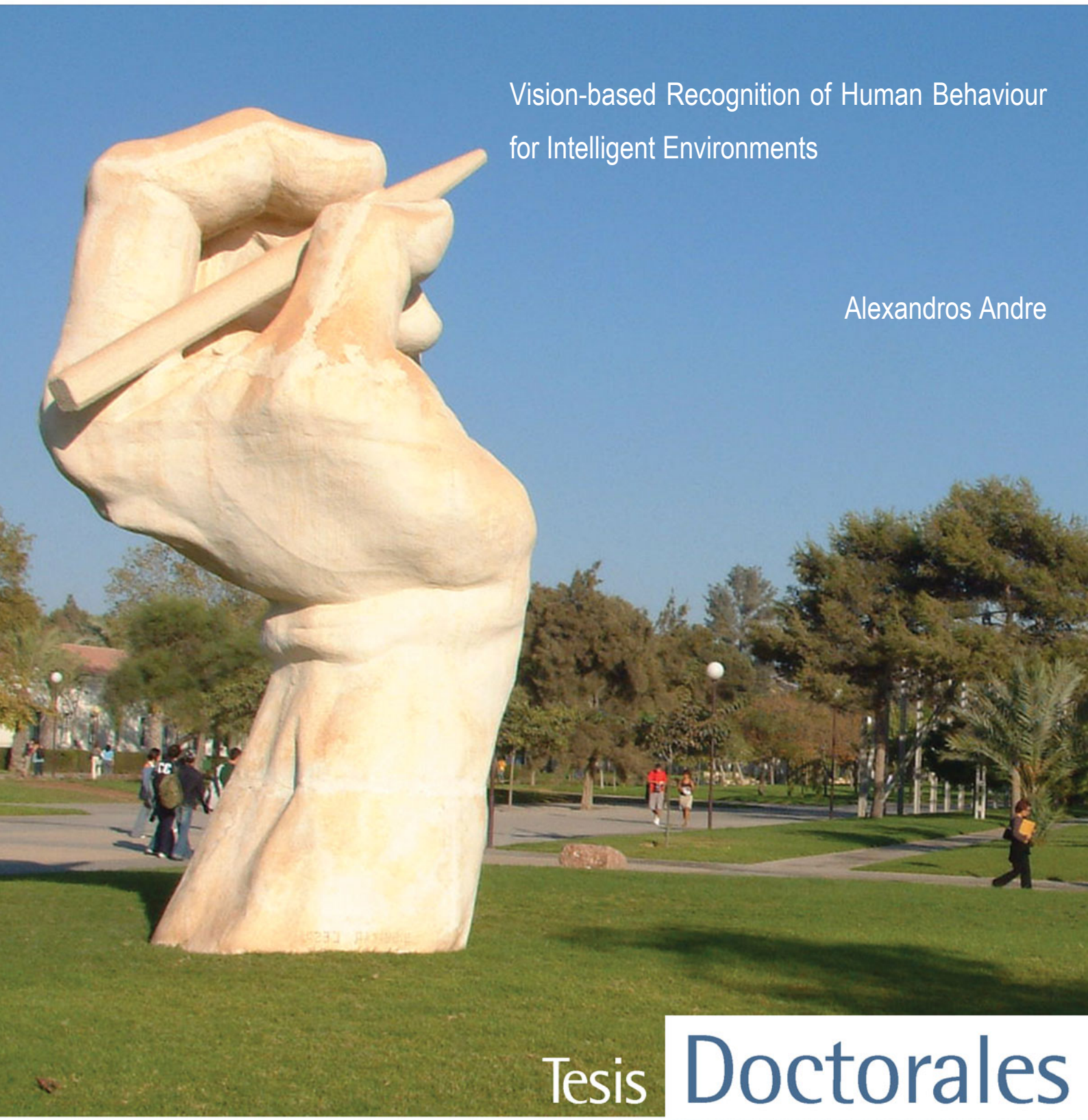
Universitat d'Alacant
Universidad de Alicante

Vision-based Recognition of Human Behaviour for Intelligent Environments

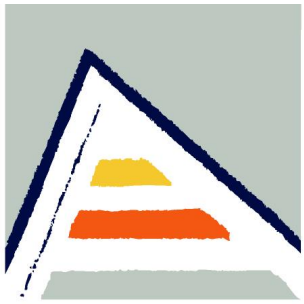
Alexandros Andre



Tesis Doctorales

www.eltallerdigital.com

UNIVERSIDAD de ALICANTE



Universidad de Alicante

Departamento de Tecnología Informática y Computación

Vision-based Recognition of Human Behaviour for Intelligent Environments

Alexandros Andre Chaaraoui

Disertación presentada para la obtención del Grado de Doctor
(Doctor of Philosophy)

Director
Dr. Francisco Flórez Revuelta

Enero 2014

This thesis has been approved
by the following reviewers for the
International PhD Honourable Mention

Dr. Jean-Christophe Nebel
(Kingston University, UK)

Dr. Jesús Martínez del Rincón
(Queen's University of Belfast, UK)

Author's contact details

Alexandros Andre Chaaraoui
alexandros@ua.es
www.alexandrosandre.com

Departamento de Tecnología Informática y Computación
Universidad de Alicante
Carretera San Vicente del Raspeig s/n
E-03690 San Vicente del Raspeig (Alicante) - Spain

Für Melania,
sie ist ein Geschenk Gottes und
mein persönlicher Glücksbringer

Agradecimientos

En primer lugar, quiero agradecerle a mi tutor, Paco, el gran esfuerzo que ha realizado. Desde que empezamos a trabajar juntos hace tres años, siempre ha estado disponible y receptivo ante cualquier duda y preocupación. Ha sabido dar las respuestas adecuadas a mis preguntas, o apuntar hacia el sitio correcto para resolverlas. Destacaría su optimismo y complicidad, cualidades que se han visto reflejadas en el trabajo diario y en los resultados alcanzados. Quiero agradecerle especialmente que no solo me haya guiado hasta este punto culminante, sino que también haya sabido contagiarme la pasión por la investigación. Me considero afortunado de haber sido su alumno.

También debo agradecerle a Juanma sus consejos y guía. Siempre ha estado dispuesto a compartir su experiencia y sabiduría como profesor y como persona. Tanto en la docencia como en la investigación, resulta muy gratificante poder aprender de él.

Este trabajo tampoco habría sido posible sin los investigadores y estudiantes del grupo de Domótica y Ambientes Inteligentes: mi amigo y colega Jose, porque han sido tantos años y parece que algo debemos de estar haciendo bien; Pau, que aunque haya dejado de ser compañero, se ha convertido en un valioso amigo; y todos con los que he compartido tantos días en metalTIC, DAI Lab y el laboratorio: Javi, Rafa, Vicente R. y Mario. No me olvido tampoco de mis compañeros de departamento: Sergio, Vicente M., Marcelo, Óscar, Felipe y Paco. Esos cafés no estarían igual de ricos sin su grata compañía.

Durante mi estancia en Kingston he podido hacer muchos amigos, y quiero agradecer a los compañeros de KU cómo me han recibido y apoyado: Paolo, JC, Dimitris, Gordon, Raphael, Reyhaneh, Spyros, James, Vicky, Matthaíos, Bastien...

Quiero aprovechar la oportunidad para dar las gracias también a todos los profesores que han impartido alguna de las clases a las que he asistido. Sin duda,

desde primaria hasta el máster en investigación, ha sido por ellos que pueda llamarme ingeniero y futuro doctor. Lo intento, aunque me dejaré a muchos en el tintero: Herr Müller, María Dolores, Juanjo, Carmen, Pepe, Joan, Merche, Carlos Villagrá, David Tomás, Eva Gómez, Antonio Botía, Patricia, Francisco Ortiz, Pablo Gil, Juan Antonio, Dani, Jorge, Jerónimo, Virgilio, Paco Maciá...

No obstante, a los que debo todo es a mis padres y a mi familia. Me han dado un apoyo incondicional y siempre han intentado enseñarme qué es importante en la vida. Aunque no hayan escrito ni una línea de este trabajo, se llevan la gran mayoría del mérito por haber estado siempre ahí de todo corazón. Espero haberme llevado algo del ingenio creador y la tenacidad de mi padre, así como del afecto y la mentalidad abierta de mi madre. Sé que mi abuelo Arie se alegra mucho por esta tesis, y sus ánimos e interés siempre han sido valiosos.

Uno es afortunado de poder seguir formándose y aprender trabajando, pero aún así, hay momentos en los que lo mejor es desconectar, y eso requiere verdaderos amigos que saben que sigues existiendo aunque lleves meses sin llamarlos. No puedo mencionarlos porque me dejaría a muchos, pero ellos saben quiénes son, ya tomemos *Weizen*, cañas o *pints*.

Dejo para el final a la más importante, Melania. No hace falta mencionar que sin ella no habría sido posible escribir esta tesis en inglés, pero no solo quiero agradecerle la revisión, sino también su amor y ayuda constante que llenan mi vida de alegría.

Este trabajo ha sido financiado parcialmente por el Ministerio de Ciencia e Innovación bajo el proyecto "Sistema de visión para la monitorización de la actividad de la vida diaria en el hogar" (TIN2010-20510-C04-02) y la beca predoctoral de la Conselleria d'Educació, Formació i Ocupació de la Generalitat Valenciana (ACIF/2011/160). Estas entidades no asumen responsabilidad alguna por la información publicada en el presente documento.

Summary

Although it can easily be overlooked, nowadays artificial intelligence can be found almost everywhere. We ask for smartness or intelligence in the technological devices we employ, and our environment is also becoming an active part that can assist us and make our daily living easier. A critical requirement for achieving ubiquity of artificial intelligence is to provide intelligent environments with the ability to recognise and understand human behaviour. If this is achieved, proactive interaction can occur and, more interestingly, a great variety of services can be developed.

Taking into account the current demographic ageing, a main concern is how to extend the independent living at home of elderly and impaired people. In ambient-assisted living, services are developed to help people at home providing care and safety through the application of ambient intelligence. By means of home automation, homes can be equipped with sensors that provide data about the current state of the environment and the people. Applying artificial intelligence techniques, patterns can then be detected leading to the recognition of human behaviour.

In this thesis, we aim to support the development of ambient-assisted living services for smart homes or senior assisted living facilities with advances in human behaviour analysis. Specifically, visual data analysis is considered in order to detect and understand human activity at home. As part of an intelligent monitoring system, single- and multi-view recognition of human actions is developed, along several optimisations and extensions.

The first contribution to the state of the art is the analysis of related work that has been carried out. A taxonomy is presented which embraces existing definitions and defines four different levels of human behaviour analysis: motion, action, activity and behaviour. This classification is then employed to categorise and detail the existing research. A learning method and a visual feature for the

recognition of human actions are presented, taking into account specific restrictions of our application scenario and focusing especially on the combination of multiple views. We define a bag-of-key-poses model and perform classification, modelling the temporal evolution of the human posture over time. Experimentation shows outstanding results in terms of recognition rate and speed.

This thesis also deals with optimisation of human action recognition. Assuming that the intelligent monitoring system is installed in a specific scenario, huge improvements can be achieved by adjusting the method to specific actions and subjects. For this purpose, evolutionary algorithms are proposed to perform selections of feature subset, training instances and parameter values. Furthermore, an evolutionary adaptive learning approach is proposed to support scenarios in which the learning needs to continue during the execution of the system. In these proposals, significant improvements are obtained leading to very efficient and robust results across different scenarios and tests.

Finally, the presented human action recognition method has been extended to support the continuous recognition of actions in video streams. Consequently, real-time and online processing is supported addressing action detection and recognition.

Therefore, this thesis proposes several contributions to perform and improve the recognition of human actions in intelligent environments, with careful consideration of specific requirements demanded by ambient-assisted living services. The present work may pave the way for more advanced human behaviour analysis techniques, such as the recognition of activities of daily living, personal routines and abnormal behaviour detection.

Contents

List of Figures

List of Tables

Part I

Chapter 1

Introduction

1.1 Motivation

> "We can only see a short distance ahead, but we can see plenty there that needs to be done." (Turing, 1950)

In 1950, Turing raised the question of whether electronic digital computers that had been invented a few years earlier were able to 'think' similarly to humans. For this purpose, he proposed the well-known imitation game in which the machine has to fool the interrogator into believing that it is a human. Over half a century has passed since, and computers have made it into everyone's homes and pockets. Nowadays, we are seeing enormous advances in the information and communication technologies (ICT), with a fast deployment in a wide variety of social layers. This has been mainly due to the reduction of cost and size, and the simultaneous increase in computational capacity. But it also has been significantly influenced by the implantation of the Internet and the recent extension of mobile technologies, which have allowed to develop an extensive amount of services and applications. Even if the artificial intelligence research community agrees that there is still a long way until machines may reach human intelligence, great success has been obtained in specific sub-problems as data mining, pattern recognition or computer vision.

