacles.

Cricket is a localization system combining both of RF (Radio Frequency) and ultrasonic (US) waves, as shown in Fig.2.5. Each "Cricket beacon" deployed on the ceiling transmits both RF and US signals simultaneously. The receiver node (Cricket listener), attached to an object or a person to be localized, receives these signals, and estimates the distances to the different beacons to calculate its current position. Each Cricket beacon listener distance is estimated by measuring the difference between the propagation time of the RF and US waves. This parameter is called Time Difference of Arrival or TDoA [10]. The accuracy of the Cricket system localization is about 10 cm.

The Ubisense [3] localization system uses UWB technology, as shown in Fig.2.6. The system consists of (1) an array of sensors, called Ubisensors, placed at known locations in the area to be covered; each Ubisensor sen-

[3] http://www.ubisense.net/en/

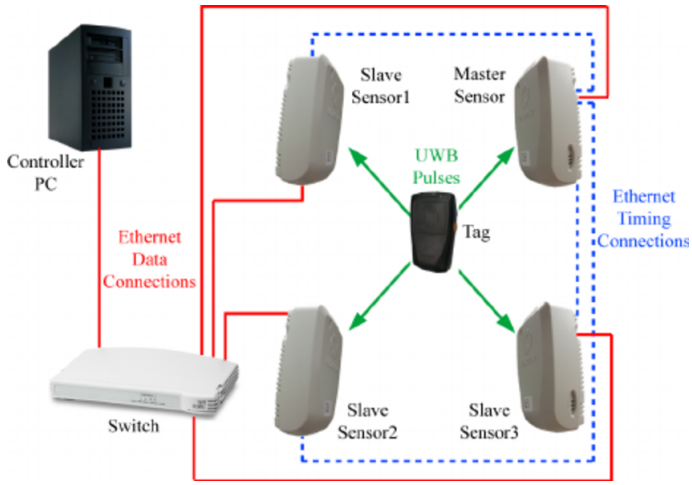



Figure 2.6 – Ubisense localization system

sor consists of an RF transmitter/receiver (2.4 GHz), and an array of UWB receivers (6-8 GHz); (2) a set of active beacons (called Ubitags), carried by persons or objects to be localized. Each Ubitag beacon has an RF transmitter/receiver and a UWB transmitter. The Ubisense system uses the Time Difference of Arrival (TDoA) and Angle of Arrival (AOA) corresponding to the direction of the received signal to calculate the position of each Ubitag beacon using triangulation. Although the Ubisense system is very accurate (accuracy of the order of 15 cm), nevertheless it is relatively expensive, imposes a high deployment cost to cover large areas (3 sensors are needed to cover a rectangular cell), and also requires special cabling.

Recently, The development of LIFI (LIight-FIdelity) technology as an alternative to Wi-Fi [11] becomes noticeable. LIFI is based on the principle of turning a light-emitting diodes (LED) light on and off several thousand times per second to transmit information using different encoding systems. Consequently, it is possible to remotely transmit different multimedia contents such as video, sound, and geolocations to a tablet or a smartphone. In addition to high-speed data transmission, LIFI technol-

ogy allows the use of LED light sources as localization sensors.

## 2.4 Connected objects

Connected objects are intelligent embedded devices that are typically composed of the following components: a processing unit, random access memory, storage space, wireless communication interfaces, and sensors. These devices can be integrated into any everyday object and can be used to assist users in their daily activities [12]. Equipped with sensors, they are capable of perceiving changes in the ambient environment, reacting to these changes, and possibly cooperating with other devices. Some connected objects are able to interact with users using different interfaces such as integrated display screens, smartphones, or tablets applications. Four types of connected objects applications can be distinguished: social communication and web browsing, home automation, well-being, and e-health.

- Social communication and web browsing: The usage of connected objects is more easier and comfortable than smartphones or tablets, which are bulkier. As shown in Fig.2.7, several models of smartwatches are already commercially available, such as the Apple Watch or Samsung Gear, which, for example, are equipped with applications for displaying updates from social networks, e-mails, SMS and calendar applications. Currently, these watches can be used as communication peripherals for smartphones or tablets. They can manage incoming or missed calls and control the music stored on a smartphone (playing, pausing, and selecting songs). Besides, these watches can be integrated with more advanced features to be used as remote control devices based on hand gestures detection, or voice commands recognition. Recently, Google glass was in-

(a) Samsung Galaxy Gear (b) Apple Watch (c) Google glasses

Figure 2.7 – Accessories for social communication and web browsing

troduced as the first commercial smart glass in the market. Google glass is equipped with a built-in camera, a microphone, a touchpad, mini-screen, and internet access using WiFi or Bluetooth. Google Glass was designed to provide different services such as real-time communication using social networks and web browsing, answering phone calls, and showing navigation maps. Google is now moving towards the development of a new generation of Google glasses to offer a different kind of services such as assisting physicians during surgeries by displaying the historical patient records, medical images related to the surgery, the patient's vital signs as well as a range of alert messages to reduce the risk of error.

- Home Automation: Many projects were aiming to develop various connected objects to support different home automation applications. Most of these projects have been turned into products that are now commercially available. Besides, The contributions of giant companies such as Google, Apple, Samsung, etc., to the home automation market augurs well for a strong presence of home automation technologies in our daily lives. In the following, we present a few examples of products that show this evolution. For example, the Joshfire[4] company offers a sofa that de-

[4]http://www.joshfire.com/

tects an RFID chip installed in a phone or wallet to identify the user, and to provide a personalized on-screen display. The Nest[5] thermostat, originally produced by Nest Labs, recognizes and stores the user's preferred temperatures, reduces the heating in the user's absence and can be controlled remotely. NextBulb[6] is a connected bulb technology designed to reduce the consumption of electricity. It allows you to know in which room a user is located to automatically manage the switching on and off of the lamps. NextBulb uses Bluetooth Low Energy (BLE) technology to measure the distance from the bulb to the user's smartphone. SafeLight[7] is a system consisting of a color LED bulb that can be controlled (color, intensity, white temperature) using a smartphone or tablet application. This bulb also has a smoke sensor to detect the fires and warn the users. Recently, connected scales became commercially available such as Countertop[8] or Tefal Cooking Connect[9]. These scales are equipped with embedded applications for suggesting recipes, giving instructions, etc. Besides, These scales can monitor the user's nutrient requirements and provide him/her with nutritional advice based on his/her activities. Finally, These scales can recommend the user to use certain kitchen utensils, such as a blender or a casserole to facilitate the preparation of meals.

- Well-being: In addition to smartwatches, which are increasingly used for monitoring well-being, other several sensors have been recently developed to provide similar functionalities. These sensors are generally implicitly integrated into everyday objects or accessories. For example, cloth-

[5]https://nest.com/fr/thermostat/meet-nest-thermostat/
[6]https://www.indiegogo.com/projects/nextbulb-smart-bulb-that-knows-your-location
[7]http://www.awox.com/connected-lighting/safelight/
[8]https://techcrunch.com/2015/04/07/countertop/
[9]https://www.tefal.fr/Ustensiles-de-cuisine/Balance-de-cuisine/Cooking-Connect/p/2100093072


Figure 2.8 – E-textile system

ing, jewelry, or bed linen, which can be used to measure physical activity or monitor vital signs. Smart blankets, that adapt its temperature based on the temperature of the ambient environment and the human body. Balluga[10] smart bed, which can monitor the different vital signs (body temperature, heart rate) in real-time and take into account the user's preferences in terms of comfort and treatment options. The mattress of Balluga smart bed can produce a vibration massage to relax the muscles and help the user wake up in the morning. Dreem[11] is a connected headband, which is capable of analyzing the user's brain waves and producing small sounds through bone conduction to improve sleep quality.

- E-health: Several connected objects have been recently developed for medical use. As smartphones or bracelets, they are dedicated to fall detection and monitoring of vital signs. In this context, for example, Apple Watch consists of a heart rate sensor, a GPS, and an accelerometer. It can monitor the heart rate and detect different sports and physical ac-

[10] https://www.kickstarter.com/projects/684490728/balluga-the-worlds-smartest-bed
[11] https://dreem.com/

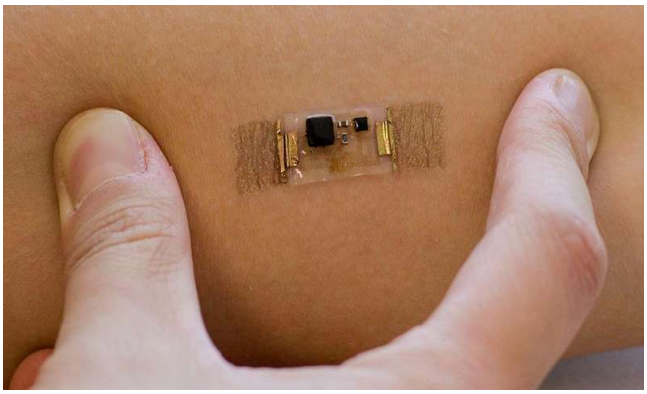


Figure 2.9 – Electronic epidermal system

tivities. Smartex[12] has developed a prototype of an intelligent garment (e-textile system) for the remote and continuous monitoring of physiological and movement data. Flexible embedded wireless sensors are integrated into the fabric to record electrocardiographic (ECG) data related to cardiac activity, and electromyographic (EMG) data related to muscle activity. Finally, a wireless epidermal electronic system developed through a collaboration between the University of Illinois, the Institute of High-Performance Computing in Singapore, and Dalian University of Technology in China. As shown in Fig2.9, the system was designed as a flexible, ultra-thin, electronic patch that sticks to the skin like a tattoo. Different types of electrodes and LEDs are integrated into the patch to measure, depends on the body location, the heart rate, brain activity, or muscle activities [13]. In the field of diabetes, in France, Diabeloop[13] is a project to develop, an artificial pancreas to treat type-1 diabetics. The system consists of a sensor that continuously measures blood sugar levels, a pump patch that automatically delivers insulin, and a smartphone application that calculates the required insulin doses to be injected.

[12]http://www.smartex.it/en/our-products
[13]https://www.diabeloop.fr/produits

## 2.5 Video analytics for ambient intelligence

As previously discussed, a great number of sensors are nowadays available, this has stressed the need for new approaches to merge low-level measurements to realize what facts they refer to in the real environment. Ambient Intelligence (AmI) techniques exploit information about the environment state to adapt the environment itself to the users' preferences. Even if traditional sensors allow a rough understanding of the users' preferences, ad-hoc sensors are required to obtain a deeper comprehension of users' habits and activities. Consequently, the technologies for video analytic becomes a need to complement/replace this huge number of heterogeneous sensors. Videos are considered as a main source of information compared to the previously mentioned technologies for many reasons. Firstly, visual information can be acquired at reasonable costs using small and cheap cameras with high-quality resolution. Consequently, many surveillance cameras are installed in many public places (train stations, intersections, public parks, airports, etc.) and inside homes. Secondly, visual information processing becomes much easier because of the availability of the powerful hardware to process and transfer the data over the wired and wireless networks available in most of the cities. Finally, visual information includes the biggest amount of information related to the objects and people populating the environment.

In this context, human activity monitoring system was developed by CEA [14] to recognize ambient objects, localize and track humans, and predict their actions based on video analytics, see Fig.2.10. The smart apartment consists of 3 different Microsoft Kinect depth sensors to monitor the kitchen and living room area, besides integrating a number of ambient sensors such as RFID, light, and limit-switch sensors. For each Kinect sensor, 4 different streams are

[14] https://www-mobilemii.cea.fr

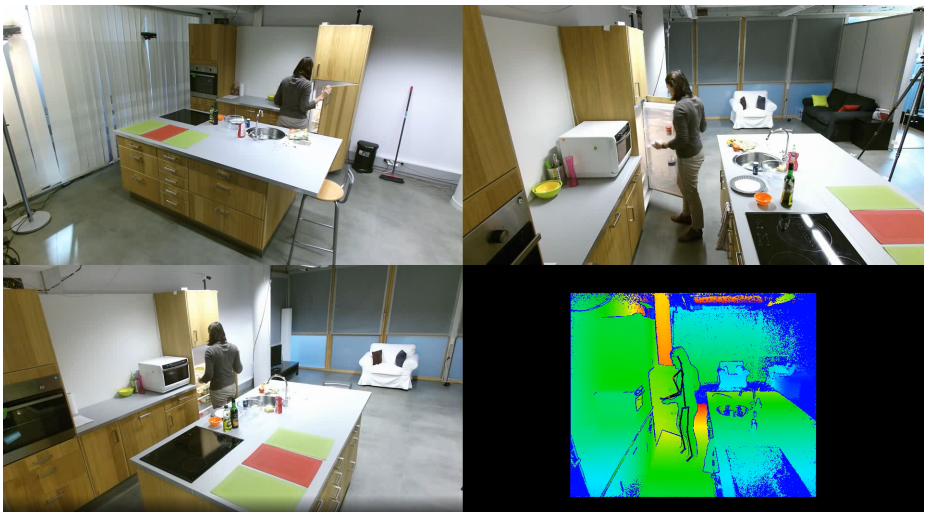


Figure 2.10 – Video-based human activity monitoring system

produced such as RGB videos, depth maps, skeleton data, and body indices. The system was able to detect a set of 7 different daily activities such as cooking, washing dishes, eating, clearing table, working, housework, and laying. Besides, The system exploited the 3D locations of human skeleton-joints in the 3D point space to detect the associated object for each action.

To support Alzheimer patients, a monitoring system was developed in the Greek Alzheimer's Association for Dementia and Related Disorders (GAADRD) in Thessaloniki, Greece [15], see Fig.2.11. The system succeeded to fuse many modalities such as audio, RGB videos, Depth videos, and wearable physiological sensors to track and detect the activities of different people suffering from conditions ranging from Mild Cognitive Impairment (MCI) to mild dementia, and in a few cases full-blown Alzheimer's Disease (AD). The system was able to detect different activities such as reading article, watering Plant, preparing drug box, preparing drink, turning on radio and talking on phone.

[15] https://team.inria.fr/stars/en/demcare-gaadrd-dataset/


Figure 2.11 – Alzheimer patients activity monitoring system

## 2.6 Physical Assistance Vs Cognitive Assistance

In ambient intelligence, physical assistance focuses on the use of ambient actuators to automatically control the user's living spaces to increase their comfort and guarantee their safety. With the development of assistive robots, such as robotic wheelchairs or exoskeletons, a new form of physical assistance can be provided to facilitate the mobility of users. Unlike physical assistance, cognitive assistance does not necessarily require the use of physical actuators, but it rather focuses on the verbal and gestural interaction with the user. In the context of cognitive assistance, on one hand, DOMUS laboratory at the University of Sherbrooke in Canada developed a system that can transform a home into a cognitive orthosis. DOMUS's system can provide a personalized, contextual, and adaptive cognitive support, which is capable of compensating the cognitive deficits of its user (problems of attention, memory, planning, etc.). On other hands, many other projects exploited the concept of companion robot to create cognitive assistance services such as PEIS-Ecology [14], CompanionAble [15], and Robot-Era[16] projects.

[16] www.robot-era.eu


Figure 2.12 – Assistive robots for mobility

### 2.6.1 Assistive robots

In the following, we present two main categories of assistive robots for performing daily tasks or activities: mobility assistive robots and domestic assistive robots. These prototypes of robots developed from various research projects, which are still at the experimental and research phases to improve their performance and make them economically viable.

#### Mobility assistive robots

Assistive robots are promising solutions to improve the autonomy of elderly people in daily physical activities such as: standing up, sitting down, walking, going upstairs/downstairs, etc. Currently, many research projects are being carried out worldwide to develop robotic systems to improve the quality of mobility services such as robotic chairs, smart canes, walker robotics, and portable robots or exoskeletons. As shown in Fig.2.12, (a) the CARRIER wheelchair robot, developed by the University of Applied Arts at the Industrial Design Studio in Austria. It is a mobility and standing robot; (b) the robotic arm wheelchair

developed by Kinova Robotics[17]; (c) the iCane intelligent cane developed at Nagoya University [16] ; (d) the c-Walker robot developed within the framework of the European DALI project[18]; (e) the E-ROWA (Exoskeletal Robotic Orthosis for Walking Assistance) lower limb exoskeleton from the LISSI laboratory, for supporting the daily activities such as walking, sitting, standing, stairs-up, and stairs-down. Exoskeletons are portable mechatronic robots that support and enhance the motor skills of the patients; they can also be used for neuromuscular rehabilitation.

**Domestic assistive robots**

Currently, domestic assistive robotics is one of the most important sectors of robotics with numerous research studies and new consumer products in the market, such as iRobot's Roomba cleaning robot. In this context, PARO sensory awareness robot was originally developed by the AIST in Japan to derive therapeutic benefits from the interaction through direct contact between patients and the robot, as shown in Fig.2.13. The robot was able to provide many cognitive services such as emotional support, improvement of patient's moods, decrease in anxiety, stress level, depressive symptoms, blood pressure, pain. Several therapeutic studies, conducted in different hospitals and long-stay centers for people with dementia or Alzheimer's disease, have shown the benefits of the PARO robot in terms of improving the patient sociability. Generally, Sensory awareness robots are equipped with tactile sensors, microphones, and speakers to react differently to various touches and to communicate verbally with patients, respectively. Besides, They have built-in emotional intelligence engines that allow them to communicate and interact emotionally and in a personalized way with different patients.

[17] http://www.kinovarobotics.com/
[18] http://www.ict-dali.eu/dali/

Domestic robots can be considered as connected mobile objects with (1) cognitive assistance capabilities such as physical and intellectual stimulation support, (2) communication capabilities such as e-mail or social interaction (social networks, relatives, caregivers, etc.). Generally, such robots consist of a mobile-base equipped with wheels and a head. Besides, they are equipped with an autonomous navigation system, voice recognition and synthesis engines, and different network interfaces. Some examples of domestic robots are shown in Fig.2.14, such as Ava from iRobot, Pepper from SoftBank Robotics, Buddy from BlueFrog Robotics.

The development of multitasking service robotics for domestic applications is still at the experimental and research stage, as evidenced by the numerous studies, particularly in humanoid robotics. The humanoid robot ROMEO was intendedly developed to explore and deepen research into assistive robotics for elderly and frail people, see Fig.2.14 (e). The PR2 robot from Willow Garage is probably the most sophisticated personal assistive robot. It has a wheeled mobile-base to navigate, and two arms each with 7 degrees of freedom to perform handling tasks, see Fig.2.14 (d). However, given its price, this robot is considered more as an experimental robot for research rather than as a product for daily use. Finally, the Care-o-bot 4 robot was developed by the Fraunhofer IPA in Germany to provide the same capabilities of the PR2 robot Fig.2.14 (f).

### 2.6.2 Ubiquitous Robotics

The evolution of ubiquitous computing and ambient intelligence paradigms, and the recent development of service robotics, have led to the development of a new branch of robotics research, called ubiquitous robotics or networked robotics [17, 18]. The challenges surrounding the ubiquitous robotics are numerous in terms of fields of applications. One of the most important applica-

Figure 2.13 – PARO sensory awareness robot

tions is the home care of the elderly and dependent people in the context of the Silver economy (or senior economy).

The integration of service robotics in ambient intelligence environments aims to create cyber-physical spaces that provide, anywhere and anytime, a wide range of services to improve the quality of life, the physical and mental state, and the social well-being of users. The objective is to create a unified ecosystem exploiting all connected objects/entities in the environment (sensors, actuators, smartphones, smart TV, digital tablets, smartwatches, service robots, etc.) to create intelligent services and spaces (web and physical) according to the vision of the Web of Objects (Web of Things).

The notion of service as defined in ubiquitous computing or ambient intelligence can be used identically in ubiquitous robotics. Therefore, the interaction of the robot with the ambient environment is totally reconsidered, since it is no longer based on predefined instructions. A ubiquitous robot is expected to be interoperable and compatible with any connected object/entity it discovers in its ambient environment instead of being statically pre-programmed for specific actions. Besides, it should be able to use its own objects/entities (sensors and actuators) as well as interacting with other connected objects/entities


(a) Ava (b) Pepper (c) Buddy

(d) PR2 (e) Romeo (f) Care-o-bot 4

Figure 2.14 – Different domestic robots

in the ambient environment. The main objective is to migrate from a centralized computation system where only the robot own the intelligent capabilities such as Perception, Decision Making by Reasoning, Actuation and Communication (PRAC), to the concept of service robotics where each object, robot or system in the ambient environment is equipped with PRAC capabilities. In this case, the physical robot, by itself, is only a part of the ambient and ubiquitous intelligence; this intelligence is embodied in the different connected objects populating the ambient environment as well as being centralized on the physical robot which, by its autonomy, ensures the guarantees the service operation, and maintenance. An essential property of the ambient intelligent system, according to this new vision, is the ability to dynamically adapt the robotic services provided according to the context. In this case, the intelligent system should be sensitive to the context. A context-sensitive intelligent system is a system that has the ability to react, adapt, and reconfigure itself according to the detected environmental changes. In order to provide context-sensitive services, the ambient intelligence system should have advanced cognitive capabilities to interpret the context, recognize the user's emotions, activities, and intentions to proactively take the most appropriate decision and to provide the most relevant assistance services for the situation.

The recent development of cloud computing has led to the rapid evolution of ubiquitous robotics and the development of the field of cloud robotics. By leveraging the high-performance computing infrastructure of the cloud, robots will be able, in the future, to improve their cognitive abilities in terms of performing the perception and reasoning tasks that require intensive computation, which is impossible to be executed by the on-board embedded computers of the robots [19, 20, 21].

Chapter 3

# Review of human activity recognition/anticipation and emotion recognition approaches

## 3.1 Introduction

The objective of this chapter is to present and also analyze the state of the art of human activity recognition, human intention anticipation, and human emotion estimation in an ambient environment. Firstly, we we reviewed and analyzed the literature of the human action recognition approaches. Secondly, a detailed analysis of the state of the art human intention anticipation approaches is presented. Finally, a comparative study of different emotion estimation approaches is presented and analyzed. In this chapter, the analysis of both human activity recognition and human intention anticipation is split into: data-driven, knowledge-driven, hybrid approaches. In other hand, the analysis of human emotion recognition is split into classical emotion recognition and deep-learning based emotion estimation. Finally, the major obstacles facing the recognition

of both human activities/emotions as well as the anticipation of human intentions are analyzed, besides the objectives of the proposed thesis are detailed.

## 3.2 Review of human activity recognition

Human activity recognition is a challenging task which can be exploited in different domains such as AAL systems [22, 23, 24], and rehabilitation [25, 26]. Recently, human activity recognition has attracted the attention in the AAL domain because of its vital impact on enhancing individual's life style/quality. In the context of ambient intelligence, wearable-sensors based approaches are proposed to exploit the mobile sensors such as Inertial Measurement Units (IMUs), smart watches, and mobile phones to recognize simple human activities such as walk, stand, sit, and running. On other hands, visual-based approaches are proposed to exploit different visual modalities such as monocular vision, stereo vision, depth maps, and points cloud to recognize complex human daily activities [27] such as cleaning, drinking, and eating. Finally, multi-sensor based approaches are proposed based on fusing ambient sensors, mobile sensors, and visual-based modalities to recognize more complex and interlaced activities such as preparing breakfast, and washing clothes. In this part, we will review the approaches that exploit both wearable based and visual based approaches to recognize human activities. These approaches can be summarized into three categories [28]: (1) data-driven approaches [29], [30], [31], [32],[33] (2) knowledge-driven approaches [34], [35], [36],[37], as well as (3) hybrid approaches [38, 39, 22].

### 3.2.1 What is an activity?

Aggarwal and Ryoo [40] present 4 categories of human activities classified according to their complexity: gestures, actions, interactions and group activities.

1. Gestures: Gestures are the "elementary movements" of the human body. They present the atomic components that give a meaningful description of the person's movements; such as "raising a hand" or "lifting a foot".

2. Actions: Actions present a simple activity of a person. Actions are composed of a set of temporally organized gestures; such as "walking" composed of "lifting a foot" and "get-down a foot".

3. Interactions: Interactions are human activities that involve several people and/or objects. For example: two people shaking hands is an interaction that involves two people, and a person writing letter is an interaction that involves a person, paper, and pen.

4. Group activities: Group activities are a set of activities carried out by a groups of several people and/or objects; such as "football match" that involves many people and a ball, or "a group of people have a meeting" that involves group of people, and group of objects (chairs, papers, pens, projector, etc).

In our work, in the context of ambient intelligence, an activity can be recognized based on a set of associated elementary events, which are partially ordered in time and comply with specific constraints. These elementary events describe the gestures, states, or actions of one or more persons/objects and captured by the ambient sensors such as cameras, smart-plugs, wearable sensors, etc. Consequently, in this thesis, an "activity" can be defined as:

1. A set of actions by a single person such as "walking and hand-waving" at the same time is an activity consisting of two actions that happen in the same time.

2. An interaction or set of interactions of a person with one or more objects such as "preparing a coffee" or "taking a medication".

3. An interaction or set of interactions between several people and/or objects such as "group of people having lunch together".

4. A group activity such as "a meeting" can be broken down into: one person enters, another person enters, they shake hands.

### 3.2.2 Activity recognition

Activity recognition is the process of detecting human activities ***after*** their execution [41]. Consequently, the activity recognition systems infer the activity labels after the complete termination of those activities, which can be useful in non-critical scenarios such as meal preparation assistive services. The earlier studies were mostly interested in recognizing simple activities such as walking, standing, and running in a controlled experimental environment [42, 43, 44]. Nevertheless, the recent studies have progressively evolved towards recognizing more complex activities in real-time from heterogeneous input modalities such as RGB-D cameras [45, 41], and wearable devices [46, 47].

### 3.2.3 Data-driven approaches

Data-driven methods rely heavily on data, which is prone to various flaws such as imprecision, uncertainty, confusion, incompleteness, disagreement, and so on. Imprecision is concerned with the substance of the data and quantifies its quantitative deficiency, while uncertainty is concerned with the truthfulness of the data and quantifies its degree of adherence to fact [48], [49]. Confusion refers to unclarified or unspoken knowledge that can be interpreted in a variety of ways [48], [49]. Incompleteness refers to a lack of knowledge regarding a

specific feature of the problem. Disagreement happens when two or more bits of evidence are incompatible, resulting in inconsistent and thereby incorrect meanings.

A dataset is generally required to train models in data-driven approaches. These methods are mainly focused on a probabilistic or predictive interpretation of data. Most of the traditional data-driven approaches are based on hand-crafted features, which are known as shallow features. In one hand, in terms of werable sensors, the most commonly used features for human activity recognition are hand-engineered features [50], time domain features such as mean, median, variance, skewness, kurtosis and range [47] and frequency domain features such as temporal fast Fourier transform (tFFT) [51], Discrete Fourier Transform (DFT) and Power Spectral Density (PSD) [47]. To recognize human activities using the hand-crafted features, a set of classification approaches are proposed such as Hidden Markov Models [52], Artificial Neural Network [53], Support Vector Machine (SVM) [54] and naive Bayes classifiers [55]. On other hands, in terms of camera-based systems, the most commonly used features for human activity recognition are based on representing the human skeleton joint's dynamics manually such as skeleton-joints trajectories co-variance matrix [56], skeleton-joints relative positions [57], or different body-parts translations and rotations [58]

In [59], to recognize low-level human tasks, an ensemble-based SVM classifier is used. The model was able to recognize sitting, standing, walking, going up and coming down the stairs. The accelerometer and gyroscope data were obtained from Samsung Galaxy S2 mobile phone. An SVM is also used in detecting several activities patterns, especially in [60], [61], [62], [63], [64], [65]. In [60], an SVM model is exploited to recognize nine different tasks, including, climbing stairs, walking forward, walking right, walking left, hopping up, lying,

standing, sitting, and going down. In [61], SVM model was used as a social activity identification system. The model was able to recognize social actions as shaking hands, pushing, touching, and kicking. In [62], The SVM model is used to identify eleven key human daily tasks such as cleaning, prepare food, doing laundry, and work.

In [66], both NB and K-NN models are exploited to detect human activities using a data obtained from a single three-axial accelerometer. The preprocessing phase consists of extracting the most relevant features using two different methods: Relief-F and Sequential Floating Search (SFFS). Finally, Both NB and K-NN models explited the extracted features to identify six low-level movements including walking, running, hopping up, standing, sitting and standing. Additionally, [67], [68] employ the K-NN model. In [68], a K-NN model was developed to identify eight low-level movements such as standing, sitting, biking, ascending and descending steps, and teeth brushing. Besides, the NB model is exploited in [69], [70], [71]. In [69], The NB model was developed to identify five low-level human activities such as: running, walking, standing, sitting, and lying. In [70], The NB model was developed to classify twelve different high-level daily tasks, such as using a computer, learning, listening to music, and serving hot tea.

In [72], a DT model was developed to identify human actions using data obtained from the eWatch smartwatch. The DT model was able to identify six different actions, such as put on wrist, put in backpack, use it as a belt. The model was able to identify the six common everyday human movements, like sitting, standing, walking, going up and down the stairs, as well as jogging. Besides, in [73] eight human movements are identified using the K-Means model, which is trained on a dataset obtained from a wristwatch: sitting, lying, standing, jogging, walking, biking, ascending and descending stairs. In [74], The K-

means model detects seven simple low-level tasks, such as cooking, cleaning, sleep, taking out the garbage, and vacuuming. In [75], RF-based model was developed to identify six general human movements, including standing, sitting, reclining, walking, hiking, walking up/down the stairs, and lying on the ground using a data collected from a wearable IMU sensor. In [71], an RF model was developed to identify seven high-level activities, including lying, sleeping, washing, walking slowly, walking naturally, nordic walking, and racing. In [76], an RF model is used to classify eleven low-level movements such as jumping in place, jumping jacks, and bending.

In [68], a variety of classifiers were applied, including K-NN, DT, SVM, and meta-level classifiers to recognize human actions. Eight low-level actions such as standing, sitting, running, ascending and descending stairs, and brushing teeth are recognized in this analysis using data obtained from a wearable accelerometer. In [77], HMMs and CRFs models exploited the data collected from wearable accelerometers, gyroscopes, and magnetometers to recognize human movements such as walking, standing, jogging, sitting, and biking. As described in [78], 6 simple tasks, such as standing, sitting, walking up stairs, walking down stairs, and running, were identified using HMM and CRF algorithms. In [79], a HMM model was developed to exploit the environmental ambient sensors to recognize a vast variety of human activities such as loading a medication dispenser, moving chairs, watering trees, and cooking dinner. Using GPS info, in [80] employs the CRF model was used to identify six different activities such as working, sleeping, picking up, leisure exercise, visiting, and turning off/on a vehicle. In [81], CRF model was developed to identify five high-level activities including dressing, doing laundry, showering, cooking a dinner, and washing dishes. In [82], using a set of wearable sensors and ambient sensors, a CRF model was able to identify eight low-level human actions such as: walking, run-

ning, using stairs, using elevator, sitting, standing, and brushing teeth.

In [83], a spatial-temporal hierarchical model combined with a linear dynamic system (LDS) is proposed to encode the skeleton-joint sequences and to learn the dynamic structure of human motions respectively. The model was able to recognize seven different human activities such as walking, sitting, standing, lying, running, and ascending/descending stairs. In [58], human actions are modeled as curves in groups using Lie algebra, then an SVM model is exploited to classify the 3D relationships between different body parts using rotations and translations. The proposed approach was able to identify six high-level activities such as eating, cleaning, reading, washing dishes, working, doing laundry. In [84], the 3D joint coordinates histograms are exploited as a representation of human postures. Then, a Hidden Markov Model (HMM) is proposed to classify the temporal dynamics in the skeleton motion activity sequences. The HMM model was able to identify five different postures including standing, sitting, lying. In [85], a graph-based model combined with the shortest path graph kernel (SPGK) based SVM is proposed to represent the motion sequences and to detect human actions respectively. The model was able to recognize the following activities standing, walking, ascending stairs, descending stairs, running, shopping, taking bus, and moving.

In [86], a novel fusion approach for ensemble classifiers is introduced for human activity recognition. It is based on the idea of a Generalized Fuzzy Soft Set (GFSS). Besides, a weighted aggregate operator is used to weighting each classifier output and to create a more stable fused classifier. To solve the GFSS decision-making problem, a new ranking algorithm was proposed. The approach suggested was assessed by a CRF and HMM ensemble classification for identifying different high-level action recognition such as idleness, leaving home, showering, brushing teeth, breakfast preparation, dinner preparation,

snacking and drinking.

One of the first attempts to recognize human basic activities with a deep learning model is proposed in [32]. The proposed model is based on Restricted Boltzmann Machines (RBM) which allows automatically extracting features from accelerometer raw data. The proposed model was used to recognize the low-level activities from Opportunity dataset [87]. In [31], a convolutional neural network (ConvNet) model is proposed for recognizing basic gestures from the accelerometer and gyroscope raw data. This ConvNet model outperformed the other state-of-the-art models in gesture recognition such as Dynamic Time Wrapping (DTW) and Hidden Markov Model (HMM). In [88], for the recognizing human activities, a range of state-of-the-art, deep models are proposed such as CNN, and LSTM deep-learning models. The proposed models were evaluated using different public including Opportunity dataset [87], PAMAP2 dataset[89], and Daphnet Gait datasets [90]. In [29], a novel deep CNN model for human activity detection based on data obtained from a smartphone's accelerometer and gyroscope sensors is proposed. The proposed model enables identification of seven different low-level human actions including standing, sitting, ascending and descending stairs, running, shopping, taking the subway, and walking. In [91], A more advanced body-measurement deep model incorporating convolutional and LSTM units is proposed. The model was able to recognize 21 different activities from a wearable IMU unit such as running, standing, opening/closing a refrigerator, and washing tables. In [88], a CNN model was developed to identify 21 different activities, including standing, walking, lying, sitting, opening and closing doors, and opening and closing refrigerator. In [92], an LSTM-based model was developed to identify nine different high-level actions such as leaving, toileting, bathing, eating, eating, snacking, and watching television. In [93], an LSTM-based model has been used for recognizing six

different low-level actions, including jogging, cycling, standing, lying, hiking, and going down.

As shown in Table 3.1, the most important data-driven methods are detailed, including classical machine-learning and deep learning approaches, combined with their key features, including the classifier used, detected activities, precision, and sensors used.

As mentioned previously, classical data-driven approaches depend heavily on hand-crafted features, which affects their overall performance. Deep learning models, on the other hand, are able to automatically extract more relevant features. Recently, The researchers focusing on human activity recognition [91],[88],[93] gained considerable interest in deep learning models.

The researchers in [94] and [88] addressed the unlimited capabilities of the deep learning models by conducting more detailed studies using both ambient and wearable sensors approaches. Besides, these studies discussed the limitations of the deep learning models, such as the fact that these models often need a large amount of computing power and thus cannot be conveniently implemented on battery-constrained devices like smartphones and smartwatches for real-time activity recognition [94]. Furthermore, the success of deep learning models is still extremely dependent on huge amount of labelled data. Since obtaining appropriate huge amount of labeled data is costly and time- consuming, more reliable unsupervised activity recognition using deep learning models remains an open research problem. Another difficulty for human activity recognition using deep learning, is developing models that are scalable, flexible and adaptable to recognize human high-level overlapped activities. These activities are hard to examine at the higher levels because they involve a lot of semantic and contextual knowledge. Current models often neglect the latent similarities between the input modalities, which leads to not achieving high

recognition rates [94].

### 3.2.4 Knowledge-driven approaches

In the context of knowledge-driven approaches, prior assumptions as a basis integrated with a domain knowledge are used to create an activity model. Domain knowledge is a human experience which is recorded in a formal representation. In contrast to data-driven approaches, domain knowledge does not need large data-sets to create an activity model. In general, knowledge-driven approaches can be divided into logical and ontological approaches. In logical approaches as in [34][35], the knowledge representation formalisms are used to encode the activities into a logical structure. Besides, a knowledge-based inference engine combined with the logical structure is used to recognize the activities. In ontological approaches as in [95] [36], a hierarchy of concepts and classes is developed to model the domain concept. The concepts and classes are connected using relationships with applied restrictions besides having a set of properties. The ontological approaches have more flexibility and re-usability compared to the classical logical approaches because of the more generic representation of the domain is used. Compared to the data-driven approaches, the knowledge-driven approaches don't require a data-set to create the activity model which makes them reusable and complete independence from the environment/scenario. Besides, the knowledge driven approaches have a clear construction of the activity model which depends on an explicit formalization of the domain. In contrast, the data-driven approaches have the ability to represent the uncertainty and the temporal variables which are difficult in the knowledge-driven approaches.