Currently, an emerging issue is the development of home automation and intelligent environments. How can the application of the ICT to the user environment be able to interact with people in a way which is unobtrusive and context-aware in order to act proactively? Here, the ambient intelligence (AmI)

paradigm arises. An intelligent environment requires the use of sensors and actuators, distributed in an ubiquitous way, to be able to interact naturally, observing and interpreting the actions and intentions of people. The system has to learn and adapt to the needs of each person to improve his or her daily living (Nakashima et al., 2009). This kind of technology is not only able to serve entertainment and comfort purposes, but more importantly, people requiring certain types of assistance can benefit from intelligent environments. In ambient-assisted living (AAL), ambient intelligence is applied to the promotion and extension of independent life at home of elderly or impaired people. In this sense, AAL can give diverse type of support to ensure people's health and safety, and increase their autonomy and well-being, by means of providing services from automatic supervision of medication to intelligent monitoring (Kleinberger et al., 2007; Sun et al., 2009). These services are nowadays in great demand due to the rapidly ageing populations. The European Statistical Office projects that by 2060, the ratio between working and retired people will have passed from four-to-one to two-to-one in the EU. Taking into account that indeed EU Member States spend approximately a quarter of their GDP on social protection, a main concern is if these achievements can be maintained in the current economic and demographic context (EC, 2012). For this reason, allowing people to stay active and independent as they grow older is key to tackle the challenge of demographic ageing.

In order to support these AAL services, ICT may also be exploited to provide intelligent environments with the necessary infrastructure. Sensors are considered the base of intelligent environments. A wide range of these exists: from simple binary sensors, as motion detectors or pressure mats, to sensors that record sound or images. Based on this sensory knowledge and by means of machine learning, patterns that allow to infer human activities can be detected. This is the subject of study of human behaviour analysis (HBA). This field is currently attracting a great deal of interest, since HBA can also be applied to a wide range of other application fields, from human-computer interaction (HCI) for natural user interfaces or gaming to video surveillance. Visual data is considered to be very valuable for HBA, since it provides rich sensor information with multiple cues about people's behaviour and their environment, and at the same time, it is less invasive than body-worn sensors (Nakashima et al., 2009).

In this sense, several research projects, covering mainly empirical research, have secured public funding in this field recently. This has been seen in the AAL Joint Programme, which provided a budget of 600 million euro in the time frame from 2008 to 2013. Due to its success, a follow up to this programme has been proposed in the upcoming EU Framework Programme 'Horizon 2020' (EC, 2013). Specifically, the 'TALISMAN+' research project — System for Intelligent Monitoring and Improvement of Personal Autonomy — is funded by the Spanish Ministry of Science and Innovation (TIN2010-20510-C04-02). This project aims to provide AAL services based on scientific and technological innovation regarding the following areas: 1) sensors, 2) intelligent monitoring, 3) reasoning, 4) service management and 5) safety and security. The projects involves four research groups of different Spanish universities (Technical University of Madrid, University of Deusto, University of Castile-La Mancha and University of Alicante). Its main goal is to provide the scientific basis and technical infrastructure required to support personal autonomy of elderly and impaired people, fostering the development of the Spanish industry in this area. The Domotics and Ambient Intelligence Group of the University of Alicante is responsible for the 'vision@home' subproject, which is in charge of the vision-based monitoring and recognition of human behaviour at home taking into account privacy. Accordingly, the present work is carried out under this project.

In this regard, this thesis aims to support AAL scenarios by means of providing advances to HBA. Vision is considered as the main source of information in order to analyse and understand human activities of daily living at home. In this way, the goal is to provide intelligent environments with the capacity to detect and understand human activity, by means of taking advantage of computer vision (CV) techniques.

Last but not least, although the intrinsic limitations are not yet clear, the abilities of *seeing* and *understanding* are certainly part of those steps that have to be taken in order to come closer to the aforementioned goal of providing machines with human intelligence.

1.2 Thesis objectives

The work presented in this thesis pursues the following objectives:

(a) To study vision-based approaches for the understanding of human activities in order to join and classify existing approaches, establish a theoretical framework and identify the specific subject of study that should be addressed next to support HBA.

(b) To propose a method for the indoor monitoring of people's activities that is able to infer knowledge from visual observation. Human behaviour should be monitored in order to make possible both the detection of relevant events and the collection of statistics about the human behaviour (*e.g.* for health care monitoring or detection of abnormalities).

(c) To satisfy specific demands of AAL services. This means that scenario-specific constraints need to be taken into account, like continuous and real-time execution and the handling of multiple cameras.

(d) To reach and verify the robustness required to recognise human activities in a wide variety of circumstances, considering that these change among subjects, subjects' conditions and environments.

1.3 Structure of the thesis

This thesis is divided into two parts and eight chapters. The first part is concerned with defining the context of this thesis in terms of its theoretical framework and the application that motivates its development. Afterwards, the specific tasks and goals to be pursued are detailed. In the second part, the technical proposals and their implementation, along with experimental results and the corresponding discussions and conclusions are presented.

In chapter 2, the state of the art of vision-based HBA is reviewed. A taxonomy is defined which joins existing definitions and interpretations, and establishes different HBA levels in terms of the involved time and degree of semantics. Then, state-of-the-art works are reviewed going through these HBA levels. The available benchmarks and datasets are also summarised. Finally, conclusions are drawn.

In chapter 3, a proposal is defined for vision-based human action recognition. Specific requirements of the method to be developed are detailed, as well as the assumptions that are made. The architectural framework in which the work of this thesis is placed is explained.

In chapter 4, a method for recognition of human actions in video sequences is presented. Based on a bag-of-key-poses model, representative human poses are learnt from images provided by multiple cameras. Temporal cues are modelled using sequences of key poses. Recognition is then performed by means of sequence matching. This method allows to employ different algorithms in each of its processing steps.

In chapter 5, an implementation of the method from chapter 4 is detailed. A visual feature is presented for pose representation. This representation is based on human silhouettes and results in a very characteristic and low-dimensional feature vector. Two approaches are proposed for the learning from multiple views relying on feature and model fusion techniques. Extensive experimentation is performed and a further improved technique for feature fusion is also introduced. The presented results achieve real-time performance and exceed the state-of-the-art recognition rates. Finally, data acquisition is discussed and an alternative method based on RGB-D sensors is validated.

In chapter 6, three different optimisation techniques for the presented method are introduced. Based on evolutionary algorithms, selection of feature subset, training instances and parameter values is performed. Highly accurate results are obtained by learning the specific constraints of a classification scenario. An adaptive learning approach based on an evolving bag of key poses is also presented to support the incremental learning of previously unknown data in execution time. This approach behaves significantly better than static incremental learning.

In chapter 7, the proposed method is extended to support the recognition of continuous video streams. The concept of action zones is introduced to identify and learn the most discriminative segments of actions. These are then matched during recognition using a sliding window approach. A suitable continuous evaluation scheme is also presented, and experimental results are included.

Last but not least, chapter 8 discusses and concludes this thesis. Open issues and future research lines seen in previous chapters are summarised. For reference purposes, three complementary appendices follow. Chapter A details the characteristics of the datasets that have been employed as benchmarks. Please

consult this appendix to see sample images, a complete list of action classes and technical details about image resolution, segmentation and camera setups. In chapter B, the related publications in which the contributions of this thesis have been published are listed. Finally, in chapter C an extended summary of this work is detailed in Spanish.

Chapter 2

Related work

Human behaviour analysis is being of great interest for computer vision and artificial intelligence researchers. Its main application areas, like video surveillance, human-computer interaction and ambient intelligence, have been in great demand in recent years. This chapter provides a review on human behaviour analysis for ambient-assisted living and ageing in place focusing especially on vision techniques. First, a clearly defined taxonomy is presented in order to classify the reviewed works, which are consequently presented following a bottom-up abstraction and complexity order. At the motion level, pose and gaze estimation, as well as basic human movement recognition, are covered. Next, the mainly-used human action and activity recognition approaches are presented with examples of recent research works. Finally, by increasing the degree of semantics and the time interval involved in the recognition, the behaviour level is reached. Furthermore, existing datasets and benchmarks that are helpful for evaluation are analysed.

2.1 Introduction

Human behaviour analysis —and understanding (HBU)— involve a wide range of investigation fields, from motion detection and background extraction to expert systems and high-level abstraction behaviour models. Although both terms are used interchangeably, human behaviour understanding may be seen as a subtask of the analysis where a semantic meaning is attributed not only to what people are doing, but also to their routines and lifestyles. This chapter targets two purposes: On the one hand, existing works need to be categorised assuming a common taxonomy and a clear differentiation basis so as to appropriately define

the context of the work presented in this thesis. On the other hand, although some areas, like video surveillance, are covered thoroughly, more recent areas, like AAL and ageing in place at smart home scenarios, present a lack of unifying works and recent state-of-the-art reviews. This makes initiation in these areas difficult. Therefore, this chapter intents to ease this task.

Since the following state-of-the-art review of HBA and HBU techniques is approached from an ambient intelligence point of view, we deal especially with indoor and real-time scenarios. These techniques are therefore suitable for AAL purposes. This way, recognition of human actions and activities of daily living (ADL) covers the main interest of this chapter. Nevertheless, it is necessary to first establish the different HBA levels in order to place the aforementioned types of classification into the appropriate levels and consider the related constraints and pre-processing tasks.

To avoid common difficulties present in vision-based systems (such as occlusions, view-dependent features, lighting conditions, etc.), occasionally these systems are enhanced with other sensors. Mostly binary sensors and RFID labels are used. Therefore, although vision will be focused on mainly, other complementary sensors involved will also be discussed briefly.

The remainder of this chapter is organised as follows: Section 2.2 goes through taxonomies which are applied by other authors, and presents an abstraction-, degree-of-semantics- and time-oriented classification which is used in the rest of the review and as theoretical framework in this thesis. Section 2.3 summarises existing reviews in vision-based human behaviour analysis. Section 2.4 deals with the lowest level of HBA, *i.e.* motion, pose and gaze estimation. These elements are used as *action primitives* in section 2.5, where human actions are recognised based on image or video data and alternative or complementary data sources. Section 2.6 focuses on activity recognition methods which are of special interest in AAL: ADL in indoor environments, like *cooking* and *personal grooming*, are recognised with different approaches detailed in that section. Finally, section 2.7 deals with behaviour recognition methods that establish the highest addressed degree of abstraction. Section 2.8 summarises the available datasets which are most used in the reviewed works.

2.2 HBA taxonomies

In this section, different human behaviour analysis taxonomies from some of the most recent and relevant research works are discussed in order to point out differences and converge towards a well-defined classification.

Moeslund et al. (2006) defined an action taxonomy that has been adopted in later works and subsequent surveys. From lower to higher degree of abstraction, three levels are defined:

- Basic motion recognition derives in so called *action* or *motor primitives* representing the atomic entities out of which actions are built. Therefore, as stated in Poppe (2010), an action primitive is an atomic movement that can be described at the limb level.
- A set of different or repetitive *action primitives* make up an *action*.
- The actual *activity* can be recognised involving a larger scale of events, the context of the environment and the interacting objects or humans.

This way, when making a cup of tea, single movements of arms and hands would be *action primitives*; placing the kettle on the stove or grabbing a cup from the cupboard would be *actions*; and finally, the whole process would make up an *activity*, since different actions and interaction with several objects are involved.

Although this taxonomy is clearly defined and quite often referenced in HBA-related papers, most researchers use their own taxonomy, as usefulness depends on research goals and application areas. Since this classification is particularly focused on actions, it is difficult to adapt to higher level approaches where the main targets are ADL and behaviour analysis.

In the review paper of Aggarwal and Ryoo (2011), *activities* are categorised into four different levels: limb-level atomic movements are called *gestures*; a set of multiple temporally organised gestures performed by a single person is considered an *action*; if two or more persons and/or objects are involved, this is called an *interaction*; and when persons and objects form conceptual groups, this is classified as *group activities*. This definition is also adopted in other reviews (Borges et al., 2013; Vishwakarma and Agrawal, 2013). Nonetheless, in Borges et al. (2013) also higher levels of analysis are considered. Combining motion and appearance with contextual information provided by the scene leads to recognising *behaviours*.

So-called *interactions* are present when other subjects or objects are relevant to interpret a person's behaviour.

In the work of Wu et al. (2007), *activities* are defined as the combination of *actions* and *objects*. Whereas actions are recognised by a set of *verbs*, objects or places are recognised by a set of *nouns* which are targets of actions. Instead of recognising the actions, object recognition is tackled in order to infer human activities. Turaga et al. (2008) distinguish between actions and activities by defining that activities involve coordinated actions among a small number of people.

Regarding behaviour analysis, Ji et al. (2008) define behaviours as human motion patterns involving high-level description of actions and interactions. In contrast to Moeslund et al. (2006), dependence on the context of the environment, objects and human interaction are taken into account at the behaviour level. In Monekosso and Remagnino (2010), behaviours are understood as patterns in a sequence of observations of activities or events. Repeatable patterns and anomalies are detected by tracking activities such as *cooking*, *eating*, *watching TV* or *no detectable activities*, as well as events from the environment, which are emitted by binary sensors installed in smart homes.

Traditionally, in video analysis also the term *event* can be found (Hongeng et al., 2004; Suriani et al., 2013), even though its use is not consistent among authors. Hongeng et al. (2004) employ the term as a superclass to small scale actions (gestures, facial expressions and poses) and large scale activities. Other authors refer to activities and actions, respectively as events and sub-events. But in general, it can be agreed that the task of event recognition is more closely related to video annotation and scene analysis (involving contextual information) than to human behaviour analysis, *i.e.* it is answering the question *What is happening?*, instead of *What is the person doing?*

In this thesis, HBA tasks are classified into *motion, action, activity* and *behaviour* levels regarding the degree of semantics and the amount of time involved in the analysis. Figure 2.1 shows that both the time frame taken into account and the degree of semantics (DoS) involved in the recognition and classification process grow as we reach a higher level of the pyramid.

At the *motion* level, tasks such as movement detection, and background extraction and segmentation are faced (Hu et al., 2004; Moeslund et al., 2006; Porle et al., 2009). Using a time frame of units of frames, a lot of research has been

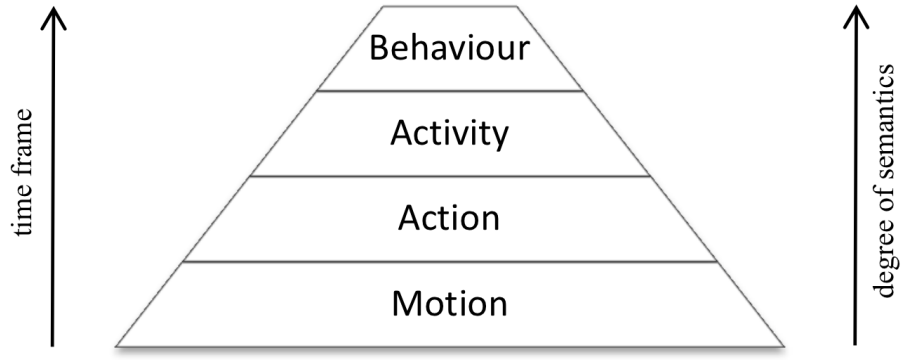


Figure 2.1: Human Behaviour Analysis levels – Classification.

done in the field of pose and gaze estimation (Launila and Sullivan, 2010; Ozturk et al., 2009; Reale et al., 2010a,b; Rybok et al., 2010; Shimizu and Poggio, 2003).

At the *action* level, human body motion is not only detected, but also recognised in order to establish what a person is doing or the objects the person is interacting with. For this reason, human behaviour understanding is performed from this level onwards. In a time frame of units of seconds, simple human activities, like *sitting*, *standing* or *walking* (Bao and Intille, 2004; Chung and Liu, 2008; Lester et al., 2006; Liu et al., 2010; Zhou et al., 2008), can be recognised. We consider *gestures* a specific type of actions which are minor in scale and usually specific to the motion of arms, hands and head.

At the *activity* level, a set of multiple actions is classified in order to understand human behaviour in a time frame from tens of seconds to units of minutes. ADL, like *cooking*, *taking a shower* or *making the bed*, are recognised, as these require tracking and classification of a sequence of actions in a particular order. This way, the sets of actions are understood as activities, where these activities are either the goals or the results of their involving human actions.

At the *behaviour* level, highly-semantic comprehension comes into play. Within a time frame ranging from days to weeks, ways of living, personal habits, and timetables and routines of ADL can be analysed. At this point, abnormal behaviours and anomalies can be detected, for instance, in order to be able to detect senile dementia prematurely (Karaman et al., 2010; Mihailidis et al., 2008, 2004).

Table 2.1 summarises the different degrees of semantics considered by the taxonomy, along with some examples. Not only the time frame and the semantic

Table 2.1: Classification of tasks according to the degree of semantics involved.

DoS	Time-lapse	Description
Motion	frames, seconds	Movement detection, background subtraction and segmentation, and gaze and head-pose estimation are performed.
Action	seconds, minutes	The objects the person is interacting with are established. Simple human primitives are recognised (sitting, standing, walking, etc.).
Activity	minutes, hours	Tasks that consist of a sequence of actions in a particular order. ADL are recognised (*e.g.* cooking, taking a shower or making the bed).
Behaviour	hours, days, ...	Highly-semantic comprehension comes into play (ways of living, personal habits, routines of ADL).

degree grow at higher levels of this hierarchy, but also the complexity and the computational cost. This leads to heavy and slow recognition systems, as each level requires most of the previous level's tasks to be done too. For this reason, level abstraction is key in order to analyse only the necessary parts and avoid redundant processes. Human tracking is the best example because it can be approached at least at the first three levels, having different tracking targets and using different kinds of features from the underlying levels. Therefore, tracking will not be discussed in this chapter on its own, but tracking approaches from the analysed works will be mentioned when significant.

2.3 Summary of related reviews

There are some reviews which analyse and describe 'human motion', 'activity recognition' or 'human behaviour understanding' (Aggarwal and Ryoo, 2011; Holte et al., 2012; Hu et al., 2004; Jaimes and Sebe, 2007; Moeslund, 2001; Moeslund et al., 2006; Poppe, 2007, 2010; Suriani et al., 2013; Wang, 2003). Naturally, earlier works review lower level techniques (*e.g.* the work by Gavrila (1999)), and later works also review further abstraction levels, approaching more to what is understood as *behaviour* by the taxonomy employed in this thesis.

Gavrila's work deals with a great amount of techniques which are aimed at providing background about the first steps that have been taken in advanced human motion analysis based on vision. Works without an explicit shape model

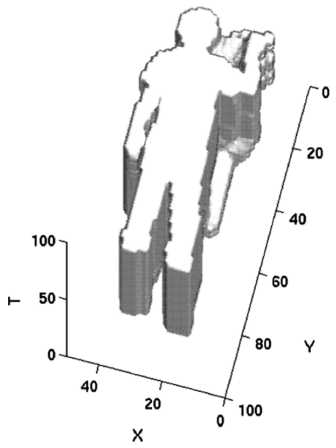


Figure 2.2: Gorelick et al. (2007) XYT volume (reprinted from Turaga et al. (2008)).

are included. These either use segmentation (as background subtraction or skin detection) in order to obtain a region of interest or aim to describe directly parts of the image by means of Haar wavelets or principal component analysis (PCA). In this case, low-level intensity features are commonly employed. When shape models are used, XYT volumes (see figure 2.2) (Gorelick et al., 2007) or stick figure models are built. Others use a 'blob finder': each blob is defined as the shirt, pants, hands and head of a person, and these are found in an image (Wren et al., 1997). There is also a subsection dedicated to 3D body modelling. When it comes to action recognition, the document presents a variety of techniques that can detect actions (at the level defined in the taxonomy employed throughout this document). Most techniques are based on dynamic time warping (DTW), as well as hidden Markov models (HMM).

In the work by Moeslund (2001), bigger emphasis is given to the recognition of motion and actions, whereas activities and behaviours are treated to a minor extent. In a later work by the same author (Moeslund et al., 2006), activity and behaviour recognition are dealt more widely. Nevertheless, the techniques revealed in Moeslund (2001) allow a greater understanding of current methods. Thus, both this and Gavrila (1999) can be used to introduce the early phases of behaviour analysis, which is widely seen as a post processing or a step to follow after prior segmentation and low-level analysis.

The work by Moeslund et al. (2006) can be divided into two major parts: pose and motion capture, and action recognition. In the first part, a series of works are introduced and three different phases are presented: 1) model initialisation,

a **b** **c**

Figure 2.3: Upper-body part detector results as stated in Ferrari *et al.*'s work (reprinted from Ferrari et al. (2008)).

2) tracking and 3) pose estimation. Model initialisation, which captures prior knowledge of a specific person in order to constrain tracking and pose estimation, is presented from different points of view. Different kinematic structures and initialisations can be used (*e.g.* a skeleton with a certain number of joints with specified degrees of freedom —DoF—). The subject's shape can be approximated with different techniques, either using simple shape primitives such as cylinders, cones, etc., or a polygonal mesh. The appearance is also considered based on the detected skin, clothing or body parts (limbs or trunk). See, for instance, figure 2.3. Similar techniques have been seen in several other works (Andriluka et al., 2009; Bourdev et al., 2010; Eichner and Ferrari, 2009; Ferrari et al., 2008; Park and Kautz, 2008; Sapp et al., 2010; Shotton et al., 2011). Figure-ground segmentation techniques are then introduced according to the used approach. Six different families are seen: 1) background subtraction; 2) motion-based segmentation; 3) appearance-based segmentation; 4) shape-based segmentation; 5) depth-based segmentation; and 6) temporal correspondences. Pose estimation techniques are introduced next, and these are classified into three groups as in Moeslund (2001), according to the presence and use of an explicit model: model-free, indirect models or direct models.

With regard to model-free approaches, research about 'body plans' and combinations of body part detectors are presented (Wu and Nevatia, 2005). More recent works include Andriluka et al. (2009); Bourdev et al. (2010); Eichner and Ferrari (2009); Ferrari et al. (2008) and Sapp et al. (2010). Furthermore, techniques that are example-based are also presented. These use either a representation of the mapping from 2D silhouette sequences in image space to skeletal motion in 3D pose space (Rosales and Sclaroff, 2000) or direct lookup of silhouette sequences for recognition (Agarwal and Triggs, 2004; Howe, 2004; Shakhnarovich et al., 2003;

Sminchisescu et al., 2005). The main drawback of this family of techniques is that when using 2D silhouettes as key poses, a constraint on the point of view is added, which usually limits recognition to that exact point of view. To overcome this, different viewpoints can be added into the database, which may lead to worse inter-class recognition. In turn, techniques based on indirect models use direct reconstruction of both model shape and motion from the visual-hull (Cheung et al., 2003; Mikić et al., 2003). Finally, some methods rely on direct models, such as multi-view 3D pose estimation using gradient descent techniques, and more recently particle filtering (a state space search reduction is used). An evolution of Moeslund's work can be seen in the papers by Bandouch et al. (2008); Beetz et al. (2010), which are commented later on. Regarding direct models, monocular 3D pose estimation and learnt motion model methods are also presented.

A more recent survey is the one by Hu et al. (2004). This paper is highly comprehensive, as it starts with motion detection and object classification and tracking, but also deals with HBU. It divides the approaches to this task based on the used techniques (DTW, finite state machines, HMM, time-delay neural networks, syntactic techniques, non-deterministic finite automata and self-organising maps). Then, it also reveals techniques aimed at describing behaviours using natural language. After that, it goes on to personal identification by means of different methods including biometrics (limb lengths, speeds, gait, height, weight, face recognition, etc.). The fusion of data from multiple cameras and future developments are also taken into account.

Multi-view recognition is also the main interest of the work by Holte et al. (2012, 2011). Action recognition approaches are divided into 2D and 3D modelling, and a very valuable comparison of results is given for the INRIA Xmas Motion Acquisition Sequences dataset (IXMAS) (Weinland et al., 2006). They conclude that 3D reconstructed data is promising, although its quality highly depends on the available foreground segmentation and possible self-occlusions, and is restricted to cases where multiple views of the same field of view are available. However, as stated in the review, this may be solved with depth sensors as time-of-flight range cameras, or the Microsoft Kinect device.