Table 3.1 – State of the art comparison data-driven approaches for human activity recognition

| **Reference** | **Year** | **Model** | **Dataset** | **Number of Activities** | **Level of Activities** | **Accuracy** | **Sensors** |
|---|---|---|---|---|---|---|---|
| [70] | 2009 | NB | custom | 12 | high-level | 92.41% | environmental, wearable |
| [79] | 2010 | HMM | custom | 15 | high-level | 73.15% | environmental |
| [60] | 2011 | SVM | custom | 9 | low-level | 93.10% | wearable |
| [64] | 2012 | SVM | custom | 6 | low-level | 93.50% | wearable |
| [65] | 2014 | SVM, HMM | custom | 5 | low-level | 90.80% | smartphone |
| [66] | 2014 | NB, K-NN | custom | 6 | low-level | 98.40% | wearable |
| [96] | 2015 | HMM | Kast | 7 | high-level | 97.00% | environmental |
| [96] | 2015 | HMM | Aruba1 | 11 | high-level | 92.00% | environmental |
| [96] | 2015 | HMM | Adlnormal | 5 | high-level | 98.00% | environmental |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [47] | 2015 | k-Means,<br>GMM,<br>HMM,<br>K-NN,<br>RF,<br>SVM,<br>GMM | custom | 12 | low-level | 83.89%<br>72.95%<br>75.60%<br>99.25%<br>98.95%<br>95.55%<br>85.05% | wearable,<br>environmental |
| [62] | 2015 | Binary SVM | Aruba1 | 11 | high-level | 91.88% | environmental |
| [62] | 2015 | Binary SVM | Kasteren | 7 | high-level | 94.30% | environmental |
| [62] | 2015 | Binary SVM | Kyoto1 | 5 | high-level | 96.67% | environmental |
| [62] | 2015 | Binary SVM | Kyoto2 | 5 | high-level | 94.98% | environmental |
| [62] | 2015 | Binary SVM | Kyoto8 | 11 | high-level | 95.04% | environmental |
| [62] | 2015 | Binary SVM | Tulum1 | 10 | high-level | 99.28% | environmental |
| [62] | 2015 | Binary SVM | Tulum2 | 16 | high-level | 84.92% | environmental |
| [62] | 2015 | Binary SVM | Milan | 15 | high-level | 95.20% | environmental |
| [62] | 2015 | Binary SVM | Cairo | 13 | low-level | 94.17% | environmental |
| [69] | 2016 | HMM, NB | custom | 5 | low-level | 95.40% | wearable |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [97] | 2016 | Neural network | custom | 6 | low-level | 92% | wearable |
| [78] | 2016 | AdaBoost,<br>CRF | custom | 6 | low-level | 93% | wearable |
| [91] | 2016 | DeepConvLSTM | Skoda | 10 | low-level | 95.80% | wearable |
| [88] | 2016 | CNN<br>RNN<br>LSTM | custom | 4 | low-level | 59.1%<br>69.8%<br>74.5% | wearable |
| [88] | 2016 | CNN<br>RNN<br>LSTM | PAMAP2 | 12 | low-level | 93.7%<br>88.2%<br>86.8% | wearable<br>environmental |
| [98] | 2016 | shapelet-based | custom | 18 | low-level | 96.54% | wearable,<br>environmental |
| [29] | 2016 | CNN | custom | 4 | low-level | 95.75% | smartphone |
| [91] | 2016 | DeepConvLSTM | custom | 4<br>17 | low-level<br>middle-level | 86.4% | wearable |
| [99] | 2016 | K-NN,<br>DT | custom | 7<br>6 | low-level<br>high-level | 95% 92% | smartwatch,<br>smartphone |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [88] | 2016 | CNN<br>RNN<br>LSTM | Daphnet<br>Gait | 3 | low-level | 68.4%<br>76.0%<br>74.1% | wearable |
| [75] | 2016 | RF | custom | 6 | low-level | 100% | smartphone |
| [100] | 2016 | interval-based<br>model,<br>Allen's interval<br>relations | custom | 5 | high-level | 98%<br>96% | wearable,<br>object,<br>environmental |
| [101] | 2017 | Allen's interval<br>relations,<br>Bayesian<br>network-based | custom | 5 | high-level | 98% | wearable,<br>object,<br>environmental |
| [102] | 2017 | Ameva<br>algorithm | custom | 8 | low-level | 95% | smartphone |
| [103] | 2017 | MCODE | custom | 6 | low-level | 73% | smartphone |
| [103] | 2017 | MCODE | custom | 5 | high-level | 86% | smartphone |
| [103] | 2017 | MCODE | custom | 3 | low-level | 88% | smartphone |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [74] | 2017 | k-means clustering | custom | 7 | low-level | 81% | environmental, wearable |
| [63] | 2017 | RF,<br>NN,<br>J48 | custom | 6 | low-level | 98.28%<br>97.84%<br>97.26% | smartphone, environmental |
| [63] | 2017 | RF,<br>NN,<br>J48 | custom | 6 | low-level | 95.06%<br>94.01%<br>93.62% | smartwatch, environmental |
| [104] | 2017 | regression tree | custom | 8 | low-level | 95% | environmental, wearable, smartphone, camera |
| [71] | 2017 | Random Forest | custom | 7 | high-level | 87% | smartphone, wearable |
| [73] | 2017 | LDA,<br>K-mean | custom | 9<br>8 | high-level<br>low-level | 89% | smartphone, smartwatch, wearable |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [77] | 2017 | LDA, AdaBoost, HMM, CRF | Smartphone Dataset | 7 | low-level | 97.24% | wearable |
| [77] | 2017 | LDA, AdaBoost, HMM, CRF | UCI HAR | 6 | low-level | 91% | wearable |
| [86] | 2017 | HMM, CRF, NB | Van Kastersen | 10 | high-level | 70.50% | environmental |
| [86] | 2017 | RF K-NN | Opportunity | 10 | high-level | 94% | wearable |
| [76] | 2018 | RF | custom | 7 | low-level | 94% | wearable, environmental |
| [76] | 2018 | RF | Berkeley MHAD | 11 | low-level | 99.50% | wearable, environmental |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [76] | 2018 | RF | custom | 4 | low-level | 92.50% | wearable, environmental |
| [92] | 2018 | LSTM | Ordonez | 9 | high-level | 99.47% | environmental |
| [76] | 2018 | RF | UTD-MHAD | 26 | low-level | 98% | wearable |
| [92] | 2018 | LSTM | CASAS | 8 | high-level | 97.68% | environmental |
| [105] | 2019 | CNN | custom | 6 | low-level | 95.70% | smartphone |
| [93] | 2020 | Lightweight RNN<br>LSTM | WISDM | 6 | low-level | 95.78% | smartphone |
| [106] | 2020 | bidirectional LSTM | custom | 6 | low-level | 96.34% | wearable, environmental |
| [107] | 2020 | LSTM,<br>CNN,<br>RNN | MHEALTH | 6<br>6<br>6<br>5 | low-level | 94.05%<br>83.42%<br>81.32%<br>85.92% | wearable, environmental |
| [108] | 2020 | LSTM | custom | 4 | low-level | 93.20% | wearable |

The use of increasingly advanced logical hypotheses of activities such as Event Calculus (EC) as part of knowledge driven approaches is an important subject for the problem of human activity recognition. Event Calculus (EC) is an event-driven language in which events and their consequences are presented and reasoned [109]. A domain is formalized in EC using events, fluents, and predicates. The events are represented as occurrences that occur or take place, while fluents are defined as any domain property that can change over time. Predicates are characterized in terms of their connections to events and fluents. In [34], the authors proposed an EC-based approach for representing high-level activities that require the execution of several events sequentially or concurrently. Sensor data is processed in this approach to classify the presence of events; the detected events are then used by a reasoning engine to identify the individual high-level activities such as eating breakfast and reading the newspaper. In [110], an advanced EC-based approach for recognizing long-term human activities is proposed. Long-term activities are defined in this context by placing temporal restrictions on short-term activities. For example, the long-term activity immobile is identified by using the short-term activity inactive for a period of time which is greater than a specified value, where the long-term activity immobile is characterized as a person sitting on a chair or on the floor, or has fallen on the floor. In [111], the authors proposed a more robust EC implementation for computing the variations of long-term activities. Additionally, this research demonstrates how fragmented short-term activity narratives, ambiguous annotation of short and long-term events, and a small dictionary of short-term activities and contextual variables both influence the performance of the activity recognition. Finally, in [112], the authors summarized the most groundbreaking methods and algorithms for dynamic event detection, such as EC-based systems, and explores the key conceptual relations and distinctions

between all these models.

Due to their strong cognitive interpretation of the ambient environment and reasoning capacities, ontologies have been commonly used to infer human activities [113]. In [114], the authors employs an ontology to model both sensors and human activities as ontological classes, taking into account the entity-based and location-based definitions of the activities. In [115], [116], [117], [118], multiple ontologies are proposed to represent various human activities classes; these classes are then exploited by different reasoning engines to infer different human activities. The advantages of the ontological-based approaches over data-driven approaches are discussed in [119]. Additionally, this work demonstrates that the use of ontologies in case-based reasoning methods enables data-driven approaches to address the overfitting issues in case of limited size training data [120]. To express the relationships between established concepts in an ontology, the ontology can be combined with a rule-based language such as SWRL [121]. The success of ontology-based methods is strongly dependent on the kind of ontology used and its relationship to the rule-based language chosen. OWL-DL is used in [118] to classify high-level activities by using contextual features that allow the representation of particular aspects of the physical ambient environment. This research demonstrates how expressive, rational, and conclusive algorithms can be used to identify differences between the contextual features and to recognize human activities.

Due to the fact that the expressiveness shortcomings of OWL 1 have been recognised in different fields, this language has been enhanced and developed, but in the meantime it's still keeping the DL decidability components. Consequently, The language OWL 2 has been evolved as a product of this expansion. In contrast to OWL 1, OWL 2 adds additional functionality; these new features are proposed to make it easy to read or express the represented knowledge,

such as disjoint union of classes, while others include new expressivity, such as richer data types, keys, property chains, and disjoint property. In [117], the authors investigated the use of OWL 2 for classifying different high-level activities.

Due to the complex dynamic nature of AAL systems, there are many conditions/requirements for encoding temporal relationships between events occurs in the ambient environment using these systems. In this context, certain knowledge-based approaches allow the provision of a set of strategies for dealing with these temporal representation. In [122], a combination of ontological and temporal knowledge representation formalisms is proposed. In this research, the authors proposed a series of rules based on the Allen temporal relations to infer high-level dependencies between activities and thus the ongoing high-level activities such as cooking and making tea. The method suggested in [122] is strengthened in [123] by taking into account two properties of the Time concept including: temporal relation and model of transformation. The temporal relation determines the exact moment or time period which a low or high-level action occurs, while the model of transformation shows that a high-level activity is composed of two or more low-level events in a given temporal inference.

In [124], an ontological-based approach was proposed for recognizing high-level activities. The proposed ontology enables the use of DL to encode temporal operators such as Allen's Interval Algebra. This approach enables the identification of high-level activities and events with significant temporal associations in their context. In [125], more advanced ontological-based approach is proposed for understanding human high-level activities using SPARQL queries and OWL 2 interaction patterns. SPARQL is used as a uniform declarative language for aggregating, translating, and enriching low-level contextual RDF in-

formation with higher-level derivations. The SPARQL-based reasoning engine enables the proposed approach to employ both of temporal reasoning and dynamic declaration of entities to recognize the performed activities. The authors of [126] and [127] proposed a SPARQL-based architecture for recognizing high-level activities by combining SPARQL CONSTRUCT graph patterns with OWL ontologies. In [128], the authors proposed an OWL 2 ontology to represent human daily activities and the specifications of sensors used to recognize these activities. The proposed ontology takes into account the measuring properties and characteristics of sensors platform, including their position. The aim of this analysis is to determine the most suitable sensors for recognizing real-world activities. This ontology is based on the W3C-standard Semantic Sensor Network (SSN) ontology for representing sensor networks. The SSN Ontology lacks some sensor definitions, including sensor types, characteristics, properties, units of measurement, and locations, which must be described in an external ontologies. In [128], an extended ontology for representing these definitions is proposed by expanding the SSN ontology.

In [129], The authors proposed an OWL 2 ontology to represent the AAL domain's goals, measurements, roles, and sensors. The ontology facilitates the representation of criteria, such as essential activities, for achieving a goal, such as human activity recognition. The Goal-oriented Requirement Language (GRL), a language that enables goal-oriented representation and reasoning about goal requirements, is used to do goal analysis. The authors constructed the proposed ontology using existing AAL ontologies, including OntoiStar [130], DogOnto [131], KPIOnto [132], and SSN [133]. In [134], The authors proposed a goal-based ontology to describe the conviction, desire, and intention concepts, where the conviction concept denotes a user's perception of his/her world, the desire concept denotes the user's goals, such as "making a coffee," and the in-

tention concept denotes the actions taken by the user to achieve his/her goal. For example, the action "reaching for a cup" is performed in order to fulfill the intention "making coffee".

The authors in [135] provide an overview of ontologies for recognizing human activities. Various ontologies are discussed, including CoBrA-Ont, CONON, Delivery Context Ontology, SOUPA, mIO!, Pervasive Information Visualization Ontology (PiVOnm), CoDAMoS, and situation ontology. The aim of this analysis is to compare various ontology-based approaches by calculating the fatal flaw rate in ontology-based reasoning. The most popular flaws include having unconnected ontology objects, omitting annotations, omitting domain or range of properties, recursive meanings, and using various naming conventions throughout the ontology. In [37], a novel fuzzy ontology is proposed for gait-cycle recognition from skeleton-joint 3D coordinates. Besides, in [136], a 3D depth sensor ontology called Kinect ontology is proposed to model the information related to the user movement and the object interaction. Finally, in [137], a novel fuzzy ontology is proposed to represent the human activities, and reason about incomplete, and uncertain knowledge.

Table 3.2 outlines the strengths and weaknesses of various knowledge representation and reasoning formalisms, including RDF, various variants of OWL, SPARQL, and SWRL [138].

Table 3.2 – State of the art formalisms used for knowledge representation and reasoning.

| Engine | Description | Pros | Cons |
|---|---|---|---|
| RDF | A basic and main version of reasoning | Powerful reasoning | Reduced expressiveness |
| OWL Lite | Limited version of OWL DL | Powerful reasoning | Reduced expressiveness |
| OWL DL | DL-based language with an RDF syntax/ All of RDF documents are not OWL DL ontologies (Pellet, Her miT, FaCT++, RacerPro) | High expressiveness | High computational cost |
| OWL Full | Extended version of RDF, similar to OWL DL but it can support all RDF documents | High expressiveness | Unresolvable |
| OWL 2 EL | Reduce expressivity for providing light-weight ontologies (CEL, SHER, snorocket, ELLY (extension of IRIS)) | Low computational cost | Reduced expressiveness |
| OWL 2 QL | Scalabe reasoning for large dataset using SQL, (Owlgres, QuOnto, Quill) | Easy to use | Reduced expressiveness |
| OWL 2 RL | A reasoning version with simplified modeling and implementation, Fast and scalable implementation(OWLIM, Jena, Oracle OWL) | High expressiveness Easy to use | Reduced scalability |
| SPARQL | SPARQL can express queries across diverse data sources (query language for RDF) | W3C standard | Lack of tools (eg. editors) |
| SWRL | Combination of OWL Lite and the OWL DL with the Unary/Binary Datalog RuleML (Rule Markup Language) | High expressiveness | Undecidable |
| DL-Safe rules | A decidable fragment of SWRL | Resolvable | Reduced expressiveness |

### 3.2.5 Hybrid approaches

Hybrid approaches merge data-driven and knowledge-based approaches in order to overcome their shortcomings while maximizing their benefits. In [139], the authors proposed a hybrid approach to recognize human activity by focusing on the relationships between human activities and their associated objects. The research makes use of commonsense knowledge of how objects can be used during the daily activities. Then, by combining HMM and DBN, a probabilistic model of object usage, physical actions, and behaviors is used to understand high-level activities such as preparing tea. In [140], a hybrid approach based on an MLN is proposed for recognizing human activities. In order to recognize high-level daily activities with the potential to handle data uncertainty, the authors used MLN to combine the commonsense knowledge with a probabilistic model. In that study, the data uncertainty corresponds to the lost sequences of events occurring during an activity due to sensor data errors. In [141], the authors expanded analysis done in [140] by including probabilistic ontology-based activity recognition. The proposed probabilistic ontology-based activity recognition system represents the domain knowledge using an ontology, which simplifies the probabilistic reasoning based on MLN. The probabilistic ontology enables the provision of coherent knowledge by allowing the representation of complexity, time, uncertainty, and high-level activities in a unified system.

In [142], the authors proposed an MLN-based approach, the MLN soft rules were used to model the low-level activities by learning the context of the activities, such as spatio-temporal contextual information. To model the high-level activities, the MLN hard rules were used. In [143], to recognize human daily activities and manage the sensor's uncertainty, a hybrid approach is used. In that study, the sensor data and the associated uncertainty are represented us-

ing the ontological modeling. In other hands, the production rules are used to estimate the activity and associated uncertainty score from sensor data. Each production rule is composed of a set of clauses connected by logic operators, where a clause is represented by an RDF triple to model the sensor data with its associated uncertainty using the ontology. The Dempster Shafer hypothesis is used to classify the activities depending on the detected actions. In [144], the authors proposed a hybrid approach for human activity recognition that combines both of probabilistic and symbolic reasoning based on log-linear DL. The log-linear DL, implemented in [145], combines DL [146] and probabilistic log-linear models [147] in order to handle the data uncertainty. In [148], Th authors proposed a probabilistic EC-based framework for recognizing human activities. This framework enables the management of data uncertainty by exploiting the linear-time algorithm, where data uncertainty corresponds to the missing sequences as a result of sensor failure. In this research, to recognize the high-level activities, the linear-time algorithm is exploited to calculate all maximal temporal intervals that meet a predefined probability threshold. In [137], a fuzzy OWL 2 ontology is exploited to model the uncertain domain events. It enables the representation of uncertainty and the temporal relationship between different based on fuzzy state machine model. Compared to classical ontologies, the fuzzy ontologies are able to handle the implicit uncertainty in real-world domains. Besides, the fuzzy ontologies are far more realistic and have a more coherent world view than classical ones [137]. By exploting the fuzzy ontologies, the system is able to include the relative results with the exact matching of the queries using advanced search algorithms. Besides, fuzzy ontologies benefit from their semantics, which makes them more adaptable to be mapped between various ontologies [137].

In [39], a novel hybrid approach is proposed for recognizing human high-

level activities by using a pattern learning algorithm and MLN. The pattern learning algorithm is exploited to learn the relations between the probabilistic interval of different events. By combining the pattern learning algorithm with MLN, the approach was able to infer the uncertain high-level activities based on the learned relations. Seven different Allen's temporal relations and one hierarchical relation are considered in this analysis. Finally, In [149] , the authors proposed a hybrid approach for human activities recognition that combines HMM and symbolic logic. The HMM algorithm is used to identify the atomic activities. Using an symbolic logic reasoning , the approach was able automatically to recognize high-level human daily activities.

## 3.3 Review of human activity anticipation

Activity anticipation is the process of inferring human activities ***during or before*** of their execution [41]. Consequently, the activity anticipation systems anticipate the activity labels before the termination of those activities, which can be useful in critical emergency scenarios such as the detection of drinking detergent instead of water. In [150], two types of activity anticipation categories are distinguished: physical-activity anticipation [151, 152, 153, 154, 155] and mind-intention anticipation [156, 157, 158, 150].

### 3.3.1 Physical-activity vs Mind-intention anticipation

Physical-activity anticipation studies are interested in recognizing the physical actions with the objects populating the ambient environment. The physical-activity anticipation can be divided into two main categories, short-term and long-term activity (physical-intention) anticipation [41]. On one hand, short-term activity anticipation focuses on the prediction of relatively short period

activities, which may take several seconds. On the other hand, long-term activity (physical-intention) anticipation focuses on inferring human activities in the far future based on the currently observed actions.

Unlike physical-activity anticipation, mind-intention anticipation studies are interested in the interpretation of the mental desire behind "**why**" a human is looking to a specific object in a given place. In the earlier cognitive studies [159], four main fixation roles were defined to describe human mind intentions: locate, direct, guide, and check. These roles were extended in [150] to analyze human intentions in complex daily activities: 1) locate: the person recognizes the location of an object; 2) direct: the person's hands move toward an object to do an action; 3) guide: the person guides an object toward another one; 4) check: the person checks the object state.

### 3.3.2 Human activity anticipation approaches

Human activity/action anticipation and human intention anticipation are often used interchangeably [135], [160], [161], [162], [163]. In [161], a GMM and Gaussian regression models are proposed to anticipate human actions from wearable sensors. In [162], a combination of CNN model and MLP neural network are proposed to extract automatically the spatial features and to anticipate human daily activities. In [164], a unified framework for anticipating human short-term intentions based on DBN is proposed. In [163], the authors proposed an approach for anticipating human short-term activities based on fuzzy inference. This approach enables the anticipation of human actions by integrating the inferred knowledge about currently recognized activities with predictions based on previously observed activities. The fuzzy inference is used to merge these two forms of knowledge in order to anticipate short-term activities. In [165], the authors proposed a hybrid approach for anticipating hu-

man daily activities activities. To learn the user preferences from ontologies, a bottom-up method is proposed in this approach . A Restricted Boltzmann Machine uses the learned preferences to anticipate human daily actions. In [166], The authors proposed a hybrid framework for anticipating human long-term activities by combining both the Planning Domain Definition Language (PDDL) and DBN model. The PDDL is a predicate-based language commonly used for action monitoring, planning, and execution. The framework modeled different human activities are modeled based on the PDDL language. Then, the DBN model is used to anticipate the future activities by calculating the probability distribution for different activities.

The previously-mentioned research studies considered a small range of human activities. This is because the difficulties in obtaining vast amounts of labelled data for anticipating human activities to train data-driven approaches. In knowledge-driven approaches, as mentioned previously, they failed to address the data uncertainty and ambiguity, which reduces their capabilities to anticipate complex human activities. Anticipating different human activities is not well investigated and evaluated using hybrid approaches due to the difficulty of obtaining a dataset that combines both the contextual knowledge and multi-modal sensors for human activity anticipation.

## 3.4 Emotion recognition

### 3.4.1 What are emotional expressions?

Emotional expressions are the behaviors that communicate our emotional state or attitude to others. They are expressed through verbal and non-verbal communication. Complex human behavior can be understood by studying physical features from multiple modalities; mainly facial, vocal and physical gestures.

Recently, spontaneous multi-modal emotion recognition has been extensively studied for human behavior analysis.

Darwin concluded through his observations and descriptions of human emotional expressions that emotions adapt to evolution, are biologically innate, and universal across all human and even non-human primates ([167]). Formal, systematic research studies have since been realized on the universality of emotions. This work demonstrated: (i) the universality of six basic emotions (anger, disgust, fear, happiness, sadness and surprise) and (ii) the cultural differences in spontaneous emotional expressions ([168]).

### 3.4.2 Emotions vs Affects

A human's emotion resulting from an interaction with stimuli is referred to as an *affect*. In psychology, an affect refers to the mental counterparts of internal bodily representations associated with emotions. In fact, humans express affect through facial, vocal or gestural behaviors. The notion of affect is subjective, and in the literature it is represented by two alternative views: **the categorical view** where affects are represented as discrete states with a wide variety of affective displays and **the dimensional view**, where we suppose that affects might not be culturally universal and alternatively, should be represented in a continuous arousal-valence space, see figure 3.1. Recently, a trend in the scientific community has emerged towards developing new technologies for processing, interpreting or simulating human emotions through Affective Computing or through Artificial Emotional Intelligence. Consequently, a broad range of applications have been developed in Human-Computer Interaction, health informatics and assistive technologies.

### 3.4.3 Human emotions / affects recognition approaches

In recent years, many researchers have focused on the creation of human-like social robots. Only a few studies are interested in providing robots with human emotion recognition capabilities [169, 170, 171, 172]. Some researchers have focused on unimodal affect/emotion recognition where only single source is used such as facial expressions [169, 171, 173, 174, 175, 176], voice [177], textual expressions [178, 179], and body language [180]. Other researchers have investigated multimodal techniques using visual and spoken information [181, 182, 183, 170, 172]. Using multimodal sources allows, on the one hand, to increase robustness and performance in terms of emotion recognition due to the complementarity and diversity of information when multiple modalities are available, and on the other hand, to overcome resources unavailabilities when one modality isn't available by using the remaining modalities [184].

### 3.4.4 Data driven approaches

Although data-driven approaches are deeply linked with training data, they have been explored in the above studies to recognize emotions. Among the used models, Artificial Neural Networks (ANN) [175], Hidden Markov Models (HMM) [176], and Bayesian Networks [172]. These approaches allow only the recognition of observed emotions which can be summarized in the following seven affective/emotion states: *sad, fear, anger, happy, neutral, surprise, and disgust.*

More recently however, work in affective computing has paid more attention to multimodal emotion recognition by developing approaches to multimodal data fusion. Research on affect recognition has seen considerable progress as the focus has shifted from the study of laboratory-controlled databases to

databases covering real-world scenarios. In traditional emotion recognition databases, subjects posed a particular basic emotion in laboratory-controlled conditions. In more recent databases, videos are obtained from real-life scenarios with *in-the-wild* environmental conditions and less constrained settings, which exhibit characteristics like illumination variation, noise, occlusion, non-frontal head poses, and so on. Today, automatic emotion recognition of the six basic emotions in acted visual and/or audio expressions can be performed with high accuracy. However, in-the-wild emotion recognition is a more challenging problem due to the fact that spontaneously occurring behavior varies more widely in its audio profile, visual aspects, and timing. Multimodal fusion for emotion recognition concerns the family of machine learning approaches that integrate information from multiple modalities in order to predict an outcome measure. Such is usually either a class with a discrete value (e.g., happy vs. sad), or a continuous value (e.g., the level of arousal/valence), as shown in figure 3.1. Several literature review papers survey existing approaches for multimodal emotion recognition ([185, 186, 187, 188]). There are three key aspects to any multimodal fusion approach: (i) which features to extract, (ii) how to fuse the features, and (iii) how to capture the temporal dynamics.

**Extracted features:** several handcrafted features have been designed for audio-visual emotion recognition (AVER). These low-level descriptors concern mainly geometric features like facial landmarks. Meanwhile, commonly-used audio signal features include spectral, cepstral, prosodic, and voice quality features. Recently, deep neural network-based features have become more popular for AVER. These deep learning-based approaches fall into two main categories. In the first, several handcrafted features are extracted from the video and audio signals and then fed to the deep neural network ([189, 190, 191]). In the second category, raw visual and audio signals are fed to the deep network

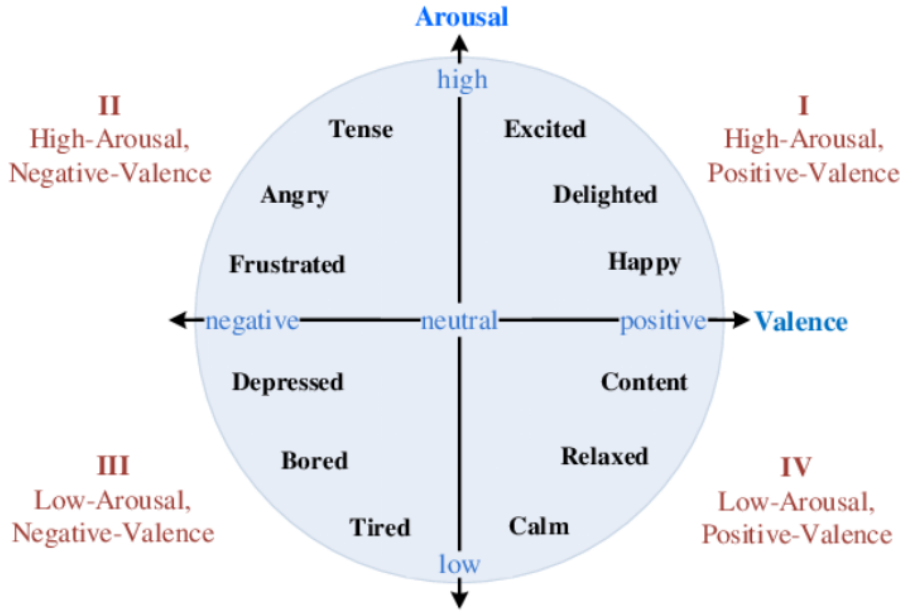



Figure 3.1 – Discrete vs Continuous emotion mapping

([192, 193, 194]). Deep convolutional neural networks (CNNs) have been observed as outperforming other AVER methods ([185]).

**Multimodal features fusion:** An important consideration in multimodal emotion recognition concerns the way in which the audio and visual features are fused together. Four types of strategy are reported in the literature: feature-level fusion, decision-level fusion, hybrid fusion and model-level fusion ([195, 188]). *Feature-level fusion* also called *early-fusion* concerns approaches where features are immediately integrated after extraction via simple concatenation into a single high-dimensional feature vector. Such is the most common strategy for multimodal emotion recognition. *Decision-level fusion* or *late fusion* concerns approaches that perform fusion after an independent prediction is made by a separate model for each modality. In the audio-visual case, this typically means taking the predictions from an audio-only model, and the prediction from a visual-only model, and applying an algebraic combination rule of the multi-

ple predicted class labels such as 'min', 'sum', and so on. *Score-level fusion* is a subfamily of the decision-level family that employs an equally weighted summation of the individual unimodal predictors. *Hybrid fusion* combines outputs from early fusion and from individual classification scores of each modality. *Model-level fusion* aims to learn a joint representation of the multiple input modalities by first concatenating the input feature representations, and then passing these through a model that computes a learned, internal representation prior to making its prediction. In this family of approaches, multiple kernel learning ([196]), and graphical models ([186, 197]) have been studied, in addition to neural network-based approaches.

**Modelling temporal dynamics:** audio-visual data represents a dynamic set of signals across both spatial and temporal dimensions. [185] identify three distinct methods by which deep learning is typically used to model these signals: *Spatial feature representations:* concerns learning features from individual images or very short image sequences, or from short periods of audio. *Temporal feature representations:* where sequences of audio or image inputs serve as the model's input. It has been demonstrated that deep neural networks and especially recurrent neural networks are capable of capturing the temporal dynamics of such sequences ([198]). *Joint feature representations:* in these approaches, the features from unimodal approaches are combined. Once features are extracted from multiple modalities at multiple time points, they are fused using one of strategies of modality fusion ([189]).

### 3.4.5 Knowledge driven approaches

To deal with issues related to emotions and affects, some studies have investigated knowledge-driven approaches as an alternative to data-driven approaches. Different ontologies were proposed in the literature, such as the semantic lex-

icon of feelings proposed in [199] and *WordNetAffect*, the extension of the lexical database *WordNet*, proposed in [200]. Recently, an ontology, called *Emotion Ontology*, covering all aspects of emotion, affect and mental states from the neuroscience point of view, was proposed in [201]. In [202], Gil et al. propose an upper ontology, called *EmotionsOnto*, to describe emotions and their recognition systems. Developing cognitive robots with the capability of managing natural interactions with humans according to their emotional contexts needs that robots and all entities populating the ambient environment share semantically the same knowledge. In this perspective, an ontology covering all commonsense concepts of human mental states and entities populating the environment is needed. The ontologies proposed recently to model human emotions and affects are mostly dedicated to specific applications such as social networking or e-learning. Furthermore, most of these ontologies are not flexible enough to be applied for assistive robotics. Therefore, to better address human-robot interaction scenarios, it is important to build a flexible ontology that allows an easier extension with commonsense knowledge and concepts through other existing ontologies.

## 3.5 Discussion

In this chapter, we presented a comprehensive review of human activity recognition, activity anticipation, and emotions recognition approaches. In the context of ambient intelligence, the purpose of activity recognition / anticipation is to analyze, and interpret the different events and actions generated by the ambient objects and people to understand the context of their behaviour or even anticipate their intentions. The different challenges facing the recognition and anticipation of complex activities/intentions are detailed as follows :

- Recognition of concurrent activities: In some cases, activities can be carried out simultaneously. A person can perform several activities at the same time, he/she can walk while hand-waving to someone. This kind of activities should be recognized with specific kind of approaches, which are different from sequential activities recognition models. The recognition of these concurrent activities can be affected by the racing between activities, and thus it leads to non-deterministic and undesirable recognition behaviours.

- Recognition of interlaced activities: Some activities can be suspended or interlaced to make another one. For example, if a person is preparing a coffee and the doorbell rang, he will pause the task of preparing the coffee and go to check who is there, then go back to prepare the coffee.

- The ambiguity in the intentions interpretation: Interpretations of similar activities may differ from one intention to another. For example: taking a spoon "activity" may be part of several different "intentions" such as "cooking", "making coffee" or "taking a medicine".

- Recognition of activities involving several people (Group activities): It is possible to have several people present in an activity and each of these people can do another activity in parallel (for example, in a meeting one person speaks and another writes) or several people doing the same activity (for example, "a choreography").

The reviewed approaches are divided into three different categories: data-driven, knowledge-driven, and hybrid approaches. The primary shortcomings of data-driven methods is their failure to take the ambient context and the expert knowledge into account. The majority of data-driven approaches of human activity recognition and anticipation ignore the contextual information of

human actions and activities, such as spatio-temporal attributes, objects of interest, the orders of activities, level of activity, and length of activity. Besides, additional contextual knowledge are rarely considered in the literature, such as relationships between atomic actions and low-level activities, or between low-level activities and high-level activities. These relationships are critical for understanding the contextual human activities and help to anticipate them. However, the ability to represent and model the knowledge, check the activity properties during axiomatization, as well as verifying the characteristics of human action and activities, is the key benefit of knowledge-driven approaches. However, knowledge-based approaches are not able to handle both data and sensors uncertainty during the recognition and anticipation of human activities. To get the best of both cases, the hybrid approaches are proposed to combine both data-driven and knowledge-driven approaches. The majority of hybrid approaches employ data-driven part to recognize the uncertain input data as atomic actions, and low-level activities, while the knowledge-driven part is used to infer the implicit relationships between the detected activities, apply both spatial and temporal reasoning to anticipate human actions, and to recognize the longer-term activities based on the inferred relations.

Regarding the emotion recognition, despite the considerable advancements in the field, multi-modal emotion recognition has remained a challenging task. Most existing methods still lack generalisation ability across datasets acquired under different conditions. The majority of modern approaches in the emotion recognition literature either learn or fine-tune an end-to-end network that can only be used for a specific dataset, and/or use more general features as input to a more basic model. To overcome these limitation, it is essential to learn an independent feature extractor for each modality, that could be employed for any dataset. This approach could achieve a good level of generalisation by training

on multiple labeled datasets.

In the AAL domain, one of the main challenges is developing a unified framework to perceive and represent the contextual knowledge associated with the entities populating an ambient intelligent environment (human, objects, robots, sensors, actuators, etc.). In fact, two types of contextual knowledge can be distinguished here: (1) those are directly observable from sensors or human-machine interfaces (ambient sensors, localization systems, verbal interaction interface, etc.) and (2) those are non-observable but can be inferred (user emotion, activity/intention, abnormal activities, etc.). To achieve those objectives, it is necessary to develop a unified, generic, and expressive framework, allowing the perception and the representation of the spatial and temporal dimensions of the ambient environment and taking into account the multi-modal nature of AAL systems. Consequently, this framework allows to obtain a richer, finer, and more coherent modeling of the spatial and/or temporal knowledge of the context, thus providing a better adaptation of the user assistance services. To tackle the previously mentioned challenges, the contributions of this thesis can be summarized as follows:

- A unified framework combines both of human egocentric and environmental vision perspectives that allows the recognition and proactively anticipation of human daily activities.

- A novel Multi-Modal approach that allows the recognition of human emotions from visual and auditory data and taking into account the contextual attributes of human reactions.

- A novel Hierarchical Multichannel Deep Residual Network (HMResNet) model that allows the recognition of low-level human daily activities from uncertain wearable inertial sensors.

- A novel 3D Spatio-Temporal Joint based Convolutional Neural Network (STJ-CNN) that allows the recognition of high-level human daily activities from skeleton depth information.
- ConceptNet [203] based ontology that allows the modelling of the ambient environment and to infer the implicit relations between the objects populating the ambient environment and the human activities.
- Two novel datasets for human hand detection in the cluttered environment are proposed: DHA-$11_{\mathrm{TH}}$ (Diverse Hands) and SKNS (skin/non-skin texture) datasets
- Validating the proposed approaches through the implementation of different personal assistance scenarios using the living lab experimentation environment at the LISSI laboratory.

Chapter 4

# Narrative Knowledge Representation and Reasoning

## 4.1 Introduction

In this chapter, we present the foundations of the Narrative Knowledge Representation Language (NKRL) used for context representation in AAL environments. First, we present the definitions used throughout this chapter. Then, we describe, on the one hand, the foundations of the modelling of narrative knowledge by exploiting the HClass and HTemp ontologies, and on the other hand, the NKRL reasoning mechanisms.

## 4.2 Fundamentals of NKRL

This paragraph presents the main concepts used in NKRL:

- Narrative event: It corresponds to an event reported in a story, which is composed as sequence of events related by means spatial and temporal modulators. There are two types of narrative events: The fictional nar-

rative and non-fictional (or factual) narrative. The former allows representing events from a simulated world and the second, events that occur effectively in the real world.

- Elementary event: This is an event in which a change in the state of the world occurs at a given time. The change can be localized spatially and temporally, such as "open the window of the living room", "call a friend in the evening", "prepare breakfast in the early morning", etc.

- Static entity: It may either refer to a material object (wall, washbasin, etc.) or an immaterial object (piece of music, e-mail, etc.); a static entity has the characteristic of not changing during the entire life of the application.

- Dynamic event: It is defined as a set of elementary events characterizing the behavior of an entity (person, object, etc.), such as: An elderly person uses his remote control to turn on his air conditioner, the system sends an alarm to the emergency services as soon as it detects that an elderly person has fallen, etc. ;

- Concept: In this work, the notion of concept is similar to that of concept or class in the semantic web. Properties or attributes can be associated with a concept. There are two types of concepts: Concepts that can be instantiated directly (sortal_concept) and concepts that cannot be instantiated directly (non_sortal_concept).

- Instantiated concept: For example, *CHAIR_125, BED_2012, ROBINET_45*, etc. are instances of the concepts *chair_, bed_, faucet_*.

- Non-instantiated concept: For example, from the concept *color_*, it is not possible to directly create an instance; *RED_120* cannot be considered as an instance since it makes no sense if used independently. However,

using the concept *color_* appearance, a specialization of the instantiated concept physical *_appearance*, it is then possible to associate a color to an instantiated concept, such as *RED_PLATE*.

- Predicate: It allows the representation of non-fictional events. There are seven predicates to model an event, an action, a state, a situation, etc., table 4.1.

- Role: There are seven conceptual roles: *SUBJECT , OBJECT , SOURCE, BENIFICIARY , MODAL, TOPIC and CONTEXT* . A role defines a functional relationship between a predicate and its arguments, and can be used to identify the actors of an event.

- AECS operators: There are four operators to define the relationships between the attributes of a role argument: ALTERNATIVE(A), ENUMERATION(E), COORDINATION(C) and SPECIFICATION(S). These operators are used to define. A relationship can be: disjunctive, distributive, collective or attributive, table 4.2.

NKRL was basically designed to represent and process natural language knowledge contained in mainly non-fiction documents, such as reports, memos, etc. It is based on two N-aires ontologies that guarantee an expressive representation of narrative knowledge, including elementary events and temporal relationships between these events. The concept ontology HClass allows the representation of portions of narrative knowledge using concepts. The event ontology HTemp allows the representation of dynamic events. Besides, NKRL allows combining predicates and conceptual roles to model narrative information, and to establish semantic links (causality, purpose, etc.) between elementary events. The inference mechanisms associated with the NKRL allow to establish implicit or explicit relationships between knowledge.

Table 4.1 – Knowledge representation basic predicates

| Predicate | Description |
|---|---|
| BEHAVE | This predicate makes it possible to model knowledge relating to the role or attitude that an entity (person, object, etc.) can adopt for a specific purpose, for example, an elderly person taking his or her medication, two friends watching a football match together, and so on. |
| EXIST | This predicate can be used to represent the fact that an entity may be present in a given place, e.g. an elderly person is in their bedroom, an elderly person has fallen at the entrance of a store, etc. |
| EXPERIENCE | This predicate is used to represent the fact that an entity may be affected by an event. For example, observing the increase in temperature in a given place, an elderly person having a heart attack, etc., is a good example. |
| MOVE | With the help of this predicate, it is possible to represent events related to entity movements, such as moving a piece of furniture, transmitting a message, etc. |
| PRODUCE | This predicate can be used to represent the execution of a task or activity by an entity, for example, an inertial measurement unit produces information related to the movements of an entity. |
| OWN | This predicate makes it possible to model the notion of ownership between entities, for example, a person owns an apartment, the room where the elderly person has fainted is part of the house, and so on. |
| RECEIVE | With the help of this predicate, it is possible to represent events related to the reception of information, such as, for example, a person receives a text message. |

Table 4.2 – AECS Operators

| Operator | Description | Logic description |
|---|---|---|
| ALTERNATIVE | Disjunction | OR |
| ENUMERATION | Distribution | Generalization |
| COORDINATION | Conjunction | AND |
| SPECIFICATION | Attribution | Specialization |

## 4.3 Dynamic knowledge modeling in NKRL

Representation of narrative knowledge is based on four main components:

- Definitional Component: This component is used to represent concepts such as: component, robot, actuator, sensor, etc. Concepts can range from the most general concepts, such as a human being *(human_being)* to the most specific concepts, such as a table. According to the naming rules imposed in the language, the identifier of a concept must be a string of lowercase characters ending with the underscore character "_".

- Enumerative component: This component can be used to represent an enumeration of the instances of the concepts in the definitional component. Naming rules for instances require that the identifier of an instance must be a string of uppercase characters ending with the underscore character "_". As examples, *STEVE_* and *ERIC_* are instances of the *human_being* concept and *ROBOT_PEPPER* is an instance of the *robot_*-concept.

- Descriptive component: This component allows representing the structures of the elementary events. An elementary event is modeled using a predicate and one or more roles; each role can have one or more arguments. The general structure of a predicate template, given in table 4.3, consists of several elements :

– PREDICATE: Corresponds to one of the seven basic predicates of the language: *PRODUCT, MOVE, EXPERIENCE, EXIST, OWN, RECEIVE, BEHAVE*.

The instantiation of a predicate occurrence follows the following rule:

$$(\mathrm{L}(\mathrm{P}(\mathrm{R}_1 a_1)(\mathrm{R}_2 a_2).....(\mathrm{R}_n a_n))) \tag{4.1}$$

where :

* L: A semantic label or symbol uniquely identifying a predicate occurrence, i.e. verifying the (UNA) Unique Name Assumption [1];
* P: predicate ;
* $\mathrm{R}_k (k = 1, ..., n)$: The set of roles associated with the occurrence ;
* $a_k (k = 1, ..., n)$: The set of arguments associated with the roles.

Table 4.3 – General structure of a predicate template

| PREDICATE |
|---|
| SUBJ { <argument > : [location] } |
| OBJ { <argument > : [location] } |
| SOURCE { <argument > : [location] } |
| BENF { <argument > : [location] } |
| MODAL { <argument >} |
| TOPIC { <argument >} |
| CONTEXT { <argument >} |
| [ modulators ] |
| [ temporal attributes ] |

[1] Unique Name Assumption is a commonly accepted assumption in most model-driven tools. It consists of assuming that different names will always denote different elements in the model. This is usually not true in DL reasoners because of the essential nature of knowledge integration problems.