The recent works of Poppe (2010) and Aggarwal and Ryoo (2011) are also focused on vision-based action and activity recognition. Poppe makes a distinction between image representation (global, local and application specific) and action classification works (exemplar- and model-based techniques), and also in-

cludes action detection approaches. Aggarwal and Ryoo take an approach-based taxonomy and distinguish between space-time approaches (space-time volumes, trajectories or features), sequential approaches (exemplar- or state-based) and hierarchical ones (statistical, syntactic or description-based).

In comparison with these review works, the present chapter aims to provide a two-fold contribution: 1) join the different vision-based HBA levels under a well-defined taxonomy in order to bring together existing definitions and, at the same time, cover low- to high-level HBA techniques from motion to behaviour recognition; and 2) establish the necessary contextualisation of the work presented in this thesis.

2.4 Motion, pose and gaze estimation

Motion recognition is the basis for estimation of human pose and gaze direction (also referred to as focus of attention) and for further HBA tasks. Motion can be seen as a series of poses along time. The models used in Andriluka et al. (2009); Moutzouris et al. (2011); Sapp et al. (2010) assume that the human body is an articulated system of rigid segments connected by joints. In this sense, human motion is often considered as a continuous evolution of the spatial configuration of the segments or the body posture (as stated in Li et al. (2010) and exploited in Andriluka et al. (2009); Moutzouris et al. (2011); Sapp et al. (2010)). On the other hand, the gaze can either be seen as a line in the 3D space or a cone. If only the horizontal plane is employed, this can be simplified to a direction and an angle.

2.4.1 Motion and pose estimation

In the literature, very different approaches (Mcivor, 2000; Toyama et al., 1999) for foreground/background segmentation can be found, that are very popular in computer vision tasks. Such techniques are aimed at determining the position of the moving objects in a scene. Frame differencing might be the earliest of these techniques, followed by other background subtraction approaches based on a wide variety of models, such as statistical models as adaptive background mixture

models (using mixture of Gaussians) (Stauffer and Grimson, 1999), Wallflower (Toyama et al., 1999), or others (Mcivor, 2000; Toyama et al., 1999).

Segmentation algorithms, if not accompanied by further techniques, provide only very basic pose estimation (as a silhouette or area), that can be only used as information on *where* the subject is. Depending on the task to be done, this could be enough; otherwise, further estimation refinement techniques are at hand. For example, Lv and Nevatia (2007) take a single silhouette from a video, and with it, they are able to recognise actions from a set of previously learnt examples.

Other ways to determine the position of objects in a scene are based on object detection without segmentation. In the work by Dalal and Triggs (2005), descriptors (histograms of oriented gradients) are generated for image windows. These provide a means for recognition of certain shapes, like the human body. These descriptors are then fed to a classifier for training. After that, in the test stage, the classifier determines whether or not each sample in a window is of the object class. Applying this method, object presence in a scene can be confirmed. It can also be applied to determine the kind of object present in the bounding box obtained from a previously applied segmentation algorithm. In Bourdev (2011), so called 'poselets' are proposed for the detection of parts of the human body, such as a frontal or profile face, or a head-and-shoulder pose. These image patches are described based on the local configuration of key points. Each poselet is trained to recognise a specific visual pattern.

Segmentation-free pose estimation techniques are based on human models (or object models in general), like complex anthropomorphic 2D or 3D models (*i.e.* van der Meulen and Seidl (2007), used in Bandouch et al. (2008)), or approximations (cylinders or ovals (van der Meulen and Seidl, 2007), skeletons and stick figures (Carranza et al., 2003), etc.). The main difficulty that arises when no segmentation is used is how to find a way to determine where the joints of people are (Tao et al., 2007).

Nevertheless, those models can also be used with segmentation; with that coarse knowledge extracted from the previous phase where silhouettes are estimated, a finer pose estimation can be inferred. To do so, Boulay et al. (2006) propose a system which, using a single camera, is able to determine the pose from a set of poses. Using the data from the segmentation phase, they proceed to estimate 3D cues, such as depth and others. With these, and using a virtual 3D environment with the same size and camera position, they calculate how the

person might appear as a 2D silhouette depending on the pose in those exact coordinates. After that, the virtually generated poses and the actual pose are compared with a histogram comparison algorithm, which yields the most probable pose as a result. The whole workflow runs in real time (4 to 15 fps). Further work applies the described technique in an AAL scenario (Zouba et al., 2008). In this way, there are works which choose to cover the blob with an elliptical model in order to extract conclusions based on the direction of the major and minor axis of the ellipse and the length ratio between them (Nait-Charif and McKenna, 2004). For instance, this method enables to detect if the individual is *standing*, *sitting* or *lying*. As a result, the system is able to detect possible falls.

Anderson et al. present a 3D–volume approach which uses 2D silhouettes extracted from calibrated cameras to mount what is called a 'person voxel' (Anderson et al., 2009). They further apply fuzzy logic to classify three different body poses (upright, in-between and on-the-ground) and detect falls.

More recently, other works based on similar approaches have appeared (Bandouch et al., 2008; Beetz et al., 2010). These allow a very detailed pose estimation. Their method is based on the use of three cameras, instead of one, without prior knowledge of the room model (no calibration). The only requisite is that the cameras are set so that they see the object from different perspectives (ideally orthogonally, with non-overlapping views). With such a scheme, the *MeMoMan* project (Bandouch et al., 2008) has been able to determine very fine pose estimation in real time. The technique is similar to the one from Boulay et al. (2006); Zouba et al. (2008). Using the silhouettes (three in this case), a hierarchical 3D human model is applied, that 'fits' into the observed silhouettes. To reduce the dimensionality, mathematical models, which allow faster pose estimation, are used. Furthermore, the system has been applied in a kitchen scenario for ADL analysis (Beetz et al., 2010).

2.4.2 Gaze estimation

So far, only works related to body pose estimation have been discussed. In the field of AAL, other body cues are interesting too, such as gaze direction (or focus of attention), which provides further information related to what is being done in the scene (Marin-Jimenez et al., 2011). Gaze estimation can also be used to detect distraction or abnormal situations. On the other hand, different researchers treat gaze differently; it can be understood as a line in the 3D space

Figure 2.4: Focus of attention estimation of a group of people (reprinted from Canton-Ferrer et al. (2008)).

(a beam-like approach) or a cone (Canton-Ferrer et al., 2008); or, if working only in the horizontal plane (Launila and Sullivan, 2010; Ozturk et al., 2009), a direction and an angle (compass-like approach). Some of the reviewed works use gaze estimation for different purposes. Canton-Ferrer et al. (2008), for instance, use gaze estimation for attention analysis in smart classrooms or offices (see figure 2.4). Doshi and Trivedi (2010a,b) use gaze estimation to detect driving styles and distractions.

In Marin-Jimenez et al. (2011), gaze estimation is seen as an additional cue for video annotation and understanding. The authors present various techniques with different degrees of precision; by either using simple information about *left* or *right* head orientation, or yaw and pitch angles along with inferred depth (z–axis) information. For the upper-body and head detectors described, the model of Felzenszwalb et al. (2010) is used.

Head pose estimation can be understood as a complementary task for the detection of the focus of attention of a specific person, as the gaze of a person is determined to some extent by the pose of his or her head (Marin-Jimenez et al., 2011). In Launila and Sullivan (2010), different colour and shape properties are combined in order to estimate the pose of the players' heads in a soccer match. This work is part of a greater project whose final objective is to reconstruct a whole soccer match in 3D and in real time. This would make it possible to watch the match from any point of view and solve conflictive situations at refereeing. Within the field of head pose estimation, in Ozturk et al. (2009) gaze direction

is obtained in steps of 22.5 degrees in wide indoor areas such as airports and shopping centres. They develop the first solution to the head pose estimation problem using only the data proceeding of a single 2D camera. A two level particle filter made up of colour and edge histograms is used for tracking. Afterwards, the individual silhouette is matched to one of 16 patterns by using shape descriptors and scale-invariant feature transform (SIFT) points. An extensive review on head pose estimation can be found in Murphy-Chutorian and Trivedi (2009).

2.5 Action recognition

After the initial step of motion detection and pose or gaze estimation, the next step to take in order to reach the HBA level of *action recognition* consists in modelling a richer degree of semantics related to the person's motion over a time period of a few seconds. Human actions can be understood as a series of motions detected, either in the whole or in some parts of the subject's body, as arms, legs, head, etc. Such actions can be recognised based on the different body poses involved and their variation through short periods of time.

In order to understand the difference between an action and an activity, not only the time-lapse is taken into account; the objects and people involved are important too. For instance, a person manipulating an object is performing an *action* (say, opening a lid). While several of such *actions*, performed with different objects and in a specific temporal order, compose what is called an *activity* (*e.g. cooking a meal*). This section is centred in the former, and as such, it will describe the works and techniques which are either based on or aimed at recognising *actions*. In the following, human action recognition (HAR) methods which are categorised by how action modelling is approached will be seen, separating those works which focus on additional goals besides the recognition, as *view invariance*. Section 2.5.2 summarises approaches based on alternatives to regular RGB images.

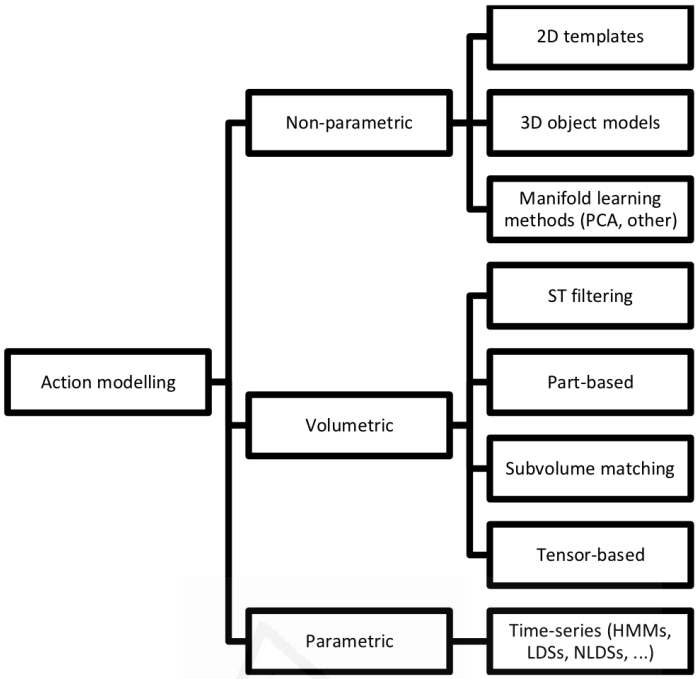


Figure 2.5: Action modelling classification of Turaga et al. (2008).

2.5.1 Human action recognition methods

Action modelling

Various categorisations have been found for action recognition methods (Li et al., 2010; Turaga et al., 2008; Weinland et al., 2007). These works classify action recognition based on the approaches used for action modelling. According to Turaga et al. (2008), approaches for action modelling can be categorised into three major classes: non-parametric, volumetric and parametric approaches (see figure 2.5). In non-parametric modelling techniques, frame-wise features are extracted out of the learning data. These are then used as templates to perform classifications. Parametric models clearly define temporal motion dynamics, learning the required parameters from the training data. In between these two options, we find volumetric approaches that consider the video as a 3D volume of which spatio-temporal data can be extracted and matched.

Among the first, 2D-template-based approaches stand out. The work by Bobick and Davis (2001) presents two mechanisms for temporal template processing: motion-energy image (MEI) and motion-history image (MHI). A sample of these is shown in figure 2.6. The MEI is a binary image which represents where motion

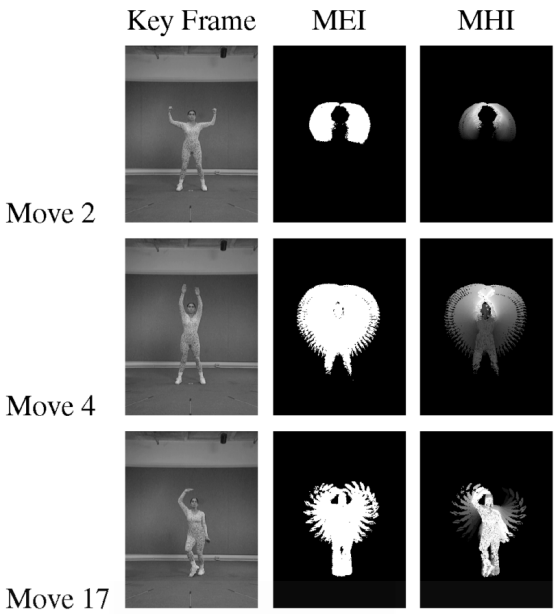


Figure 2.6: Examples of MEI and MHI images (reprinted from Bobick and Davis (2001)).

has occurred, while the MHI is a scalar-valued image where intensity of the pixels is a function of the recentness of motion. Also under this class, Ben-Arie et al. (2002) reveal a technique in which data representing the angle of the limbs is extracted from the silhouettes. Only a few representative poses are taken into account for the learning process and stored in a database. Indexing is performed so that searches of only one specific limb can be performed over the action model database. For recognition, after extracting each limb's pose, a voting scheme that returns the most likely action is used.