- Argument: Represents the attributes associated with each role: *SUBJ, OBJ, SOURCE, MODAL, TOPIC, CONTEXT*.
- Location: Represents the space where the event occurs. The location and temporal attributes parameters are used for the temporal representation of knowledge as we will see later.

- Factual component: This component allows modeling instances of elementary events. An event instance corresponds to an occurrence of predicate (predicative occurrence).

In the MOVE predicate template, the NL Description symbol is used to specify a natural language description of the predicate, see table 4.4. Note that a role or variable, specified in square brackets, is optional. Concerning the MOVE predicate template, the roles *SUBJ, OBJ* as well as the variables *var1* and *var3*, must be specified; the roles *BENF, MODAL and CONTEXT*, as well as the variables *var2, var4, var5, var6 and var7*, are considered as optional. The variables *var1, ..., var7* are variables are associated with constraints which are used to check that the value assigned to each variable at the instantiation of an occurrence is consistent with the terms (concept, instances) contained in the definitional component. Thus, the constraints specified in the templates of the event ontology HTemp are associated with the concepts contained in the same ontology. Consequently, the knowledge consistency check relies on the concept ontology HClass to establish a hierarchy of concepts and instances according to the generalization/specialization principle.

### 4.3.1 Representation of temporal knowledge in NKRL

The representation of temporal knowledge in NKRL is based on the use of temporal annotations, also called temporal modulators. These annotations make

Table 4.4 – Template of MOVE predicate

| |
|---|
| NL Description : 'Transmit a Structured Information'<br>PREDICATE : MOVE<br>SUBJ var1 : [(var2)]<br>OBJ var3<br>[BENF var4 : [(var5)]]<br>[MODAL var6]<br>[CONTEXT var7 ]<br>{[modulators] !=abs}<br>{[temporal attributes]}<br>var1 = <human_being> \| <artefact_><br>var2 = <location_><br>var3 = <symbolic_label><br>var4 = <human_being> \| <artefact_><br>var5 = <location_><br>var6 = <media_> \| <information_support><br>var7 = <situation_> |

it possible to specify the beginning and/or end of an event, or to specify that an event occurred on a given date, table 4.5.

**Example 1 :**

Consider the following situation: John, an elderly person living alone, stayed in his living-room between the dates date-1 =10/21/2020/ 6:33 and date-2 =10/21/2020/ 7:06. This event is modelled by the occurrence aal1.c2 of the predicate EXIST, see table 4.6.

**Example 2 :**

The occurrence aal1.c4 of the predicate EXPERIENCE, shown in table 4.7, indicates the increase (the growth_ property) of the temperature at date date-1= 21/10/2020 8:53, but gives no indication on the end of this event. The temperature_ property is used to specify that it is a temperature; the LIVING_ROOM_2 instance represents the location where the event occurred.

**Example 3 :**

Table 4.5 – Time modulators

| Acronym | General description |
|---|---|
| begin | Start date of the event |
| end | End date of the event |
| obs | This modulator is used if no time information about the beginning or end of an event is available. For example, the system detected a temperature increase at 19:23. |

The instance of the predicate *PRODUCT*, shown in Table 4.8, is a representation of the following event: the house control system, described by the symbol *HOME_CONTROL_SYSTEM_1*, has detected (property detection_) the presence, in the kitchen (KITCHEN_1), of a person named John, at the time specified in the date attribute-1.

Table 4.6 – Example of occurrence of predicate EXIST.

| aal1.c2) EXIST SUBJ JOHN_ : (LIVING_ROOM_2)<br>date-1 : 21/10/2020/6 :33<br>date-2 : 21/10/2020/7 :06 |
|---|

Table 4.7 – Example of occurrence of predicate EXPERIENCE.

| aal1.c3)<br>EXPERIENCE SUBJ : HOME_CONTROL_SYSTEM_1 :<br>(SPECIF temperature_ LIVING_ROOM_2)<br>OBJ : growth_<br>{obs}<br>date-1 : 21/10/2020/8 :53<br>date-2 : |
|---|

Table 4.8 – Example of occurrence of the predicate PRODUCE.

| aal1.c4)<br>PRODUCE SUBJ HOME_CONTROL_SYSTEM_1<br>OBJ detection_ : (KITCHEN_1)<br>TOPIC JOHN_ date-1 : 21/10/2020/17 :19<br>date-2 :<br>On 21/10/2020/5:19 p.m.,<br>the system detects that John is present in the kitchen |
|---|

### 4.3.2 Binding occurrences

Binding occurrences can be used to create links between elementary events. Unlike predicate occurrences, these structures do not use the *PREDICATE* and ROLE symbols to specify semantic links. The creation of binding occurrences using operators such as *COORD, GOAL, CAUSE*, etc. must conform to the syntax below :

$$(operator[arg_1 arg_2 ... arg_n]) \tag{4.2}$$

where $(arg_1, ..., arg_n)$ represent occurrences of predicates.

**Example:** Consider the following statement: "The robot moves towards Steve to give him a medicine". This statement has two elementary events: i) "The robot moves towards Steve"; and ii) "The robot gives medicine to Steve". The following occurrences of the predicate MOVE represent these two events:

| Sent1.C2)<br>MOVE SUBJ ROBOT_1 : location_1<br>OBJ STEVE :location_2<br>date-1 :<br>date-2 :<br>Move : AutonomousPersonDisplacement |
|---|

| Sent1.C3) | |
|---|---|
| | MOVE SUBJ ROBOT_1 : location_2 |
| | OBJ medicine_ |
| | BENEF STEVE |
| | date-1 : |
| | date-2 : |
| | Move : TransferObject |

The *GOAL* link operator can, for example, be exploited to create a link between these two occurrences, and model the fact that the event represented by the occurrence Sent1.C3, represents the objective (or goal) associated with the occurrence Sent1.C2. Formally,

$$S_{ent}1.C1 : \text{GOAL}(S_{ent}1.C2; S_{ent}1.C3) \tag{4.3}$$

## 4.4 Concept and Event Ontologies

Knowledge representation in NKRL is based on two ontologies: (i) the concept ontology HClass represents a hierarchy of concepts where the definitional and enumerative components are defined, and (ii) the event ontology HTemp, a hierarchy of predicate templates allowing the representation of events, where the descriptive and factual components appear.

### 4.4.1 ConceptNet-based Ontology

The concept ontology HClass is a binary high-level ontology that describes general concepts of commonsense knowledge in most fields. The concepts and instances composing these branches are structured according to a taxonomy based on the subsumption relationship, also called hierarchy link (isA) and represented by the operator $\subseteq$. Formally, the concept, sortal_concept is defined in

the concept ontology as follows :

$$sortal\_concept\_ \subseteq h\_class \cap sortal\_concept \equiv entity\_; situation\_ \quad (4.4)$$

The concept ontology HClass is based on ConceptNet knowledge-graph [203], which is a developed knowledge graph version of the Open Mind Common Sense (OMCS) project [204], a general upper ontology that represents the knowledge of the basic things that any person may know. In addition, ConceptNet combines the OMCS knowledge with 1) Multilingual Information extracted from Wikitionary; 2) common knowledge collected based on "games with a purpose" [205, 206]; 3) a multilingual Japanese dictionary (JMDict) [207]; 4) a small set of facts which extracted from Wikipedia info-boxes (DBpedia) [208]; 5) multilingual WordNet [209] and linked data representation of WordNet [210]. Consequently, ConcepNet consists of over 8 million nodes and over 21 million edges. Besides, the ConceptNet knowledge graph is connected through 36 relations divided into asymmetric (directed) and symmetric (undirected) relations as follows :

- **Asymmetric relations**: Part-Of, At-Location, Has-A, Capable-Of, Is-A, Made-Of, Used-For, Causes, Causes-Desire, Created-By, Defined-As, Derived-From, Desires, Entails, ExternalURL , Form-Of, Has-Context, Has-First-Subevent, Has-Last-Subevent, Has-Prerequisite, Has-Property, Instance-Of, Manner-Of, Motivated-By-Goal, Obstructed-By, Receives-Action, Sense-Of, and Symbol-Of

- **Symmetric relations**: Antonym, Distinct-From, Near-To, Synonym, Related-To, and Similar-To.

For example, the concept ontology exploits a set of asymmetric relations such as At-Location, Is-A, and Used-For to define the notion of contextual af-

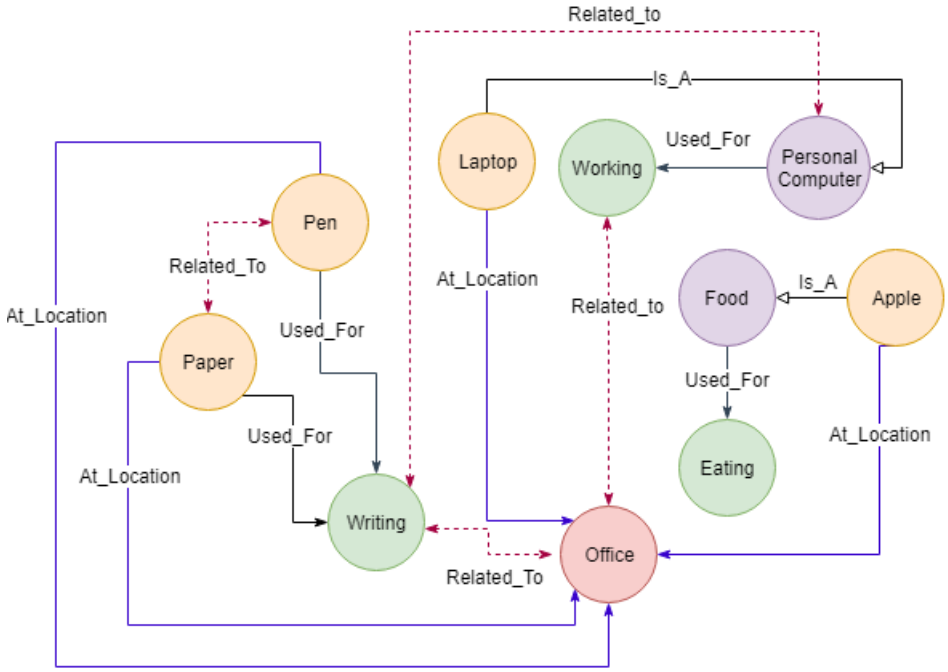



Figure 4.1 – An Example of the ambient-environment representation based on the concept ontology HClass

fordances as a relationship between the action classes and the object instances. Besides, the Related-To and Near-To symmetric relation are used to infer the implicit relations between the ambient-objects and the probable actions (activities), as shown in Fig.4.1

### 4.4.2 Event Ontology

The event ontology HTemp is a hierarchy of templates allowing to represent structured/dynamic knowledge of logically and semantically coherent and temporally ordered elementary event sets. It includes seven predicates:

1. BEHAVE: This predicate allows modeling the actions carried out by one or more dynamic entities or their behaviors. The predicate occurrence below represents the fact that Steve is a neighbor of John since date-1=

21/10/2020, see table 4.9.

Table 4.9 – BEHAVE Template example

| BEHAVE SUBJ STEV E_ |
| --- |
| OBJ (SPECIF neighbour_ JOHN_) |
| date-1 : 21/10/2020 |
| date-2 : |

2. EXIST: This predicate allows modeling an event concerning the presence of an entity in a given place. The statement "John is present in the living room at date1=21/10/2020/17:35", can be represented by the following occurrence of the predicate EXIST, see table 4.10.

Table 4.10 – EXIST template example

| EXIST SUBJ JOHN_ :(LIVING_ROOM_2) |
| --- |
| date-1 : 21/10/2020/17:35 |
| date-2 : |

3. EXPERIENCE: This predicate is used to represent an event affecting an entity (success, accident, etc.). For example, the following occurrence allows us modeling the fact that John suffers from respiratory failure since the date date-1=21/10/2020/17:35, see table 4.11.

Table 4.11 – EXPERIENCE template example

| EXPERIENCE SUBJ JOHN_ |
| --- |
| OBJ respiratory_distress |
| date-1 : 21/10/2020/17:35 |
| date-2 : |

4. MOVE: This predicate allows modeling events related to actions, such as moving a subject, sending an e-mail, etc. As an example, the following

occurrence models the fact that the robot ROBOT_PEPPER moved from the LOCATION_1 position to the LOCATION_2 position, at the date date-1=21/10/2020/11:36, see table 4.12.

Table 4.12 – MOVE template example

| MOVE SUBJ ROBOT_PEPPER : LOCATION_1 |
|---|
| OBJ ROBOT_PEPPER : LOCATION_2 |
| date-1 : 21/10/2020/11 :36 |
| date-2 : |

5. OWN: This predicate is used to represent the notion of ownership between entities or the state of an entity. The representation of the fact that the front door, represented by the symbol FRONT_DOOR_2, is unlocked since the date date-1=21/10/2020/17:30, is written as shown in table 4.13.

Table 4.13 – OWN template example

| OWN SUBJ FRONT_DOOR_2 |
|---|
| OBJ property_ |
| TOPIC unlocked_ |
| {obs} |
| date-1 : 21/10/2020/17:30 |
| date-2 : |

6. PRODUCE: This predicate is used to represent the execution of a task or activity by an entity. The instance of the predicate PRODUCE below is used to model the fact that the control system of the house, represented by the symbol HOME_CONTROL_SYSTEM has detected that John is sitting in the wheelchair represented by the symbol WHEELCHAIR_1, see table 4.14.

7. RECEIVE: This predicate allows representing events related to the recep-

Table 4.14 – PRODUCE template example

| PRODUCE SUBJ HOME_CONTROL_SYSTEM<br>OBJ detection_<br>TOPIC SPECIF(JOHN_ SPECIF<br>(sitting_activity WHEELCHAIR_1))<br>date-1 : 21/10/2020/12 :03<br>date-2 : |
|---|

tion of information. Using the predicate RECEIVE, the representation of the statement: "John received a phone call in the living room on date-1=21/10/2020/12 :06", can be written as shown in table 4.15.

Table 4.15 – RECEIVE template example

| RECEIVE SUBJ JOHN_ : (LIVING_ROOM_2)<br>OBJ (SPECIF<br>information_content PHONE_CALL_1)<br>SOURCE HOME_CONTROL_SYSTEM<br>date-1 : 21/10/2020/12 :06<br>date-2 : |
|---|

## 4.5 NKRL-based reasoning

The NKRL reasoning engine is based on the question(query)-answer principle. The processing of a question (query) is done by triggering a matching mechanism provided by the Filter Unification Module (FUM), and a mechanism for inferring hypothesis rules and transformation rules.

### 4.5.1 Matching mechanism

Based on the concept ontology HClass, the FUM module is in charge of matching (unifying) the occurrences present in the knowledge base with a query rep-

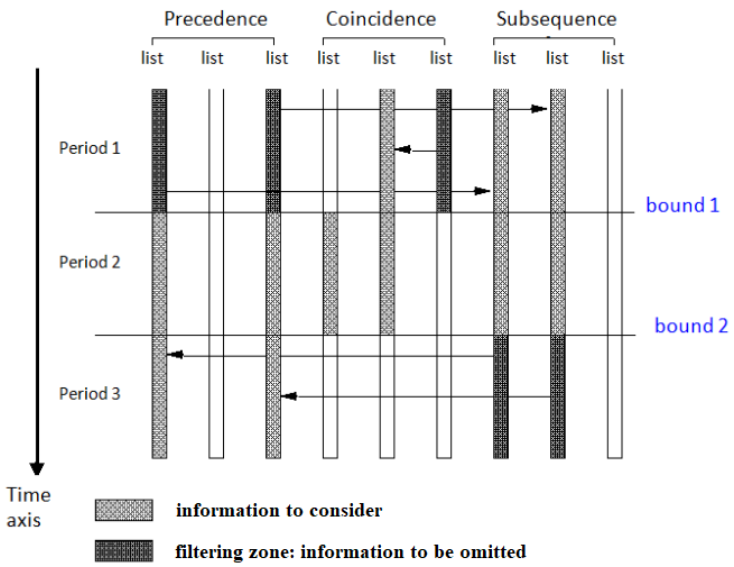



Figure 4.2 – Indexing algorithm

resented as a predicate occurrence. The generalization/specialization principle is used between concepts/instances associated with the concept ontology. Thus, any predicate occurrence matched with the query is a plausible answer to the query. The FUM module relies on two main functions: filtering and unification.

### Knowledge filtering

The knowledge filtering is based on a temporal reasoning algorithm which skips occurrences in the knowledge base whose dates do not fall within the interval specified in the query by the date-1 and date-2 attributes. In the knowledge base, each predicate occurrence is referenced using a pair of attributes: the label of the occurrence and the temporal attributes: date-1 and date-2.

To illustrate the principle of the reasoning algorithm used by the FUM module, consider the following statement: "John is a person who was recovering. He went home on date date-1= 21/10/2020. Five days later he was re-hospitalized",

see Figure 4.2. The occurrence below allows us modeling the fact that John was present at his home from date date-1= 21/10/2020, see table 4.16.

Table 4.16 – John was present at his home from date date-1= 21/10/2020

| aal1.c1) EXIST SUBJ JOHN_ :(HOME_)<br>[begin]<br>date-1 : 21/10/2020<br>date-2 : |
|---|

Now let's assume that for John's medical follow-up, we want to know if he was present at the hospital between 21/10/2020 and 29/10/2020. This request is written as follows, see table 4.17.

Table 4.17 – Query to check if John was present at the hospital between 21/10/2020 and 29/10/2020

| [query1] = EXIST SUBJ JOHN_ :(HOSPITAL_)<br>(21/10/2020, 29/10/2020) |
|---|

This request only allows selecting the occurrences whose dates are in the interval [21/10/2020- 29/10/2020].

In terms of the classification of events in the knowledge base, there are nine lists of dates; these lists are grouped into three sets of three lists each, see Figure 4.2: precedence, coincidence, subsequence. The precedence set groups the events that took place before the date specified in the date-1 attribute. The subsequence set groups events that took place after date-2. Finally, the coincidence set allows, as for it, to represent the events by using the obs modulator. Each list has three periods: period 1, period 2 and period 3 bounded by bound 1 and bound 2. Thus, period 1 represents the period before the bound 1 terminal, period 2 the period between the two terminals (bound 1 and bound 2), and period 3 the period after the bound 2 terminal.

In the above example, bound 1 corresponds to the lower bound (21/10/2020) of the interval [21/10/2020- 29/10/2020], and bound 2 to the upper bound (29/10/2020) of the interval [21/10/2020- 29/10/2020]. The predicate occurrence aal1.c1 only selects those occurrences whose date corresponds to the second bound 2, i.e. those occurrences whose date-1 attribute corresponds to the concidence set. Occurrences whose date corresponds to the period period 1 are excluded.

**Unification**

Unification is used to match queries as predicate occurrences with the occurrences in the knowledge base. The first step is to match the predicate of the query with the predicate of each occurrence in the knowledge base. If the matching is successful, i.e. both occurrences have the same predicate, then the matching of each query role with the corresponding role of the knowledge base occurrence is carried out. For example: Suppose Frank, John's nurse, was on duty during the night of the following interval: [date-1 = 25/10/2020/22:00 , date-2 = 26/10/2020/06:00]. The representation of this fact can be written as follows:

```
aal1.c2)
        BEHAVE SUBJ SPECIF (FRANCK_ nurse_)
                : (HOME_1)
        [begin]
        date-1 : 25/10/2020/22 :00
        date-2 : 26/10/2020/06 :00
```

Suppose that for the purposes of John's medical follow-up, we want to know which nurses were on duty between the following dates: date-1 = 25/10/2020/12:00 and date-2 = 26/10/2020/12:00. This query can be written as follows:

In the query "query2", the nurse is represented by the role SUBJ. The matching of the SUBJ role of the query with the SUBJ role of the occurrence "aal1:c2"

```
[query2] =
        BEHAVE SUBJ SPECIF (human_being nurse_)
                :(HOME_1)
        date-1 : 25/10/2020/12 :00
        date-2 : 26/10/2020/12 :00
```

involves matching the attribute "SPECIF (human_being , nurse_):(HOME_1)" with the attribute "SPECIF (FRANCK_ , nurse_):(HOME_1)". These two attributes are represented by the specification operator SPECIF and the location parameter defined by the instance HOME_1.

To match the instance FRANK_ with the concept human_being, we use the concept ontology HClass. Assuming successful matching, the occurrence aal1.c2 is selected as the response to the query2 query.

### 4.5.2 Hypothesis-Transformation Rules

The expressiveness of NKRL language is exploited, through inference rules, to automatically establish relationships between events or between characteristics of the same event. An inference rule is composed of a predicate, role(s), attribute(s) and a set of constraints represented by variables. It is formally defined as follows:

$$\mathrm{X}(var_i) \longrightarrow \mathrm{Y}(var_j); var_i \subseteq var_j \tag{4.5}$$

where X denotes the query represented as a predicate occurrence or link structure. Y denotes the set of inferred occurrences. In addition, the constraint $var_i \subseteq var_j$ imposes that the $var_i$ variables included in the X query are in the set of $var_j$ variables included in the set Y of inferred occurrences.

There are two types of inference rules: transformation rules and hypothesis rules. A transformation rule includes an antecedent corresponding to a generalization of the initial query, and one or more consequents. A hypothesis rule

is composed of a premise and one or more conditions, each corresponding to a reasoning step. In the reasoning process, a hypothesis rule can be used to search for the causes that led to a given context by reconstructing the links between the events that generated this context. Suppose, for example, that we want to explain why the air conditioning system stopped working on a summer evening in a given time interval, despite a high ambient temperature. The idea here is to show the existence, in the knowledge base, of a predicate occurrence proving that a person has stopped the air-conditioning system.

## 4.6 Conclusion

In this chapter, we presented the basic principles of NKRL language and the associated reasoning mechanisms. Knowledge representation is based on two N-aires ontologies that guarantee an expressive representation of knowledge. The concept ontology HClass is a hierarchy of general concepts of commonsense allowing to represent portions of narrative knowledge. The event ontology HTemp allows the representation of dynamic events. NKRL allows combining predicates and conceptual roles to model narrative knowledge, and also to infer semantic links (causality, purpose, etc.) between events. The reasoning engine is based on the question (query)-answer principle. The processing of a query relies on a mechanism for matching and filtering occurrences in the knowledge base, and a mechanism for applying hypothesis rules and transformation rules. In the context of ambient intelligent environments, the objective of this thesis is to exploit the high expressivity of NKRL to improve the ambient environment representation in terms of context recognition, and also to model human-robot interaction using verbal natural language.

Chapter 5

# Hybrid approach for contextual emotion recognition

## 5.1 Introduction

In this chapter, a hybrid approach for contextual emotion recognition for cognitive assistance services in ubiquitous environments is proposed. The proposed approach is able to recognize accurate explicit discrete emotions using a multilayer perceptron neural network fusion model combined with possibilistic logic. Besides, the proposed approach is able to recognize non-directly observable emotions using expressive emotional knowledge representation and reasoning model. The first section of the chapter discuss the motivation behind the proposed approach. The second section present the data-driven components of the proposed approach. Then, the contextual knowledge representation and reasoning are detailed in the fourth section. Finally, in the last section the evaluation of the proposed approach using different datasets is detailed.

## 5.2 Motivations

Considering *emotional* or *affective* aspects is fundamental for natural assistive interaction where robots act as companion entities that can support conversation, understanding, and responses [172]. Endowing ubiquitous robots with cognitive capabilities for recognizing emotions, sentiments, affects and moods of humans in their context is an important challenge to discern the meanings that a facial expression or a natural language dialogue can have, and decide how to react to a given situation.

Emotion is defined in [211] as an immediate cognitive, behavioral and physiological reaction following an event; the author differentiates emotion terms from other tightly coupled terms such as sentiments, affects and moods.

In recent years, many researchers have focused on the creation of human-like social robots. Only a few studies are interested in providing robots with human emotion recognition capabilities [169, 170, 171, 172]. Some researchers have focused on unimodal affect/emotion recognition where only single source is used such as facial expressions [169, 171, 173, 174, 175, 176], voice [177], textual expressions [178, 179], and body language [180]. However, these data-driven techniques that are generally highly dependent on learning data can be insufficiently effective to recognize non-directly observable emotions. The latter cannot be interpreted in an accurate way without considering the user's context [212].

In this chapter, *emotion contextual recognition* refers to the recognition of emotions of humans considering their context. To address the above-mentioned challenges, multimodal recognition and conceptual representation of the emotional knowledge in AAL systems must be powerful enough to supply a general description of the environment where the user evolves and thus better recognize non-directly observable emotions such as stressed, motivated, depressed,

..etc. In terms of applications, this chapter focuses on the cognitive assistance of people in a smart spaces by providing them relevant and timely information based on their needs.

## 5.3 Proposed hybrid approach

For more accurate emotion recognition in daily living environments, a hybrid model-based emotion contextual recognition approach for cognitive assistance services in ubiquitous environments is proposed. This model is based on: (i) a hybrid-level fusion exploiting a multilayer perceptron neural network model and the possibilistic logic; (ii) an expressive emotional knowledge representation and reasoning model to recognize non-directly observable emotions; this model exploits jointly the emotion upper-ontology (EmUO) and the n-ary ontology of events HTemp supported by NKRL language. The architecture of the proposed emotion recognition approach is shown in Figure 5.1.

### 5.3.1 Features extraction

For an accurate recognition of human' emotions, three modalities are taken into account in this study for the expressiveness of the information they contain: text, audio, and face. From the face modality, human's eye gaze, lips, brows, and face muscles motions and positions appear as good features enabling the recognition of facial expression. In the text modality, the words used, their syntactic structure, and their meaning represent the main keys of an emotion recognition. However, considering only the text can decrease significantly the recognition accuracy when relevant information in the audio are not available. The main relevant features in the audio modality are: the manner with which utterances are produced, the intensity and quality of the voice. To im-

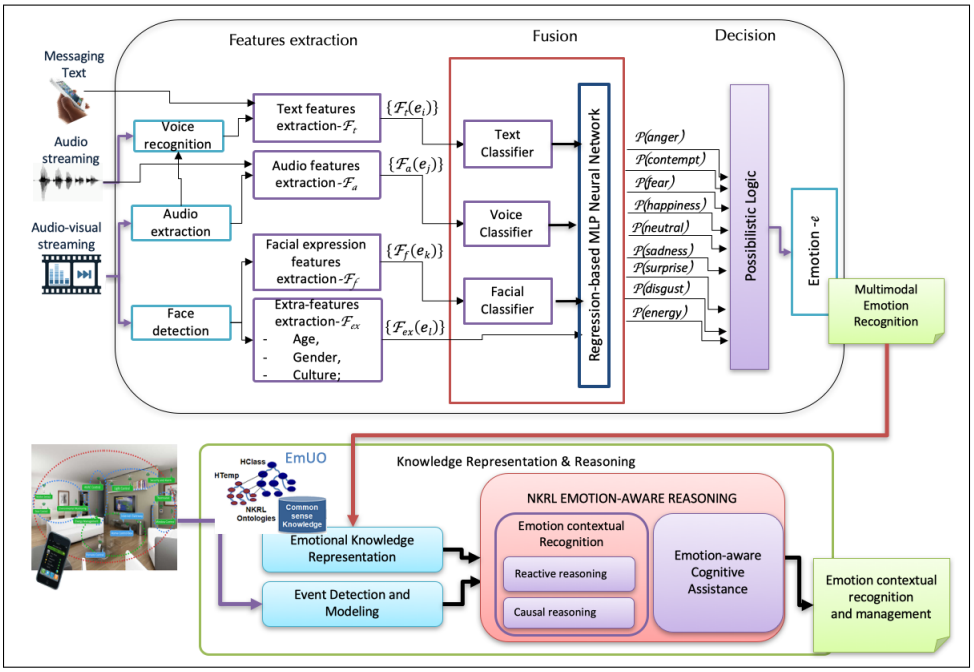



Figure 5.1 – Hybrid emotion recognition approach

prove the emotion recognition, extra-features are considered in this study for their influences in the emotional meaning as demonstrated in the literature. These features extracted from audio-visual data are: age [213], gender [214], and culture [215].

### 5.3.2 Unimodal Classifiers

In this work, a deep neural network-based model for a finer recognition of user's emotion is presented. The proposed model is a fusion of three deep neural networks: (i) a deep CNN model, trained with knowledge distillation, for facial emotion recognition (FER), (ii) a modified and fine-tuned VGGish model for voice emotion recognition (VER), and (iii) state-of-the-art deep network for text emotion recognition (TER) .

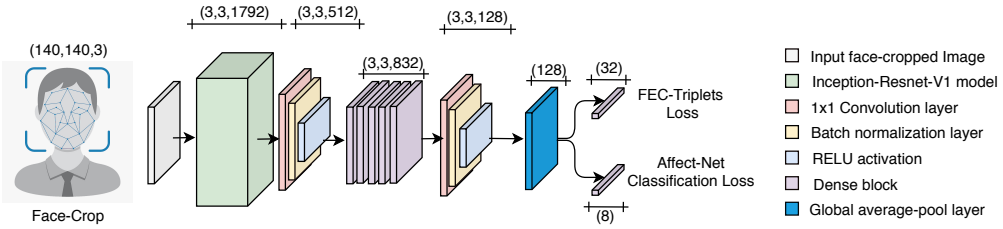



Figure 5.2 – Facial expression recognition neural network architecture, before distillation.

### Facial expression classifier

It analyses a facial expression in an image, and returns a set of expressions with their probabilities. These expressions are: *anger, contempt, disgust, fear, happiness, neutral, sadness,* and *surprise*. The first component of the proposed multimodal architecture is a deep convolutional neural backbone network (CNN) for facial expression recognition. The input to this network is a single RGB face image, detected and cropped using Multi-Task Cascaded Convolutional Networks (MTCNN) ([216]). The output of this network is a compact vector of dimension $D_{face}$.

In this work the 'facial expression embedding network' term is used to refer to the backbone CNN deep network, and its trained using *knowledge distillation* ([217]). Knowledge distillation is a two step process whereby a *teacher* network is trained on the task of interest, and then a (typically smaller) *student* network is trained on predictions made by the teacher. Specifically in this work, the benefits of *self-distillation* are leveraged, where the student network is the same size as (or at least, is not smaller than) the teacher network. It has been widely observed that distilling the knowledge of a teacher network to an equivalent student network leads to a regularization effect that improves performance on held out test data ([218]). Self-distillation is exploited to improve the performance of the facial expression embedding network. The training procedure for

this network thus consists of two phases:

1. Training a teacher model: the teacher model is a fine-tuned FaceNet ([219]), trained simultaneously on two different visual facial expression recognition datasets (Section 5.3.2).

2. Training a student model: a second CNN is additionally trained to mimic the outputs of this fine-tuned FaceNet (Section 5.3.2).

#### The teacher network

The starting point for the teacher model is a pre-trained FaceNet ([219]). This pre-trained model is fine-tuned for emotion recognition using two datasets:

- **AffectNet** ([220]), which consists of around 440,000 in-the-wild face crop images, each of which is human-annotated into one of eight facial expression categories (Neutral, Happy, Sad, Surprise, Fear, Disgust, Anger and Contempt).

- **Google Facial Expression Comparison (FEC)** ([221]), which consists of around 700,000 triplets of unique face crop images. Annotations denoting the most similar pair of face expressions in each triplet are provided. The goal is to train a model that places the similar pair closer together in a learned embedding space.

The teacher model's architecture (Fig. 5.2) is almost identical to the model proposed in [221], the only difference being that we add an additional output head for the AffectNet loss. A pre-trained FaceNet[1] is taken up until the Inception 4e block. This is followed by a 1x1 convolution and a series of five untrained DenseNet ([222]) blocks. Another 1x1 convolution followed by global

[1]In this, a FaceNet pretrained on the VGGFace2 dataset is used. The pre-trained FaceNet model architecture and weights were obtained from https://github.com/timesler/facenet-pytorch.

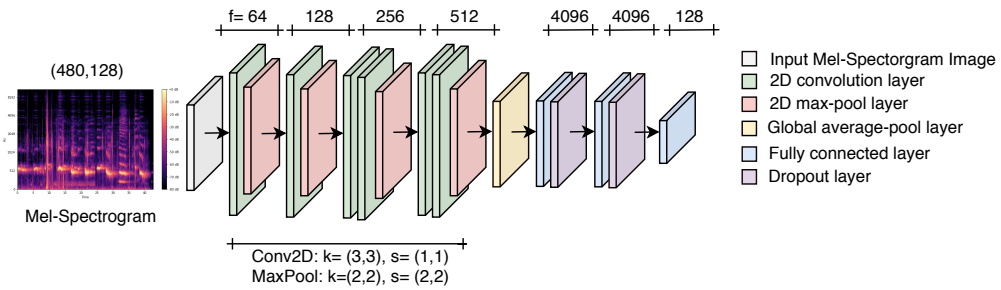



Figure 5.3 – Modified VGGish backbone feature extractor for Speech Emotion Recognition

average pooling is used to reduce this representation to a single $D_{face}$ dimensional vector. After pooling, two independent linear transformations serve as output heads. These heads take the $D_{face}$-dimensional facial expression representation vector as input and make separate predictions for the AffectNet and FEC tasks. A 32-dimensional embedding is used for the FEC triplets task, while an 8-dimensional output head produces class logits for AffectNet (which has 8 classes). The teacher network training procedure is detailed in Algorithm 1. To improve the regularisation effects of self-distillation through model ensembling, two teacher networks are trained, and their outputs are concatenated to serve as distillation targets (see Section 5.3.2 for details). The only difference between the two teacher networks are the random seeds used for weight initialization, and the penultimate layer dimensionalities: $D_{face} = 128$ for the first teacher network, and $D_{face} = 256$ for the second are used.

#### Student network

The student network is a DenseNet201 pretrained on ImageNet.[2] The student network training procedure is essentially the same as described in Algorithm 1, except that sample batches of unlabeled data from an internal dataset are

[2] the implementation and pretrained ImageNet weights provided in the `torchvision` Python package are used.

**Algorithm 1** Visual model: training the teacher network.
$f_\Theta$: Given feature extractor network,
$g_\phi$: Google FEC output head,
$h_\theta$: AffectNet output head,
N: number of training steps,
α: AffectNet loss weight.

**for** *iteration in range(*N*)* **do**
  $(\mathbf{X}_{\text{FEC}}, \mathbf{y}_{\text{FEC}}) \leftarrow$ batch of Google FEC triplets and labels
  $(\mathbf{X}_{\text{Aff}}, \mathbf{y}_{\text{Aff}}) \leftarrow$ batch of AffectNet images and class labels
  $\mathbf{e}_{\text{FEC}} \leftarrow f_\Theta(\mathbf{X}_{\text{FEC}})$ ▷ Face embeddings for FEC images
  $\mathbf{e}_{\text{Aff}} \leftarrow f_\Theta(\mathbf{X}_{\text{Aff}})$ ▷ Face embeddings for AffectNet images
  $\mathbf{v}_{\text{FEC}} \leftarrow g_\phi(\mathbf{e}_{\text{FEC}})$ ▷ Predict vectors for triplet loss
  $\mathbf{p}_{\text{Aff}} \leftarrow h_\theta(\mathbf{e}_{\text{Aff}})$ ▷ Predict class probabilities for AffectNet
  $\text{L}_{\text{FEC}} = \texttt{triplet\_loss}(\mathbf{v}_{\text{FEC}}, \mathbf{y}_{\text{FEC}})$
  $\text{L}_{\text{Aff}} = \texttt{cross\_entropy\_loss}(\mathbf{p}_{\text{Aff}}, \mathbf{y}_{\text{Aff}})$
  $\text{L} = \text{L}_{\text{FEC}} + \alpha * \text{L}_{\text{Aff}}$ ▷ Total loss for training step
  Obtain all gradients $\Delta_{\text{all}} = (\frac{\partial \text{L}}{\partial \Theta}, \frac{\partial \text{L}}{\partial \phi}, \frac{\partial \text{L}}{\partial \theta})$
  $(\Theta, \phi, \theta) \leftarrow \texttt{SGD}(\Delta_{\text{all}})$ ▷ Update feature extractor and output heads' parameters simultaneously
**end**

used, which is called *PowderFaces*. The PowderFaces dataset was created by downloading approximately 20,000 short, publicly-available videos from various online sources. MTCNN face detection was then applied to the extracted frames from those videos, producing approximately 1 million individual face crops. The sampled batches of face crops from the Google FEC, AffectNet, and PowderFaces datasets are passed through the proposed two teacher networks. Each of the two teacher networks produces predictions for the Google FEC task (32-dimensional) and AffectNet class logits (8-dimensional). These four vectors (i.e., two vectors from two teacher networks) are individually L2-normalised. The four normalised vectors are then concatenated, producing one long vector of dimension 80. A knowledge distillation loss ('Relational Knowledge Distillation' ([223]) is then calculated by comparing the output of a third output head in the student network to this 80-dimensional target vector. This knowledge

distillation loss is then added to the standard AffectNet and Google FEC losses, which are calculated as per the teacher network training procedure.

**Voice emotions classifier**

It classifies the input voice into the following 11 emotions: such as *supremacy, hostility, criticism, self-control, leadership, creativeness, friendliness, love, loneliness, sadness,* and *defensiveness emotions*.

The proposed deep model for audio-based emotion recognition is based on a modified version of the VGGish model ([224]). The starting point is the original VGGish model, pre-trained on the Audio Set dataset ([225]). The VGGish backbone consists of 6 convolutional layers that output 64, 128, 256, and 512 feature maps ($f$) respectively. For each convolution layer, a kernel ($k$) with size 3x3, and stride ($s$) of 1x1 is used. A max pooling layer with a kernel ($k$) of size 2x2, and stride ($s$) 2x2 is then applied. This VGGish backbone is exploited, but its last convolution and max pooling layers are replaced with a global average pooling layer. The resulting model produces an output vector of dimension 256. Finally, three randomly-initialized fully-connected layers are added, with output dimensionalities of size 4096, 4096, and 128. The aim of these layers is to extract a standard embedding vector with size of 128 that reflects the emotional characteristics of the input audio segment.

The expanded VGGish backbone architecture is exploited and fine-tuned on the RECOLA dataset. Two separate VGGish networks are fine-tuned: one to predict arousal and the other to predict valence. The inputs of size $[480, 128]$ is passed to the VGGish model, which are the mel-spectrogram representations of 30 seconds of audio from one of the videos in the RECOLA dataset. The target used for fine tuning is then the average ground truth arousal or valence for the target values corresponding to the input 30 seconds of audio. The target label

is predicted by passing the 128-dimensional audio representation through a fully-connected layer $f_\phi$ with a `tanh` activation. The training procedure for fine tuning the audio feature extraction model is detailed in Algorithm 2.

**Algorithm 2** VGGish fine-tuning algorithm for predicting arousal.
$f_\Theta$: Given the VGGish feature extractor network,
$f_\phi$: arousal prediction head,
N: number of training steps.

**for** *iteration in range(*N*)* **do**
  $(\mathbf{X}, \mathbf{y}) \leftarrow$ batch of RECOLA spectrograms and targets
  $\mathbf{e} \leftarrow f_\Theta(\mathbf{X})$ ▷ Calculate VGGish embeddings for batch
  $\mathbf{p} \leftarrow f_\theta(\mathbf{e})$ ▷ Predict arousal for all elements in batch
  $Loss = -$`concordance_correlation_coeff`$(\mathbf{p}, \mathbf{y})$
  Obtain all gradients $\Delta_{\text{all}} = (\frac{\partial Loss}{\partial \Theta}, \frac{\partial Loss}{\partial \theta})$
  $(\Theta, \theta) \leftarrow$ `Adam`$(\Delta_{\text{all}})$ ▷ Update VGGish model, output head
**end**

#### Text emotions classifier

it uses linguistic analysis of the input text to detect a set of sentiments as *joy, fear, sadness, anger, analytical, confident* and *tentative sentiments.* The model allows estimating the probability of the general sentiment of the input text and the probability of the sentiment for each sentence;

### 5.3.3 Fusion based on Multilayer Perceptron (MLP) neural network

In this study, a Multilayer Perceptron (MLP) neural network is exploited to estimate emotion probability from the combination of the multimodal prediction of the above classifiers associated with the features: age, gender and culture. As shown in Fig.5.1, the MLP Neural Network-based regression model for emotion fusion aggregates the output probabilities from the unimodal classifiers (video, audio, and text) into a single dimension input features vector and classifies it

into 9 standard emotions. Besides, in case of modality absence (audio, video, or text), the classification outputs of the MLP are replaced by zeros to maintain the regression fusion network input size.

The Multilayer Perceptron (MLP) neural network consists, in this study, of five layers of neurons which are fully connected. The input layer consists of a set of neurons representing the prediction of each classifier, such as emotion in text prediction, emotion in audio prediction, facial expression prediction, age, culture, and gender. The output layer consists of the neurons representing the emotions classes considered as outputs. In this chapter, nine classes of observed emotions are considered: *anger, contempt, fear, happiness, neutral, sadness surprise, disgust, and energy.* The *MLP neural network* model consists, also, of three hidden layers where each hidden layer has 100 nodes. To prevent the model from over-fitting on the training data, the Dropout technique is exploited on 10% of the hidden nodes. The weights initialization process is based on *small normal distribution* with zero mean and 0.05 standard deviation. The neurons output/input computing process exploits the forward propagation algorithm with a nonlinear function, the Rectified Linear Units (ReLU) and Softmax functions. Errors-computing and weights-updating process is based on the loss function. This function exploits mean squared error functions. In order to minimize the loss function, the Adaptive Moment Estimation (Adam) optimization algorithm [226] is exploited to optimize the values of neural network weights.

### 5.3.4 Decision based on Possibilistic Logic (PL)

In this study, a categorical model is used for emotion representation. Therefore, the proposed hybrid fusion model based on a MLP neural network infers a set of emotions with their probabilities. In this context, managing uncertainty

consists of managing the ambiguity of emotion over complex emotions, and filtering out emotions which are unlikely to correspond to real-world overlapped emotions.

The inference step takes as input the probability of emotions predicted from the multimodal recognition classifiers at each time $t$, and infers the overlapped emotions. The possibilistic logic [227] is exploited over the most likely overlapped emotion to decide whether its probability corresponds to a real-world overlapped emotion. Possibilistic logic provides an efficient way to find the probability level from which a multimodal emotion recognition has sufficient evidence to recognize a real-world overlapped emotion.

A possibilistic interpretation is a mapping $\pi : \mathscr{I}_\Phi \longrightarrow [0,1]$ where $\pi(\mathrm{I})$ is the degree to which the world I is possible. In particular, every world I such that $\pi(\mathrm{I}) = 0$ is impossible, while it is totally possible when $\pi(\mathrm{I}) = 1$. The possibility of an emotion $\phi$ in a possibilistic interpretation $\pi$, denoted by $\mathrm{P}oss(\phi)$, is defined by $\mathrm{P}oss(\phi) = \{\mathrm{M}ax(\pi(\mathrm{I}))|\mathrm{I} \in \mathscr{I}_\Phi, \mathrm{I} \models \phi\}$. Intuitively, the possibility of $\phi$ is evaluated in the most possible world when $\phi$ is true. The dual notion to the possibility of an emotion $\phi$ is the necessity of $\phi$, denoted by $\mathrm{N}ec(\phi)$, which is defined by $\mathrm{N}ec(\phi) = 1 - \mathrm{P}oss(\neg\phi)$. In possibilistic theory, for all possibilistic interpretations $\pi$ and emotions $\phi$ and $\psi$, the following relationships hold:

- $\mathrm{P}oss(\phi \wedge \psi) \leq min(\mathrm{P}oss(\phi), \mathrm{P}oss(\psi))$;
- $\mathrm{P}oss(\phi \vee \psi) = max(\mathrm{P}oss(\phi), \mathrm{P}oss(\psi))$;
- $\mathrm{P}oss(\neg\phi) = 1 - \mathrm{N}ec(\phi)$;
- $\mathrm{P}oss(\bot) = 0$;
- $\mathrm{P}oss(\top) = 1$;
- $\mathrm{N}ec(\phi \wedge \psi) = min(\mathrm{N}ec(\phi), \mathrm{N}ec(\psi))$;

- $\mathrm{N}ec(\phi \vee \psi) \geq max(\mathrm{N}ec(\phi), \mathrm{N}ec(\psi))$;
- $\mathrm{N}ec(\neg\phi) = 1 - \mathrm{P}oss(\phi)$;
- $\mathrm{N}ec(\bot) = 0$;
- $\mathrm{N}ec(\top) = 1$;

## 5.4 Contextual emotion recognition

The contextual emotion recognition and management proposed in this study consists of exploiting jointly the emotions predicted from the multimodal emotion recognition model, and the events characterizing the user's context, in a knowledge representation and reasoning model to infer the non-directly observable emotions and trigger the assistance service adapted to the emotional context of the user.

### 5.4.1 Emotional knowledge representation

The Narrative Knowledge Representation Language (NKRL) is exploited in the description of emotions and human states information in ambient intelligent environments. In this study, the commonsense knowledge based on the extended HClass ontology (E*m*UO) allow describing static knowledge such as feeling, sentiment, passion, pleasure, etc.

As emotion is defined as immediate reaction following an event, it is important to take into account what is happening in the ambient environment around the user. To deal with this important aspect, dynamic knowledge representation based on the HTemp ontology of NKRL is exploited to describe the observed or inferred emotion in a specific context, such as, *"having a positive/negative experience"*.

**Emotion upper-ontology (EmUO)**

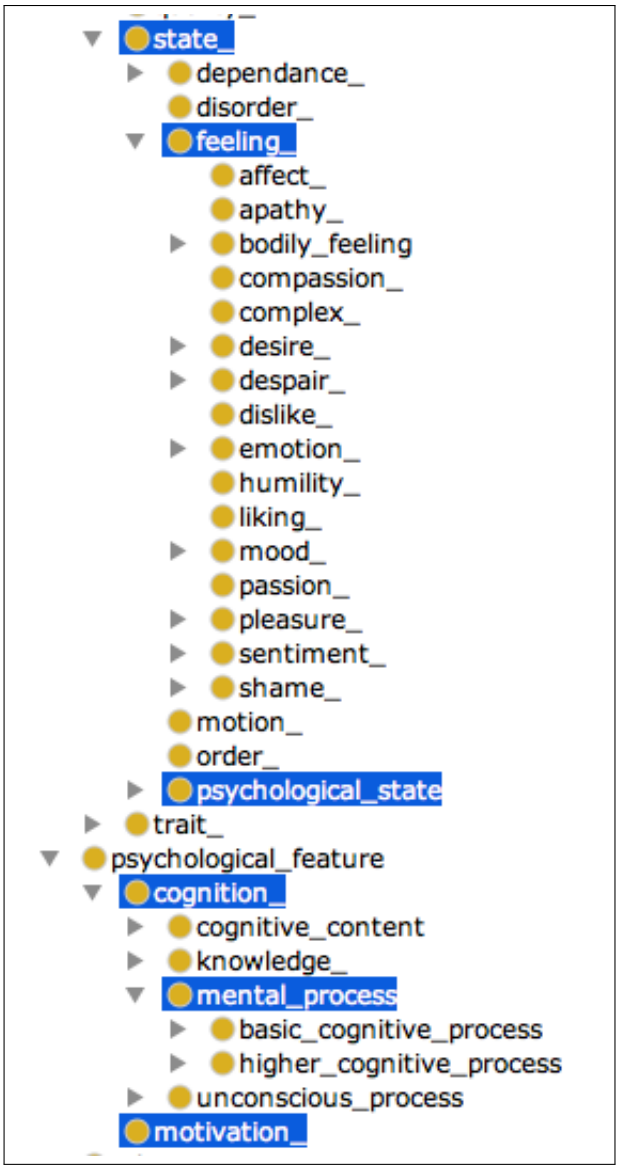



Figure 5.4 – A segment of the *emotion upper ontology (EmUO) describing human states taxonomy*

The *Emotion upper-ontology (EmUO)*, extension of the *HClass ontology*, is proposed in this study to represent the emotional context attributes of the user in an ambient intelligent environment. In particular, it covers all commonsense

concepts of human states such as affects and emotions, cf. figure 5.4. *EmUO ontology* combines different "sentiment and emotion lexica" such as *SenticNet3* and *WordNet-Affect*.

**Emotional knowledge and the** HT*emp* **ontology of NKRL**

To represent contextualized emotions, the HT*emp* ontology is exploited. This ontology is designed as formal representation of generic classes of events such as "threatening someone with violence", "having a positive/negative experience", and "evaluating an artefact", etc. The NKRL templates allow representing the full emotional context of human where contextual attributes like culture, religion, and other social rules should be taken into account.

In terms of emotional context representation, the 'BEHAVE' templates are well suited to represent emotional context features. Particularly, the "*Behave: Focus*" template allows representing emotional context features such as desire, intention, etc. Besides, the "*Behave: Attitude*" template can be exploited to represent situation or behavior of a person, a social body, or a situation/activity.

Some specific templates such as *Experience*, *Produce* and *Receive* are important from *emotion analysis* point of view. An experience of a human-being or social body, such as success, richness, illness, racism, violence, etc. can be considered as *positive*, *negative*, or *neutral* and can be represented respectively using the following templates: *Experience: PositiveSituation*, *Experience: NegativeSituation*, and *Experience: GenericSituation*.

The *Produce* templates, in the context of emotion analysis, are *Produce: Acceptance/Refusal* and *Produce: CreateCondition/Result*. The templates *Produce: PositiveCondition/Result* and *Produce: NegativeCondition/Result* are used to represent an emotion with respect to a given action/situation. Furthermore, the *Receive: DesiredAdvice* template, one of the templates *Receive*, is important

in emotional context representation.

Table 5.1 – Representation of the emotional context: Matthew is happy

| |
|---|
| E.occ2: **EXPERIENCE**<br>**SUBJ** MATTHEW :<br>SPECIF(ADP_building SHOWROOM_1)<br>**OBJ** SPECIF (feeling_ joy_)<br>**date-1:** 14/08/2017 11:23:15<br>**date-2:**<br>Experience:Human/Social |

Let us consider the example where a companion robot, called *Pepper* which perceives that a visitor, called *Matthew*, is happy. The semantic description of this emotional context is given in table 5.1 in the symbolic label of the predicative occurrence (L = "E.$occ2$"). In this example, the conceptual predicate (P = "EXPERIENCE") denotes the experience of the entity which is here a human. This human entity is represented with the role (R = "SUBJ") and its argument ($a$ = "MATTHEW"), an instance of the HC$lass$ concept "$human\_being$". The emotion is represented by the role "OBJ" and its argument *SPECIF( feeling_-joy_)*.

The question *"what is the meaning of someone smiles?"* could be answered that it's about *'a happiness'* or *'embarrassed smiling'*. Depending on the user's context, an embarrassed smiling can then be differentiated from a happiness smiling.

The novelty of this study lies in the description of emotional knowledge using n-ary ontologies that makes possible the contextual recognition of emotions, such as negative surprise and embarrassed smile, and the recognition of non-directly observable emotions such as *curiosity* and *attention*.

### 5.4.2 NKRL-based contextual emotion inference

To better understanding human emotions in their context, the NKRL inference engine is exploited to infer implicit relations between predicative occurrences of emotions and the events characterizing the user's context. Two kinds of high-level rules, transformations and hypotheses, are systematically used in the NKRL inference engine.

The transformation rules are defined as implications expressed as in formula 5.1. The constraint $var_i \subseteq var_j$ corresponds to the variables defined in the antecedent that must also appear in the relative consequent according to the emotional commonsense knowledge.

$$\mathrm{A}(var_i) \Rightarrow \mathrm{C}s_i(var_j); var_i \subseteq var_j. \tag{5.1}$$

**Emotion contextual recognition**

In ambient intelligent environments, certain emotions such as upset, confusion, curiosity, are not directly observable since they are context-dependant and exploiting only data-driven approaches may not be sufficient to predict them. To recognize a non-directly observable emotion, the rich expressiveness of the n-ary ontologies and powerful inference mechanisms of the NKRL are exploited in this study. The recognition of this kind of emotions allows a better decision-making in terms of assistance services. Three methods are considered in this study to recognize non-directly observable emotions:

- The first method is based on transformation rules and reactive reasoning that allow inferring humans feelings following the occurrence of an event characterizing the user's context. Formally, by exploiting a transformation rule, denoted by $\mathscr{TR}_j$, an event $e$ makes true a situation (i.e feeling)

$f$ starting from an instant $t$ such as:

$$(f,t) \Leftarrow (e,t) \wedge \mathscr{T}\mathscr{R}_j$$

. The feeling of *"confusion"* is a non-directly observable emotion that can be recognized by analyzing the meaning of a query of the user during a discussion using the NKRL language. The native feature of this language aims to represent and reason on the meaning of natural language sentences allowing therefore a finer recognition of non-directly observable emotions by contextualizing them. An example of transformation rule allowing to infer that a human feels confusion is shown in table 5.2. In this rule, the recognition of the feeling *"confusion"* can be triggered by a misunderstanding of a topic during a discussion. It consists of the aggregation of different situations: process information and general topic of discussion that are represented respectively by the NKRL templates *Behave:Questioner*, and *Behave:Participant*. In this study, the representation of the topic of each utterance allows to check if two topics have the same main topic using the specification/generalization relationship in the EmUO ontology.

- The second method is based on binding occurrences defined using axioms. It consists of causal law, formally, a predicative occurrence X causes a predicative occurrence Y:

$$\mathrm{CAUSE(X,Y)}$$

For example, an abnormal activity such as eating many sweets at the same time can be caused by a happiness or sadness feeling after the occurrence of such events. In this case, the emotion is triggered from the

| |
|---|
| $Antecedent$ : |
| COORD($e_1$ $e_2$) |
| $e_1$: BEHAVE SUBJ var1 : |
| OBJ conversation_ |
| CONTEXT var2 |
| date-1 |
| date-2 |
| var1:human_being |
| var2: topics_ |
| Behave: Participant |
| $e_2$: BEHAVE SUBJ var1 : |
| MODAL SPECIF(quering_ information_) |
| CONTEXT var3 |
| date-1 |
| date-2 |
| var1:human_being |
| var3: topics_ |
| G(var3): ! var2 |
| Behave: Questioner |
| $Consequent$ : |
| $f_1$: BEHAVE SUBJ var1: |
| OBJ SPECIF (feeling_ confusion_ ) |
| date-1 $t$ |
| date-2 |
| var1:human_being |
| Behave:NegativeConcrete |

Table 5.2 – Example of transformation rules $\mathscr{TR}$

observed event "eating many sweets".

- Causality-based reasoning consists of identifying the relationship between a cause and its effect. This model aims to identify a possible cause of an observed emotion in order to improve the recognition in terms of accuracy. To deal with that, the NKRL inference engine is exploited by using hypothesis rules. The latter are defined using axioms that consist of state constraints and can be used to express the following assertion: a

predicative occurrence X can be caused by the predicative occurrences $Y_1, Y_2, ..., Y_n$. Formally, a hypothesis rule can be expressed as follows:

$$X \; if \; Y_1, Y_2, ..., Y_n$$

For example, the hypothesis rule, given in table 5.3, is used to get a possible cause of the observed emotion "surprise". Two possible causes are described in this rule: receiving money and receiving medical process. If the first cause is proved, the transformation rule, given in table 5.4, allows inferring a new emotion such as *positive-surprise*.

| |
|---|
| P*remise*: |
| $f$: EXPERIENCE SUBJ var1: |
| OBJ SPECIF (feeling_ surprise_ ) |
| date-1 $t$ |
| date-2 |
| var1:human_being |
| Experience:Human/Social |
| A*ntecedent*: |
| $e_1$: RECEIVE SUBJ var1: |
| date-1 $t$ |
| date-2 |
| var1:human_being |
| Receive:GetMoneyl |
| $e_2$: RECEIVE SUBJ var1: |
| date-1 $t$ |
| date-2 |
| var1:human_being |
| Receive:MedicalProcess |

Table 5.3 – Example of hypothesis rule

**Emotion-aware cognitive assistance**

Endowing ambient intelligent environments with the ability to provide emotion-aware services is the main objective of the used reasoning techniques. In this

| |
|---|
| A*ntecedent*: |
| *e*: RECEIVE SUBJ var1: |
| date-1 *t* |
| date-2 |
| var1:human_being |
| Receive:GetMoneyl |
| C*onsequent*: |
| *f*: EXPERIENCE SUBJ var1: |
| OBJ SPECIF (feeling_ positive-surprise_ ) |
| date-1 *t* |
| date-2 |
| var1:human_being |
| Experience:PositiveHuman/Social |

Table 5.4 – Transformation rules N 2

chapter, the states of humans populating an ambient intelligent environment, such as *user feeling*', are considered in the context modeling.

To reason about the emotions needed for performing an action, each task should be described based on the emotional state of the user. In this study, a task selection model is used to choose the appropriate task for the recognized user emotion. Formally, a task selection consists of a tuple $\Pi = \langle \mathrm{K}, \mathrm{E}, \mathrm{T} \rangle$ where K is a set of tasks, and E is a set of emotions evolving over the time T. Selecting a task $k$, having a score $s$, according to a particular emotion $e_t$ at specific time $t$ consists of determining the normalized score $\mathrm{N}_k$:

$$\mathrm{N}_k = \frac{k^s}{\sum {k^s}_i}; \tag{5.2}$$

The task score $\mathrm{K}^s$ is a factor which is evolving over time based on various parameters: the prerequisites tasks score $p_k$ of the next planned task, the current and previous emotion score respectively $e_t$ and $e_{t-1}$, and the emotion transi-

tion factor $\phi_e$ where $\phi_e \in [-1, 1]$.

$$K^s = \frac{p_k}{e_t} + \phi_e \tag{5.3}$$

$$\phi_e = e_t - e_{t-1} \tag{5.4}$$

## 5.5 Experiments

The proposed approach for contextual emotion recognition is evaluated as follows: (i) the proposed model for multimodal emotion recognition is evaluated through empirical experiments on real-world dataset. Results are reported for validation and test sets of a 10-fold cross-validation scheme. These results are compared to those obtained with two baseline models: Multiple Kernel Learning [228] and MLP Neural Network [229]; (ii) the emotion contextual recognition and management is evaluated through experiments in real world scenario dedicated to the cognitive assistance of visitors in a smart devices showroom. A qualitative feedback about satisfaction of each user was collected after these real world experiments.

### 5.5.1 Multimodal emotion dataset description

In this study, the YouTube Opinion Dataset, proposed in [230], is exploited for its richness to train the multimodal emotion recognition proposed model based on visual, textual, audio and culture features. This dataset is a collection of opinion videos collected from the social media website YouTube. It consists of 47 videos sampled into 3149 frames of people expressing their opinions about a variety of topics. All the videos show the frontal view of the subject's face that is recognizable with face-tracking software. The videos were found using the fol-

lowing keywords: opinion, review, product review, best perfume, toothpaste, war, job, business, cosmetics review, camera review, baby product review, I hate, I like. The final video set includes 20 female and 27 male speakers randomly selected from youtube.com, with their age approximately in the range of 14 – 60 years old. Although of different ethnic backgrounds (e.g., Caucasian, African-American, Hispanic, Asian), all speakers expressed themselves in English. The videos are converted to .mp4 format with a standard size of 360x480. The length of the videos varies from 2 – 5 minutes. All videos are pre-processed to address the following issues: introductory titles and multiple topics. Many videos on YouTube contain an introductory sequence where a title is shown, sometimes accompanied with a visual animation. As a simple way to address this issue, the videos are manually segmented until the beginning of the first opinion utterance. All video clips are manually transcribed to extract spoken words as well as the start time of each spoken utterance. The *Transcriber* software was used to perform this task. The transcription is carried out using only the audio track of each video clip. All 47 video clips are annotated by three annotators who were shown videos in three different random sequencing orders, so as to reduce the compound effect. It is important to note that the dataset is not annotated by the sentiment felt by the person watching the video. The annotation task is to associate the input data with a sentiment label that best summarizes the opinion expressed in the YouTube video. To ensure equitable evaluation, the dataset is divided into 10 randomly stratified folds: 9 folds for training and 1 fold for testing each time. 10% of the training set are considered for validation to tune the model hyperparameters.

### 5.5.2 Implementation

#### General description

The contextual emotion recognition approach is implemented using Python and Java as a cloud service. The communication service is based on XMPP and REST protocols. To extract features from audio-visual data, a set of services are exploited. Both of the proposed audio-visual deep learning models combined with the textual expressions extraction, [3] are exploited. The MLP neural network is implemented based on Keras framework with tensor-flow as a backend. The possibilistic logic is implemented based on *SAT4J* library.

#### Cognitive assistance of visitors in a smart devices showroom scenario

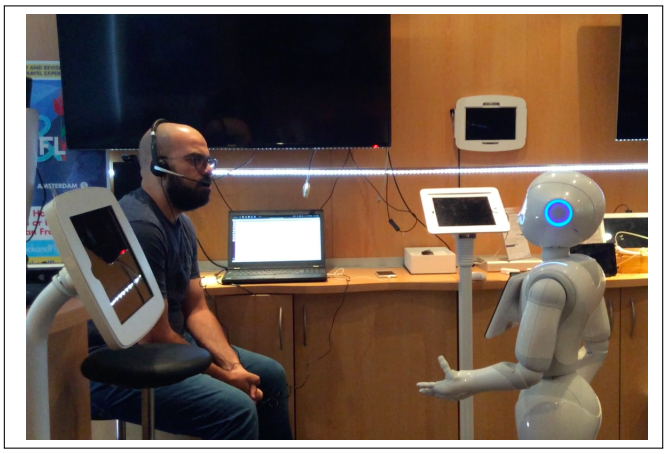


Figure 5.5 – Scene extracted from the smart devices showroom

A real world scenario dedicated to the cognitive assistance of visitors in a smart devices showroom is proposed to validate the proposed approach. In this scenario, Matthew is a visitor who needs help from a robot companion (Pepper robot) that will act as a tour guide to explain and show Matthew the smart

[3] www.alchemyapi.com/

devices showroom, cf. figure 6.10. Pepper can detect and monitor Matthew's emotions continuously based on the proposed approach to adapt the discussion based on the detected emotions. One of the most important issues is to make Pepper reacting more naturally. The proposed scenario shows how the proposed approach makes the Human-Robot interaction in an intelligent ambient environment more natural and more intelligent.

## 5.6 Evaluation

### 5.6.1 Baselines

The first baseline model consists of the classifier Multiple Kernel Learning (MKL) that learns from multimodal heterogeneous fused feature vectors to recognize sentiments [228]. These vectors are extracted from the combination of uttered words, facial expressions, and speech sounds. This model is evaluated through a traditional training/testing evaluation by dividing the dataset into 70% for training and 30% for testing. The second baseline model consists of the (MLP) Neural Network proposed in [229]. This model is learned from different classifiers such as facial, vocal and textual expression classifiers. It is intended to recognize nine emotions: *anger, contempt, fear, happiness, neutral, sadness, surprise, disgust, and energy.* The evaluation is carried out over 5 folds cross validation.

### 5.6.2 Unimodal emotion recognition

The performance of the proposed uni-modals has been evaluated using the REmote COLlaborative and affective interactions (RECOLA) corpus ([231]). In RECOLA, participants' spontaneous interactions were collected while being en-

gaged in a remote discussion that aimed to manipulate their moods. Then, six annotators measured the emotional state present in all sequences continuously on the valence and arousal dimensions. 27 audio-visual recordings of 5 minutes of interaction – 9 for training, 9 for validation, and 9 for testing – are made publicly available. The visual feature extraction network is evaluated on held-out evaluation data from the AffectNet and Google FEC datasets.

**Visual facial expression embedding network performance**

the visual facial expression embedding network was evaluated on held-out data from the two static image datasets it was trained on:

1. **AffectNet:** for AffectNet, which requires classifying face into eight discrete facial expression classes, a logistic regression model is trained on the features extracted by the student network for the entire AffectNet training set.[4] This model achieves state-of-the-art results on the AffectNet validation set, with an accuracy of 61.6% (Table 5.5).

2. **Google FEC:** following [221], the model is evaluated by exploiting triplet accuracy on the Google FEC test set. Using this metric, the proposed model is substantially outperforming the state-of-the-art on the FEC test set with an accuracy of 86.5% (Table 5.6).

### 5.6.3 Audio-Visual evaluation on RECOLA

To evaluate the performance on RECOLA of visual and audio feature extractors separately, the same fusion architecture is retrained, but disable either the audio or visual feature inputs and the associated pre-transform network.

[4] the classes in the AffectNet training set are re-weighted so that they have equal representation, as per the validation set.

Table 5.5 – Performance of the proposed visual facial expression embedding network on the AffectNet validation set compared to existing state-of-the-art models

| **Models** | **Accuracy (8 facial expressions)** |
|---|---|
| [232] | 59.6% |
| [233] | 59.3% |
| Proposed (Teacher model) | 61.3% |
| Proposed (Distilled student) | **61.6%** |

Table 5.6 – Triplet prediction performance of the proposed visual facial expression embedding network on the Google FEC test set compared to existing state-of-the-art models

| **Models** | **Accuracy** |
|---|---|
| [221] | 81.8% |
| Proposed (Teacher model) | 84.5% |
| Proposed (Distilled student) | **86.5%** |

**Visual-only:** feeding the embeddings from the visual feature extractor into the visual-only version of the fusion model performs well on the RECOLA dataset (Table 5.7). Such reaches a CCC of 0.55 for predicting valence and 0.57 for predicting arousal on the validation set, while on the test set the CCC reaches 0.66 for valence and 0.57 for arousal. This result illustrates the robustness of the visual feature extractor. In terms of valence prediction, the model outperformed the state-of-the-art models , even though *only visual* features are used as input.
**Audio-only:** the results (Table 5.7) show that the modified VGGish backbone feature extractor for audio segments performs well, CCCs of 0.52 and 0.70 for valence and arousal, respectively on the RECOLA test set. The achieved results for arousal prediction match the existing state-of-the-art models when only audio features are used (Table 5.8).

Table 5.7 – RECOLA dataset results (in terms of CCC) for predicting arousal and valence on train, development and test sets.

| | **Valence** | | | **Arousal** | | |
|---|---|---|---|---|---|---|
| **CCC** | Train | Dev | Test | Train | Dev | Test |
| Visual only | .6 | .55 | .66 | .49 | .57 | .57 |
| Audio only | .55 | .46 | .52 | .78 | .80 | .70 |
| Audio-visual | .69 | .63 | .74 | .78 | .81 | .72 |

Table 5.8 – Performances of the proposed audio embedding network on the RECOLA dataset comparing to existing state-of-the-art models. In parenthesis are the performances obtained in the development set. —— : no results reported in the original papers.

| **Models** | **Arousal** | **Valence** |
|---|---|---|
| [192] | .70 (.75) | .31 (.41) |
| [234] | .67 (.76) | .36 (.48) |
| [190] | ——(.80) | ——(.40) |
| Proposed | .70 (.80) | .52 (.46) |

### 5.6.4 Multimodal emotion recognition

In this section, the dataset described in section 5.5 is exploited, to evaluate the performance of the proposed multimodal emotion recognition model in terms of F-1 score, precision and recall metrics. This performance is compared to those obtained with baseline models: Multiple Kernel Learning (MKL) [228] and MPL Neural Network [229].

The first evaluation regards the MLP Neural Network-based regression model for the multi-modal emotion fusion/recognition. This model predicts the probabilities of 9 emotions considered in this study: happiness, sadness, anger, neutral, energy, contempt, disgust, fear and surprise. The results of the MLP model are obtained from the evaluation of 10 models. In these tests, a mean square error of 0.024 and a mean absolute error of 0.095 are obtained.

The second evaluation of the emotion recognition based on decision-level fusion exploiting the MLP neural network and the possibilistic logic. Table

Table 5.9 – Multimodal emotion recognition performance versus selected features

| Mean $F_1$ score | Without Features | Age | Gender | Culture | All features |
|---|---|---|---|---|---|
| Anger | 0.38 | 0.55 | 0.52 | 0.64 | 0.89 |
| Contempt | - | 0.00 | 0.00 | - | 0.00 |
| Disgust | 1.00 | 0.86 | 0.67 | 1.00 | 1.00 |
| Energy | 0.50 | 0.55 | 0.64 | 0.67 | 0.91 |
| Fear | 0.67 | 1.00 | 1.00 | 1.00 | 1.00 |
| Happiness | 0.69 | 0.72 | 0.76 | 0.74 | 0.97 |
| Neutral | 0.80 | 0.75 | 0.83 | 0.85 | 0.98 |
| Surprise | 0.55 | 0.67 | 0.57 | 0.46 | 0.90 |
| Sadness | - | - | - | - | - |

5.9 shows the performance of the proposed multimodal emotion recognition model based on the selected features such as age, gender and culture. The reported results show that including the feature culture improves significantly the recognition of the emotions fear and anger whereas the feature gender has an impact on the recognition performance of the emotion happiness. One can also observe that the feature age slightly improves the recognition performance of the emotion surprise. Including all features significantly improves the performance of the proposed multimodal emotion recognition in terms of $F-1$ score from 0.67 to 0.95.

The performance of the emotion recognition based on decision-level fusion is reported in Table 5.10 by varying the number of modalities from one to three. One can observe that the multimodal emotion recognition model outperforms all other modalities in all detected emotions. The modality face improves significantly the recognition of the emotions happiness, anger, neutral and surprise. The emotions disgust and fear are recognized in the multimodal emotion recognition with $F_1$ score equal to 1.0 where the emotions happiness and neu-

Table 5.10 – Emotion recognition performance using the test set

| mean F-1 score | MLP Neural Network + Possibilistic Logic | | | | | | |
|---|---|---|---|---|---|---|---|
| | Unimodal | | | Bimodal | | | Multimodal |
| Emotions | T | A | F | T+A | T+F | F+A | A+T+F |
| Anger | 0.42 | 0.24 | 0.48 | 0.79 | 0.69 | 0.55 | 0.89 |
| Contempt | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Disgust | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | 0.00 | 1.00 |
| Energy | 0.67 | 0.37 | 0.67 | 0.76 | 0.67 | 0.77 | 0.91 |
| Fear | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| Happiness | 0.29 | 0.40 | 0.70 | 0.52 | 0.61 | 0.58 | 0.97 |
| Neutral | 0.56 | 0.51 | 0.75 | 0.71 | 0.79 | 0.78 | 0.98 |
| Surprise | 0.53 | 0.42 | 0.57 | 0.48 | 0.62 | 0.63 | 0.90 |
| Sadness | - | - | - | - | - | - | - |

T:text, F:face, A:Audio

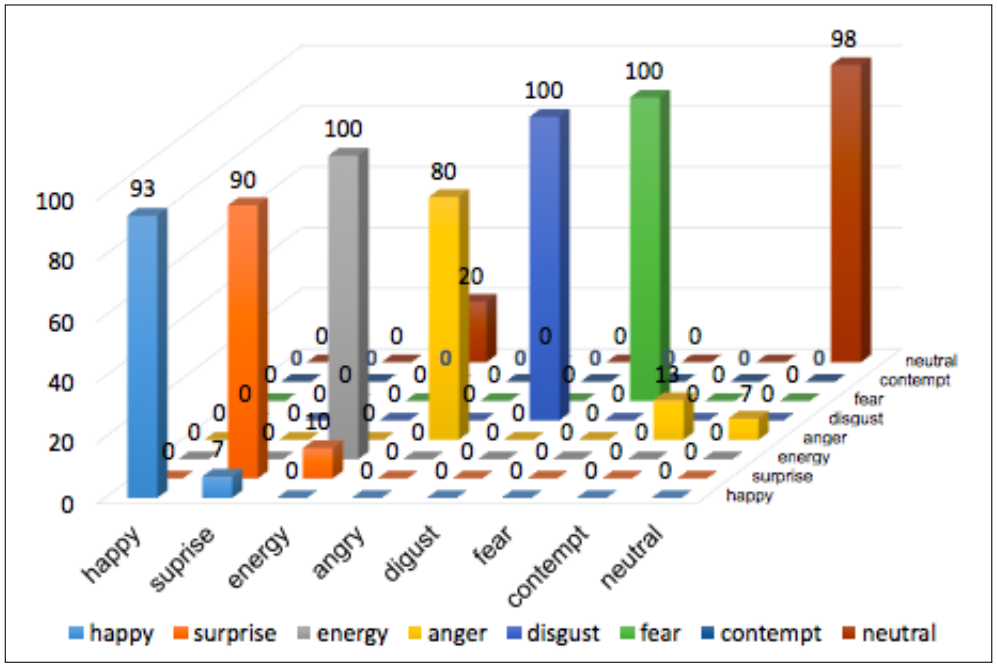



Figure 5.6 – Normalized confusion matrix of the multimodal emotion recognition at the low level

tral are recognized with $F_1$ score equal to, respectively, 0.97 and 0.98. The con-

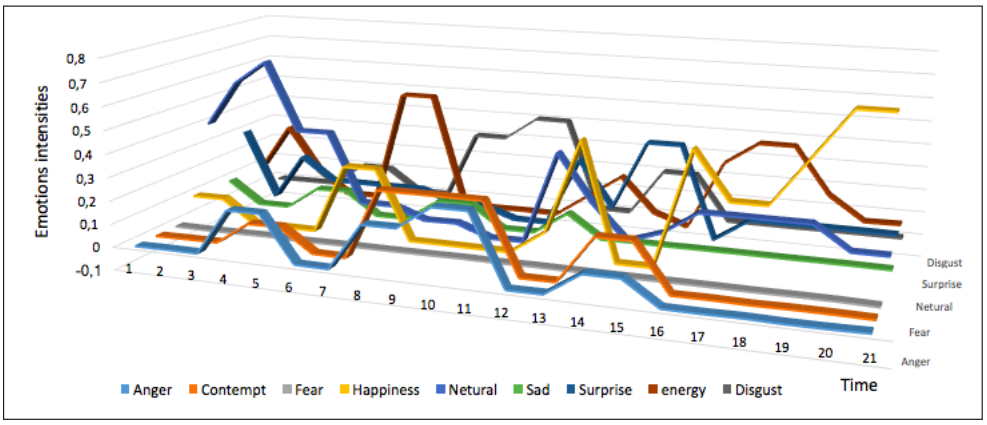



Figure 5.7 – Emotions intensities transition during interactions

fusion matrix of the multimodal emotion recognition shows that happiness is recognized as surprise, and surprise is recognized as energy, see. figure 5.6. The positive emotions (happy, energy, surprise) are slightly misclassified by the proposed multimodal emotion recognition model, due to the fact that the ground-truth labels are strongly coupled such as a good surprise might cause happiness and energy emotions. Consequently, the proposed approach exploits the n-ary ontologies to discern the coupled emotions. The performances obtained with

Table 5.11 – Comparison to baseline models

| Accuracy | Multimodal emotion models | | |
|---|---|---|---|
| | Baselines | | The proposed model |
| | MKL[228] | MLP-NN [229] | |
| Multimodal Techniques | 88.60% | 85.45% | 93.75% |

the proposed multimodal emotion recognition model and the baseline models (MLP neural network-MLP-NN and Multiple Kernel Learning-MKL) are shown in Table 5.11. These results show that the proposed multimodal emotion recognition model outperforms all the baseline models. Exploiting jointly possibilistic logic and the MLP neural network allows improving the performance of the

multimodal emotion recognition with an improvement of 8.3% in terms of the mean accuracy in comparison with the model proposed in [229].

The average processing time to recognize emotions for each fold, during the 10 folds validation set, is 0.35$s$ which is compatible with the dynamic nature of an ambient intelligent environment for real time recognition.

### 5.6.5 Contextual emotion recognition

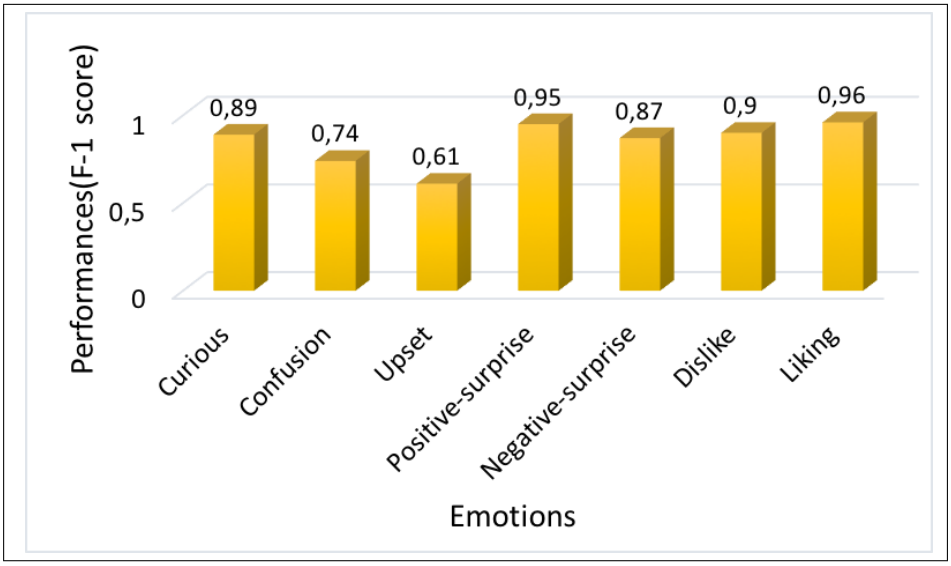



Figure 5.8 – Non-directly observable emotion recognition

First evaluation concerns the non-directly observable emotion recognition, cf. figure 5.8. In the proposed scenario, two emotions are recognized using the proposed approach: curious and confusion. For instance, the emotion "confusion" is recognized by analyzing the meaning of a user's query during a discussion using NKRL language. Such types of emotions cannot be recognized using only data-driven approaches or logic-based approaches.

Second evaluation focuses on the humans emotions management. Figure 5 shows the transition of the visitor's emotions during the tour-guide experi-

ments. One can observe that the interaction starts with the visitor person having neutral feeling. The proposed human emotions management approach reduces significantly the neutral emotion to transit it to happy and surprise.

### 5.6.6 Evaluation of the hybrid approach

Real world experiments of the tour-guide scenario have been carried out at the showroom of the *ADP (Aeroport De Paris)* company with the participation of ten (10) persons of 15 *min* each. The participants have different cultures, different accents and their ages vary from 22 years to 55 years. To evaluate the performance of the proposed approach, three metrics were considered: the size of the ontology, the runtime performance in terms of of the services response time, and the qualitative feedback about satisfaction of each participant involved in the experiments. In terms of runtime performance, three services were evaluated: emotion recognition and representation service, attention recognition and representation service, and reasoning service.

#### Quantitative evaluation

Table 5.12 – Response Time of each service

| Services Runtime (ms) | Emotion recognition & representation | Attention recognition | Reasoning service |
|---|---|---|---|
| Minimum time | 4 | 20 | 0 |
| Maximum time | 247 | 33 | 17 |
| Average time | 75.25 | 27.33 | 6.55 |

- **Knowledge base size:** The HClass ontology includes, nowadays, more than 9.000 concepts describing commensense knowledge covering a wide range of concepts about human-centric applications, ambient environments, and everyday activities. This ontology can be extended automati-

cally by other upper ontologies such as wordnet. The knowledge base of the Pepper robot consists of instances of HTemp ontology whose number increases during tests;

- **Runtime performance:** The different proposed services were evaluated during experimentations on macintosh computer (3i core, 8 Go). Table 7.6 shows that the average response time of the reasoning service is less than 7 $ms$. Moreover, the average time needed for Emotion recognition and representation service is less than 76 $ms$. During the experimentation, the attention recognition and representation service runtime varies between 20 $ms$ and 33 $ms$.

#### Qualitative evaluation

A feedback is collected about the participants satisfaction after real world experiments. The results showed that 98% of them were satisfied about the overall assistance services in terms of intelligent guide tour. Besides, 80% of them were satisfied about the response time of the robot during interactions. Participants describe their experiences as enjoyable.