Cherla et al. (2008) use DTW along with a two–component feature vector, whose elements are: the width profile, calculated from the silhouette's bounding box and some spatio-temporal features, such as the displacement of the centroid (in X and Y), and the standard deviation (also in both axes). Because of the dimensionality reduction applied, actions performed sidewise (*e.g. walking*) can be detected better than actions which take place in a frontal manner (except for *hand waving* and other actions which involve sidewise movements of the limbs).

A very popular technique used among non-parametric action modelling is the bag-of-words model (BoW). Known from natural language processing, this technique relies on the frequency of appearance of a sparse collection of words. A codebook is generated that contains a set of relevant words, which are feature or

template instances, commonly obtained using clustering algorithms as K-means or K-mediods. The matching words can be found over a collection of video frames or over a single frame when dealing with a set of local features. The results of word frequencies in form of histograms can then be obtained using nearest neighbour or more sophisticated classifiers. Recently, such a technique has been extended in Jain et al. (2013). Relying on optical flow, first they separate the dominant and the residual motion in order to compensate camera motion, and then kinematic features are extracted based on local first-order differential scalar quantities computed on the flow field. These are employed to extract local descriptors over dense trajectories, as histograms of oriented gradients (HOG) and optical flow (HOF) among others. Besides a traditional bag-of-words model, a vector of local aggregated descriptors (VLAD) is employed. This technique is very similar, but instead of using the frequency of appearance of the matched words, it models the differences of each visual word to the sample. The resulting differences matrix is then classified like the BoW histograms using a linear support vector machine (SVM) with a one-against-all scheme. This approach outperforms the conventional one especially when different types of descriptors are combined.

Optical flow is also employed in Fathi and Mori (2008). This work relies on low-level features extracted from optical flow. Weighted combinations of mid-level motion features are built covering small spatio-temporal cuboids from which the low-level features are chosen. Image differencing is employed in Sadek et al. (2012) in order to extract very simple local and global motion features with respect to the centre of motion of the blob. Action classification is performed by means of one-vs-all SVM classifiers, and near real-time recognition is achieved.

3D object or pose models, which have also be seen in the last section, typically assume a kinematic model, and actions are represented in a joint or parameter space. This approach shows difficulties to estimate the pose correctly without the use of markers or other means of easing the task (see section 2.5.2). In contrast to this, in the latter approach, the use of spatio-temporal shapes as action templates reduces the burden. As it has been seen, actions are modelled using information retrieved from the images (such as silhouettes, optical flow, etc.). For recognition, a comparison between the observation–inferred templates and learnt templates is required. The limitation of this technique is that both learning and recognition need to be done under similar camera configurations. The approach proposed by Weinland et al. (2007) takes advantage of the template-based methods, and

tries to avoid the problem of consistent learning and recognition camera setups. Whereas 3D exemplars are reconstructed and learnt during training using multiple viewpoints, action recognition is performed only from 2D cues. Multiple 3D exemplars are chosen based on forward selection so as to account for different actor-related differences as body proportions and clothing.

In volumetric approaches, in contrast, features are not extracted on a frame-by-frame basis, but instead the whole video is considered as a 3D volume of pixel intensities, and standard image features are extended to deal with the 3D case (Turaga et al., 2008). For instance, Laptev (2005) proposed an spatio-temporal (ST) generalisation of the Harris interest point detector to model and recognise actions in space–time. Niebles et al. (2008) used such interest points in a bag-of-words model in order to represent actions. Clustering of the features and different classifiers, such as SVM and probabilistic graphical models can be applied afterwards. Based on similar ideas, Ryoo and Aggarwal (2009) propose a kernel function to measure similarity between pairs of videos.

Three works by Oikonomopoulos et al. (2008, 2005, 2009) describe techniques aimed at recognising basic actions. The earlier work (Oikonomopoulos et al., 2005) reveals a technique for aerobic exercises action recognition by means of a sparse representation based on spatio-temporal salient points; these are obtained using a method presented by Kadir and Brady (2003) that takes scale variance into account, among other considerations. A distance measure between two points, which is based upon the Chamfer distance, is also introduced. In Oikonomopoulos et al. (2008), the authors deal with the use of *B–splines* as a means of describing the movement of points along the time axis; these descriptors will feed a codebook afterwards, that allows further recognition to be performed. The later work (Oikonomopoulos et al., 2009) is based on ST shape model. The goal is to detect actions without prior temporal segmentation of the video stream. In the paper, a ST segmentation is performed and a voting scheme is used. The technique is robust against partial occlusions, and multiple activities can be detected if performed simultaneously.

Recently, Zhu et al. (2013) proposed a volumetric descriptor called spatio-temporal Laplacian pyramid coding (STLPC) for holistic representation of human actions. By means of a pyramid model, dense spatio-temporal features are detected at different scales. These are then further processed with a bank of 3D Gabor filters and max pooling within filter bands and ST neighbourhoods. Therefore,

structural and temporal information are simultaneously captured. A linear SVM is employed for classification. Results suggest that this approach outperforms the popular HOG3D and SIFT3D descriptors. However, as these are not holistic features, the spatio-temporal volume has to be divided into equally sized cubes, obtaining the final descriptor as the concatenation of the locally obtained ones.

Finally, parametric action modelling can be useful in order to recognise complex actions which need to be followed step by step. In this sense, the HMM, which is widely used in natural language processing (NLP), are very suitable, since the poses or action parts involved in a human action could be seen as different words that make up a sentence. Such techniques have been used extensively, for instance, in gait and sports action recognition (Liu and Sarkar, 2006; Yamato et al., 1992), also including extensions as *a priori* beliefs of state duration (Hongeng and Nevatia, 2003). Within the Prometheus (FP7) project, a work by Quintas et al. (2011) describes the use of concurrent hidden Markov models (CHMM) for the detection of ADL in a smart home environment. Although the work talks about behaviours, the events described along the paper correspond to what is classified as actions by our taxonomy. Thus, this research falls into the category treated in the current section. Similarly, in Ángeles Mendoza and Pérez de la Blanca (2007), a left-right continuous HMM with mixed Gaussian output probability is learnt for each action. A silhouette contour-based feature is extracted and used in order to consider transitions between states. Good results are achieved relying on the discriminating signal of successive contour shapes of an action cycle. Martínez-Contreras et al. (2009) use an HMM to learn the successive winning neurons that are activated in a self-organising map (SOM). The SOM is employed in order to learn the manifold of MHI templates grouping spatial and temporal information. Every new MHI pattern activates a winning neuron which is then used as discrete observation for the HMM. Conditional random fields (CRF) are used for action recognition in Song et al. (2013). They present a hierarchical sequence summarisation approach that learns discriminative feature representations at multiple temporal granularities. The CRF is used for sequence learning, whereas for sequence summarisation, observations which are similar in the latent space are grouped together. These steps are used recursively. Improvements in classification can be observed from two to four iterations.

Parametric time-series models require parameter estimation. This is usually performed by means of optimisation techniques using the training data. However,

this is computationally expensive and subject to local maxima. Furthermore, optimisation algorithms commonly rely on random initialisations, which means that several search paths need to be explored until a satisfactory cross-validation accuracy is reached. In order to avoid this, in Moghaddam and Piccardi (2013), a one-off training initialisation method for HMM is proposed and applied to HAR. Different types of action classes and visual features have been tested, achieving better accuracy than the average of multiple random initialisations.

Last but not least, considering action recognition in real-world scenarios, depending on the location and each specific situation (*i.e.* the context), not all types of actions can happen at any particular moment. Martínez del Rincón et al. (2013); Santofimia et al. (2012) make use of this thought by modelling plausible action stories based on common-sense knowledge. Both world- and domain-specific knowledge is taken into account in order to discard invalid action recognitions and consider corrections based on expected actions. The five best matching action classes are returned by the computer vision system, which is based on a BoW model of salient points and a SVM classifier. The results are then fed into the reasoning system that is based on Scone (Fahlman, 2010). This kind of techniques can also ease the recognition of an increased number of actions classes in challenging real-world scenarios where traditional vision-based systems reach their limits (Nebel et al., 2011).

View-invariant action recognition

Great advances have been achieved lately in view-invariant human motion analysis and cross-view action recognition (Ashraf et al., 2013; Holte et al., 2012; Junejo et al., 2011; Lewandowski et al., 2010; Liu et al., 2011; Lv and Nevatia, 2007; Weinland et al., 2010, 2006). Regarding the mentioned MHI and MEI of Bobick and Davis (2001), in Weinland et al. (2006) this idea has been extended to a free-viewpoint representation, *i.e.* a model generated from view-invariant motion descriptors which are obtained from different viewpoints. The feature descriptor is based on Fourier analysis in cylindrical coordinates. Pursuing a similar objective, Lv and Nevatia (2007) obtain 684 different tilt and pan angle views for each pose using the *POSER* software to generate synthetic human figures. A graph model called *action net* is employed to learn the possible transitions between poses across and among action classes. Action recognition can then be performed from any given view finding the most likely sequence of nodes within

the graph. Robustness to occlusions and viewpoint changes is targeted in Weinland et al. (2010). A 3D histogram of oriented gradients is computed for densely distributed regions and combined with temporal embedding to represent an entire video sequence. Hierarchical SVM classifiers are used for recognition. In the work of Lewandowski et al. (2010), a view- and style-invariant action descriptor is proposed based on an action manifold with a torus shape. These action manifolds represent actions independently from camera view and actor style, using bidirectional nonlinear mapping for embedded space projection and temporal Laplacian Eigenmaps for dimensionality reduction. In Junejo et al. (2011), images sequences are represented with temporal self-similarity-based descriptors. These describe the similarity between pairs of low-level features (motion trajectory points or HOG features). Interestingly, these self-similarities show to be almost view-invariant while retaining class discrimination capabilities. In other words, whereas both the features and the self-similarities change among different classes, only the features change among different views of the same class and the self-similarities remain stable. A similar concept is used in Liu et al. (2011), where view-wise BoW models are used to find shared high-level features based on the co-occurrences of pairs of words among different views. Holte et al. (2012) propose a sophisticated method based on 4D (3D and time) spatio-temporal interest points (STIP) and optical flow histograms. Histograms of optical 3D flow (HOF3D) are extracted in the neighbourhood of each 4D STIP. Different techniques like circular bin shifting are tested in order to make these features view invariant. Recognition is performed with a BoW model and a SVM classifier, achieving outstanding performance both on multi-view and cross-view recognition. In this case, a synchronised and calibrated multi-view setup is required in order to perform 3D reconstructions based on 3D mesh models. Finally, in the recent work of Ashraf et al. (2013), the human body is represented with a set of line segments. The motion of these segments is compared relying on *fundamental ratios*. These are based on the fundamental matrices that encode the correlation among pairs of frames, and are invariant to rigid transformations of the camera. Although these kind of methods commonly present an increased computational cost, they are key in scenarios with extremely unrestricted viewpoints as, for instance, in mobile robots. A more extensive review can be found in Ji et al. (2010).

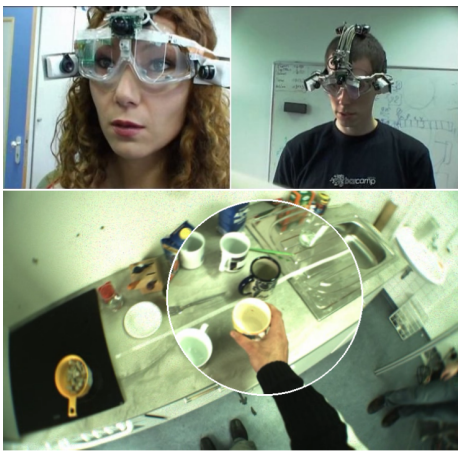


Figure 2.7: A gaze directed camera used in Sun et al. (2009); the camera itself is shown in the upper images. The lower image shows a superimposed camera view generated by the device (reprinted).

2.5.2 Using alternative or complementary data

Other works base their motion recognition techniques on other kinds of sensing, which can be more intrusive than the methods described up to this point. These are based either on other kinds of vision or other kinds of sensors.