## 5.7 Conclusion

In this chapter, a hybrid contextual emotion recognition approach is proposed for cognitive assistance services in ubiquitous environments. Its principle consists of, on the one hand, the multimodal emotion recognition based on hybrid-level fusion exploiting a multilayer perceptron neural network model and the possibilistic logic, and on the other hand, the expressive NKRL knowledge representation and reasoning exploiting both HClass and HTemp ontologies for representing both commonsense knowledge and the ambient environment dy-

namics, respectively. The performance of multimodal emotion recognition based on hybrid-level fusion was enhanced in terms of *F-1 score* from 0.67 to 0.95 by including the selected features: age, gender, and culture. The proposed approach is able to recognize perfectly the 9 observed emotions considered in this study: happy, sad, anger, neutral, energy, contempt, disgust, fear, and surprise. The proposed emotion contextual recognition and management approach is able to recognize the non-directly observable emotions by contextualizing the observed emotions. Seven non-directly observable emotions were considered in this study: curious, confusion, upset, dislike, liking, positive surprise, and negative surprise; these emotions are recognized accurately with an average *F-1 score* 0.84. The scenario dedicated to the cognitive assistance of visitors of smart devices showroom showed promising results in terms of relevance and usefulness of the provided service.

Chapter 6

# Hybrid approach for human activities recognition

## 6.1 Introduction

In this chapter, a novel hybrid framework for human activity recognition is proposed. A combaniation of Spatio-Temporal Joint based Convolutional Neural Network (STJ-CNN) and Hierarchical Multichannel deep Residual Network (HMResNet) is proposed to recognize human activities from both skeleton keypoints and multichannel IMUs's raw data. Besides, a novel representation and inference based on NKRL HClass and HTemp ontologies are proposed to represent and combine the detected human activities with the ambient events, and infer the semantic context of the detected activity. The proposed approach combines both the IMUs-based and the skeleton-based activity recognition to overcome the misclassification error caused by sensor instability, visual occlusions, and visual perspective changes.

## 6.2 Proposed hybrid approach

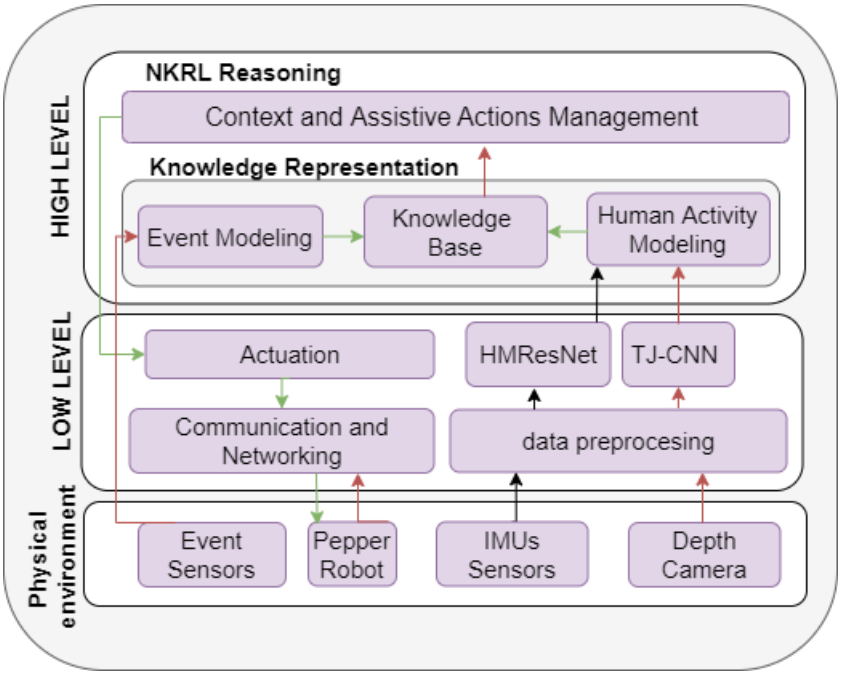



Figure 6.1 – Hybrid approach for human activity recognition

### 6.2.1 Data Preprocessing

In this chapter, both the skeleton joints (keypoints) and IMUs raw-data are preprocessed in three main steps: data selection, missing values linear interpolation, and temporal segmentation. In the data selection step, the human skeleton is divided into five main parts: human's head, right arm, left arm, right leg, and left leg, such that each part consists of multiple joints as shown in Fig.6.2. Besides, both of the linear and angular motion based on accelerometer and gyroscope raw data are selected from the IMUs raw data. In the segmentation step, for each IMUs channel and each skeleton joint coordinate (i.e., X,Y,Z), to create temporal-shifted windows, the overlapping sliding window technique is exploited to partition the continuous input frames into fixed-size window of 1

sec and overlapped with 0.25 sec. Both of the window size and the overlapping ratio are defined based on the empirical experiments carried out on different datasets [235, 236]. Finally, for each IMUs channel and each skeleton joint, the missing data points are filled using polynomial interpolation.

### 6.2.2 STJ-CNN for skeleton-based activity recognition

In this chapter, To recognize human daily activities from skeleton joints (key-points) in an AAL system, a Spatio-Temporal Joint based Convolutional Neural Network (STJ-CNN) is proposed. Compared to the previous approaches [237, 238, 45, 27], STJ-CNN model consists of two different levels of fusion: Joint-level and Body-Part level fusion networks. On one hand, for each joint , a novel temporal multi-channels residual network (TRN) is exploited to combine the temporal 3D (X,Y,Z) coordinates and extract low-level features from that combination. On the other hand, for each body part (i.e., arm), an inverted pyramid convolutional fusion layer (CFL) is used to combine the low-level features of the joints (i.e., shoulder, elbow, hand) relative to that body part. Finally, a fully connected network is exploited to fuse the different body parts features and to recognize the human activity. The hierarchical fusion networks enable the proposed approach to extract more complex features and help to train more sparse and deep architectures.

#### Skeleton joint level fusion

At this level, for each skeleton joint coordinate (i.e., X,Y,Z), as shown in Fig.6.3, the convolutional network building block (CBB) of the STJ-CNN model is proposed. The CBB module consists of a 1D convolution layer followed by Batch Normalization [239] and Rectified Linear Unit (R*e*LU) layers, which are exploited to prevent the vanishing gradient problem, accelerate the model convergence,

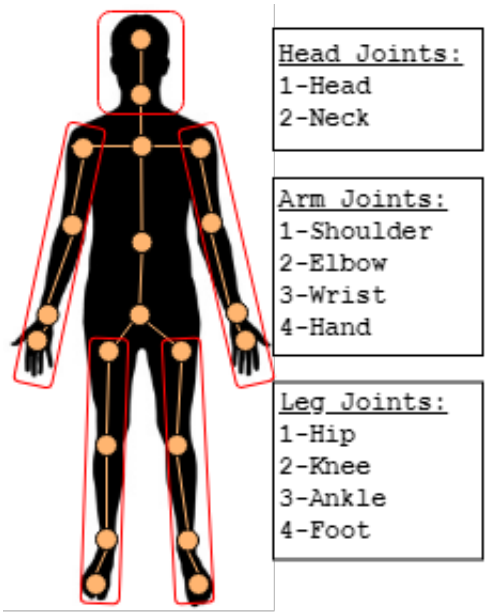



Figure 6.2 – Human Body Parts (red boxes) and Skeleton Joints (orange dots) Selection

and improve the learning generalization. To build the coordinates feature extractor (CFE), a set of 3 CBB modules are created with incremental kernel sizes (KS), and without strides to learn 64, 128, and 265 feature maps respectively, as shown in Fig.6.3. To develop a deeper feature extractor (DFE), inspired by ResNet convolution network [240], multiple CFE modules combined with residual shortcut connections are stacked together. The features transferred using the residual shortcut connections are concatenated with the latent-space features to create a new feature vector for the incoming convolutional layers. Finally, a Global Average Pooling layer is used to minimize the model over-fitting, as shown in Fig.6.3. The optimal number of feature maps and kernel sizes are defined based on the experimental trials applied on different datasets.

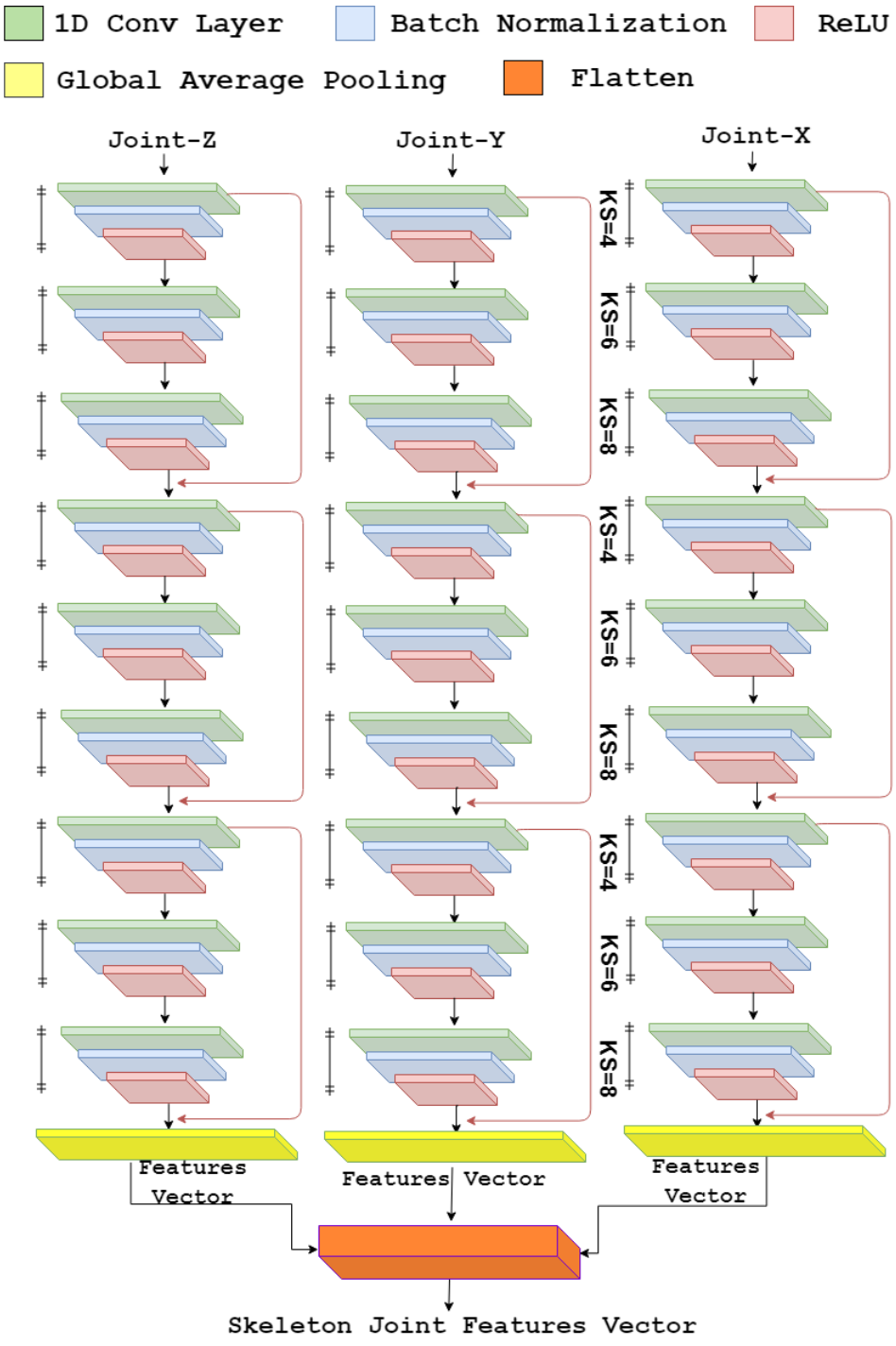



Figure 6.3 – Skeleton Joint Deep Convolutional Residual Feature Extractor Network

**Body parts level fusion**

At this level, for each body part (i.e, arm), after extracting the feature vectors of the relative skeleton joints (i.e, wrist, hand, elbow, shoulder) for that part, the inverted pyramid convolution fusion layers (CFL) are applied on the concatenated features of the relative joints to extract high-level features, and reduce the number of learning parameters without degrading the overall accuracy [241], as shown in Fig.6.4. The proposed inverted pyramid architecture exploits the down-sampling technique without increasing the number of feature maps (Fm) to represent the long-term association of different joints features, while reducing the computation of the successive blocks.

As shown in Fig.6.4, the inverted pyramid architecture consists of 3 building blocks; each one consists of 3 convolution layers (conv) followed by max-pooling (Mp) of kernel size 4 and stride (S) with 2 steps. The max-pooling operation is applied to create a new internal representation from the concatenated features of the relevant joints for that body part. Finally, the stride operation is applied to reduce the size of the generated representation by half.

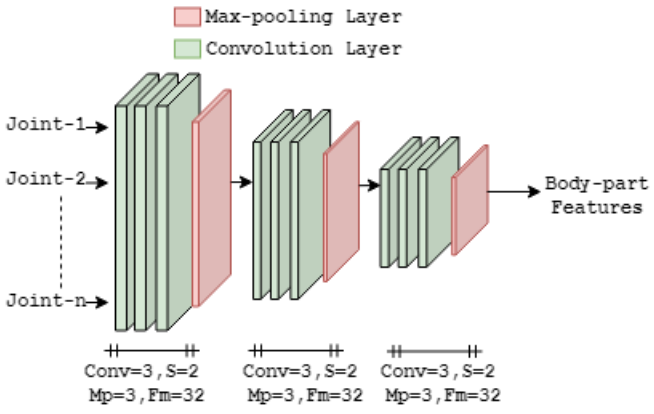



Figure 6.4 – Inverted Pyramid Convolutional Network for body part feature extraction

**Human activity classification**

As shown in Fig.6.5, a bottleneck MLP (Multi-Layer Perceptron) network is exploited to classify the human activities based on the extracted features from different body-parts fusion networks.

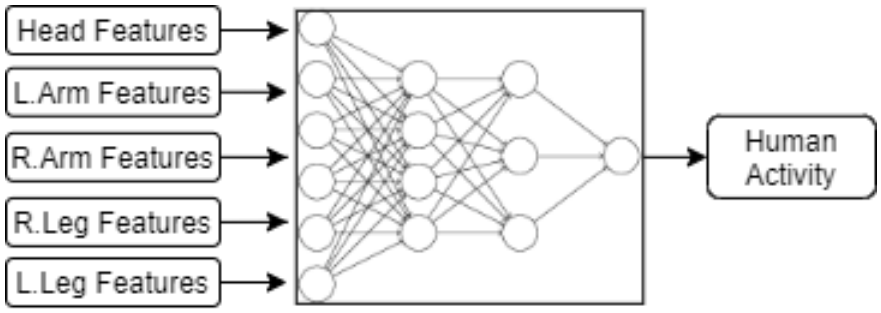



Figure 6.5 – Human Activity Classification Network

### 6.2.3 HMResNet for IMUs-based activity recognition

In this chapter, to recognize human daily activities based on IMUs raw-data, a novel Hierarchical Multichannel deep Residual Network (HMResNet) is proposed. Compared to the state of the art deep learning based approaches [242, 29, 30, 243], the proposed model is based on multilevel fusion layers, with residual shortcut connections, exploits the multichannel raw data to accurately recognize human daily activities. At the features fusion level, a Multichannel 1D Deep Residual Network combined with a Bottleneck MLP neural network is used, for each sensor channel, to automatically extract features from raw data. At the decision fusion level, a multi-sensor fusion layer based on deep 1D ResNet followed by a fully connected MLP neural network is exploited for recognizing daily human activities.

### Feature Level Fusion

At the feature level fusion, a 1D deep ResNet is exploited to extract features automatically from the preprocessed raw data, followed by a Bottleneck MLP neural network to extract sensor level features as shown in Fig.6.6.

### Deep Residual Network (ResNet)

Basically, the Deep Residual Network (R*es*N*et*) developed by Microsoft research labs, is exploited in [240] for image recognition. ResNet got the first place in the five main tracks of COCO and ImageNet competitions, which cover challenges on object recognition, image classification, and semantic segmentation. Afterwards, many studies started to evaluate ResNet performance in different fields such as speech recognition [244], and question answering systems [245]. However, to our knowledge, a single attempt was proposed in [246] to use ResNet for time-series classification.

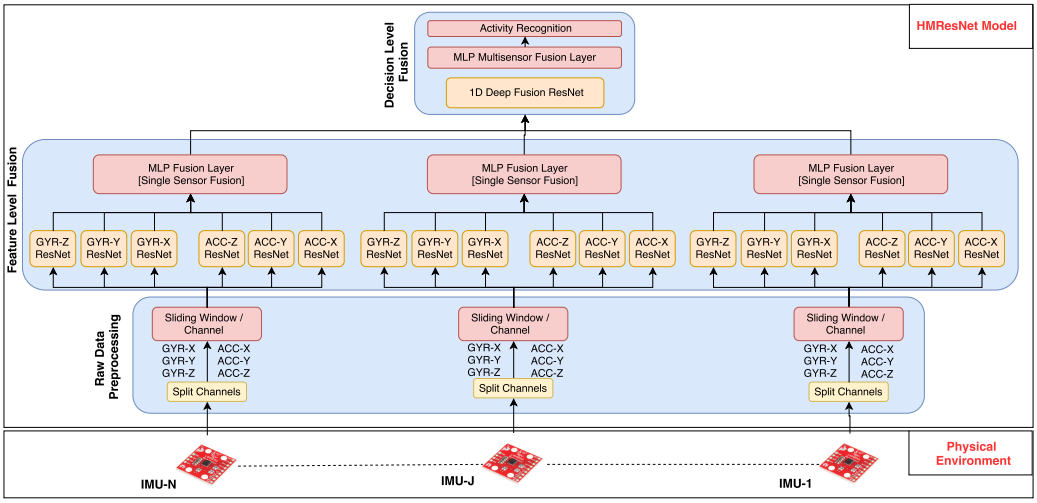



Figure 6.6 – Daily Human Activity Recognition based-Hierarchal Multichannel Deep Residual Network Model for robotic systems Exploiting N IMUs

In the proposed model, the basic block of ResNet is 1D convolutional layer

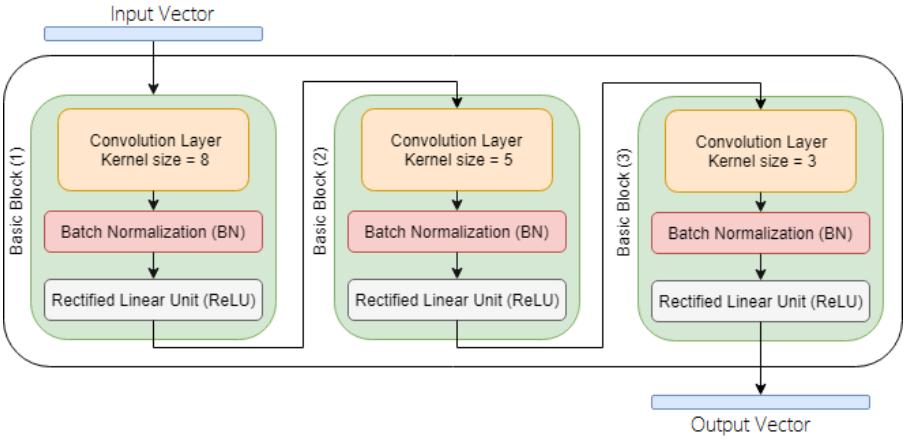



Figure 6.7 – Multilayer Convolution Feature Extractor Unit (MCFEU)

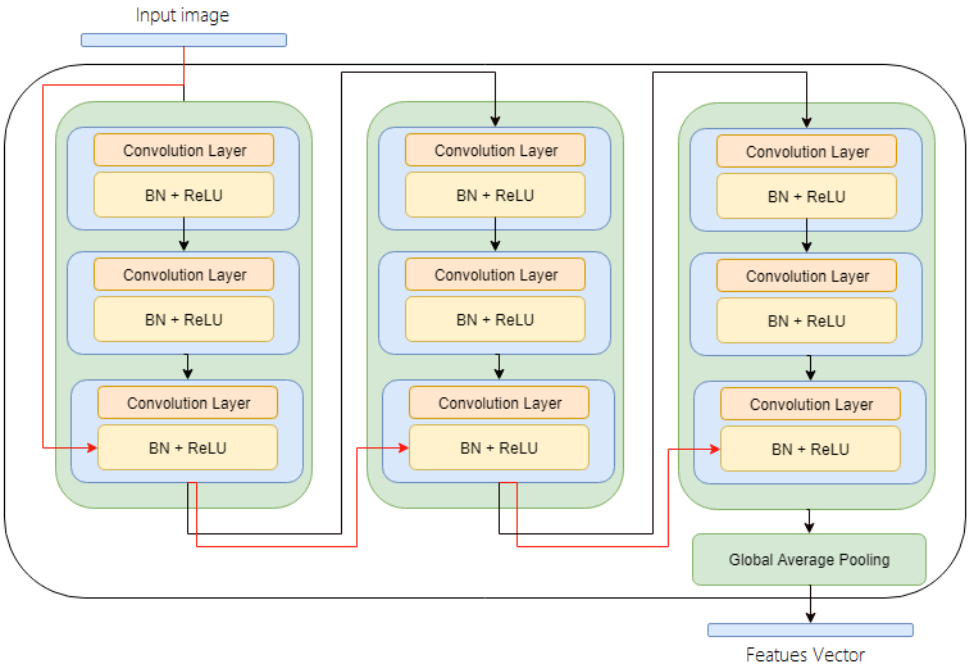



Figure 6.8 – Deep Residual Network Based on Stacked MCFEU units

with kernel ($W_n$) of size ($s$) followed by Batch Normalization (BN) [239] and Rectified Linear Unit (R*e*LU) layers. To avoid the problem of the vanishing gradient, ReLU activation function is used. The Batch normalization (BN) is applied to speed up the model convergence and improve the model generalization. Inspired by ResNet152 deep model [240], a plain network based on 3 basic blocks is developed with different 1D kernel sizes, without strides, with 32, 64, and 64 feature maps respectively to create Multilayer Convolution Feature Extractor Unit (MCFEU) as shown in Fig.6.7. Both the kernel sizes and the number of feature maps have been chosen based on the empirical experiments which were conducted on different datasets [29, 47]. MCFEU is exploited to extract multilevel features from time-series preprocessed raw data. The complete deep ResNet model is developed by stacking multiple MCFEU units, besides adding residual shortcut connection between the MCFEU units as shown in Fig.6.8. The shortcut connections are exploited to ensure that every MCFEU unit is learning more complex features and to solve the problem of the vanishing gradient for deep networks [240]. Before feeding the extracted features to the decision level fusion layer, a Global Average Pooling (GAP) layer [247] is used to minimize the model overfitting by reducing the total number of the learned parameters.

**MLP Neural Network for Sensor Level Fusion**

In this chapter, for every single IMU, a bottleneck MLP neural network is exploited as a fusion layer for the sensor channels as shown in Fig.6.6. The bottleneck MLP neural network acts as a nonlinear dimension reduction module, used to extract low-dimensional features from the integrated deep ResNet output features. Finally, the outputs of all bottleneck MLP neural networks are integrated into a single feature vector, which is fed to the decision level fusion

layer. In this work, the bottleneck MLP network includes two fully connected hidden layers where each layer consists of 1000 nodes. The Dropout algorithm [157] is applied on 30% of the nodes to prevent the network from overfitting the training features. To break the symmetry of the neurons performance, The network weights are randomly initialized with small values close to zero based on normal distribution. For every node, the Rectified Linear Unit (ReLU) is exploited as an activation function.

**Decision level fusion**

At decision level fusion, both the 1D deep ResNet and the Bottleneck MLP neural networks are exploited to recognize daily human activities as shown in Fig.6.6. The Bottleneck MLP neural network is composed of three fully connected layers. In the hidden layers, each layer consists of 1000 nodes which are based on ReLU activation function. The output layer consists of a number of nodes equal to the total number of target activities. Besides, the softmax activation function is used for the output layer. During the training phase, the categorical cross entropy cost function is exploited to calculate the difference between the target labels and the predicted labels. This difference is exploited by the back-propagation algorithm [248] to update the parameters to be learned of both the feature level and decision level layers during the training phase. Finally, The Adam algorithm [249] is used for optimizing the MLP categorical cross-entropy cost function.

### 6.2.4 Contextual activity detection

To enable an AAL system to serve humans in an environment populated with different objects, it should perceive both the ambient environment status and the human activities. Therefore, a knowledge representation engine is required

to model different environmental events (i.e. on/off coffee machine, on/off shower) and human activities (i.e. washing dishes, drink, walk). Besides, a reasoning engine is required to infer the contextual activities with respect to the observed situation. In this section, NKRL knowledge representation is exploited to model both the ambient events and the human activities. Besides, NKRL reasoning engine is used to exploit the represented event and activities to detect the contextual activities.

**Daily activities knowledge representation**

In this work, an ontological model based on the Narrative Knowledge Representation Language (NKRL) [250] is proposed to model both the $static$ and $dynamic$ characteristics of any entity populating the AAL environment. The *HClass* ontology is used to model the static commonsense knowledge for human-centered applications such as *Human*, *Activity*, *Object*, etc. This ontology is a binary ontology characterized by the common properties relationships of semantic web ontologies to describe the generalization / specialization between concepts. Besides, The *HTemp* ontology is exploited to model the dynamic and temporal ambient events that can be inferred or observed by the AAL system. Besides, The *HTemp* ontology is used to represent the human activities detected from different modalities (IMUs and skeleton key-points).

**NKRL daily activities-aware reasoning**

The NKRL reasoning engine exploits both the *"transformation and the hypothesis rules"* to infer, on the one hand, the implicit relations between the different predicative occurrences, on the other hand, to infer the chronological/semantic context. In other words, the NKRL reasoning engine allows exploiting the ambient environment temporal representation (i.e. Fridge is opened at 10:00 am,

coffee machine is switched on at 10:02 am, floor cleaning robot is fully charged at 10:15 am) and combining it with the temporal representation of the recognized activities from the HMResNet and STJ-CNN models (i.e. Person is cooking at 10:09 am) to infer a new activity (i.e. Person is preparing breakfast).

In HTemp ontology, the time interval ($[00h, 12h]$) are associated with a predefined time marks (*"morning_"*) which are predefined as *HClass* concepts. Consequently, based on the temporal-axioms (i.e., After, Before), the NKRL engine allows inferring the chronological/semantic contexts by binding different predicative occurrences, see Eq.6.1, which represents the fact that a predicative occurrence Y occurred after a predicative occurrence X.

$$\text{AFTER}(X, Y) \tag{6.1}$$

In this study, the contextual human activities can be predicted by inferring the chronological order of the associated predicative occurrences. For example, a daily habit for a human called "Steve", who is usually cooking before cleaning the kitchen. As shown in table.6.1, the NKRL represents the main temporal events such as *"Every morning, Steve is cooking"* and *"Every morning, Steve is cleaning the kitchen"* with the predicative occurrences **"E.occ1"** and **"E.occ2"**.

| |
|---|
| E.occ1: EXPERIENCE<br>SUBJ STEVE_1 : KITCHEN_1<br>OBJ COORD (SPECIF(cooking_)<br>SPECIF(during_ morning_))<br>date-1: 15/09/2019 10:00:00<br>date-2: |

Consequently, as shown in Eq. 6.2, the NKRL reasoning engine allows inferring the binding occurrences "Steve is cleaning the kitchen after cooking" from the chronological order of the observed events and activities. Besides,

| E.occ2: EXPERIENCE<br>SUBJ STEVE_1 : KITCHEN_1<br>OBJ COORD (SPECIF(cleaning_)<br>SPECIF(during_ morning_) )<br>date-1: 15/09/2019 10:23:13<br>date-2: |
|---|

Table 6.1 – NKRL Predicative Occurrences Representation

the NKRL reasoning engine allows inferring the implicit activity "preparing his breakfast" from the temporal information associated with the predicative occurrences **"E.occ1"**.

$$\text{AFTER}(\textbf{E.occ1}, \textbf{E.occ2}) \tag{6.2}$$

Finally, to provide an assistive service for a human such as cleaning the kitchen floor, the NKRL transformation rules are used to infer the contextual action to be taken by the robot. As shown in table 6.2, the NKRL inference engine used the detected activities *cooking, cleaning* to infer the currently required action to be taken by the robot *" clean_kitchen_"*. In this example, a robot will wait for a human to finish *cooking* activity. Then, once the human starts *cleaning* the floor, the robot will start *"clean_kitchen_"* task to assist the human. The human behavior *"cleaning at morning, after cooking"* is represented using the predicate *AFTER*. The operator *COORD* appears in the antecedent of the transformation rule, to aggregate and combine the detected activities (*cooking, cleaning*) and the observed ambient event that occurred (*"battery_charged_"*). Finally, the consequent part consists of the required action to be taken by the robot (*"clean_kitchen_"*).

**Antecedent:**
COORD($b_1$ $e_3$)
$b_1$ AFTER($e_1$ $e_2$)
$e_1$: EXPERIENCE SUBJ var1 :
OBJ COORD (SPECIF(cooking_ )
SPECIF(during_ morning_) )
var1:human_being
$e_2$: EXPERIENCE SUBJ var1 :
OBJ COORD (SPECIF(cleaning_)
SPECIF(during_ morning_) )
var1:human_being
$e_3$: EXPERIENCE SUBJ var1 :
OBJ SPECIF(battery_charged_)
date-1: var2 date-2:
var1:robot_
var2:morning_
**Consequent:**
$a_1$: PRODUCE SUB var1:
OBJ SPECIF(task_clean_kitchen)
var1:robot_

Table 6.2 – Transformation rule to provide an assistive service for human

## 6.3 Experiments

The proposed hybrid approach for human activity recognition is evaluated in three phases: 1) on one hand, the STJ-CNN model is evaluated using Daily Home Life Activity dataset (DAHLIA) [235] and NTU RGB+D dataset [251]. ; 2) The HMResNet model is evaluated using HARS dataset [252] and PH12-ARI Sensors dataset [47].; 3) on the other hand, to validate the contextual activities detection in real-time scenarios, an assistive service scenario is implemented.

### 6.3.1 Datasets

#### Daily Home Life Activity dataset (DAHLIA)

Daily Home Life Activity Dataset (DAHLIA) [235] is the most comprehensive public dataset for recognizing human daily living activities. DAHLIA dataset consists of 44 different subjects participated in recording 51 different sessions in a fully monitored kitchen with 3 different Kinect depth sensors. For each session, RGB videos, depth maps, skeleton data, and body indices are recorded for 40 minutes. Besides, a set of 7 different daily activities (cooking, washing dishes, eating, clearing table, working, housework, and laying) are labeled. In this study, the skeleton data, which consists of 3D locations of 25 human skeleton-joints in the 3D point space are used.

#### NTU RGB+D dataset

NTU RGB+D dataset [236] consists of RGB videos, infrared videos , depth maps, and 3D skeleton data of 60 different activities captured by 3 different Kinect depth sensors. These activities are split into three main categories: daily, mutual, and health related activities. The activities were performed by 40 different subjects, which produced 56,880 activity samples in total. In this study, the skeleton data are used, which consists of 3D locations of 25 human skeleton-joints in the 3D point space. The main evaluation protocols proposed in [251] are exploited for both cross-subject and cross-views. Regarding the cross-subject evaluation, the activities performed by the first 20 subjects (50%) are used for training, while the activities performed by the remaining 20 subjects (50%) are used as a testing set. Regarding the cross-view evaluation, the activities captured by the first 2 Kinect sensors are used for training, while those activities captured by the last Kinect are used as a testing set.

**HARS Dataset**

This dataset contains data from six activities, collected using a smartphone with a built-in tri-axial accelerometer/gyroscope inertial sensor. Activities were performed by 30 volunteer subjects holding a smartphone in a pocket tight around their waist. The activities are a mix of periodic and static activities such as *WALKING, WALKING UPSTAIRS, WALKING DOWNSTAIRS, SITTING, STANDING, and LAYING*. The data were sampled at 50Hz and divided into fixed length windows of 128 samples with 50% overlap. Butterworth low-pass filter is used to separate body acceleration and gravity from the accelerometer raw data. The data were separated into a training set with 7352 windows from 21 randomly selected subjects, and testing set of the remaining 2947 windows.

**PH12-ARI dataset**

This dataset contains data collected from three IMUs sensors placed on the chest, the right thigh and the left ankle of the subject. Each IMUs sensor has built-in inertial sensor with a tri-axial accelerometer, a tri-axial gyroscope and a tri-axial magnetometer. The dataset consists of 12 different activities performed by 6 volunteer subjects. The activities are a mix of periodic and static activities with transitional activities such as: *A1:WALKING DOWNSTAIRS, A2: STANDING, A3:SITTING DOWN, A4:SITTING, A5:FROM SITTING TO SITTING ON THE GROUND, A6:SITTING ON THE GROUND, A7:LYING DOWN, A8:LYING, A9:FROM LYING TO SITTING ON THE GROUND, A10:STANDING UP, A11:WALKING, and A12:WALKING UPSTAIRS*. The dataset was sampled at 25Hz and no sliding window was applied to the raw data.

### 6.3.2 STJ-CNN model evaluation

In this section, the performance of the proposed STJ-CNN model is compared with those obtained with different state-of-the-art models on both DAHLIA and NTU RGB+D datasets.

**Daily Home Life Activity Dataset (DAHLIA)**

| | **View-1** | **View-2** | **View-3** | **Cross-View** |
|---|---|---|---|---|
| **DOHT [235]** | 0.58 | 0.60 | 0.71 | 0.31 |
| **ELS [235]** | 0.18 | 0.26 | 0.55 | 0.32 |
| **Plain-CNN** | 0.624 | 0.667 | 0.751 | 0.305 |
| **STJ-CNN [ours]** | **0.823** | **0.831** | **0.847** | **0.382** |

Table 6.3 – Performance comparison of the proposed STJ-CNN model against the state-of-the-art models on DAHLIA dataset in F-score

The F-score results obtained with the Deeply Optimized Hough Transform (DOHT) [235] model, Efficient Linear Search (ELS) [235] model, Plain-CNN, and the proposed STJ-CNN model, are shown in Table 6.3. To ensure a fair comparison with the baseline models, all models are trained only using the skeleton raw data from DAHLIA dataset. From the F-score results, the STJ-CNN model substantially outperforms the baseline models and leads to obtain the best F-score for both single and cross views configurations. As shown in Table 6.3, compared to the best classical baseline model (DOHT), the F-score is improved by 7.2%, 13%, 23%, and 24% for Cross-View, View3, View2, and View1 respectively. Compared to the STJ-CNN model, the Plain-CNN model input vector is a concatenation of all skeleton joints without any body-parts fusion layers. Consequently, the proposed STJ-CNN model allows obtaining an improvement of 7.7%, 9.6%, 16.4%, and 19.9% for Cross-View, View3, View2, and View1 respectively. The F-score obtained with the Cross-View testing configuration is sub-

stantially decreased for all models, because the visual features learned from different views are not identical.

**NTU RGB+D dataset**

| Model | Cross-View | Cross-Subject |
|---|---|---|
| HBRNN-L [253] | 0.64 | 0.591 |
| Part-aware LSTM [251] | 0.703 | 0.629 |
| Trust Gate ST-LSTM [236] | 0.777 | 0.692 |
| Two-stream RNN [254] | 0.795 | 0.713 |
| STA-LSTM [255] | 0.812 | 0.734 |
| Ensemble TS-LSTM [256] | 0.813 | 0.746 |
| Visualization CNN [257] | 0.826 | 0.760 |
| VA-LSTM [258] | 0.876 | 0.794 |
| ST-GCN [27] | 0.883 | 0.815 |
| HCN [259] | 0.911 | 0.865 |
| MS-G3D Net [260] | **0.962** | **0.915** |
| STJ-CNN [ours] | 0.928 | 0.876 |

Table 6.4 – Performance comparison of the proposed STJ-CNN model against the state-of-the-art models on NTU RGB+D dataset in accuracy

As shown in Table 6.4, the proposed STJ-CNN model outperforms the-state-of-the-art models and achieves the best accuracy of 92.8% and 87.6% for both cross-views and cross-subjects respectively. These results show the robustness of the STJ-CNN model compared to the baseline models, which are based on different deep architectures such as Recurrent Neural Network (RNN) based models [Two-stream RNN [254], Part-aware LSTM [251], Trust Gate Spatio-Temporal LSTM (SP-LSTM) [236], and Ensemble Temporal Sliding LSTM (TS-LSTM) [256]], and Convolution Neural Networks (CNN) based models [Visualization CNN [257], Spatio-Temporal Graph Convolutional Networks (ST-GCN) [27], Hierarchical Co-occurrence Network (HCN) [259]]. The HCN [259] was designed to learn the global co-occurrences from the 3D skeleton joints data. As the pro-

posed STJ-CNN model, the HCN [259] handles each skeleton joint as a separate input channel. In contrast to the proposed model, a plain convolution layer is exploited as a fusion layer and to learn the global co-occurrence features for all joints. As shown in Table 6.4, compared to HCN [259], the proposed STJ-CNN model allows obtaining an improvement of 1.7% and 1.1% for cross-view and cross-subject evaluations respectively. The results shows that the hierarchical features learned from the different fusion layers are more relative and representative than the global co-occurrence features learn by the HCN model. Finally, compared to the state-of-the-art Disentangling and Unifying Graph Convolutions Network (MS-G3D Net), the proposed model achieves comparable performance in terms of cross subject evaluation. Based on the latter results, the STJ-CNN model shows a good performance on two different datasets from two different domains, which reflects the stability of the proposed model. Besides, the ability of the proposed model to classify different daily activities, from multiple individual and cross views for different subjects, reflects the robustness of the proposed model against the occlusion and noisy skeletal joint data. The multi-level fusion for different skeleton joints and for different body parts combined with residual shortcut connections allows the extraction of more complex features than hand-crafted, plain CNN, and global co-occurrence features. Finally, the STJ-CNN model shows better performance than both the classical and deep learning baseline models for recognizing human daily activities based on skeleton-joints data.

### 6.3.3 HMResNet model evaluation

The HMResNet model is evaluated against the following baseline models : (i) k-NN with time domain and frequency domain features [47] using the PH12-ARI dataset. (ii) Convnet combined with MLP neural network applied to raw

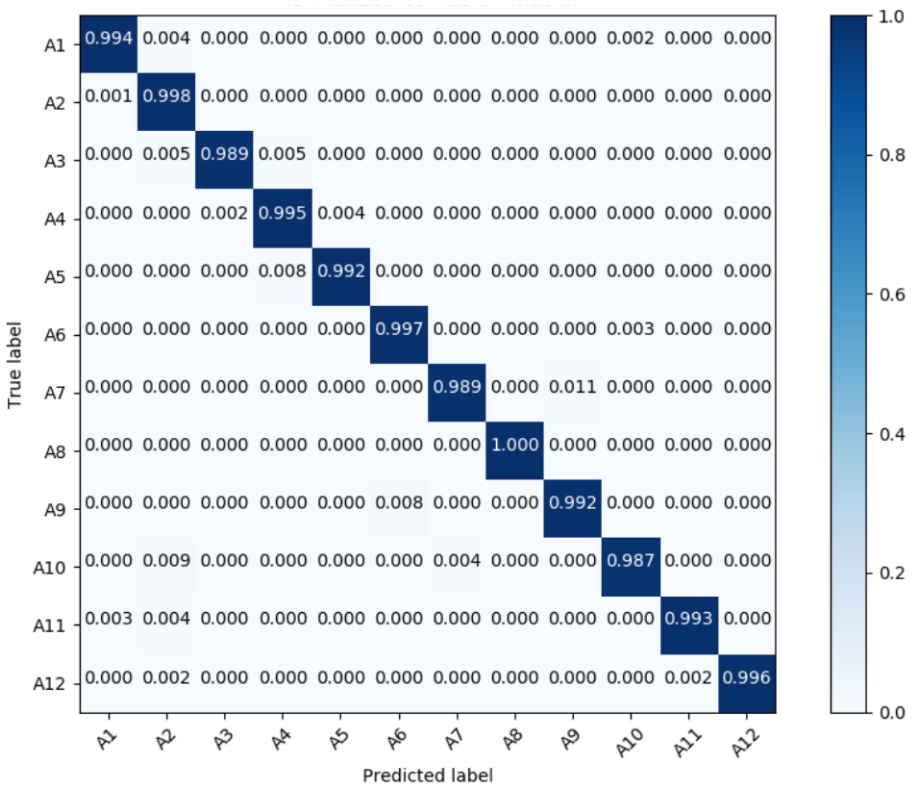



Figure 6.9 – Confusion Matrix obtained by HMResNet using PH12-ARI dataset

data [29], and Convnet with tFFT features [29], using the HARS dataset. The architecture shown in Fig.6.6 is used for both datasets, the only differences are the number of input sensors and the number of output classes since the input and output of the datasets are different. Both of the features level deep networks and the decision level deep networks are trained together to ensure the consistency of the learning process.

The obtained results show that the proposed HMResNet model significantly outperforms the baseline models and achieves better classification accuracy performance compared to the best baseline model (ConvNet). From confusion matrices, one can observe a significant improvement in performance obtained with the proposed model for static activities (SITTING, STANDING, and LAYING) which constitute a major impediment for the best baseline model to

Table 6.5 – HARS dataset Accuracy Evaluation

| Method | Accuarcy (%) |
|---|---|
| **Baseline Models [29]** | |
| PCA+MLP | 57.10 |
| HCF+NB | 74.32 |
| HCF+J48 | 83.02 |
| SDAE+MLP(DBN) | 87.77 |
| HCF+ANN | 91.08 |
| HCF+SVM | 94.61 |
| **Deep Learning Models [29]** | |
| Convnet (inverted pyramid archi)+MLP | 94.79 |
| tFFT+Convnet ($(J(L_1))$=200) | 95.75 |
| **Proposed Model** | |
| **Hierarchal Multichannel Deep ResNet** | **97.619** |

classify them correctly. In the case of LAYING activity, the number of the correctly classified classes is improved by 12.3% with no misclassified classes. For STANDING activity, the number of the correctly classified classes is improved by 4.8%, Besides the number of misclassified classes decreased by 4.5%. For SITTING activity, the number of the correctly classified classes is improved by 0.4%, Besides the number of misclassified classes is decreased by 0.6%.

These results show that, for both static and periodic activities, the proposed model shows the hierarchal architecture with multilevel fusion layers combined with residual shortcut connections allows extracting more relevant features than both the hand engineered ones and the plain CNN learned ones. In addition, the proposed model is more accurate than both the traditional machine learning and deep learning baseline models for recognizing human daily activities based on a single tri-axial accelerometer/gyroscope inertial sensor.

To evaluate the extendability and the robustness of the proposed model regardless of the hardware configuration, the proposed model was benchmarked against another baseline models using PH12-ARI dataset. In terms of average

Table 6.6 – HAR Using HARS dataset Confusion Matrix Evaluation

| | ConvNet Predicted classes | | | | | | Proposed Model Predicted classes | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Actual class** | **W** | **WU** | **WD** | **Si** | **St** | **L** | **W** | **WU** | **WD** | **Si** | **St** | **L** |
| **Walking** | **491** | 3 | 2 | 0 | 0 | 0 | **496** | 0 | 0 | 0 | 0 | 0 |
| **W. upstairs** | 0 | **471** | 0 | 0 | 0 | 0 | 3 | **468** | 0 | 0 | 0 | 0 |
| **W. downstairs** | 0 | 0 | **420** | 0 | 0 | 0 | 2 | 0 | **417** | 0 | 1 | 0 |
| **Sitting** | 0 | 0 | 0 | **436** | 34 | 21 | 0 | 1 | 0 | **438** | 52 | 0 |
| **Standing** | 0 | 1 | 0 | 24 | **496** | 11 | 0 | 0 | 0 | 10 | **522** | 0 |
| **Laying** | 0 | 0 | 0 | 43 | 23 | **471** | 0 | 0 | 0 | 0 | 0 | **537** |

precision-recall values. The performance of the proposed model in terms of precision and recall metrics and those obtained with the eight base line models are shown in Table 6.7.

Among the baseline models, the K-nearest neighbor (k-NN) model applied to time-domain, and frequency-domain features achieves the best performance in terms of average recall, and precision values, followed by Random Forest model (RF), then k-NN model without features. Finally, the Supervised Learning Gaussian Mixture Model (SLGMM) without features achieves relatively the worst results. As shown in Table 6.7, the proposed model outperforms the baseline models which are evaluated on raw data as well as on the hand-crafted features. The obtained results show that the proposed model improves the values of the precision and recall to be 99.22% and 98.88% respectively, while those obtained with the baselines methods vary from 69.88% to 98.85% and from 69.99% to 98.85%. The best results are highlighted in bold for in the table. The proposed model achieves almost perfect results as shown in the confusion matrix in Fig.6.9.

Because of the small size of the dataset, the obtained results show a slight difference in performance when comparing the the proposed (HMResNet) model to the best baseline (k-NN with features) model. Even though the small difference, the extraction of features phase requires integrating additional models and algorithms to the baseline models. Besides, the feature extraction phase

needs extra computation time, which is not practical for real-time applications.

From the latter empirical evaluation experiments, one can observe the proposed HMResNet model outperforms the baseline models, extract more relevant features, and recognize periodic, transitional, and static daily human activities from single IMU up to 3 IMUs sensors, which shows the robustness of the proposed model regardless of the hardware configuration.

Table 6.7 – PH12-ARI dataset Evaluation

| Model | Precision(%) | Recall(%) |
|---|---|---|
| **Without features [47]** | | |
| KNN | 94.62 | 94.57 |
| RF | 83.46 | 82.28 |
| SVM | 90.33 | 90.98 |
| SLGMM | 69.88 | 69.99 |
| **With features [47]** | | |
| KNN | 98.85 | 98.85 |
| RF | 98.25 | 98.24 |
| SVM | 92.90 | 93.15 |
| SLGMM | 73.61 | 74.44 |
| **Proposed Model** | | |
| **Deep Multichannel ResNet** | **99.22** | **98.88** |

## 6.4 Usecase: Cognitive daily exercises coaching

To validate the proposed approach for real-time activity recognition, a use-case of cognitive daily exercises coaching for a diabetic person is studied, see Fig.6.10. This use case consists of a robot, called Pepper, that is acting as a training coach of a diabetic person, called Alice. Indeed, Pepper recognizes and guides the daily exercises which were prescribed by a doctor for *Alice*. This work is reported in a multimedia video that is available on *LISSI*'s Website [1].

[1] http://www.lissi.fr/videos/HMResNet.php

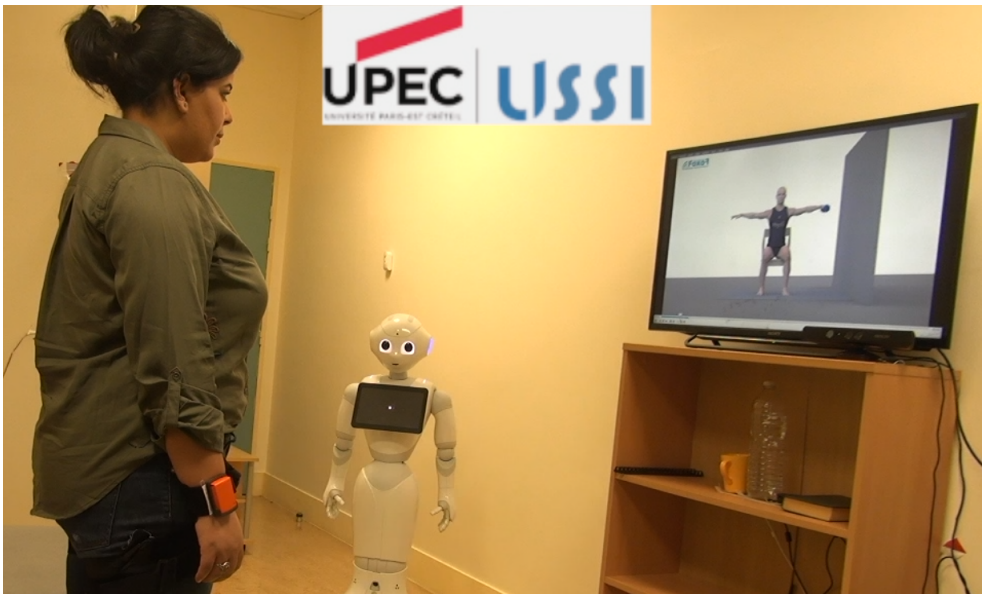



Figure 6.10 – Scene extracted from the smart home environment

During the experiment, The proposed approach was evaluated by streaming 6974 sliding windows of 1.5 seconds and the average processing time to recognize a single activity was 0.2 seconds. Therefore, The processing time is reasonably fitting the constraints of real-time activity recognition. This use case is a part of MEDOLUTION European project [2] which is funded by ITEA3 Research, Development and Innovation (RDI) program.

## 6.5 Conclusion

In this chapter, a novel hybrid framework for human activity recognition is proposed. A combaniation of Spatio-Temporal Joint based Convolutional Neural Network (STJ-CNN) and Hierarchical Multichannel deep Residual Network (HMResNet) is proposed to recognize human activities from both skeleton keypoints and multichannel IMUs's raw data. Besides, a novel representation and

[2] https://itea3.org/project/medolution.html

inference based on NKRL HClass and HTemp ontologies are proposed to represent and combine the detected human activities with the ambient events, and infer the semantic context of the detected activity. The proposed approach combines both the IMUs-based and the skeleton-based activity recognition to overcome the misclassification error caused by sensor instability, visual occlusions, and visual perspective changes. Compared to the baseline models, the performance of the STJ-CNN model shows a significant improvement up to 24% in terms of F-score on DAHLIA dataset. The performance of the daily human activity recognition based on HMResNet model is shown through two activity datasets. The proposed automatic features extraction model is more relevant than both the hand engineered features and the plain CNN learned features. It is able to recognize perfectly, in terms of precision, the static activities: *SITTING, STANDING,* and *LAYING*. The obtained results demonstrated that the proposed model outperforms baseline methods exploiting the same datasets Finally, a use-case for coaching a dependent person is carried out.

Chapter 7

# Hybrid approach for human activity anticipation

## 7.1 Introduction

In this chapter, a novel hybrid approach for human activity recognition and intention anticipation from partially observed environment is proposed. The proposed hybrid approach combines both of human ego-centric and third-person vision perspectives to accurately recognize and proactively anticipate human daily activities. At the low-level, a modified YOLO (You Only Look Once) deep model combined with GLCM-CNN ( Gray-Level Co-Occurrence Matrix - Convolution Neural Network) deep Auto-Encoder model are proposed for precise hand detection in the cluttered environments. At the high-level, ConceptNet-based [203] ontology combined with probabilistic reasoning are proposed to represent the ambient-environment and infer the implicit relations between the ambient-objects and the human activities. To evaluate the proposed approach, we collected two novel datasets for human hand detection in the cluttered environment: DHA-$11_{\mathrm{TH}}$ (Diverse Hands) and SKNS (skin/non-skin tex-

ture) datasets.

## 7.2 Proposed approach

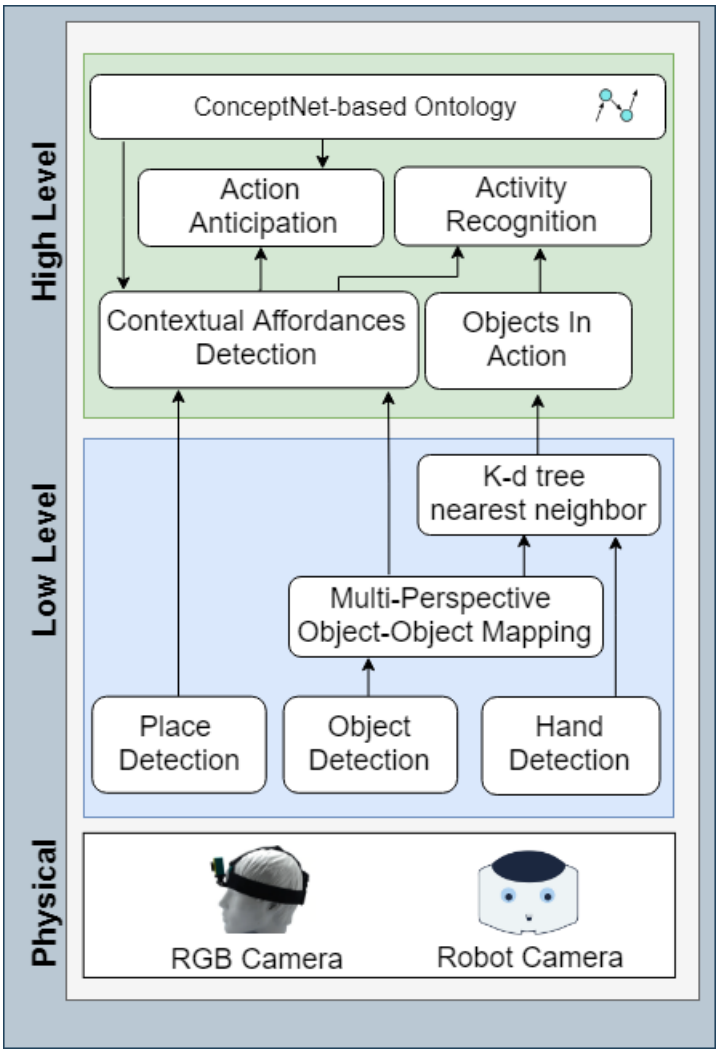



Figure 7.1 – Proposed hybrid approach for huamn activity recognition and intention anticipation

In this chapter, as shown in Fig.7.1, a novel approach for human activity recognition and intention anticipation is proposed. The role of the physical layer is

to monitor the ambient environment from two different perspectives: human ego-centric and third-person vision perspectives. An the low level layer, the output of the physical layer is processed to detect indoor place, ambient objects populating the environment, and the precise hand locations. Besides, a Muti-Perspectives Object-to-Object mapping model is used to associate the similar objects recognized from different perspectives to solve the problem of visual perception uncertainty. Finally, a Hand-to-Object mapping model is used to analyze the spatial relationships between human hands and ambient objects to solve the problem of the partial observations of an environment. At the high level layer, the human activity and the object contextual affordances are predicted using a probabilistic reasoning engine combined with ConceptNet-based ontology.

### 7.2.1 Human Hand Detection in Cluttered Environments

In the proposed approach, the problem of detecting human hands from an ego-centric vision camera worn by the user consists of two main steps: 1) the initial detection of human hands 2D bounding boxes; 2) the automatic false-positive correction based on the elimination of non-skin-texture bounding boxes as shown in Fig.7.2.

YOLOv2 (You Only Look Once version 2) [261] model is a unified end-to-end object detection deep learning architecture that is extremely fast compared to the state of the art models [262, 263]. Compared to the state-of-the-art object detection models such as RCNN [262] and Faster-RCNN [263], in YOLOv2, the problem of object detection and localization was framed as an end-to-end regression problem starting from the pixel-based processing to the bounding boxes localization. Consequently, the YOLOv2 network architecture has successfully solved the problems encountered when using regions-proposal based

models [262, 263] and more precisely 1) the need of generating the region-proposals boxes thereafter execute independently the classification stage; 2) the need of a post-processing stage to refine the detected boxes, eliminate the duplication, and re-score the boxes based on the ambient objects in the scene; 3) the partial view of the entire image during the training makes the region-proposal based models miss the global contextual information of the classes besides their appearance. Based on this analysis, the YOLOv2 model is exploited in this chapter as a backbone model to detect human hands in cluttered environments. However, as shown in Fig.7.2.(c), YOLOv2 generates a lot of false-positive human hand boxes because of the variety of skin colors, cluttered and skin-like backgrounds, and the highly deformable shapes of human hands. As shown in Fig.7.2.(e), to eliminate the false positive boxes, a GLCM-CNN deep Auto-Encoder is proposed to verify whether the proposed bounding box contains skin-texture or not.

To enable the YOLOv2 model to detect human hands of different races, genders, and ages, the last convolution layer of the model is replaced by a novel Supervised GLCM-CNN Auto-Encoder network. latter allows classifying the skin-texture of the proposed human hands bounding boxes. The GLCM-CNN network consists of 3 main blocks: 1) GLCM texture extraction network 2) CNN-based Encoder-Decoder network for extracting latent space features; 3) Multi-layer perceptron (MLP) fully connected network for classifying the latent space features to skin or non-skin labels as shown in Fig.7.2.(e).

The Gray-Level Co-Occurrence Matrix (GLCM) features were originally proposed in [264] to extract textural features from remotely sensed satellite images. GLC matrix is a second-order statistical representation that describes the relationship between gray level neighboring pixels by calculating their joint probabilities [264]. In other words, a GLC matrix represents the occurrence of the

pixel-pairs of an image in a horizontal, vertical, or diagonal direction. In [264], a set of 14 different GLCM features was proposed such as contrast, homogeneity, dissimilarity, correlation, and energy. In this study, the input bounding box image is resized to $224x224$ pixels image, then a $7x7$ pixels sliding window with a stride of 3 pixels is used to create the GLCM-CNN input sub-image patches. For each sub-image in the input patch, the GLCM pooling layer is applied to extract the homogeneity (see Eq.7.1), and the energy (see Eq.7.2) feature maps. In one hand, homogeneity is the closeness of the distribution of elements in the GLCM to the GLCM diagonal. In other hands, energy is also known as uniformity, uniformity of energy, and angular second moment. Finally, both of the original RGB bounding box image and the GLCM homogeneity, and the energy feature maps are combined through a series of convolutional layers, as shown in Fig.7.3.

$$homogeneity = \sum_{i,j} \frac{\mathrm{P}(i,j)}{1 + \|i-j\|} \tag{7.1}$$

$$energy = \sqrt{\sum_{i,j} \mathrm{P}^2(i,j)} \tag{7.2}$$

In this thesis, a new loss function combining both YOLOv2 and GLCM-CNN loss functions is proposed to train both of YOLOv2 and the GLCM-CNN Auto-Encoder simultaneously, see Eq.7.3. On one hand, the YOLOv2 loss function (see Eq.7.4) consists of multi-part loss equations as follow: 1) the loss of the predicted bounding box position $(x,y)$, where: $\mathbb{1}_{ij}$ denotes that the $j^{th}$ bounding box in cell $i$ is responsible for the object prediction. ; 2) the loss of the predicted bounding box width and height $(w,h)$; 3) the loss of the predicted confidence score where C denotes the confidence score, $\hat{\mathrm{C}}$ the IOU of the predicted bounding box with the ground truth, and $\mathbb{1}_{ij}^{noobj}$ the inverse of $\mathbb{1}_{ij}$; 4) the object

classification loss where $\mathbb{1}_i^{obj}$ equals 1 when there is an object in the cell, and 0 otherwise. Consequently, the classification error is only penalized if an object exists inside the cell. Finally, the λ parameters are exploited as weighting parameters of the entire loss function to increase model stability where $\lambda_{coord} = 5$ and $\lambda_{noobj} = 0.5$ [261]. On the other hand, the GLCM-CNN Auto-Encoder loss function (see Eq.7.5) consists of two-parts loss equations as follow: 1) the classification categorical cross-entropy loss function of the MLP neural network where: $lbl_i$ denotes the ground truth skin-label of the bounding box image and $\hat{lbl}_i$ denotes the predicted skin-label of the bounding box image; 2) the reconstruction loss function of the decoder network where the mean-squared error between the bounding box image pixels $img^i_{[r,c]}$ and the reconstructed image pixels $\hat{img}^i_{[r,c]}$, is calculated. For each detected bounding box image from the YOLOv2 model, $\mathbb{1}_i^{Hand}$ equals 1 if the bounding box contains a human hand image and 0 otherwise.

In this study, both of YOLOv2 and GLCM-CNN models are pre-trained independently on both EgoHands [265] and Skin/Non-Skin data-sets respectively. Besides, the entire model is fine-tuned on both Viva-Hands [266] and DHA-$11_{Th}$ (Diverse Hands) datasets to enable the model to detect human hands in the cluttered environments. Finally, the model is used to localize human hands from both the human ego-centric and third-person vision perspectives, as shown in Fig.7.1.

### 7.2.2 Indoor-Place and Ambient-Objects Detection

In this paper, a Residual deep neural network (ResNet) [155] is used to recognize the indoor places such as living-room, bed-room, kitchen, and office. The ResNet basic building block (residual block) concatenates 4 (3x3) convolution layers; each one consists of 64 feature maps of (7x7) pixels used. These feature

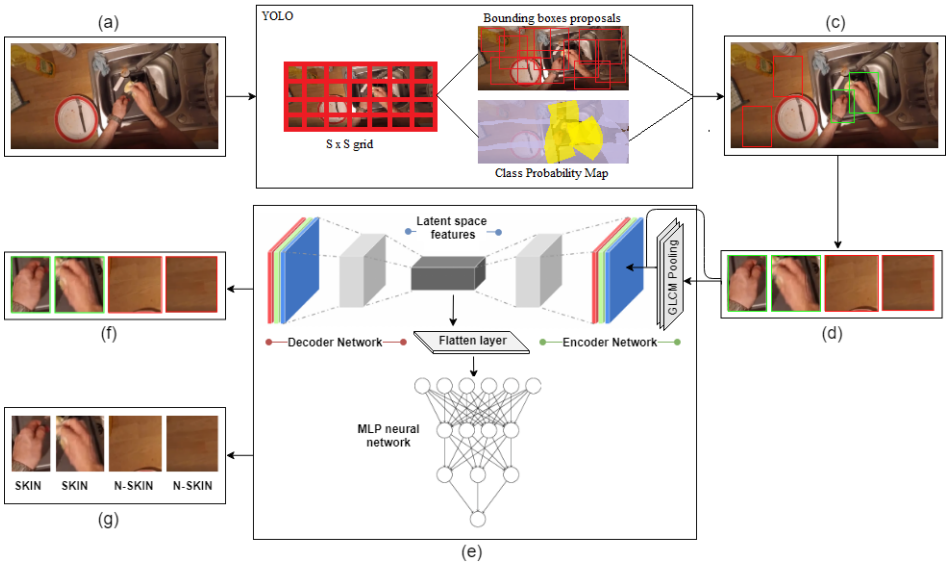



Figure 7.2 – GLCM-YOLO model: texture-based hand detection deep learning model. a) Input RGB video frame. b) YOLOv2 state-of-art object detection model for hand localization. c) Proposed regions of interest. d) Content of each bounding box. e) Stacked auto-encoder combined with MLP neural network for texture classification. f) Reconstructed image from the decoder network. g) Classification of bounding boxes labels

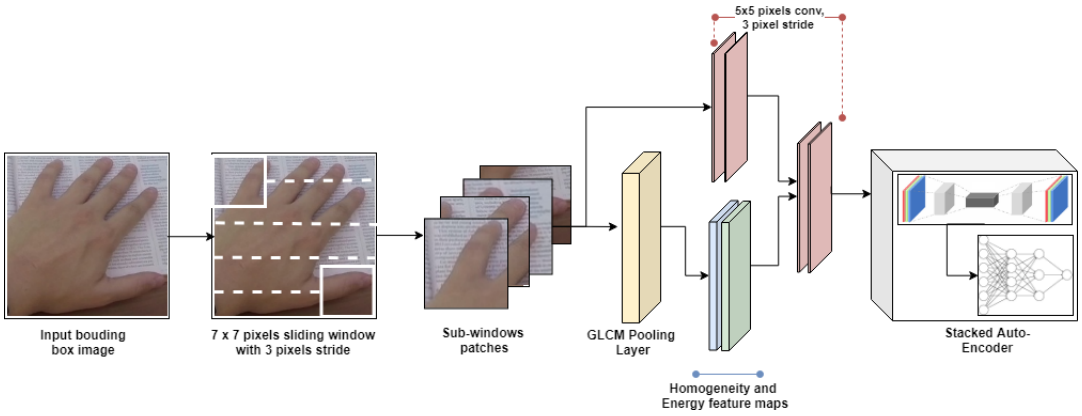



Figure 7.3 – GLCM skin-texture verification deep network architecture

maps are used as low-level feature extractors and are followed by (3x3) pixels max-pooling layer with a stride of 2 pixels to down-sample the number of the extracted features. The ResNet model concatenates 16 basic building blocks (residual blocks) to extract latent space features. To reduce the total number of learned parameters, a Global Average Pooling (GAP) layer [247] is used. Finally, a fully connected neural network with 2 hidden layers is used as a classification network. In this study, the ResNet model is pre-trained on the Place365 dataset [267] which consists of 2 million images of indoor and outdoor places. Finally, The resulting model is used to recognize the indoor places from both the human ego-centric and third-person vision perspectives, as shown in Fig.7.1.

To detect the ambient objects, YOLOv2 [261] deep neural network is used. The object detector is pre-trained on Microsoft COCO [268] (Common Objects in Context) dataset which consists of 2.5 million images of 80 different objects. Finally, the ambient objects detector is used to recognize objects populating the environment from both the human ego-centric and third-person vision perspectives, as shown in Fig.7.1.

$$Loss_{\text{YOLO}-\text{GLCM}} = Loss_{\text{YOLO}} + Loss_{\text{GLCM}} \tag{7.3}$$

$$
\begin{aligned}
\mathrm{L}oss_{\mathrm{YOLO}} =& \lambda_{coord} \sum_{i=0}^{\mathrm{S}^2} \sum_{j=0}^{\mathrm{B}} \mathbb{1}_{ij}^{obj} [(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2] \\
&+ \lambda_{coord} \sum_{i=0}^{\mathrm{S}^2} \sum_{j=0}^{\mathrm{B}} \mathbb{1}_{ij}^{obj} [(\sqrt{w_i} - \sqrt{\hat{w}_i})^2 \\
&+ (\sqrt{h_i} - \sqrt{\hat{h}_i})^2] \\
&+ \sum_{i=0}^{\mathrm{S}^2} \sum_{j=0}^{\mathrm{B}} \mathbb{1}_{ij}^{obj} (\mathrm{C}_i - \hat{\mathrm{C}}_i)^2 \\
&+ \lambda_{noobj} \sum_{i=0}^{\mathrm{S}^2} \sum_{j=0}^{\mathrm{B}} \mathbb{1}_{ij}^{noobj} (\mathrm{C}_i - \hat{\mathrm{C}}_i)^2 \\
&+ \sum_{i=0}^{\mathrm{S}^2} \mathbb{1}_{i}^{obj} \sum_{c \in classes} (p_i(c) - \hat{p}_i(c))^2
\end{aligned}
\tag{7.4}
$$

$$
\begin{aligned}
\mathrm{L}oss_{\mathrm{GLCM}} =& \sum_{i=0}^{\mathrm{B}} \mathbb{1}_{i}^{\mathrm{H}and} [-lbl_i \log(\hat{lbl}_i) \\
&+ \frac{1}{2m} \sum_{r=0}^{m} \sum_{c=0}^{m} (img_{[r,c]}^{i} - \hat{img}_{[r,c]}^{i})^2]
\end{aligned}
\tag{7.5}
$$

### 7.2.3 Multi-Perspectives Object-to-Object Mapping

In this thesis, a ResNet-Siamese deep network deep network is proposed to map and associate the ambient objects detected from both human ego-centric and robot vision perspectives, as shown in Fig.7.4. The ResNet-Siamese model allows measuring the visual similarity of the detected objects and associate similar ones. Basically, the Siamense network model was proposed in [269] for hand-written signature verification. The input module of the Siamense network consists of two identical neural networks to extract latent space features from the input image-pairs. Thereafter, the outputs of these two neural networks are combined to calculate the euclidean distance between the extracted

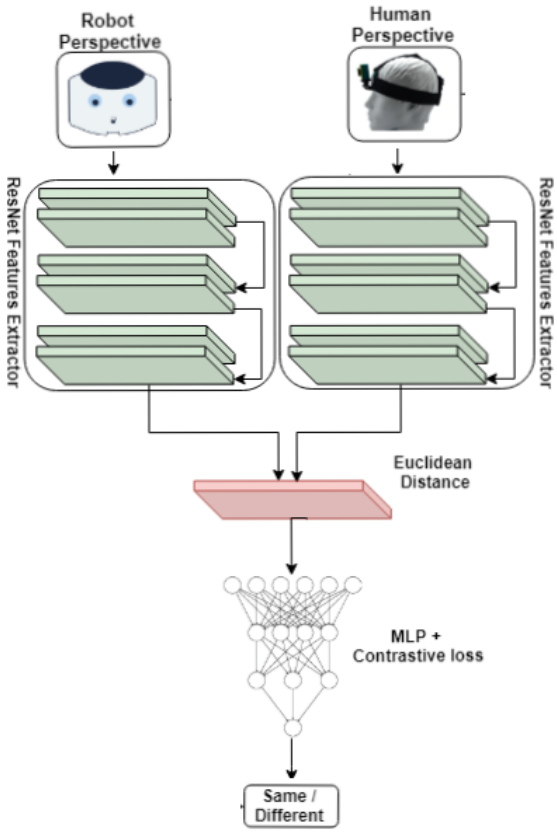



Figure 7.4 – ResNet-Siamense Network for Objects Similarity Measurement

feature vectors to measure the similarity of the input image-pairs. Compared to Siamense-CNN model [270], the proposed ResNet-Siamense Network (RSN) model integrates the residual shortcut connections within the Siamense CNN modules to extract more representative features, and solve the problem of the vanishing gradient encountered with deep networks [240].

The basic building block of the RSN model consists of a 2D convolution layer followed by Batch Normalization [239] and Rectified Linear Unit (R*e*LU) layers, which are used to prevent the vanishing gradient problem, accelerate the model convergence, and improve the learning generalization. The backbone unit of the RSN model consists of 3 basic building blocks with different

kernel sizes to extract 32, 64, and 128 feature maps respectively. To develop a deep feature extractor, multiple backbone units combined with residual shortcut connections are stacked together and cascaded with a Global Average Pooling layer to minimize the model over-fitting. During the model training, for each image patch with $n$ images, the euclidean distance $d$ between the features $(\mathrm{F}_1, \mathrm{F}_2)$ extracted from identical ResNet features extractors is calculated, see Eq.7.6. Finally, the contrastive loss function $\mathrm{L}_c$ [271] is used to learn the similarity distance between the input image-pairs in a way that similar features are becoming closer and non-neighbors are separated away, see Eq.7.7, where $y_i$ is the image target label and $r$ is the euclidean distance threshold.

$$d(\mathrm{F}_1, \mathrm{F}_2) = \sum_{f=0}^{n_f} \|\mathrm{F1}_i - \mathrm{F2}_i\| \tag{7.6}$$

$$\mathrm{L}_c = \frac{1}{2\mathrm{N}} \sum_{i=0}^{\mathrm{N}} (y_i.d^2 + (1-y).max(r - d(\mathrm{F}_1, \mathrm{F}_2), 0)^2) \tag{7.7}$$

ResNet-Siamese model is pre-trained on Core50 [272] (Common Objects in Context) dataset which consists of 50 different objects belong to 10 categories: plug adapters, mobile phones, scissors, light bulbs, cans, glasses, balls, markers, cups and remote controls. Finally, the obtained model is used to map and associate the detected objects from human ego-centric and third-person vision perspectives, as shown in Fig.7.1.

### 7.2.4 Inference of Ambient-Objects Contextual Affordances

For a human being, the contextual visual reasoning is one of the most important cognitive skills that give the ability to infer, from a little sneak peek, the objects names, colors, textures, and what actions (activities) may the objects afford. The notion of object affordances was defined in [273] as the **object**

**properties** that determine what **actions** a human can perform on them. In this thesis, the notion of the "contextual object affordances" is defined as the set of **"activities"** that an **"object"** afford in a given **"place"**. In other words, the same object may afford different actions (activities) in different locations such as a cup may afford **"drink"** in the **("kitchen, office, living-room")**, while the same cup may afford **"hold-toothbrushes"** in the **("bath-room")**. From a pure data-driven perspective, it is possible to train a deep network to detect the ambient objects and recognize their contextual affordances. However, to the best of our knowledge, there are no datasets available for recognizing the objects contextual affordances. Consequently, an inference engine is proposed to detect the contextual affordances of the ambient objects. The inference engine exploits the detected indoor places, ambient objects, and a ConceptNet-based ontology to infer the possible activities that may be afforded by the recognized objects in the current place, as shown in Fig.7.1.

#### ConceptNet-based Ontology for Contextual Object Affordances

To endow the ability of decision-making in partially observed environments to the AAL system is one of the most important challenges in the field of artificial intelligence [152]. This capability is achievable by exploiting an abstracted semantic representation of the environments, which increases the adaptability and flexibility of the approach. To represent the semantics of the environments, a comprehensive knowledge-base is required, by carefully analyzing the domain of the environment, choosing the proper vocabularies, and integrating a reasoning engine to infer the best decision to make. In this chapter, the knowledge representation and reasoning play a key role in processing the partially observable information by combining the knowledge-base with the ongoing perception of the ambient environment. In this perspective, to infer the

contextual affordances of the ambient objects, a ConceptNet-based ontology combined with a probabilistic Description Logic (DL) reasoning model is proposed.

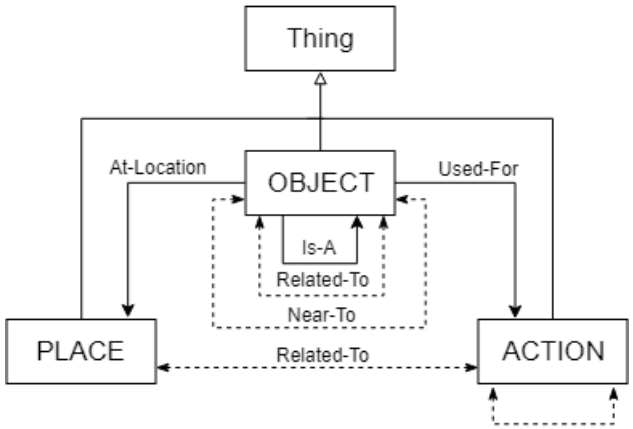



Figure 7.5 – ConceptNet-based ontology for contextual affordances representation

As shown in Fig.7.5, the proposed ontology consists of three main concepts: *OBJECT*, *PLACE*, and *ACTION*. OBJECT instances represent all of the ambient-objects populating daily living environments. PLACE instances represent the indoor places where the objects can be found. ACTION is a reified action concept whose it's instances represent the actions (activities) afforded by an object. The reification is exploited to represent the contextual affordances as first-order relations without resorting to meta-modeling subterfuges. The general schema of the Description Logic (DL) subsumption rule to infer the contextual affordances is formulated as follows:

$$\text{OBJ} \sqcap \exists \text{AtLocation.P} \sqsubseteq \exists \text{UsedFor.}\{\text{ACT}\}$$

where OBJ is a subclass of the *Object*, P is a subclass of *Place* and ACT is an instance of *Action*. In other words, as shown in Fig.4.1, if an object instance

(laptop) of a specific class (personal computer) is placed at the right place (office), it will afford a given action (work).

**Probabilistic Description Logic (DL) inference engine**

In this study, a probabilistic DL reasoner based on the probabilistic semantics for DL DISPONTE ("DIstribution Semantics for Probabilistic ONTologiEs") [274] is used to enrich the DL axioms with the probabilistic information produced by both ambient-object and indoor place detection models. In DISPONTE, the DL axioms can be annotated with the corresponding probability, to create a probabilistic knowledge-base (KB). To calculate the probability of a given query from the DISPONTE KBs, the BUNDLE (Binary decision diagrams for Uncertain reasoNing on Description Logic thEories) algorithm is used [274]. The BUNDLE algorithm exploits the DISPONTE reasoner (Pellet[1]) to infer a set of *worlds* (explanations) of a given query, where the probability of each world $w$ equals the joint probability distribution over the selected and non-selected axioms. The *worlds* (explanations) are encoded into a Binary Decision Diagram (BBD), from which the probability of the query is computed as a marginalized joint probability over the *worlds* that imply the query. To calculate the ultimate probabilities of the contextual affordances of the ambient objects in a given place, for each input video frame, the YOLOv2 object detection model produces the output vector $\text{OBJ}s = [p_1 : \text{O}_{\text{C}_1}, ..., p_n : \text{O}_{\text{C}_n}]$, where $\text{O}_{\text{C}_i}$ represents the recognized class of the $i$-th object with probability $p_i$. For the same video frame, the ResNet place detection model produces the output vector $\text{PLACE}s = [p_1 : \text{L}_{\text{C}_1}, ..., p_n : \text{L}_{\text{C}_m}]$, where PLACE$s$ is a list of the possible classes $\text{L}_{\text{C}_i}$ of the current place scene with probabilities $p_i$. To create the probabilistic KB, the DISPONTE reasoner encodes the output vectors as a set of in-

[1] https://github.com/stardog-union/pellet

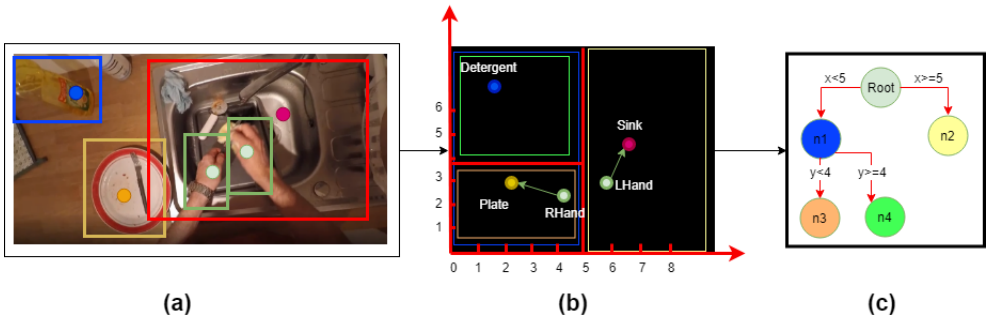



Figure 7.6 – (a) 2D locations of hands and the ambient objects; (b) 2D graph representation space of the observations; (c) Constructed search tree from the current video frame.

stances of the ontology associated with their probabilities. A place instance $place$ is created in the KB for each element $p_i : \mathrm{L}_{C_i}$ in PLACE$s$, and an axiom $p_i :: place : \mathrm{L}_{C_i}$ is added to the ontology. Similarly, an object instance $obj_i$ is created in the KB for each element $p_i : \mathrm{O}_{C_i}$ in OBJ$s$, and an axiom $p_i :: obj_i : \mathrm{O}_{C_i}$ is added to the KB. Finally, all of the object instances $obj_i$ are associated to a $place$ using *At-Location* relation. Finally, BUNDLE algorithm exploits the latter KB to infer the probabilistic contextual affordance relationships between the object instances $obj_i$ and the detected activities.

One of the most important advantages of the proposed approach is that all of the ontology axioms even the non-probabilistic ones are taken into consideration during the reasoning process. Besides, the number of inference rules is drastically reduced by aggregating the classes of the ambient-objects under a set of common super-classes to cover all the possible contextual affordances.

### 7.2.5 Human Action Recognition/Anticipation

In this study, we assume that many of human short-term actions (e.g.drink, eat) can be recognized independently of their durations by detecting the objects (Objects-in-Action) actively involved (e.g. cup, food) in the action. Conse-

quently, the Qualitative Representation (QR) of hand locations to the ambient-objects is required to identify the objects of interest and therefore anticipate the intended activity. Besides, the QR of hand locations with respect to the ambient-objects allows abstracting complex activities (e.g. working, eating, cooking) which often occur over different time durations.

In this study, The K-d tree nearest neighbor algorithm [275] is used to detect the 2D coordination of the nearest object relative to hand. For each video frame, both of Yolov2-GLCM hand detector and YOLOv2 [261] object detector will estimate the 2D locations of hands and the ambient objects to construct the search space. Finally, the nearest objects relative to hands are detected using the K-d tree search algorithm, as shown in Fig.7.6.

Finally, the probabilistic reasoner exploits the contextual affordances of the Objects-in-Action to recognize and anticipate human activities, as shown in Fig7.1.

## 7.3 Experiments

### 7.3.1 Experimental setup

The experimental environment consists of human egocentric vision RGB wearable camera, RGB robot (third-person) integrated camera, 4 wearable IMU sensors, and activity dashboard screen located see Fig.7.7. The video streams from both the wearable camera and the robot camera are used to anticipate human activities, while the IMU streams are considered for future experiments.