One remarkable example is 'wearable vision', which is a semi-intrusive scheme, in which a camera is mounted on the person's shoulder, or attached to the frame of their glasses, etc. The works that use such techniques (*e.g.* Ren et al. (2009); Sun et al. (2009); Sundaram and Cuevas (2009)) emphasise how wearable cameras avoid the body to occlude what is being managed with the hands. They also point it as a more natural approach because activities are performed looking at what is being done. These head-mounted cameras allow the researchers to work with 'first–person' images, in which they see the hands of the user and the interaction with objects. The only drawback of such a scheme is that hands provoke occlusions too, just as the body does in non-wearable cameras. Moreover, cameras of such kind need to be improved, as they can be really cumbersome for the final users.

Other sensing devices are also used in diverse works (Altun and Barshan, 2010; Bao and Intille, 2004; Maurer et al., 2006; Spriggs et al., 2009; Yang et al., 2008), as well as a combination of vision and RFID tags (Wu et al., 2007) either as a direct way to recognise the objects being manipulated or as a means for supervised learning of visual object appearance (using the correspondence between the RFID

and visual data). Yang et al. (2008) use a sensor network in order to determine the position of the body by means of a network of wireless motion sensors. In Spriggs et al. (2009), data from the wearable camera is enriched with data from the inertial measurement units (IMU), which are worn by the subjects in a form of bracelets. Furthermore, Maurer et al. (2006) recognise actions by using only one bracelet (called *eWatch*) that they attach at different body parts for comparison. In Bao and Intille (2004), bi-axial accelerometers attached to different body parts are used in order to determine the wearer's ADL (mostly actions, but some complex activities too). Altun and Barshan (2010) also use inertial/magnetic sensor units (IMSU) in order to determine the actions being performed. This last work also emphasises the fact that vision and IMSU are not exclusive. It also mentions some other papers in which vision and the mentioned sensors are used in hybrid systems (Tao et al., 2007) or as a method to check the correct classification when using only data from IMSU (Aminian et al., 1999; Najafi et al., 2003; Roetenberg et al., 2007). Kwapisz et al. (2011) propose the use of tri-axial accelerometer-equipped smart phones for the same purposes. An infra-red ceiling sensor network system is employed in Tao et al. (2012) in order to detect the existence of people under the sensors. The network is taken as an array of pixels, where the activations change the pixel values. This allows to apply fall detection preserving privacy.

Li et al. (2010) take the use of other visual sensors into account. Their technique, based on 3D point clouds, deals with the recognition of actions by means of bags of three-dimensional points. The point clouds are obtained at 15 fps using the Microsoft Kinect camera, a depth sensor that acquires the depth information through structured infra-red light. The proposed method obtains a highly reduced subset of the point cloud, based on the observation that pixels in the silhouette boundary (contours) are the most relevant ones, as they carry more information on the observed body's shape. Action recognition is then performed by means of action graphs, which are described in a previous work by the same authors (Li et al., 2008).

RGB-D data, *i.e.* RGB colour information along pixel-wise depth measurement, is increasingly being used since the Microsoft Kinect device has been released. Its low-cost and straight-forward data acquisition allows to obtain reliable depth information in indoor environments regarding a range between 0.8 and 4 m, although the range from 1.2 to 3.5 m is the official recommendation for optimal results (Livingston et al., 2012). Using the depth data and relying on an interme-

diate body part recognition process, a marker-less body pose estimation in form of 3D skeletal information can be obtained in real time (Shotton et al., 2011). This kind of data results proficient for the gesture and action recognition which is required by gaming and natural user interfaces (NUI) related applications (Chen et al., 2013). In this sense, Reyes et al. (2011) proposed a DTW-based method for real-time continuous action recognition. Weights are assigned to the skeleton's joints measuring the inter-intra class gesture variability. This is similar to the work of Sempena et al. (2011), but they employ quaternions to represent the orientation of the joints. Another representation is found in Miranda et al. (2012), where joint-angles which are invariant to the sensor's orientation are obtained. Key poses are recognised with a multi-class SVM. Temporal alignment and classification is performed using a decision forest. In Chen et al. (2013); Han et al. (2013), more detailed surveys about these recently appeared depth-based methods can be found.

2.6 At activity level: activities of daily living

At this level, the goal of the recognition is to classify a sequence of actions into the targeted *activity*. For instance, some children could be moving their legs and their arms, jumping and running, and interacting with a ball. But, *what* game are they playing? At this moment, semantics come into play and understanding the compound of actions is what gives actual value to the recognition. How can we distinguish a basketball game from a volleyball game? This is the reason why the particular order of the actions and the interacting objects are key for activity recognition. Specifically, clues rely in the type of ball, the interaction with other objects as the net or the hoop, or even the fact that in basketball the players run while bouncing the ball. This could be learnt by tracking the involved actions. In this case, we are taking into account a larger time frame and a significant higher degree of semantics than in previous HBA levels.

In particular, recognising ADL in smart homes can lead to understand what a person is doing, and allows monitoring the completeness and correctness of these activities. In this matter, Mihailidis et al. (2004) track the activity of hand washing to assist older adults with dementia. Multiple successions in the process can be correct, but not of all of them; their system is able to prompt the user if a necessary step is missing or if the order of the implied actions is

unacceptable. Vision is used as the only sensor in the developed system for two purposes: 1) tracking of hand location and 2) tracking of step-specific object locations. In a previous work (Mihailidis et al., 2000), hand washing was tracked with switches and motion sensors. This way, the system could infer whether the hands were in the sink or if soap was used. Nevertheless, the authors explain that although this data was reliable, too little was known about the user and the environment. Even if the tap is running and the motion sensor indicates that the hands are in the sink, there is no guarantee that the individual is actually washing his or her hands.

Related to this type of activity recognition, Wu et al. (2007) stand out in activity recognition based on object use. As mentioned before, these authors define activities as combinations of actions and objects, and intend to recognise and track object use in order to infer human activities. Object models are acquired automatically from video, whereas object identification is based on RFID labels. At the learning phase, the user wears a RFID bracelet which reads the RFID tags attached to the surrounding objects in a home environment. Assuming that the object being moved is always the object in use and that only one object is being moved at a time, the system learns the relationship between the segmented image and the active RFID tag by using a dynamic Bayesian network. As arms and hands move with the objects, skin filtering is applied beforehand. At the test phase, the system works without the RFID data, because objects are recognised by detecting SIFT features within the segmented area. These key points are matched based on maximum likelihood to the previously trained SIFT points. As the number of possible SIFT features is very high, clustering is applied using the K–means algorithm in order to obtain a delimited histogram of SIFT features for each object.

In Zhou et al. (2008), activity recognition is approached differently. The individual silhouette is obtained at different positions of a living room. Grouped into 10–20 prototypes each silhouette stores its centre, width and height and is manually labelled with a location. A fuzzy inference method is used to estimate the most likely physical location of test silhouettes. Location estimation and previously assigned coordinates enable average speed measurement, which is used besides location in order to recognise human indoor activities. A hierarchical action decision tree is used to classify human actions using multiple levels. At the first level, human actions are classified based on location and speed. With K–

means clustering, feature patterns are obtained; and ADL, like *walking* or *visiting the bathroom*, are recognised. This is achieved by using the K–nearest neighbour (KNN) method. At the second level, a more precise recognition of activities like *washing*, *eating* or *cooking* is achieved. From the individual silhouette, the smoothed boundary of the human body is extracted using a snake model, which is represented with Hu moment invariants (HMI) (Xu and Prince, 1998). This way, the temporal variation of the HMI values are used to measure the level of body motion and instead of tracking single actions, activities are inferred based on how active the person is. The third and last level is only used when a person remains at the same physical location while he or she is moving constantly; this happens, for instance, at exercising. In this condition, further recognition is needed and primitive visual features are used. By partitioning the video frames in blocks of 8×8 pixels, motion is analysed individually at each block. This way, for a 640×480 pixels video frame, 4 800 points are taken into account to characterise low-level motion. Locally linear embedding (LLE) is applied to reduce the dimensionality of these feature vectors into a small set of composite features. These features are handled as a trajectory and matched using distance correlation and KNN classification.

When dealing with different types of sensors in smart homes, uncertainty of sensor data needs to be considered. In Hong et al. (2009), belief in sensor data is deeply analysed. When dealing with binary sensors, like motion detectors, contact switch sensors and pressure mats, the sensor data could be erroneous due to a variety of reasons. The sensor itself could be faulty; the data could be approximate, as the exact value is impossible to be measured because of the very nature of what is being measured; or the system could have corrupted the data while reading and sending it to the upper level. Dempster–Shafer's theory of evidence (Dempster, 1968; Shafer, 1976) is considered to be able to represent ignorance due to lack of information, and to aggregate belief when new evidence is obtained. Kitchen door sensors and motions sensors are used to recognise ADL, like *making a drink* (differentiating between *hot* and *cold*) or *making breakfast* (*cereals*, *toast* or *eggs*). With multivalued mapping, rules are built in order to know which elements are involved in which actions. For instance, *making a cup of tea* necessarily implies a tea bag, but milk can be optional. This way, evidence is assigned to these rules, and it is possible to infer the most likely activity given the certainty of the current sensor data.

Nicolini et al. (2010) collected data from a couple who lived at a custom built condominium for a period of 10 weeks. Several hundreds of sensors, including audio–visual recording, collected data in order to be analysed in intervals of 30 s with 15 s of overlapping between each consecutive interval. ADL like *watching TV*, *personal grooming*, *reading* and *using the phone* are recognised with multi-labelled prediction. Besides sensor data, average activity duration is used to train SVM classifiers, one for each activity. As a refinement stage, CRF (Sutton and McCallum, 2007) are applied to model sequential observations and recognise completed activities among combinations of local on-going activities. Validation was done using *leave-one-day-out* cross validation and area under the ROC curve (AUC) as a figure of merit reaching a result from 81 to 97%.

So far, we have presented vision-based activity recognition systems which use global features, like image foregrounds or individual silhouettes, or local features, *i.e.* keypoints, as salient points or corners. The field of image analysis and processing provides a wide range of image features and types of key points; these are used in diverse application areas and present different advantages and drawbacks (Juan and Gwun, 2009; Tuytelaars and Mikolajczyk, 2007). Although most popular key points techniques as SIFT and speeded up robust features (SURF) are applied frequently at activity recognition, specific keypoint-based features for this purpose are available too. Velocity history of tracked key points is used in Messing et al. (2009). Interest points are chosen based on gradient difference and tracked using a Kanade–Lucas–Tomasi (KLT) tracker (Lucas, 1981). This way, about 500 features are tracked at a time, replacing missed points on the fly. Their velocity history is used as the basic feature. Classification is based on a generative mixture model and experimentation data is presented on the KTH dataset (see Sec. 2.8.1), as well as on an own dataset where activities like *writing a phone number on a whiteboard* or *peeling a banana* are recognised (shown in figure 2.8).

Avgerinakis et al. (2013) tackle both detection and recognition of ADL for AAL solutions. By means of a KLT tracker based on optical flow, relevant video segments are detected. In this way, motionless frames or long sequences with the same activity are ignored. These video segments are then represented with a combined HOG and HOF descriptor, encoding respectively appearance and motion characteristics, and recognised with a method based on BoW. Activities that are commonly observed in an intensive care unit (ICU) are recognised

Figure 2.8: Labelled example flows from *look up in phonebook* and *eat banana* (reprinted from Messing et al. (2009)).

in Lea et al. (2012). Using depth images, people are tracked and temporal action sequences are segmented. The employed features are based on position, orientation and multi-person interaction. Two classification approaches based on SVM and decision forest are studied reaching a success rate of 75%.