The proposed approach for human activity recognition and anticipation is evaluated in two phases: 1) the main components of the proposed approach are evaluated independently with different datasets; 2) a use-case of cognitive daily activities coaching for a dependent person is then studied to validate the

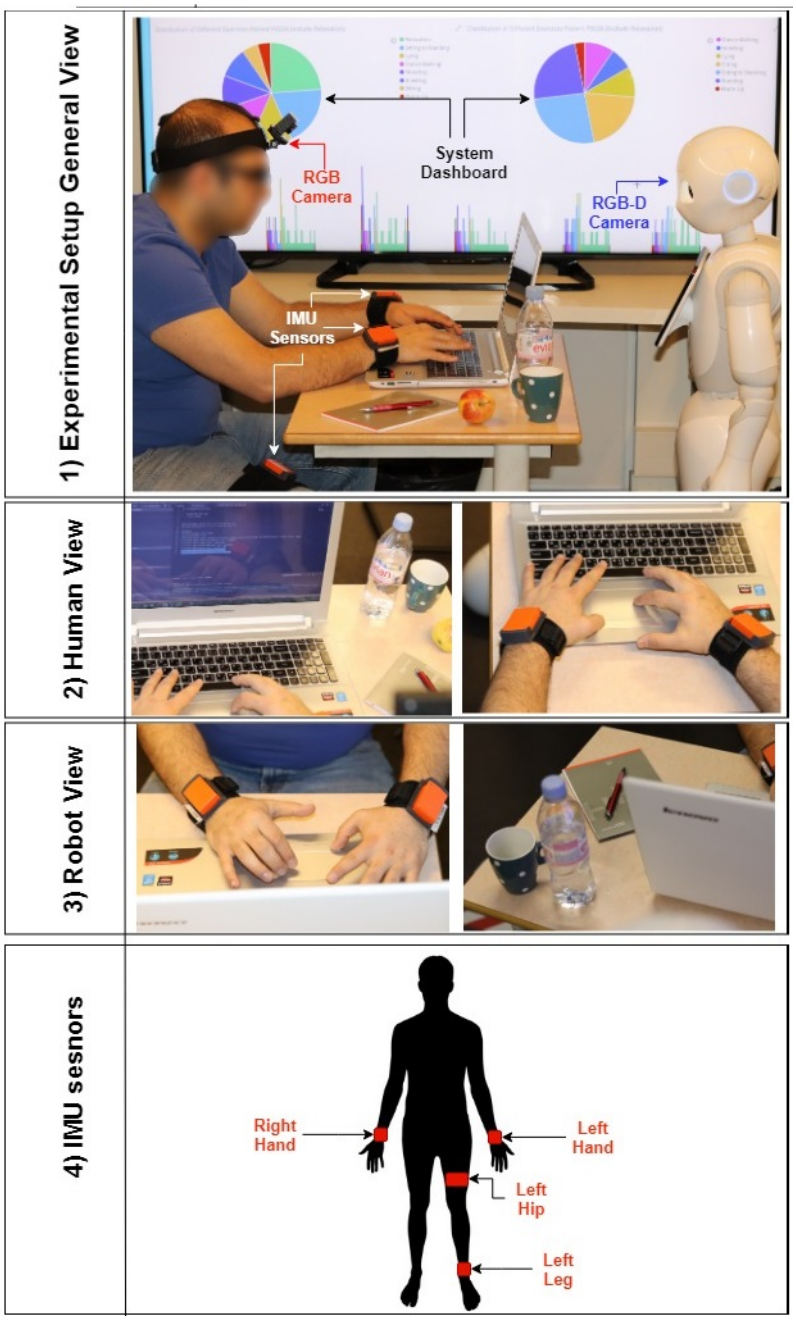



Figure 7.7 – Experimental setup of human daily activities dataset. 1) General view of the experimental environment with IMU sensors, RGB egocentric vision, and RGB-D robot camera and activity dashboard. 2) Human egocentric vision perspective. 3) Third-person vision perspective. 4) The IMU sensors location on human body

entire approach for real-time action recognition/anticipation. Finally, for each subject, satisfaction feedback is collected to evaluate the user experience after the experiments.

### 7.3.2 Datasets

Empirical experiments on real-world datasets (Microsoft COCO [276], Place365 [267], DHA-11$_{Th}$, SKNS, and CORe50 [272]) were carried out to evaluate the different components of the proposed approach, and more precisely, object detection, place detection, hand detection, skin texture detection, and object-object mapping models.

#### Microsoft COCO dataset

To evaluate the object detection model, Microsoft COCO [268] (Common Objects in Context) dataset is used. This dataset consists of 2.5 million annotated instances of 91 different objects, which are split into training-set, validation-set, and testing-set of 118K, 5K, and 41K images, respectively. In this study, the commonly used objects (keyboard, mouse, cup, microwave, etc) in daily activities are selected to evaluate the object detection model.

#### Place365 dataset

Place365 dataset includes 2 million annotated images of 365 different indoor and outdoor places, which is split into training-set, validation-set, and testing-set of 1M, 36K, and 300K images, respectively. In this study, indoor places (living-room, bed-room, kitchen, etc) are selected to evaluate the place detection model.

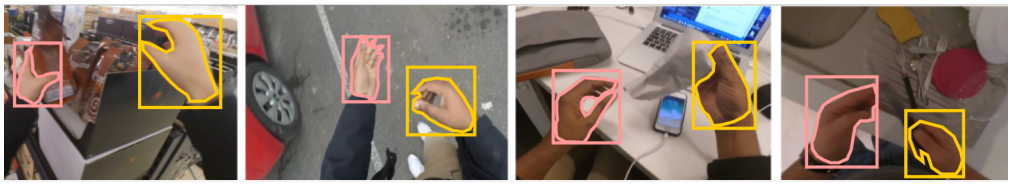


Figure 7.8 – Sample images of indoor and outdoor places extracted from DHA-$11_{\mathrm{T}h}$ dataset with hand bounding boxes and segmentation mask annotations

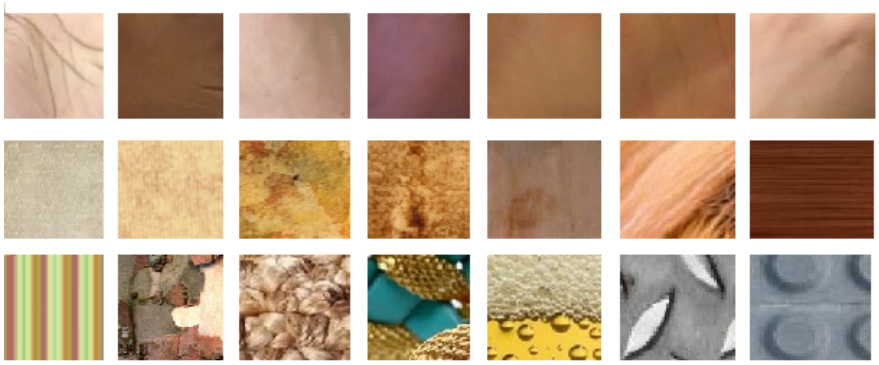


Figure 7.9 – Samples from different textures existing in SKNS textures dataset. a) The first row consists of different skin textures; b) The second row consists of non-skin textures similar to skin colors; c) The third row consists of diverse non-skin texture

**Diverse Hand Dataset**

To evaluate the hand detection model performance, a new Diverse Hands (DHA-$11_{Th}$) dataset is collected. As shown in Table.7.1, compared to the state-of-the-art datasets such as EgoHands [265], and Viva [266] datasets, DHA-$11_{Th}$ dataset consists of 11,320 egocentric vision images with more than 23,000 ground-truth labeled hands, of 6 different subjects (3 males, 3 females) aged between 21 and 30 years old of different skin colors. To create a realistic dataset, the latter is recorded in different lighting conditions, and different (indoor-outdoor) places such as supermarkets, kitchens, offices, and streets, as shown in Fig.7.8. In addition, the dataset is recorded on different days to add remarkable varieties for the recording conditions such as the subject clothes, the arrangement of the ambient-objects, and to add different noise sources as much as possible. The dataset consists of hypothetically unlimited hand positions since the subjects were freely engaged to do unscripted activities with different objects. To build the dataset, as shown in Fig. 7.7, an egocentric camera is used, which records videos of 1920x1080 pixels at 30 frames per second (fps). In the post-processing phase, the produced videos are exported to images with the Joint Photographic Experts Group (JPEG) standard format at a sampling rate of 10 fps. For each image, a set of manually annotated bounding boxes and pixel-level segmentation masks are created for each hand present in the image, as shown in Fig. 7.8. For each hand, a set of labels are created as a combination of (left-hand, right-hand), (Male, Female), (age), (Dark, Medium, Fair). To the best of our knowledge, DHA-$11_{Th}$ is the largest multi-purpose hand dataset recorded from the egocentric vision perspective. The dataset can be used in different research areas: semantic hand detection and segmentation, gender recognition, and age estimation. In this study, the dataset is randomly split into 80% for training, 20% for validation, and 20% for testing since the ratio of the activities and lo-

cations is equally distributed between the different sets. The comprehensive statistical analysis for the dataset is shown in Fig.7.10.

Table 7.1 – State of the art datasets comparison

| | Dataset | | |
|---|---|---|---|
| | EgoHands [265] | Viva [266] | DHA-11$_{Th}$ [Ours] |
| Total Images | 4,800 | 11,000 | **11,320** |
| Total Labels | 15,000 | 21,199 | **23,558** |
| Number of Subjects | 4 | - | **6** |
| Semantic Annotation | **Yes** | **Yes** | **Yes** |
| Segmentation Map | **Yes** | No | **Yes** |
| Age | No | No | **Yes** |
| Gender | No | No | **Yes** |
| Skin Color | No | No | **Yes** |
| Place | Conference Room<br>Outdoor Courtyard<br>Coffee Table in Home | Car | **Kitchen**<br>**Street**<br>**Office**<br>**Supermarket** |
| Activities | Playing Cards<br>Playing Chess<br>Solving Jigsaw Puzzle | Driving Car | **Cooking**<br>**Make Coffee**<br>**Read Newspaper**<br>**Use Laptop**<br>**Use Mobile Phone**<br>**Writing**<br>**Inspect Products**<br>**Random Gesture** |
| Resolution | 720 x 1280 | 640 x 480 | **1920 x 1080** |

**Skin/None-Skin Texture Dataset**

To evaluate the performance of the texture detection model, a new skin/non-skin dataset (SKNS) is collected. This dataset consists of approximately 800,000 skin and non-skin manually annotated texture images of 128x128 pixels (337,700 skin images and 462,200 non-skin images). To ensure the variation and the diversity of the dataset images, the skin texture images are collected by generating sub-patches from different hands dataset such as DHA-11$_{Th}$, EgoHands [265], and 11K hands [277] datasets. The data were collected from 201 different sub-

jects aged from 18 to 75, of different skin colors. The non-skin images are collected by generating sub-patches of images of material textures from different datasets such as Describable Textures [278], and ImageNet [279]. As shown in Fig.7.9, The non-skin patches consist of skin-like textures such as wood, sand, fire, and fur. The dataset is stratified splits into 80% for training, 20% for validation, and 20% for testing.

To the best of our knowledge, the SKNS dataset is the largest skin/non-skin texture dataset currently available for research; it combines 50 different textures labeled into skin and non-skin labels. In addition, SKNS dataset contains varied races and skin colors to be diverse enough to detect skin-textures pixels.

**CORe50 Dataset**

To evaluate the multi-perspectives object-to-object mapping model, CORe50 [272] dataset is used. This dataset, originally designed for continuous ambient-objects recognition, consists of images of 50 different objects belonging to 10 categories (plug adapters, mobile phones, scissors, light bulbs, cans, glasses, balls, markers, cups and remote controls). The dataset is collected in 11 different recording sessions (8 indoor, 3 outdoor) using different backgrounds and different lighting conditions. For each session and for each object, a 15 seconds egocentric vision RGB-D video has been recorded with a Kinect 2.0 sensor delivering 300 RGB-D frames. The full dataset consists of 164,866 images. To build the Core50 dataset, the objects were differently manipulated by moving and rotating them randomly using both hands. In this study, the RGB images are split into training, validation, and testing sets of 131892, 16486 and 16486 images respectively.

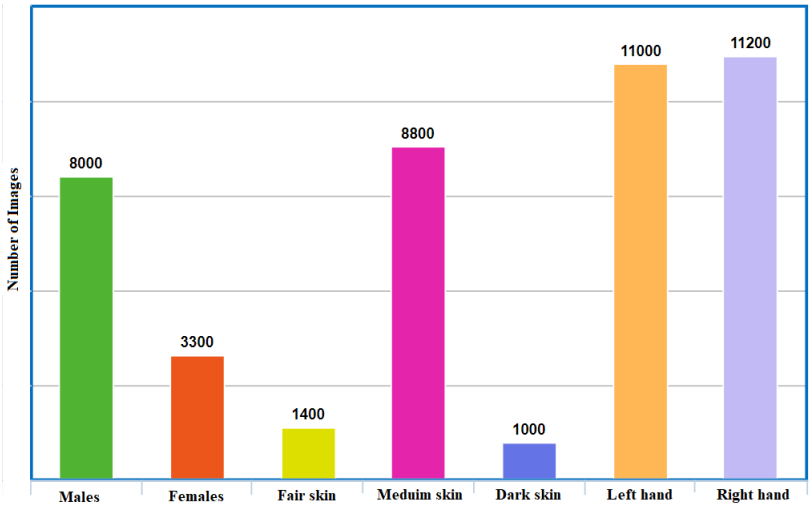



Figure 7.10 – DHA-11$_{Th}$ dataset images distribution over different categories

### 7.3.3 Implementation

The proposed approach is implemented in python 3.7 based on keras deep learning framework [2] with tensorflow as a backend. The communication between the individual modules is implemented based on MQTT [3] lightweight publish/subscribe messaging protocol. At the low-level layer, the deep learning models were run on an Intel(R) Core(TM) i7-6820HQ CPU @ 2.70GHz (8 CPUs), with 64Gb RAM combined with Nvidia M2000M GPU. At the high-level layer, the reasoner was run on an Intel(R) Xeon(R) CPU E5-1630 v4 @ 3.70GHz, 4 Core(s), 8 Logical Processor(s), with 8Gb RAM.

[2] https://keras.io/
[3] http://mqtt.org/

## 7.4 Evaluation

### 7.4.1 Place detection: ResNet model

To evaluate the performance of the place detection ResNet model, each of AlexNet [280], GoogleNet [281], and VGG [282] models were selected as baseline models. For the training phase, The indoor places images were selected from the training set of the Place365 dataset. The Top-1 and Top-5 accuracy metrics were used for the performance evaluation. [267]. The performances of the proposed model and baselines models are shown in Table 7.2. One can clearly observe that the ResNet model outperforms all baseline models in terms of the Top-5 classification accuracy, while the VGG model outperforms the proposed model in terms of the Top-1 accuracy.

With regards to the validation and testing sets, the ResNet model which has residual shortcut connections between the convolution layers allows extracting more generalized features than the plain CNN baseline models and therefore leading to more accurate results in terms of Top-5 accuracy. On the contrary, the plain VGG CNN model focuses on extracting the local features for each place, which leads to getting the Top-1 accuracy.

Table 7.2 – Evaluation results of Place365 dataset

| | Accuracy % | | | |
|---|---|---|---|---|
| | Testing-set | | Validation-set | |
| Model | Top-1 | Top-5 | Top-1 | Top-5 |
| Baseline Models | | | | |
| AlexNet | 53.3 | 82.8 | 53.2 | 82.9 |
| GoogleNet | 53.6 | 84.1 | 53.6 | 83.9 |
| VGG | **55.8** | 85.4 | **55.6** | 85.1 |
| Proposed Model | | | | |
| ResNet [Ours] | 55.2 | **87.4** | 54.5 | **86.7** |

### 7.4.2 Hand detection: YOLOv2-GLCM Auto-Encoder model

The evaluation of the hand detector model is carried out on two phases: 1) the evaluation of the GLCM-CNN Auto-Encoder for skin texture classification using SKNS dataset; 2) the evaluation of the entire YOLOv2-GLCM Auto-Encoder model using EgoHands [265] and DHA-$11_{Th}$ datasets.

To evaluate the GLCM-CNN Auto-Encoder for skin texture classification, each of plain CNN, GLCM features combined with CNN, and stacked deep autoencoder [283] were selected as baseline models. All models are trained and evaluated using SKNS dataset. As shown in Table 7.3, the proposed GLCM-CNN Auto-Encoder model achieves the highest accuracy outperforming the baseline models. Besides, the obtained results show that the integration of the GLCM features with the automatically extracted latent features improves the performance of the stacked autoencoder model in terms of accuracy. Accuracy improvements of 11.8% and 15.5% are achieved for deep autoencoder [283] and convolution neural network baseline models respectively.

To evaluate the YOLOv2-GLCM Auto-Encoder model, each of YOLOv2 [261], RCNN [262], and Faster-RCNN [263] were selected as baseline models. Besides, all models are trained and evaluated using EgoHands [265] and DHA-$11_{Th}$ datasets. The results shown in Table 7.4 demonstrates clearly that the proposed model significantly outperforms the baseline models and provides the best performance for both EgoHands and DHA-$11_{Th}$ datasets in terms of Average Precision (AP). Compared to the Faster-RCNN combined with texture features, the proposed model achieves a better AP score with an improvment ($\Delta$AP) of 6.7% and 13.1% for EgoHands and DHA-$11_{Th}$ datasets respectively. However, Table.7.4 shows that the proposed model AP performance decreases by almost 47% in the case of using DHA-11Th dataset in comparison with the EgoHands dataset. This difference can be explained by the high complexity

of the proposed DHA-$11_{Th}$ dataset, which contains more places with skin-like textures and cluttered backgrounds compared to EgoHands dataset. Finally, as shown in Table.7.4, a substantive improvement is observed by integrating the skin-texture model with the baseline models. In addition, AP improvements over baselines models of 48.6% and 15.1% are obtained for EgoHands and DHA-$11_{Th}$ datasets, respectively.

Table 7.3 – Accuracy (%) results of skin texture classification models

| | Accuracy (%) |
|---|---|
| Plain CNN | 69.5 |
| GLCM features + CNN | 73.7 |
| Deep AutoEncoder [283] | 77.4 |
| GLCM-CNN Auto-Encoder [Ours] | **89.2** |

Table 7.4 – Average Precision (%) results of hand detection models

| | Datasets | |
|---|---|---|
| Hand Detection Models | EgoHands | DHA-$11_{Th}$ |
| RCNN[262] | 31.2 | 25.3 |
| RCNN + Texture | 82.6 | 35.8 |
| Faster-RCNN[263] | 33.1 | 24.4 |
| Faster-RCNN + Texture | 86.4 | 32.6 |
| YOLOv2[261] | 44.5 | 30.6 |
| YOLOv2 + Texture [Ours] | **93.1** | **45.7** |

### 7.4.3 Multi-Perspective Object-Object Mapping

To evaluate the ResNet-Siamese model for multi-perspectives object-to-object mapping, Core50 dataset is used. In this paper, for the positive patches, each of the training, validation, and testing sets are created by constructing image-pairs of the same object with different angles. To improve the performance of the ResNet-Siamese model, a set of negative patches are created by constructing random image-pairs from different objects. As shown in Table 7.5, the

ResNet-Siamese model outperforms the CNN-based Siamese baseline model in terms of the area under ROC curve. The results show that the proposed ResNet feature extractor network allows extracting more relevant and representative features than the classical CNN network used in the baseline model, which leads to a remarkable improvement ($\Delta_{AUC}$ = 0.06) compared to the classical Siamese architecture.

Table 7.5 – Evaluation results of Multi-Perspectives Object-Object Matching Model

| Models | Area Under ROC Curve |
|---|---|
| Siamese [270] | 0.87 |
| ResNet-Siamese [Ours] | **0.93** |

### 7.4.4 Entire approach Evaluation

#### Real-World Scenario

To validate the proposed approach, a blind person assistance scenario is implemented. In this scenario, a blind person named B$ob$ is wearing a head-fixed camera and assisted by the robot companion P$epper$ in different homerooms such as office, kitchen, and living room. B$ob$ can initiate the discussion by asking P$epper$ to describe him the ambient-environment according to his hand location. The approach is used to anticipate B$ob$ intentions automatically to help him to locate the target objects correctly and informing him about the location of the related objects or the missing objects needed to complete the ongoing action. For example, if B$ob$ is targeting a S$poon$ and P$epper$ infers an E$at$ action then Pepper will inform B$ob$ about the location of the related items such as P$late$ in the scene. If there are no related items found, P$epper$ will inform B$ob$ that E$at$ action cannot be completed. The proposed scenario succeeded to validate all of the components of the proposed approach. Besides,

the scenario shows the best use of these components to make the Human-Robot interaction in daily living more symbiotic, effective, and natural. This work is reported in a multimedia video that is available on *LISSI* Website [4].

**Quantitative and Qualitative Evaluation**

To evaluate the performance of the proposed approach, three evaluation metrics are used: the proposed ontology size, the execution time of the different components, and the qualitative satisfaction feedbacks collected to evaluate the user experience after the experiments.

Table 7.6 – Response Time of iCare different Modules

| | Execution Time (ms) | | |
|---|---|---|---|
| | Minimum | Maximum | Average |
| Object Detection | 250 | 420 | 360 |
| Place Detection | 520 | 610 | 570 |
| Hand Detection | 450 | 530 | 494 |
| Object-Object Mapping | 120 | 160 | 145 |
| K-d tree | 10 | 25 | 17 |
| Affordances Inference | 150 | 520 | 354 |

- **Ontology size:** The proposed ontology consists of more than 400,000 concepts covering commonsense knowledge of the ambient environment, objects, contextual affordances, and places.

- **Runtime:** As shown in Table.7.6, for each video frame, the average execution time to recognize the indoor place, location of human hands, and all the ambient-objects is 570, 494, and 360 milliseconds, respectively. In addition, the average time consumed to map each object from human-robot perspectives, and to calculate the nearest objects to the human

[4] http://www.lissi.fr/videos/Anticipation.php

hands is 145 and 17 milliseconds, respectively. Finally, the probabilistic reasoner generates approximately 11 axioms, and the average reasoning time for inferring the contextual affordances is approximately 354 milliseconds. For each object in the frame, the probabilistic reasoner infers approximately 3 contextual affordances on average.

- **Feedback:** Based on the satisfaction feedback collected after each experiment, the results show that 83% of subjects are satisfied with the overall performance of the proposed approach in terms of cognitive assistance. In terms of Human-Robot interaction response time, 75% of them are satisfied, and the experiments were described as a truly enjoyable experience.

## 7.5 Conclusion

A hybrid approach combining deep learning models and higher-level reasoning for activities recognition and intention anticipation is proposed. In one hand, it combines different deep-learning models to detect ambient objects, indoor places, and human hands locations. Besides, the human and robot vision perspectives are aggregated using a novel deep learning model for similarity measurement combined with k-d tree search technique. In the high-level layer, a novel ontology-based on ConceptNet knowledge graph is combined with a probabilistic logic inference engine to infer human activities and intentions. Finally, the reasonable execution time of the different components allows the proposed approach to recognize and anticipate human activities in real-time. To evaluate the performance of the proposed approach, three evaluation metrics are used: The proposed ontology size, the execution time of the different components, and qualitative satisfaction feedbacks are collected to

evaluate the user experience after the experiments. Based on the satisfaction feedback collected after each experiment, the results show that 83% of subjects are satisfied with the overall performance of the proposed approach in terms of cognitive assistance.

Chapter 8

# General conclusion and perspectives

## 8.1 General conclusion

The objective of this thesis is to propose novel hybrid approaches that enable the AAL system to detect the emotional states, actions, and intentions of users, taking advantage of their context and the benefits of combining data-driven and knowledge-based techniques. The contributions of the this thesis can be summarized as follows:

- A hybrid contextual emotion recognition approach is proposed for cognitive assistance services in ubiquitous environments. Its principle consists of, on the one hand, the multimodal emotion recognition based on hybrid-level fusion exploiting a multilayer perceptron neural network model and the possibilistic logic, and on the other hand, the expressive NKRL knowledge representation and reasoning exploiting both HClass and HTemp ontologies for representing both commonsense knowledge and the ambient environment dynamics, respectively. The performance of multimodal emotion recognition based on hybrid-level fusion was enhanced by including the selected features: age, gender, and culture. The pro-

posed approach is able to recognize perfectly the 9 observed emotions considered in this study: happy, sad, anger, neutral, energy, contempt, disgust, fear, and surprise. The proposed emotion contextual recognition and management approach is able to recognize the non-directly observable emotions by contextualizing the observed emotions.

- A hybrid approach for human activity-aware AAL system is proposed. A combaniation of Spatio-Temporal Joint based Convolutional Neural Network (STJ-CNN) and Hierarchical Multichannel deep Residual Network (HMResNet) is proposed to recognize human activities from both skeleton keypoints and multichannel IMUs's raw data. Besides, a novel representation and inference based on NKRL HClass and HTemp ontologies are proposed to represent and combine the detected human activities with the ambient events, and infer the semantic context of the detected activity. The proposed approach combines both the IMUs-based and the skeleton-based activity recognition to overcome the misclassification error caused by sensor instability, visual occlusions, and visual perspective changes. Compared to the baseline models, the performance of the STJ-CNN model shows a significant improvement up to 24% in terms of F-score on DAHLIA dataset. The performance of the daily human activity recognition based on HMResNet model is shown through two activity datasets.

- A hybrid approach combining deep learning models and higher-level reasoning for activities recognition and intention anticipation is proposed. In one hand, it combines different deep-learning models to detect ambient objects, indoor places, and human hands locations. Besides, the human and robot vision perspectives are aggregated using a novel deep

learning model for similarity measurement combined with k-d tree search technique. In the high-level layer, a novel ontology-based on ConceptNet knowledge graph is combined with a probabilistic logic inference engine to infer human activities and intentions. Finally, the reasonable execution time of the different components allows the proposed approach to recognize and anticipate human activities in real-time. To evaluate the performance of the proposed approach, three evaluation metrics are used: The proposed ontology size, the execution time of the different components, and qualitative satisfaction feedbacks are collected to evaluate the user experience after the experiments. Based on the satisfaction feedback collected after each experiment, the results show that 83% of subjects are satisfied with the overall performance of the proposed approach in terms of cognitive assistance.

## 8.2 Perspectives

The following perspectives can be summarized as a result of this thesis:

- **Multimodal human emotion estimation:** emotion recognition remains a challenge for AAL systems at the moment due to the constraints of obtaining reliable results in order to provide a trustworthy interaction, as well as the time constraints associated with integrating the recognized emotion into the AAL system behavior adaptation. As demonstrated in this thesis, human emotion recognition is feasible through the collection of various types of data (video, sound, text). While some modalities have been extensively investigated as human facial expressions, others should be deeply studied as Electroencephalography (EEG), Facial Action Units (FAU), and 3D facial key-points. All of the those modalities, rep-

resent promising information sources for future developments: innovative and accessible technologies such as depth cameras, consumer-grade EEG, and smart devices, combined with deep-learning models, will accelerate the development of emotion-aware AAL systems. Besides, many of the available datasets used were collected from streaming plateform as YouTube or derived from general HMI research. Therefore, they are unsuitable for emotion recognition in real-world situations. To tackle this problem, real-world multi-modal datasets should be collected.

- **User-centric human activity recognition:** One of the key features of human activities is how the users are executing the activity. In fact the same activity may have different patterns which highly depend on the user style of executing the activity. Furthermore, activity patterns are changing over time; thus, the assumption that activity patterns of the same user remain constant over time is questionable. Additionally, users may engage in new activities, which were not included in the trained data. To address these issues, an innovative approach should be used such as: active learning or transfer learning techniques to enable the machine learning models to learn new activities without forgetting the old ones. The active learning approach is a subset of the incremental learning that enables selecting new ground truth (new activity labels) for newly selected data samples in case of data modification. As a result, active learning algorithms can distinguish between time-varying activity patterns. The transfer learning approach is a technique that enables knowledge transfer from one domain to another based on the assumption of a relationship between the source and target domains. As a result of transfer learning approaches, the approach is able to distinguish the activities of different groups of users.

- **Human intentions datasets:** By reviewing the literature, we find that there are few number human intention datasets available for recognizing short -term and long-term human intentions such as the DAHLIA dataset. Besides, the collected datasets are not realistic, for example, the user performs the identical actions/intentions everyday, which is not realistic. The process of collecting dataset, particularly for estimating human intentions, is time consuming and the annotation cost is high. To tackle those issues, we can collect more realistic datasets and exploit the crowdsource data labeling techniques, in particular the pair-wise comparison, to generate the groundtruth of the human intentions from large population of annotators.

- **Physical intentions vs social intentions:** In this thesis, human intentions are limited to physical intentions and visible behaviors, but many people already have a web presence and they are showing a social networking behavior that is becoming increasingly essential to estimate their intentions. As a result, we believe that we should broaden the proposed definition of contextual human actions to include different settings, such as the social-networking activities, and the sentiment analysis of comments.

# Bibliography

[1] A. Chibani, Y. Amirat, S. Mohammed, E. Matson, N. Hagita, and M. Barreto, "Ubiquitous robotics: Recent challenges and future trends," *Robotics and Autonomous Systems*, vol. 61, no. 11, pp. 1162 – 1172, 2013.

[2] G. D. Abowd, A. K. Dey, P. J. Brown, N. Davies, M. Smith, and P. Steggles, "Towards a better understanding of context and context-awareness," in *International symposium on handheld and ubiquitous computing*. Springer, 1999, pp. 304–307.

[3] K. Ducatel, U. européenne. Technologies de la société de l'information, U. européenne. Institut d'études de prospectives technologiques, and U. européenne. Société de l'information conviviale, "Scenarios for ambient intelligence in 2010," 2001.

[4] P. Reignier, "Ambient intelligence pro-active: Sp 'e cification 'a impl é mentation," Ph.D. dissertation, 2010.

[5] T. Deyle, H. Nguyen, M. Reynolds, and C. Kemp, "Rfid-guided robots for pervasive automation," *IEEE Pervasive Computing*, vol. 9, no. 2, pp. 37–45, 2010.

[6] A. A. Khaliq and A. Saffiotti, "Stigmergy at work: Planning and navigation for a service robot on an rfid floor," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 1085–1092.

[7] J. El Sayah, "Contribution à la modélisation, à la simulation et à l'évaluation d'applications nomades à intelligence répartie : application à l'assistance aux voyageurs aveugles dans les transports publics et les pôles d'échanges," Ph.D. dissertation, 2009, thèse de doctorat dirigée par Baudoin, Geneviève Electronique, optronique et systèmes Paris Est 2009. [Online]. Available: http://www.theses.fr/2009PEST1032

[8] R. Want, A. Hopper, V. Falcao, and J. Gibbons, "The active badge location system," *ACM Transactions on Information Systems (TOIS)*, vol. 10, no. 1, pp. 91–102, 1992.

[9] A. Harter, A. Hopper, P. Steggles, A. Ward, and P. Webster, "The anatomy of a context-aware application," *Wireless Networks*, vol. 8, no. 2-3, pp. 187–197, 2002.

[10] N. B. Priyantha, "The cricket indoor location system," Ph.D. dissertation, Massachusetts Institute of Technology, 2005.

[11] K. P. Pujapanda, "Lifi integrated to power-lines for smart illumination cum communication," in *2013 International Conference on Communication Systems and Network Technologies*. IEEE, 2013, pp. 875–878.

[12] J. Bohn, V. Coroamă, M. Langheinrich, F. Mattern, and M. Rohs, "Living in a world of smart everyday objects—social, economic, and ethical implications," *Human and Ecological Risk Assessment*, vol. 10, no. 5, pp. 763–785, 2004.

[13] D.-H. Kim, N. Lu, R. Ma, Y.-S. Kim, R.-H. Kim, S. Wang, J. Wu, S. M. Won, H. Tao, A. Islam *et al.*, "Epidermal electronics," *science*, vol. 333, no. 6044, pp. 838–843, 2011.

[14] J. Rashid, M. Broxvall, and A. Saffiotti, "A middleware to integrate robots, simple devices and everyday objects into an ambient ecology," *Pervasive and Mobile Computing*, vol. 8, no. 4, pp. 522–541, 2012.

[15] C. Schroeter, S. Mueller, M. Volkhardt, E. Einhorn, C. Huijnen, H. van den Heuvel, A. van Berlo, A. Bley, and H.-M. Gross, "Realization and user evaluation of a companion robot for people with mild cognitive impairments," in *2013 IEEE International Conference on robotics and automation*. IEEE, 2013, pp. 1153–1159.

[16] P. Di, J. Huang, K. Sekiyama, and T. Fukuda, "Motion control of intelligent cane robot under normal and abnormal walking condition," in *2011 RO-MAN*. IEEE, 2011, pp. 497–502.

[17] J.-H. Kim, Y.-D. Kim, and K.-H. Lee, "The third generation of robotics: Ubiquitous robot," in *Proc of the 2nd Int Conf on Autonomous Robots and Agents*. Citeseer, 2004.

[18] A. Sanfeliu, N. Hagita, and A. Saffiotti, "Network robot systems," *Robotics and autonomous systems*, vol. 56, no. 10, pp. 793–797, 2008.

[19] K. Kamei, S. Nishio, N. Hagita, and M. Sato, "Cloud networked robotics," *IEEE Network*, vol. 26, no. 3, pp. 28–34, 2012.

[20] G. Hu, W. P. Tay, and Y. Wen, "Cloud robotics: architecture, challenges and applications," *IEEE network*, vol. 26, no. 3, pp. 21–28, 2012.

[21] E. Guizzo, "Robots with their heads in the clouds," *IEEE Spectrum*, vol. 48, no. 3, pp. 16–18, 2011.

[22] R. Mojarad, F. Attal, A. Chibani, S. R. Fiorini, and Y. Amirat, "Hybrid approach for human activity recognition by ubiquitous robots," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 5660–5665.

[23] P. C. Roy, S. R. Abidi, and S. S. Abidi, "Possibilistic activity recognition with uncertain observations to support medication adherence in an assisted ambient living setting," *Knowledge-Based Systems*, vol. 133, pp. 156–173, 2017.

[24] A. Wickramasinghe, R. L. S. Torres, and D. C. Ranasinghe, "Recognition of falls using dense sensing in an ambient assisted living environment," *Pervasive and mobile computing*, vol. 34, pp. 14–24, 2017.

[25] D. Leightley, J. Darby, B. Li, J. S. McPhee, and M. H. Yap, "Human activity recognition for physical rehabilitation," in *2013 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 2013, pp. 261–266.

[26] K.-S. Lee, S. Chae, and H.-S. Park, "Optimal time-window derivation for human-activity recognition based on convolutional neural networks of repeated rehabilitation motions," in *2019 IEEE 16th international conference on rehabilitation robotics (ICORR)*. IEEE, 2019, pp. 583–586.

[27] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[28] A. R. M. Forkan, I. Khalil, Z. Tari, S. Foufou, and A. Bouras, "A context-aware approach for long-term behavioural change detection and abnor-

mality prediction in ambient assisted living," *Pattern Recognition*, vol. 48, no. 3, pp. 628–641, 2015.

[29] C. A. Ronao and S.-B. Cho, "Human activity recognition with smartphone sensors using deep learning neural networks," *Expert systems with applications*, vol. 59, pp. 235–244, 2016.

[30] J. Yang, M. N. Nguyen, P. P. San, X. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition." in *Ijcai*, vol. 15, 2015, pp. 3995–4001.

[31] S. Duffner, S. Berlemont, G. Lefebvre, and C. Garcia, "3d gesture classification with convolutional neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 5432–5436.

[32] T. Plötz, N. Y. Hammerla, and P. Olivier, "Feature learning for activity recognition in ubiquitous computing," in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, vol. 22, no. 1, 2011, p. 1729.

[33] L. Onofri, P. Soda, M. Pechenizkiy, and G. Iannello, "A survey on using domain and contextual knowledge for human activity recognition in video streams," *Expert Systems with Applications*, vol. 63, pp. 97–111, 2016.

[34] L. Chen, C. D. Nugent, M. Mulvenna, D. Finlay, X. Hong, and M. Poland, "A logical framework for behaviour reasoning and assistance in a smart home," *International Journal of Assistive Robotics and Mechatronics*, vol. 9, no. 4, pp. 20–34, 2008.

[35] B. Bouchard, S. Giroux, and A. Bouzouane, "A smart home agent for plan recognition," in *Conference of the Canadian Society for Computational Studies of Intelligence*. Springer, 2006, pp. 25–36.

[36] R. Hervás, J. Bravo, J. Fontecha, and V. Villarreal, "Achieving adaptive augmented reality through ontological context-awareness applied to aal scenarios." *J. UCS*, vol. 19, no. 9, pp. 1334–1349, 2013.

[37] I. Huitzil, L. Dranca, J. Bernad, and F. Bobillo, "Gait recognition using fuzzy ontologies and kinect sensor data," *International Journal of Approximate Reasoning*, 2019.

[38] D. Riboni and C. Bettini, "Cosar: hybrid reasoning for context-aware activity recognition," *Personal and Ubiquitous Computing*, vol. 15, no. 3, pp. 271–289, 2011.

[39] L. Liu, S. Wang, G. Su, B. Hu, Y. Peng, Q. Xiong, and J. Wen, "A framework of mining semantic-based probabilistic event relations for complex activity recognition," *Information Sciences*, vol. 418, pp. 13–33, 2017.

[40] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys (CSUR)*, vol. 43, no. 3, pp. 1–43, 2011.

[41] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *arXiv preprint arXiv:1806.11230*, 2018.

[42] K. Li and Y. Fu, "Prediction of human activity by discovering temporal sequence patterns," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 8, pp. 1644–1657, 2014.

[43] R. Hamid, S. Maddi, A. Johnson, A. Bobick, I. Essa, and C. Isbell, "A novel sequence representation for unsupervised analysis of human activities," *Artificial Intelligence*, vol. 173, no. 14, pp. 1221–1244, 2009.

[44] T. Duong, D. Phung, H. Bui, and S. Venkatesh, "Efficient duration and hierarchical modeling for human activity recognition," *Artificial intelligence*, vol. 173, no. 7-8, pp. 830–856, 2009.

[45] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from rgb-d videos," *The International Journal of Robotics Research*, vol. 32, no. 8, pp. 951–970, 2013.

[46] H. Abdelkawy, N. Ayari, A. Chibani, Y. Amirat, and F. Attal, "Deep hm-resnet model for human activity-aware robotic systems," *arXiv preprint arXiv:1809.07624*, 2018.

[47] F. Attal, S. Mohammed, M. Dedabrishvili, F. Chamroukhi, L. Oukhellou, and Y. Amirat, "Physical human activity recognition using wearable sensors," *Sensors*, vol. 15, no. 12, pp. 31 314–31 338, 2015.

[48] I. Bloch, "Fusion of digital information: m 'e thodological panorama," *National Journals of Robotics Research*, vol. 2005, pp. 79–88, 2005.

[49] P. Smets, "Imperfect information: Imprecision and uncertainty," in *Uncertainty management in information systems*. Springer, 1997, pp. 225–254.

[50] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine," in *International workshop on ambient assisted living*. Springer, 2012, pp. 216–223.

[51] A. Sharma, Y.-D. Lee, and W.-Y. Chung, "High accuracy human activity monitoring using neural network," in *Convergence and Hybrid Information Technology, 2008. ICCIT'08. Third International Conference on*, vol. 1. IEEE, 2008, pp. 430–435.

[52] Y.-S. Lee and S.-B. Cho, "Activity recognition using hierarchical hidden markov models on a smartphone with 3d accelerometer," in *International Conference on Hybrid Artificial Intelligence Systems*. Springer, 2011, pp. 460–467.

[53] J. Mantyjarvi, J. Himberg, and T. Seppanen, "Recognizing human motion with multiple acceleration sensors," in *Systems, Man, and Cybernetics, 2001 IEEE International Conference on*, vol. 2. IEEE, 2001, pp. 747–752.