In order to describe ADL consistently at several abstraction levels, Beetz et al. (2010) proposed so called "automated probabilistic models of everyday activities (AM-EvA)". AM-EvA consist of automated activity observation systems, interpretation and abstraction mechanisms for behaviour and activity data, as well as reasoning and query systems that enable AM-EvA to answer semantic questions about the activities. This way, these models have information about the involved actions and sub-actions, objects, agents who performed the activities, location and time. The purpose of these models is to build a knowledge-based framework to combine observations of human activities with a priori knowledge about actions, and to make the classification and assessment of actions and situations objective. Therefore, action patterns of different activities are learnt from several subjects in order to support singularities. The activities are observed with vision-based full-body motion tracking, RFID tags and magnetic sensors. The sensor data stream is segmented and classified with action classifiers which recognise movement primitives. This data is combined with time intervals and events, and represented in a first-order logic language. Probability distributions are represented either with Bayesian —or with Markov— logic networks. In conclusion, AM-EvA make it possible to train objective knowledge-based models to save meta-information from activities and query the following types of questions: 1) relational knowledge, like *Which is the whole pose sequence of a table setting activity?*, 2) action related concepts, like *Where is the table setting activity per-*

formed?, and 3) probabilistic knowledge, like *Having observed that a bowl has been taken and that an egg has been cracked, how likely is it that brownies are being baked?*

2.7 Human behaviour understanding

According to the taxonomy being employed, *behaviour* is understood as the highest level of complexity and time span. It is seen as a long lasting series of activities that *tend to* occur in a certain order. It could be seen as the observed person's daily routines. Deviations from the pattern can be seen as extraordinary, and as such, they give information about the person's evolution (health status or independence in the case of elderly people living alone (Monekosso and Remagnino, 2010)).

Under this definition, a number of works using vision as a source for human behaviour understanding are found, although many of the reviewed works in the field of behaviour analysis and AAL are based on other sensor devices (Cardinaux et al., 2008; Hara et al., 2002; Hayes et al., 2008; Jain et al., 2006; Mahmoud et al., 2011; Monekosso and Remagnino, 2010; Park et al., 2010; Virone and Sixsmith, 2008; Wood et al., 2008). For instance, Cardinaux et al. (2008); Jain et al. (2006); Monekosso and Remagnino (2010) monitor the lighting and the use of appliances. Cardinaux et al. (2008) rely on pressure mats; Hayes et al. (2008); Park et al. (2010); Wood et al. (2008) use basic motion detectors —such as door sensors or similar—; and Cardinaux et al. (2008); Hara et al. (2002); Jain et al. (2006) apply infrared sensors.

Behaviour of the people in the scene is seen in some of these works as the circadian activity rhythm (CAR); that is, the evolution of ADL throughout the day (Virone and Sixsmith, 2008; Wood et al., 2008). By learning the CAR of a person, either on a weekly or a 5–day basis, the system can recognise abnormalities in the behaviour as deviations from the previously observed routines.

From the works that do indeed include vision sources, context-aware systems (Brdiczka et al., 2009; Chung and Liu, 2008) and video annotation systems (Robertson and Reid, 2006; Robertson et al., 2006; Yamamoto et al., 2006) can be found. These systems can be classified under this section addressed to 'behaviour understanding', as the main point of these is not limited to determine the activities people are performing in each moment, but to extract and infer further

information from the recognised activities. The work by Brdiczka et al. (2009) deals with the recognition of situations (context). In the cited work, the authors present a system that tracks people in 3D using cameras. It is able to extract information about the entities' pose (*role*), speed, and interaction with other entities (these are either people, furniture or appliances), according to the distance to them. In this arrangement, people wear headsets, which detect if they are talking. Microphones are arranged so that ambient noise can be detected. Numeric codes are given to each possible permutation (single versus multiple people, with or without ambient noise, with or without people talking). Using these codes, which fuse all the collected information, different kinds of 'situations' are learnt and recognised by means of left–right HMM (these situations are *individual work*, *introduction [of various people to each other]*, *aperitif*, *siesta [of one individual]*, *presentation* or *[board] game [among multiple people]*).

As it has been mentioned in section 2.2, event recognition is closer related to scene analysis, although this necessarily involves human behaviour understanding, as it can be seen in the following work. Cheng et al. (2008), for instance, propose a system, that is used both for parking lot surveillance and indoor tracking, in which primitive actions are logged. The technique they present is based on histogram oriented occurrences (HO2), which is described as "a new feature that captures the interactions of all entities of interest in terms of configurations over space and time. HO2 features encapsulate entity tracks, inter-object relationships and the context of the environment into a spatial distribution that characterises the corresponding event." These new features allow easier multi-agent event recognition in contrast to simpler feature vectors which, for instance, in an HMM would require a larger number of rules or nodes. The proposed feature is based on the concept of histograms of oriented gradients, presented in (Dalal and Triggs, 2005; Lowe, 2004), and the shape context descriptor, described in (Belongie et al., 2000). After the HO2 features are calculated, they are then fed to an SVM classifier in order to detect the different event types (*arrivals*, *departures*, *trunk loading or unloading*, etc.).

In Robertson and Reid (2006); Robertson et al. (2006), a video annotation technique is presented which is able to extract the commentary of a tennis sequence. To do this, position, velocity, and action descriptions are fused and fed into an spatio-temporal action recogniser, which is in turn fed into an HMM which applies a smoothing process to the output using model-based scene knowledge.

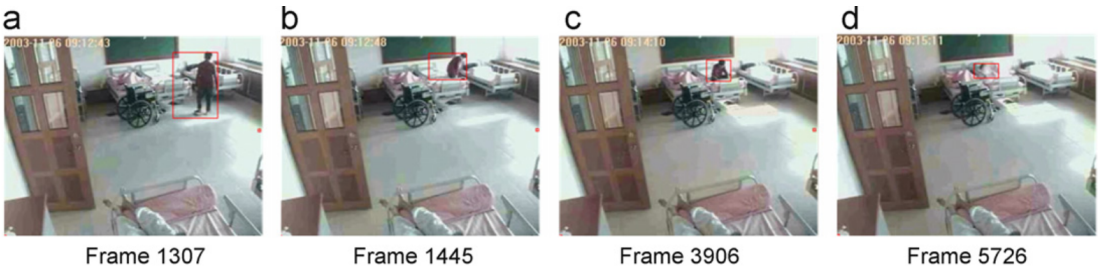


Figure 2.9: Action and activity recognition of Chung and Liu (2008) that allows further behaviour recognition. Different *actions* (a, *walking to bed*; b, *sitting on the bed*; c, *resting on bed*; d, *lying on bed*) imply an activity: *go to sleep*, which is detected by time limitation (reprinted).

This knowledge is manually modelled with the rules of the game and spatio-temporal constraints with respect to the movements of players. Their uppermost layer consists of 'behaviour HMM', which take the output of the smoothing HMM to recognise sets of activities which form a more general behaviour (*e.g. baseline-rally* or *serve-and-volley* play types in tennis). As further work, application to abnormality detection is presented.

Chung and Liu (2008) go further and propose a system which takes low-level information (such as poses, which the authors call 'activities') and combine it with other two contexts: spatial (*where* does the activity happen) and temporal (*when* does the activity happen and how long does it last). This way, using a hierarchical context hidden Markov model (HC-HMM), behaviours are learnt and later recognised from vision sources. Their techniques are applied to AAL in the context of a nursery house in which different behaviours are monitored. These include sequences of activities such as *sit down and watch TV for a while*, *go to the toilet*, *lie on bed to sleep*, *eat breakfast*, *take a walk*, etc. The normal sequences of activities are learnt from the behaviour of the monitored people in order to detect abnormality in the expected routines (either in their duration, time of day or location). In Chung and Liu (2008), results are compared using the same set of videos with other methods such as the presented in Duong et al. (2005); Nguyen et al. (2005). Duong et al. introduced what is called the switching hidden semi-Markov model (S-HSMM), which allows abnormality detection of activity duration, although the sequence of activities is restricted both in order and in the number of activities (exactly six). The space where activities happen is also restricted, as the room is divided into a discrete number of 'cells' of one square meter each, and 'hotspots' are defined according to how some appliances

and tools are arranged. Nguyen et al. present an application of the hierarchical HMM which detects three different activities, namely *have a snack*, *short meal* or *normal meal*, depending on the visited spots in each action sequence.

Other works are concerned about the recognition of behaviours that include more than one person interacting in the scene. Early works in this field include Oliver et al. (2000), in which the performance of different HMM-based techniques for the recognition of interactions among two people is compared, and it is concluded that coupled HMM (CHMM) perform better for the task. The work introduces an interesting agent framework for the synthesis of artificial behaviours that are used for training along with real-life video data. CHMM were introduced in Brand (1996) as a better approach for interaction modelling. Hongeng and Nevatia (2001) present a hierarchy of events along with a method for interaction modelling and recognition. Modelling is performed by means of action threads (associated to each actor/agent) and temporal constraints. Recognition is achieved by propagating these constraints and likelihoods in a temporal logic network (TLN).

A more recent work of Liu et al. (2010) extends the methods presented in Chung and Liu (2008) for the recognition of behaviours present in pairs or groups of people. By recognising the number of people present in the scene, switching between two different HMM-based approaches is performed by a switching module. Either individual duration HMM (IDHMM) or interaction–coupled duration HMM (ICDHMM) is used as a consequence. The methods are as well applied in AAL environments; in this case, in nursing homes.

Although vision is not used in Munstermann et al. (2012), their work shows a very well-defined approach on how recognition of ADL can be tracked over days in order to analyse the human behaviour. The personalised activity schedule of each person is modelled as daily habits in which anomalous behaviours can be detected. The daily behaviour regarding health-related ADL is then logged for the caregiver. Further tools regarding human autonomy assessment are presented.

In Martinez and Stiefelhagen (2013), a vision-based system for complementary monitoring of an ICU is proposed. Among other applications, sleep monitoring is considered based on *face agitation* and *bed occupancy* indicators. For this purpose, infra-red images are obtained from a stereo, a depth and a high-resolution camera. The last one is specifically employed for face analysis. A basic approach for behaviour recognition based on Bayesian surprise is employed in order to detect

salient events of these indicators. Nonetheless, the obtained dataset is planned to be published, which would enable comparisons and the development of more advanced techniques.

2.8 Datasets and benchmarks

Once a human behaviour analysis method as the ones described in the previous sections has been designed and implemented, an appropriate dataset is needed for the evaluation. Several technical choices have to be made, but more importantly, data with the desired HBA level from *motion* to *behaviour* needs to be recorded. In this sense, publicly available datasets are very useful tools for researchers, because they can be used as benchmarks in order to compare recognition rates, speeds and robustness between different approaches.

This section will present the most popular publicly available datasets in the field of HBA, revealing technical details and constraints, in order to ease the choice among them.

2.8.1 Datasets

When developing a new recognition system or improving an existing one, the datasets to test need to be chosen carefully. Dataset properties vary widely, and overfitting at model training can lead to illogical results. In the field of HBA, the following video datasets stand out:

- **KTH human motion dataset** (Schuldt et al., 2004): This *action* database contains six types of human actions performed by 25 subjects at four different scenarios. *Walking*, *jogging*, *running*, *boxing*, *hand waving* or *hand clapping* are performed in over 2 000 sequences. Backgrounds are homogeneous and free of clutter. Video files are classified by actions, so that unwanted actions can be excluded easily. Background segmentation is manageable with this type of images; and annotated actions can be placed at the same semantic abstraction level.

- **Weizmann human action dataset** (Gorelick et al., 2007): The Weizmann dataset constitutes the most popular dataset for human action recognition. Gorelick et al. used static front–side cameras to record single human motion from nine subjects in different environments. About 340 MB

of video sequences are available; the ten performed *actions* include *walking*, *running*, *bending*, *hand waving* and different types of *jumping*. The corresponding background sequences, with no subjects, and the subtraction masks —either with post-aligning or without it— are available too.

- **INRIA Xmas motion acquisition sequences** (Weinland et al., 2006): This dataset includes 390 × 291 px video images recorded from five different angles. 11 actors performed 14 *actions*: *check watch*, *cross arms*, *scratch head*, *sit down*, *get up*, *turn around*, *walk*, *wave*, *punch*, *kick*, *point*, *pick up*, *throw over head* and *throw from bottom up*. These actions were performed three times each, in any location and orientation in the field of view. For this reason, this dataset is being employed for multi-view and view-invariant action recognition systems. Backgrounds and illumination settings are static and free of clutter.

- **DHA–Depth-included human action video dataset** (Lin et al., 2012): The aim of this dataset is to provide depth data for actions like the ones from the popular Weizmann and KTH datasets. A total of 23 action classes, including the ten from the Weizmann dataset, have been performed by 21 actors (12 female and 9 male). RGB-D data of a resolution of 640 × 480 px is provided for a single front view. The usage of different subsets of respectively 10, 17 or all the 23 actions, is proposed. Not only general body motions are considered, but also game-oriented actions. Besides the ten actions from the Weizmann dataset, the following ones are available: *front clap*, *arm swing*, *leg kick*, *rod swing*, *side box*, *side clap*, *arm curl*, *leg curl*, *golf swing*, *front box*, *taichi*, *pitch* and *kick*.