[54] Z. He and L. Jin, "Activity recognition from acceleration data based on discrete consine transform and svm," in *Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on*. IEEE, 2009, pp. 5041–5044.

[55] X. Yang, A. Dinh, and L. Chen, "Implementation of a wearerable real-time system for physical activity recognition based on naive bayes classifier," in *Bioinformatics and Biomedical Technology (ICBBT), 2010 International Conference on*. IEEE, 2010, pp. 101–105.

[56] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations," in *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[57] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1290–1297.

[58] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 588–595.

[59] B. Romera-Paredes, M. S. Aung, and N. Bianchi-Berthouze, "A one-vs-one classifier ensemble with majority voting for activity recognition." in *Esann*, 2013.

[60] M. Zhang and A. A. Sawchuk, "A feature selection-based framework for human activity recognition using wearable multimodal sensors." in *BodyNets*, 2011, pp. 92–98.

[61] D. Chen, J. Yang, and H. Wactlar, "A study of detecting social interaction with sensors in a nursing home environment," in *International Workshop on Human-Computer Interaction*. Springer, 2005, pp. 199–210.

[62] L. G. Fahad, A. Khan, and M. Rajarajan, "Activity recognition in smart homes with self verification of assignments," *Neurocomputing*, vol. 149, pp. 1286–1298, 2015.

[63] J. D. Ceron, D. M. Lopez, and G. A. Ramirez, "A mobile system for sedentary behaviors classification based on accelerometer and location data," *Computers in Industry*, vol. 92, pp. 25–31, 2017.

[64] L. Mo, S. Liu, R. X. Gao, and P. S. Freedson, "Multi-sensor ensemble classifier for activity recognition," *Journal of Software Engineering and Applications*, vol. 5, p. 113, 2012.

[65] S. A. Antos, M. V. Albert, and K. P. Kording, "Hand, belt, pocket or bag: Practical activity tracking with mobile phones," *Journal of neuroscience methods*, vol. 231, pp. 22–30, 2014.

[66] P. Gupta and T. Dallas, "Feature selection and activity recognition system using a single triaxial accelerometer," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 6, pp. 1780–1786, 2014.

[67] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in *International conference on pervasive computing*. Springer, 2004, pp. 1–17.

[68] N. Ravi, N. Dandekar, P. Mysore, and M. L. Littman, "Activity recognition from accelerometer data," in *Aaai*, vol. 5, no. 2005. Pittsburgh, PA, 2005, pp. 1541–1546.

[69] A. Nazábal, P. Garcia-Moreno, A. Artes-Rodriguez, and Z. Ghahramani, "Human activity recognition by combining a small number of classifiers," *IEEE journal of biomedical and health informatics*, vol. 20, no. 5, pp. 1342–1351, 2015.

[70] C.-H. Lu and L.-C. Fu, "Robust location-aware activity recognition using wireless sensor network in an attentive home," *IEEE Transactions on Automation Science and Engineering*, vol. 6, no. 4, pp. 598–609, 2009.

[71] B. Cvetković, R. Szeklicki, V. Janko, P. Lutomski, and M. Luštrek, "Real-time activity monitoring with a wristband and a smartphone," *Information Fusion*, vol. 43, pp. 77–93, 2018.

[72] U. Maurer, A. Rowe, A. Smailagic, and D. Siewiorek, "Location and activity recognition using ewatch: A wearable sensor platform," in *Ambient intelligence in everyday life*. Springer, 2006, pp. 86–102.

[73] L. Peng, L. Chen, X. Wu, H. Guo, and G. Chen, "Hierarchical complex activity representation and recognition using topic model and classifier level fusion," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 6, pp. 1369–1379, 2016.

[74] H. S. Hossain, M. A. A. H. Khan, and N. Roy, "Active learning enabled activity recognition," *Pervasive and Mobile Computing*, vol. 38, pp. 312–330, 2017.

[75] M. T. Uddin, M. M. Billah, and M. F. Hossain, "Random forests based recognition of human activities and postural transitions on smartphone," in *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*. IEEE, 2016, pp. 250–255.

[76] E. Garcia-Ceja, C. E. Galván-Tejada, and R. Brena, "Multi-view stacking for activity recognition with sound and accelerometer data," *Information Fusion*, vol. 40, pp. 45–56, 2018.

[77] J. Wen and Z. Wang, "Learning general model for activity recognition with limited labelled data," *Expert Systems with Applications*, vol. 74, pp. 19–28, 2017.

[78] ——, "Sensor-based adaptive activity recognition with dynamically available sensors," *Neurocomputing*, vol. 218, pp. 307–317, 2016.

[79] G. Singla, D. J. Cook, and M. Schmitter-Edgecombe, "Recognizing independent and joint activities among multiple residents in smart environments," *Journal of ambient intelligence and humanized computing*, vol. 1, no. 1, pp. 57–63, 2010.

[80] L. Liao, D. Fox, and H. Kautz, "Extracting places and activities from gps traces using hierarchical conditional random fields," *The International Journal of Robotics Research*, vol. 26, no. 1, pp. 119–134, 2007.

[81] D. H. Hu and Q. Yang, "Cigar: Concurrent and interleaving goal and activity recognition." in *AAAI*, vol. 8, 2008, pp. 1363–1368.

[82] M. Mahdaviani and T. Choudhury, "Fast and scalable training of semi-supervised crfs with application to activity recognition," *Advances in Neural Information Processing Systems*, vol. 20, pp. 977–984, 2007.

[83] R. Chaudhry, F. Ofli, G. Kurillo, R. Bajcsy, and R. Vidal, "Bio-inspired dynamic 3d discriminative skeletal features for human action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 471–478.

[84] L. Xia, C.-C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2012, pp. 20–27.

[85] P. Wang, C. Yuan, W. Hu, B. Li, and Y. Zhang, "Graph based skeleton motion representation and similarity measurement for action recognition," in *European conference on computer vision*. Springer, 2016, pp. 370–385.

[86] S. Bouznad, F. Sebbak, Y. Amirat, A. Chibani, and F. Benhammadi, "Generalized fuzzy soft set based fusion strategy for activity classification in smart home," in *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2017, pp. 1–6.

[87] R. Chavarriaga, H. Sagha, A. Calatroni, S. T. Digumarti, G. Tröster, J. d. R. Millán, and D. Roggen, "The opportunity challenge: A benchmark

database for on-body sensor-based activity recognition," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 2033–2042, 2013.

[88] N. Y. Hammerla, S. Halloran, and T. Plötz, "Deep, convolutional, and recurrent models for human activity recognition using wearables," *arXiv preprint arXiv:1604.08880*, 2016.

[89] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in *2012 16th International Symposium on Wearable Computers*. IEEE, 2012, pp. 108–109.

[90] M. Bachlin, D. Roggen, G. Troster, M. Plotnik, N. Inbar, I. Meidan, T. Herman, M. Brozgol, E. Shaviv, N. Giladi *et al.*, "Potentials of enhanced context awareness in wearable assistants for parkinson's disease patients with the freezing of gait syndrome," in *2009 International Symposium on Wearable Computers*. IEEE, 2009, pp. 123–130.

[91] F. J. Ordóñez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.

[92] J. Medina-Quero, S. Zhang, C. Nugent, and M. Espinilla, "Ensemble classifier of long short-term memory with fuzzy temporal windows on binary sensors for activity recognition," *Expert Systems with Applications*, vol. 114, pp. 441–453, 2018.

[93] P. Agarwal and M. Alam, "A lightweight deep learning model for human activity recognition on edge devices," *Procedia Computer Science*, vol. 167, pp. 2364–2373, 2020.

[94] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognition Letters*, vol. 119, pp. 3–11, 2019.

[95] L. Chen, C. D. Nugent, and H. Wang, "A knowledge-driven approach to activity recognition in smart homes," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 6, pp. 961–974, 2012.

[96] P. Cottone, S. Gaglio, G. L. Re, and M. Ortolani, "User activity recognition for energy saving in smart homes," *Pervasive and Mobile Computing*, vol. 16, pp. 156–170, 2015.

[97] H. Guo, L. Chen, L. Peng, and G. Chen, "Wearable sensor based multimodal human activity recognition exploiting the diversity of classifier ensemble," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016, pp. 1112–1123.

[98] L. Liu, Y. Peng, S. Wang, M. Liu, and Z. Huang, "Complex activity recognition using time series pattern dictionary learned from ubiquitous sensors," *Information Sciences*, vol. 340, pp. 41–57, 2016.

[99] M. Shoaib, S. Bosch, O. D. Incel, H. Scholten, and P. J. Havinga, "Complex human activity recognition using smartphone and wrist-worn motion sensors," *Sensors*, vol. 16, no. 4, p. 426, 2016.

[100] L. Liu, L. Cheng, Y. Liu, Y. Jia, and D. Rosenblum, "Recognizing complex activities by a probabilistic interval-based model," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.

[101] L. Liu, S. Wang, G. Su, Z.-G. Huang, and M. Liu, "Towards complex activity recognition using a bayesian network-based probabilistic generative framework," *Pattern Recognition*, vol. 68, pp. 295–309, 2017.

[102] M. Á. Á. de la Concepción, L. M. S. Morillo, J. A. Á. García, and L. González-Abril, "Mobile activity recognition and fall detection system for elderly people using ameva algorithm," *Pervasive and Mobile Computing*, vol. 34, pp. 3–13, 2017.

[103] Y. Lu, Y. Wei, L. Liu, J. Zhong, L. Sun, and Y. Liu, "Towards unsupervised physical activity recognition using smartphone accelerometers," *Multimedia Tools and Applications*, vol. 76, no. 8, pp. 10 701–10 719, 2017.

[104] B. Minor and D. J. Cook, "Forecasting occurrences of activities," *Pervasive and mobile computing*, vol. 38, pp. 77–91, 2017.

[105] J. Huang, S. Lin, N. Wang, G. Dai, Y. Xie, and J. Zhou, "Tse-cnn: A two-stage end-to-end cnn for human activity recognition," *IEEE journal of biomedical and health informatics*, vol. 24, no. 1, pp. 292–299, 2019.

[106] N. T. H. Thu and D. S. Han, "Utilization of postural transitions in sensor-based human activity recognition," in *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. IEEE, 2020, pp. 177–181.

[107] K. Chen, L. Yao, D. Zhang, X. Wang, X. Chang, and F. Nie, "A semisupervised recurrent convolutional attention model for human activity recognition," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 5, pp. 1747–1756, 2019.

[108] T. Mahmud, S. S. Akash, S. A. Fattah, W.-P. Zhu, and M. O. Ahmad, "Human activity recognition from multi-modal wearable sensor data using deep multi-stage lstm architecture based on temporal feature aggregation," in *2020 IEEE 63rd International Midwest Symposium on Circuits and Systems (MWSCAS)*. IEEE, 2020, pp. 249–252.

[109] M. Shanahan *et al.*, *Solving the frame problem: a mathematical investigation of the common sense law of inertia*. MIT press, 1997.

[110] A. Artikis and G. Paliouras, "Behaviour recognition using the event calculus," in *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, 2009, pp. 469–478.

[111] A. Artikis, M. Sergot, and G. Paliouras, "A logic programming approach to activity recognition," in *Proceedings of the 2nd ACM international workshop on Events in multimedia*, 2010, pp. 3–8.

[112] A. Artikis, A. Margara, M. Ugarte, S. Vansummeren, and M. Weidlich, "Complex event recognition languages: Tutorial," in *Proceedings of the 11th ACM International Conference on Distributed and Event-based Systems*, 2017, pp. 7–10.

[113] J. Ye, L. Coyle, S. Dobson, and P. Nixon, "Ontology-based models in pervasive computing systems," *The Knowledge Engineering Review*, vol. 22, no. 4, pp. 315–347, 2007.

[114] L. Chen, C. D. Nugent, and H. Wang, "A knowledge-driven approach to activity recognition in smart homes," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 6, pp. 961–974, 2011.

[115] K. Wongpatikaseree, M. Ikeda, M. Buranarach, T. Supnithi, A. O. Lim, and Y. Tan, "Activity recognition using context-aware infrastructure ontology in smart home domain," in *2012 Seventh International Conference on Knowledge, Information and Creativity Support Systems*. IEEE, 2012, pp. 50–57.

[116] J. Ye, G. Stevenson, and S. Dobson, "A top-level ontology for smart environments," *Pervasive and mobile computing*, vol. 7, no. 3, pp. 359–378, 2011.

[117] D. Riboni and C. Bettini, "Owl 2 modeling and reasoning with complex human activities," *Pervasive and Mobile Computing*, vol. 7, no. 3, pp. 379–395, 2011.

[118] T. Springer and A.-Y. Turhan, "Employing description logics in ambient intelligence for modeling and reasoning about complex situations," *Journal of Ambient Intelligence and Smart Environments*, vol. 1, no. 3, pp. 235–259, 2009.

[119] S. Knox, L. Coyle, and S. Dobson, "Using ontologies in case-based activity recognition," 2010.

[120] J. Lester, T. Choudhury, N. Kern, G. Borriello, and B. Hannaford, "A hybrid discriminative/generative approach for modeling human activities." in *IJCAI*, vol. 5, no. 2005. Citeseer, 2005.

[121] A. Hristoskova, V. Sakkalis, G. Zacharioudakis, M. Tsiknakis, and F. De Turck, "Ontology-driven monitoring of patient's vital signs enabling personalized medical detection and alert," *Sensors*, vol. 14, no. 1, pp. 1598–1628, 2014.

[122] G. Okeyo, L. Chen, H. Wang, and R. Sterritt, "A hybrid ontological and temporal approach for composite activity modelling," in *2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications*. IEEE, 2012, pp. 1763–1770.

[123] G. Okeyo, L. Chen, and H. Wang, "Combining ontological and temporal formalisms for composite activity modelling and recognition in smart homes," *Future Generation Computer Systems*, vol. 39, pp. 29–43, 2014.

[124] F. Mastrogiovanni, A. Scalmato, A. Sgorbissa, and R. Zaccaria, "Smart environments and activity recognition: a logic-based approach," in *Activity Recognition in Pervasive Intelligent Environments*. Springer, 2011, pp. 83–109.

[125] G. Meditskos, S. Dasiopoulou, and I. Kompatsiaris, "Metaq: A knowledge-driven framework for context-aware activity recognition combining sparql and owl 2 activity patterns," *Pervasive and Mobile Computing*, vol. 25, pp. 104–124, 2016.

[126] G. Meditskos, S. Dasiopoulou, V. Efstathiou, and I. Kompatsiaris, "Sp-act: A hybrid framework for complex activity recognition combining owl and sparql rules," in *2013 IEEE international conference on pervasive computing and communications workshops (PERCOM workshops)*. IEEE, 2013, pp. 25–30.

[127] ——, "Ontology patterns for complex activity modelling," in *International Workshop on Rules and Rule Markup Languages for the Semantic Web*. Springer, 2013, pp. 144–157.

[128] C. Villalonga, H. Pomares, I. Rojas, and O. Banos, "Mimu-wear: Ontology-based sensor selection for real-world wearable activity recognition," *Neurocomputing*, vol. 250, pp. 76–100, 2017.

[129] C. Diamantini, A. Freddi, S. Longhi, D. Potena, and E. Storti, "A goal-oriented, ontology-based methodology to support the design of aal environments," *Expert Systems with Applications*, vol. 64, pp. 117–131, 2016.

[130] K. Najera, A. Perini, A. Martinez, and H. Estrada, "Supporting i* model integration through an ontology-based approach." in *iStar*, 2011, pp. 43–48.

[131] D. Bonino and F. Corno, "Dogont-ontology modeling for intelligent domotic environments," in *International Semantic Web Conference*. Springer, 2008, pp. 790–803.

[132] C. Diamantini, D. Potena, and E. Storti, "Sempi: A semantic framework for the collaborative construction and maintenance of a shared dictionary of performance indicators," *Future Generation Computer Systems*, vol. 54, pp. 352–365, 2016.

[133] M. Compton, P. Barnaghi, L. Bermudez, R. Garcia-Castro, O. Corcho, S. Cox, J. Graybeal, M. Hauswirth, C. Henson, A. Herzog *et al.*, "The ssn ontology of the w3c semantic sensor network incubator group," *Journal of Web Semantics*, vol. 17, pp. 25–32, 2012.

[134] J. Rafferty, C. D. Nugent, J. Liu, and L. Chen, "From activity recognition to intention recognition for assisted living within smart homes," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 3, pp. 368–379, 2017.

[135] N. D. Rodríguez, M. P. Cuéllar, J. Lilius, and M. D. Calvo-Flores, "A survey on ontologies for human behavior recognition," *ACM Computing Surveys (CSUR)*, vol. 46, no. 4, pp. 1–33, 2014.

[136] N. D. Rodríguez, R. Wikström, J. Lilius, M. P. Cuéllar, and M. D. C. Flores, "Understanding movement and interaction: an ontology for kinect-based 3d depth sensors," in *Ubiquitous computing and ambient intelli-*

*gence. Context-awareness and context-driven interaction*. Springer, 2013, pp. 254–261.

[137] N. D. Rodríguez, M. P. Cuéllar, J. Lilius, and M. D. Calvo-Flores, "A fuzzy ontology for semantic modelling and recognition of human behaviour," *Knowledge-Based Systems*, vol. 66, pp. 46–60, 2014.

[138] J. Ye, S. Dasiopoulou, G. Stevenson, G. Meditskos, E. Kontopoulos, I. Kompatsiaris, and S. Dobson, "Semantic web technologies in pervasive computing: A survey and research roadmap," *Pervasive and Mobile Computing*, vol. 23, pp. 1–25, 2015.

[139] S. Wang, W. Pentney, A.-M. Popescu, T. Choudhury, and M. Philipose, "Common sense based joint training of human activity recognizers." in *IJCAI*, vol. 7, 2007, pp. 2237–2242.

[140] K. Gayathri, S. Elias, and B. Ravindran, "Hierarchical activity recognition for dementia care using markov logic network," *Personal and Ubiquitous Computing*, vol. 19, no. 2, pp. 271–285, 2015.

[141] K. Gayathri, K. Easwarakumar, and S. Elias, "Probabilistic ontology based activity recognition in smart homes using markov logic network," *Knowledge-Based Systems*, vol. 121, pp. 173–184, 2017.

[142] K. Gayathri, S. Elias, and S. Shivashankar, "Composite activity recognition in smart homes using markov logic network," in *2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom)*. IEEE, 2015, pp. 46–53.

[143] H. Sfar, "Real time intelligent decision making from heterogeneous and imperfect data," Ph.D. dissertation, Université Paris-Saclay (ComUE), 2019.

[144] R. Helaoui, D. Riboni, M. Niepert, C. Bettini, and H. Stuckenschmidt, "Towards activity recognition using probabilistic description logics," *Activity Context Representation: Techniques and Languages*, vol. 12, no. 5, pp. 26–31, 2012.

[145] M. Niepert, J. Noessner, and H. Stuckenschmidt, "Log-linear description logics," in *Twenty-Second International Joint Conference on Artificial Intelligence*. Citeseer, 2011.

[146] F. Baader, D. Calvanese, D. McGuinness, P. Patel-Schneider, D. Nardi *et al.*, *The description logic handbook: Theory, implementation and applications*. Cambridge university press, 2003.

[147] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[148] A. Artikis, E. Makris, and G. Paliouras, "A probabilistic interval-based event calculus for activity recognition," *Annals of Mathematics and Artificial Intelligence*, pp. 1–24, 2019.

[149] H. Jia and S. Chen, "Integrated data and knowledge driven methodology for human activity recognition," *Information Sciences*, vol. 536, pp. 409–430, 2020.

[150] P. Wei, Y. Liu, T. Shu, N. Zheng, and S.-C. Zhu, "Where and why are they looking? jointly inferring human attention and intentions in complex tasks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6801–6809.

[151] Z. Wang, A. Boularias, K. Mülling, B. Schölkopf, and J. Peters, "Anticipatory action selection for human–robot table tennis," *Artificial Intelligence*, vol. 247, pp. 399–414, 2017.

[152] K. Ramirez-Amaro, M. Beetz, and G. Cheng, "Transferring skills to humanoid robots by extracting semantic representations from observations of human activities," *Artificial Intelligence*, vol. 247, pp. 95–118, 2017.

[153] P. Duckworth, D. C. Hogg, and A. G. Cohn, "Unsupervised human activity analysis for intelligent mobile robots," *Artificial Intelligence*, vol. 270, pp. 67–92, 2019.

[154] T. Bandyopadhyay, K. S. Won, E. Frazzoli, D. Hsu, W. S. Lee, and D. Rus, "Intention-aware motion planning," in *Algorithmic foundations of robotics X*. Springer, 2013, pp. 475–491.

[155] H. Kurniawati, Y. Du, D. Hsu, and W. S. Lee, "Motion planning under uncertainty for robotic tasks with long time horizons," *The International Journal of Robotics Research*, vol. 30, no. 3, pp. 308–323, 2011.

[156] C. Vondrick, D. Oktay, H. Pirsiavash, and A. Torralba, "Predicting motivations of actions by leveraging text," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2997–3005.

[157] D. Xie, "Inferring the intentions and attentions of agents from videos," Ph.D. dissertation, UCLA, 2016.

[158] C. Yu and L. B. Smith, "Linking joint attention with hand-eye coordination–a sensorimotor approach to understanding child-parent social interaction," in *CogSci... Annual Conference of the Cognitive Science Society. Cognitive Science Society (US). Conference*, vol. 2015. NIH Public Access, 2015, p. 2763.

[159] M. Land, N. Mennie, and J. Rusted, "The roles of vision and eye movements in the control of activities of daily living," *Perception*, vol. 28, no. 11, pp. 1311–1328, 1999.

[160] Z. Wang, K. Jiang, Y. Hou, W. Dou, C. Zhang, Z. Huang, and Y. Guo, "A survey on human behavior recognition using channel state information," *Ieee Access*, vol. 7, pp. 155 986–156 024, 2019.

[161] B. Bruno, F. Mastrogiovanni, A. Sgorbissa, T. Vernazza, and R. Zaccaria, "Analysis of human behavior recognition algorithms based on acceleration data," in *2013 IEEE International Conference on Robotics and Automation*. IEEE, 2013, pp. 1602–1607.

[162] K. Makantasis, A. Doulamis, N. Doulamis, and K. Psychas, "Deep learning based human behavior recognition in industrial workflows," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 1609–1613.

[163] G. Batchuluun, J. H. Kim, H. G. Hong, J. K. Kang, and K. R. Park, "Fuzzy system based human behavior recognition by combining behavior prediction and recognition," *Expert Systems with Applications*, vol. 81, pp. 108–133, 2017.

[164] R. H. Baxter, N. M. Robertson, and D. M. Lane, "Human behaviour recognition in data-scarce domains," *Pattern Recognition*, vol. 48, no. 8, pp. 2377–2393, 2015.

[165] N. Phan, D. Dou, H. Wang, D. Kil, and B. Piniewski, "Ontology-based deep learning for human behavior prediction with explanations in health social networks," *Information sciences*, vol. 384, pp. 298–313, 2017.

[166] G. Maeda, M. Ewerton, G. Neumann, R. Lioutikov, and J. Peters, "Phase estimation for fast action recognition and trajectory generation in human–robot collaboration," *The International Journal of Robotics Research*, vol. 36, no. 13-14, pp. 1579–1594, 2017.

[167] D. Matsumoto, *The handbook of culture and psychology*, 2001, ch. Culture and emotion, pp. 171–194.

[168] P. Ekman, W. V. Friesen, M. O'sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. LeCompte, T. Ayhan Pitcairn, P. E. Ricci-Bitti, K. Scherer, M. Tomita, and A. Tzavaras, "Universals and cultural differences in the judgments of facial expressions of emotion," *Journal of personality and social psychology*, vol. 53, no. 4, p. 712, 1987.

[169] M.-J. Han, C.-H. Lin, and K.-T. Song, "Robotic emotional expression generation based on mood transition and personality model," *IEEE transactions on cybernetics*, vol. 43, no. 4, pp. 1290–1303, 2012.

[170] L.-A. Perez-Gaspar, S.-O. Caballero-Morales, and F. Trujillo-Romero, "Multimodal emotion recognition with evolutionary computation for human-robot interaction," *Expert Systems with Applications*, vol. 66, pp. 42–61, 12 2016.

[171] X. Zhao, J. Zou, H. Li, E. Dellandréa, I. A. Kakadiaris, and L. Chen, "Automatic 2.5-d facial landmarking and emotion annotation for social interaction assistance," *IEEE transactions on cybernetics*, vol. 46, no. 9, pp. 2042–2055, 2015.

[172] A. Hong, N. Lunscher, T. Hu, Y. Tsuboi, X. Zhang, S. Alves, G. Nejat, and B. Benhabib, "A multimodal emotional human-robot interaction archi-

tecture for social robots engaged in bi-directional communication," *IEEE Transactions on Cybernetics*, pp. 1–15, 2020.

[173] Y. Zong, W. Zheng, Z. Cui, G. Zhao, and B. Hu, "Toward bridging microexpressions from different domains," *IEEE transactions on cybernetics*, pp. 1–14, 2019. [Online]. Available: 10.1109/TCYB.2019.2914512

[174] K. Mistry, L. Zhang, S. C. Neoh, C. P. Lim, and B. Fielding, "A micro-ga embedded pso feature selection approach to intelligent facial emotion recognition," *IEEE transactions on cybernetics*, vol. 47, no. 6, pp. 1496–1509, 2016.

[175] T. Makioka, Y. Kuriyaki, K. Uchimura, and T. Satonaka, "Quantitative study of facial expression asymmetry using objective measure based on neural networks," in *2016 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, Oct 2016, pp. 1–4.

[176] N. A. Gorlova and T. A. Gultyaeva, "Emotion classification on image using hidden markov models," in *2016 13th International Scientific-Technical Conference on Actual Problems of Electronics Instrument Engineering (APEIE)*, vol. 03, Oct 2016, pp. 1–1.

[177] S. Deb and S. Dandapat, "Multiscale amplitude feature and significance of enhanced vocal tract information for emotion classification," *IEEE transactions on cybernetics*, vol. 49, no. 3, pp. 802–815, 2018.

[178] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore, Maryland, USA: Association for Computational Linguistics, June 2014, pp. 655–665.

[179] J. Pang, Y. Rao, H. Xie, X. Wang, F. L. Wang, T. Wong, and Q. Li, "Fast supervised topic models for short text emotion detection," *IEEE Transactions on Cybernetics*, pp. 1–14, 2019.

[180] D. McColl, C. Jiang, and G. Nejat, "Classifying a person's degree of accessibility from natural body language during social human–robot interactions," *IEEE transactions on cybernetics*, vol. 47, no. 2, pp. 524–538, 2016.

[181] H. Meng and N. Bianchi-Berthouze, "Affective state level recognition in naturalistic facial and vocal expressions," *IEEE Transactions on Cybernetics*, vol. 44, no. 3, pp. 315–328, 2013.

[182] Y. Hu, J. S. J. Ren, J. Dai, C. Yuan, L. Xu, and W. Wang, "Deep multimodal speaker naming," *CoRR*, vol. abs/1507.04831, 2015.

[183] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of eeg signals and facial expressions for continuous emotion detection," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 17–28, Jan 2016.

[184] M. Paleari, R. Chellali, and B. Huet, "Bimodal emotion recognition," in *International Conference on Social Robotics*. Springer, 2010, pp. 305–314.

[185] P. V. Rouast, A. Marc, and C. Raymond, "Deep learning for human affect recognition: Insights and new developments," *IEEE Transactions on Affective Computing.*, 2019.

[186] T. Baltrušaitis, A. Chaitanya, and M. Louis-Philippe, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[187] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 1, pp. 39–58, 2008.

[188] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[189] F. Ringeval, F. Eyben, E. Kroubi, J. P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognition Letters*, vol. 66, pp. 22–30, 2015.

[190] L. He, D. Jiang, L. Yan, E. Pei, P. Wu, and H. Sahli, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pp. 73–80, 2015.

[191] E. Rejaibi, D. Kadoch, K. Bentounes, R. Alfred, M. Daoudi, A. Hadid, and A. Othmani, "Clinical depression and affect recognition with emoaudionet," *arXiv preprint arXiv:1911.00310*, 2019.

[192] P. Tzarakis, G. Trigeorgis, M. Nicolaou, S. Björn, and Z. Stefanos, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[193] P. Tzirakis, Z. Jiehao, and S. Bjorn W., "End-to-end speech emotion recognition using deep neural networks," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5089–5093, 2018.

[194] R. Basnet, M. T. Islam, T. Howlader, S. M. M. Rahmanb, and D. Hatzinakos, "Estimation of affective dimensions using cnn-based features of audiovisual data," *Pattern Recognition Letters*, vol. 128, pp. 290–297, 2019.

[195] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio–visual emotion recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 8, pp. 3030–3043, 2017.

[196] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Emotion recognition in the wild with feature fusion and multiple kernel learning." Proceedings of the 16th International Conference on Multimodal Interaction, 2014, pp. 508–513.

[197] T. Baltrušaitis, B. Ntombikayise, and R. Peter, "Dimensional affect recognition using continuous conditional random fields," *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pp. 1–8, 2013.

[198] D. H. Kim, W. J. Baddar, J. Jang, and Y. M. Ro, "Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition," *IEEE Transactions on Affective Computing*, vol. 10, no. 2, pp. 223–236, 2017.

[199] Y. Y. Mathieu, "Annotation of emotions and feelings in texts," in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2005, pp. 350–357.

[200] C. Strapparava, A. Valitutti, and O. Stock, "The affective weight of lexicon," in *Proceedings of the fifth international conference on language resources and evaluation*, 2006, pp. 423–426.

[201] J. Hastings, W. Ceusters, K. Mulligan, and B. Smith, "Annotating affective neuroscience data with the emotion ontology," in *Third International Conference on Biomedical Ontology*. ICBO, 2012, pp. 1–5.

[202] R. Gil, J. Virgili-Gom, R. Garca, and C. Mason, "Emotions ontology for collaborative modelling and learning of emotional responses," *Computers in Human Behavior*, vol. 51, Part B, pp. 610 – 617, 2015.

[203] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[204] P. Singh, T. Lin, E. T. Mueller, G. Lim, T. Perkins, and W. L. Zhu, "Open mind common sense: Knowledge acquisition from the general public," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, 2002, pp. 1223–1237.

[205] K. Nakahara and S. Yamada, "Development and evaluation of a web-based game for common-sense knowledge acquisition in japan," *Unisys Technology Review*, vol. 30, no. 4, pp. 295–305, 2011.

[206] Y.-l. Kuo, J.-C. Lee, K.-y. Chiang, R. Wang, E. Shen, C.-w. Chan, and J. Y.-j. Hsu, "Community-based game design: experiments on social games for commonsense data collection," in *Proceedings of the acm sigkdd workshop on human computation*. ACM, 2009, pp. 15–22.

[207] J. Breen, "Jmdict: a japanese-multilingual dictionary," in *Proceedings of the Workshop on Multilingual Linguistic Ressources*. Association for Computational Linguistics, 2004, pp. 71–79.

[208] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in *The semantic web*. Springer, 2007, pp. 722–735.

[209] F. Bond and R. Foster, "Linking and extending an open multilingual wordnet," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2013, pp. 1352–1362.

[210] G. Miller, *WordNet: An electronic lexical database*. MIT press, 1998.

[211] J. A. Jacko, *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications, Third Editiion*, 3rd ed. Boca Raton, FL: CRC Press, 06 2012.

[212] S. Kaiser and T. Wehrle, "Facial expressions in social interactions: Beyond basic emotions," *Advances in consciousness research. Animating expressive characters for social interactions. Amsterdam: John Benjamins publishing company*, 2004.

[213] M. W. Murry and D. M. Isaacowitz, "Age differences in emotion perception the effects of the social environment," *International Journal of Behavioral Development*, p. 0165025416667493, 2016.

[214] U. Hess, R. B. Adams Jr, and R. E. Kleck, "Facial appearance, gender, and emotion expression." *Emotion*, vol. 4, no. 4, p. 378, 2004.

[215] N. Lim, "Cultural differences in emotion: differences in emotional arousal level between the east and the west," *Integrative Medicine Research*, vol. 5, no. 2, pp. 105 – 109, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2213422016300191

[216] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct 2016.

[217] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[218] H. Mobahi, M. Farajtabar, and P. L. Bartlett, "Self-distillation amplifies regularization in hilbert space," *arXiv preprint arXiv:2002.05715*, 2020.

[219] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

[220] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, pp. 18–31, 2019.

[221] A. Agarwala and R. Vemulapalli, "A compact embedding for facial expression similarity," 2019, pp. 5676–5685.

[222] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[223] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3967–3976.

[224] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney,

R. Weiss, and K. Wilson, “Cnn architectures for large-scale audio classification,” *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[225] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, “Audio set: An ontology and human-labeled dataset for audio events,” 2017.

[226] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[227] D. M. Gabbay, C. J. Hogger, and J. A. Robinson, *Handbook of Logic in Artificial Intelligence and Logic Programming: Volume 5: Logic Programming*. Clarendon Press, 1998.

[228] S. Poria, E. Cambria, and A. F. Gelbukh, “Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis.” in *EMNLP*, 2015, pp. 2539–2544.

[229] N. Ayari, H. Abdelkawy, A. Chibani, and Y. Amirat, “Towards semantic multimodal emotion recognition for enhancing assistive services in ubiquitous robotics,” in *2017 AAAI Fall Symposium Series*, 2017.

[230] L.-P. Morency, R. Mihalcea, and P. Doshi, “Towards multimodal sentiment analysis: Harvesting opinions from the web,” in *Proceedings of the 13th International Conference on Multimodal Interfaces*, ser. ICMI ’11. New York, NY, USA: ACM, 2011, pp. 169–176. [Online]. Available: http://doi.acm.org/10.1145/2070481.2070509

[231] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, “Introducing the recola multimodal corpus of remote collaborative and affective interac-

tions," *In 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pp. 1–8, 2013.

[232] M.-I. Georgescu, R. T. Ionescu, and M. Popescu, "Local learning with deep and handcrafted features for facial expression recognition," *IEEE Access*, vol. 7, pp. 64 827–64 836, 2019.

[233] H. Siqueira, S. Magg, and S. Wermter, "Efficient facial feature learning with wide ensemble-based convolutional neural networks," *ArXiv*, vol. abs/2001.06338, 2020.

[234] J. Han, Z. Zhang, N. Cummins, F. Ringeval, and B. Schuller, "Strength modelling for real-world automatic continuous affect recognition from audiovisual signals," *Image and Vision Computing*, vol. 65, pp. 76–86, 2017.

[235] G. Vaquette, A. Orcesi, L. Lucat, and C. Achard, "The daily home life activity dataset: A high semantic activity dataset for online recognition," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 497–504.

[236] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *European conference on computer vision*. Springer, 2016, pp. 816–833.

[237] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[238] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-rnn: Deep learning on spatio-temporal graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5308–5317.

[239] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015. [Online]. Available: http://arxiv.org/abs/1502.03167

[240] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[241] I. Ullah and A. Petrosino, "About pyramid structure in convolutional neural networks," in *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2016, pp. 1318–1324.

[242] Y. Zheng, Q. Liu, E. Chen, Y. Ge, and J. L. Zhao, "Exploiting multi-channels deep convolutional neural networks for multivariate time series classification," *Frontiers of Computer Science*, vol. 10, no. 1, pp. 96–112, 2016.

[243] S.-M. Lee, S. M. Yoon, and H. Cho, "Human activity recognition from accelerometer data using convolutional neural network," in *Big Data and Smart Computing (BigComp), 2017 IEEE International Conference on*. IEEE, 2017, pp. 131–134.

[244] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "The microsoft 2016 conversational speech recognition system," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5255–5259.

[245] H. De Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, and A. C. Courville, "Modulating early visual processing by language," in *Advances in Neural Information Processing Systems*, 2017, pp. 6597–6607.

[246] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE, 2017, pp. 1578–1585.

[247] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[248] M. T. Hagan and M. B. Menhaj, "Training feedforward networks with the marquardt algorithm," *IEEE transactions on Neural Networks*, vol. 5, no. 6, pp. 989–993, 1994.

[249] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[250] G. Zarri, "Knowledge acquisition from complex narrative texts using the nkrl technology," I. B. Gaines and P. o. t. t. B. K. A. f. K.-B. S. W. M. Musen, editors, Eds., vol. 1, Department of Computer Science of the University of Calgary, 1995.

[251] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.

[252] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones." in *ESANN*, 2013.

[253] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1110–1118.

[254] H. Wang and L. Wang, "Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 499–508.

[255] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[256] I. Lee, D. Kim, S. Kang, and S. Lee, "Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1012–1020.

[257] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognition*, vol. 68, pp. 346–362, 2017.

[258] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2117–2126.

[259] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," *arXiv preprint arXiv:1804.06055*, 2018.

[260] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," *arXiv preprint arXiv:2003.14111*, 2020.

[261] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[262] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[263] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[264] R. M. Haralick, K. Shanmugam *et al.*, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610–621, 1973.

[265] S. Bambach, S. Lee, D. J. Crandall, and C. Yu, "Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[266] T. H. N. Le, K. G. Quach, C. Zhu, C. N. Duong, K. Luu, and M. Savvides, "Robust hand detection and classification in vehicles and in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2017, pp. 1203–1210.

[267] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[268] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312

[269] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a" siamese" time delay neural network," in *Advances in neural information processing systems*, 1994, pp. 737–744.

[270] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4353–4361.

[271] S. Chopra, R. Hadsell, Y. LeCun *et al.*, "Learning a similarity metric discriminatively, with application to face verification," in *CVPR (1)*, 2005, pp. 539–546.

[272] V. Lomonaco and D. Maltoni, "Core50: a new dataset and benchmark for continuous object recognition," *arXiv preprint arXiv:1705.03550*, 2017.

[273] J. J. Gibson, *The ecological approach to visual perception: classic edition*. Psychology Press, 2014.

[274] F. Riguzzi, E. Bellodi, E. Lamma, and R. Zese, "Reasoning with probabilistic ontologies," in *Proceedings of the 24th International Conference on Artificial Intelligence*, ser. IJCAI'15. AAAI Press, 2015, pp. 4310–4316.

[275] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, no. 9, pp. 509–517, 1975.

[276] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[277] M. Afifi, "11k hands: gender recognition and biometric identification using a large dataset of hand images," *Multimedia Tools and Applications*, 2019. [Online]. Available: https://doi.org/10.1007/s11042-019-7424-8

[278] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi, "Describing textures in the wild," in *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[279] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[280] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[281] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[282] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[283] Y. Lei, W. Yuan, H. Wang, Y. Wenhu, and W. Bo, "A skin segmentation algorithm based on stacked autoencoders," *IEEE Transactions on Multimedia*, vol. 19, no. 4, pp. 740–749, 2016.

# Publications

- Ayari, Naouel, Hazem Abdelkawy, Abdelghani Chibani, and Yacine Amirat. "Towards semantic multimodal emotion recognition for enhancing assistive services in ubiquitous robotics." In AAAI Fall Symposium Series. 2017.

- Abdelkawy, Hazem, Naouel Ayari, Abdelghani Chibani, Yacine Amirat, and Ferhat Attal. "Deep HMResNet model for human activity-aware robotic systems." In AAAI Fall Symposium Series. 2018.

- Abdelkawy, Hazem, S. Fiorini, A. Chibani, N. Ayari, and Yacine Amirat. "Deep CNN and Probabilistic DL Reasoning for Contextual Affordances." In AAAI Fall Symposium Series. 2018.

- Ayari, N., H. Abdelkawy, A. Chibani, and Y. Amirat. "Hybrid Model-Based Emotion Contextual Recognition for Cognitive Assistance Services." In IEEE Transactions on Cybernetics. 2020.

- Abdelkawy, Hazem, Naouel Ayari, Abdelghani Chibani, Yacine Amirat, and Ferhat Attal. "Spatio-Temporal Convolutional Networks and N-Ary Ontologies for Human Activity-Aware Robotic System." In IEEE Robotics and Automation Letters 6, 2020.