- **TUM kitchen dataset** (Tenorth et al., 2009): This dataset targets ADL at a kitchen scenario at a low *action* level. *Table setting* is performed by several subjects in different ways. Some subjects transport items one by one, and others behave natural, grasping several objects at once. Video images, with a resolution of 384 × 288 px captured at 25 fps, and motion capture data, extracted with a marker-less full-body tracker, are provided. Furthermore, RFID tag readings from fixed readers (at placemat, napkin, plate and cup) and sensor data from magnetic sensors (at doors and drawers) are available. Each frame has manually been labelled separately for the left hand, the right hand and the trunk of the person. Among others, actions like *carrying an*

object, *standing still*, *reaching*, *walking*, *taking something*, or *closing a door* are included.

- **HOHA - Hollywood human actions** (Laptev et al., 2008): This dataset contains video sequences from 32 movies with annotations of eight types of *actions*: *AnswerPhone*, *GetOutCar*, *HandShake*, *HugPerson*, *Kiss*, *SitDown*, *SitUp* and *StandUp*. Training and testing sets are provided, as well as an automatically labelled training set with approximately 60% correct labels. A second version is available with about 1 200 minutes of video and four new actions in addition to the existing ones: *DriveCar*, *Eat*, *Fight* and *Run*. As video clips are taken from movies, persons in the images are mainly focused and background changes are frequent. Therefore, this dataset is very useful and challenging. Nevertheless, it should not be forgotten that this type of images is difficult to obtain with regular surveillance cameras.

- **MuHAVi dataset** (Singh et al., 2010): By targeting silhouette-based multi-view human action recognition methods, this dataset includes video data obtained from multiple cameras. Images are taken with night street light illumination at a constant but uneven background. At each corner and each side of a rectangular platform, a *Schwan* CCTV camera is installed. These cameras captured, according to our taxonomy, 16 different *actions* (*WalkTurnBack*, *RunStop*, *Punch Kick*, *hotGunCollapse*, *PullHeavyObject*, *PickupThrowObject*, *WalkFall*, *LookInCar*, *CrawlOnKnees*, *WaveArms*, *JumpOverFence*, *DrunkWalk*, *ClimbLadder*, *SmashObject*, *JumpOverGap*) and one *activity* (*DrawGraffiti*) performed by 7 actors, three times each. Each frame has a 720×576 px resolution and is taken at 25 fps. However, silhouettes are annotated only at a small sub-set of the available video data.

- **UCF sport action datasets** (Rodriguez et al., 2008): Among other datasets available at UCF, this dataset stands out because it contains nearly 200 video sequences at a resolution of 720×480 px. Images are intentionally taken from real scenarios (usually from broadcast television channels), as performances recorded on purpose from actors could lead to unrealistic and laboratory-conditioned training data. On the contrary, images taken from sport broadcasting or from *Youtube*, as it happens at the UCF50 dataset, present large variations in camera motion, object appearance and

scale, viewpoint, clutter and illumination settings. These are therefore very challenging. Considering our taxonomy of HBA levels, this dataset does not only include *actions* (*walking*, *swinging*, *running*, *golf swinging*, *kicking*, *lifting*), but also activities (*diving*, *horseback riding*, *skating*).

- **CAVIAR test scenarios**: The CAVIAR project (CAVIAR Project, 2004) also published its database. Its images are taken in two different scenarios: an entrance lobby and a shopping centre. *Activities* of real scenarios are recorded (*walking alone*, *meeting other people*, *window shopping*, *entering and exiting shops*, *fighting*, *passing out* and *leaving a package in a public place*) at a resolution of 384 × 288 px. Ground–truth data is provided in XML format at frame level. Video sequences, taken from wide angle cameras installed as surveillance cameras at the ceiling corners, include several people, as well as crowd movements.

- **CMU–MMAC database** (De la Torre Frade et al., 2008): The multimodal activity database from the Carnegie Mellon University targets cooking and food preparation *activities*. Not only video data has been taken, but also audio and other sensor data (motion, accelerometers and gyroscopes). Five subjects were recorded in a kitchen while preparing five different recipes: *brownies*, *pizza*, *sandwich*, *salad* and *scrambled eggs*. Video images were taken from three high spatial resolution cameras (1024×768 px) at low temporal resolution (30 fps) and three low spatial cameras (640×480 px), two at high temporal resolution (60 fps), and a wearable one at low temporal resolution (12 fps). Audio data was recorded with five balanced microphones and a wearable watch. Motion was captured with 12 infrared cameras of 4 MP at 120 fps. Five 3–axis accelerometers and gyroscopes contributed to the rest of the data. The computers used to record the sensor data were synchronised using the network time protocol (NTP).

- **PlaceLab datasets** (Intille et al., 2006): The PlaceLab live-in laboratory provides a full home-like environment for data gathering for ubiquitous technologies and home settings studies. Two datasets are available: whereas PLIA1 is a legacy dataset, PLIA2 improves data sharing and visualisation by employing new data formats. This second dataset is also compatible with their visualisation and annotation tool called Handlense. PLIA2 includes 4 hours of video data (infra-red and RGB), in which one

Table 2.2: Comparison of dataset characteristics (from lower to higher DoS).

Dataset	DoS	#Classes	Multi-view	Reso-lution	Back-ground	Silhou-ettes	Out-/ Indoor
KTH	Actions	6	No	160×120	simple	No	*both*
Weizmann	Actions	10	No	180×144	simple	Yes	outdoor
INRIA-XMAS	Actions	14	Yes	390×291	simple	Yes	indoor
DHA	Actions	23	No	640×480	simple	Yes[a]	indoor
TUM Kitchen	Actions	10[b]	Yes	780×582	simple	No	indoor
HOHA	Actions	8/12	No	240 lines	complex	No	*both*
MuHAVi	*both*	17	Yes	720×576	complex	Yes[c]	indoor
UCF Sports	*both*	9	No	720×480	complex	No	*both*
CAVIAR	Activities	6	Yes	384×288	complex	No	indoor
CMU-MMAC	Activities	5	Yes	1024×768	simple	No	indoor
PlaceLab (PLIA2)	Activities	6	Yes	320×240	simple	No	indoor

[a] Depth data is provided.
[b] Approximately 10 annotated sub-actions of one activity: *setting the table*.
[c] They are provided in the Manually–Annotated Subset (MAS).

subject performs common household *activities* (*preparing a recipe*, *doing a load of dishes*, *cleaning the kitchen*, *doing laundry*, *making the bed*, and *light cleaning around the apartment*). Besides video data, while performing the activities, accelerometer data is recorded by so called MITes, which are attached to objects of interest (*i.e.* objects which are related to human activities) as remote controls, chairs, etc. Videos are annotated not only with the type of activity, but also with body posture, location and social context.

Table 2.2 summarises the characteristics of the reviewed datasets. For further datasets and details we refer to the recent review of Chaquet et al. (2013).

2.9 Remarks

This chapter has reviewed the different levels of HBA following an abstraction-, degree-of-semantics- and time-oriented classification. Going through recent examples of research works, the most used and promising feature types, recognition methods and system design methodologies have been detailed. From motion, pose and gaze estimation to behaviour recognition, and following an initially defined taxonomy, we have analysed vision and multi-modal-based approaches.

Clearly, it can be seen that at the motion, pose and gaze estimation level, several methods achieve high and robust success rates. Relying on tracking and/or body pose recognition, human actions can be recognised. It can be observed that human action recognition shows the most advanced techniques and methods, as researchers put great emphasis on attempting to resolve this HBA level which, due to its wide variety of applications, sparks great interest. Nevertheless, at higher levels, especially at behaviour, there is still a long way to go to achieve off-the-shelf products. Still, huge advances have been made in the last ten years. But the challenge to design and develop stable and general systems persists, as most systems only solve specific problems in very particular environments.

Especially at the field of AAL, advances in these works are very valuable since these can enormously improve personal autonomy and quality of life for elderly and cognitively impaired people, and at the same time significantly reduce care costs.

Chapter 3

Proposal

In chapter 2, the background of this thesis has been detailed. Having performed an analysis of the research field, we have been able to detect the current core subject of research of the field and the main difficulties that are being experienced. This subject is the detection and recognition of human *actions*. At present, this level of HBA is attracting the greatest interest due to its wide range of applications. It has also been seen that the level underneath, *i.e.* motion and pose estimation, is able to provide reliable information under more or less expected circumstances. On the other hand, the reviewed methods that belong to the HBA levels above actions, *i.e.* activities and behaviour, built directly upon motion and low-level sensor information, instead of taking advantage of the semantic value of action recognition. For this reason, reliable and efficient action recognition may fill this gap and ease the development of higher-level HBA.

Based on the result of motion and pose recognition, we can tackle the next step in order to classify the human motion and recognise its semantic meaning. Recently, approaches have been presented which are able to recognise certain type of actions. Mostly single-view setups are used and the issue of real-time execution has been mentioned sparingly. Furthermore, current approaches seem to work well only on specific actions or datasets. In this sense, advances in this field will be proposed to improve the robustness and adaptability to scenario-specific constraints of human action recognition. This will allow to pave the way for further HBA, like ADL recognition, since activities have been defined in section 2.2 as sequences of actions in particular orders. A specific proposal that takes into account the needs of an intelligent environment providing AAL services

is made below. Furthermore, the architectural design of the system in which this proposal will have to be deployed is detailed.

3.1 Vision-based human action recognition

Recognising human actions like *walking*, *sitting* or *falling* at people's homes can be very valuable to support a great variety of AAL services. As it has been mentioned in chapter 1, an intelligent environment needs to be aware of the events that are happening in order to assist proactively. This assistance can be of any kind, from closing a running tap to firing an alarm or calling the health care services. Not only short-term events are interesting, but also the long-term behaviour of the person. Action recognition allows to collect information about the daily living of people which can make it possible to learn routines and recognise abnormal situations, thus allowing early risk detection of health issues.

In order to perform human action recognition, vision is probably the most valuable sensor information that can be employed. Video data provides information about pose and motion along time. This allows us as humans to recognise if someone is *sitting* or *standing*, but also if interactions with other objects or people occur. Not to mention that we can even realise if the person is in a hurry, or if he or she is angry or upset, just by seeing how these actions are carried out. Therefore, visual data provides a complex and extensive set of cues related to the human behaviour that is recorded. However, the analysis of video, and more specifically of visual data of human behaviour, comes along with a few difficult hurdles. There exists a great variability related to how humans move and perform actions, and this information is observed differently depending on the capturing conditions. Although this is not perceived by a human observer, a machine can have significant difficulties in distinguishing the ways people walk or run, or when an action started and ended (Kellokumpu, 2011).

Nowadays, most application scenarios have more than one camera available. Due to the reduction of costs and the increase of popularity of outdoor and indoor cameras, there are commonly several cameras installed covering the same field of view. Especially in human action recognition, one camera can be insufficient due to partial occlusions (objects like furniture could be in the way, but also other persons) and ambiguous or unfavourable viewing angles. Several video streams can be analysed and multi-view representations can be modelled to improve ac-

tion recognition. However, this tasks still has to overcome great difficulties, as dealing with multi-view data leads to high computational cost and burdensome systems (Holte et al., 2012; Moeslund et al., 2006). The main reasons for this situation are: 1) the additional increase of difficulty in learning from multiple views, because the combination of multi-view data results in a greater data variance and complex learning models; and 2) the involved decrease of recognition speed, since at least two views need to be processed and analysed (or chosen from).

On the one hand, in this thesis, we aim to overcome these general challenges in order to provide a more robust method for HAR. On the other hand, we also consider specific constraints related to our particular application.

3.1.1 Requirements

The specified technical objectives of this proposal can be summarised as follows: we want to perform efficient and robust vision-based human action recognition from multiple views taking into account AAL scenarios. In order to accomplish this goal the following requirements have to be considered:

Recognition of a wide variety of actions The proposed method should be able to recognise motion-based human actions like *walking*, *running*, *falling*, *bending*, *standing up*, *sitting down*, etc. Furthermore, a way should be provided to customise the system to the recognition of specific actions regarding the needs of the application.

Support of scenario-related circumstances As it has been mentioned earlier, the way actions are performed and recorded depend on the subjects, their condition and on the environment. Therefore, the method will have to prove its robustness to a variety of scenarios and subjects of different age and gender. In this sense, the learning of these scenario-specific constraints should be supported by adapting the configuration of the method so as to perform better if these conditions are known and stable.

Learning from multiple views The system should be able to learn from multiple views if these are available. This means that both single- and multi-view recognition have to be supported, and that the system should be able to take advantage of further views in order to improve the recognition. Furthermore, since training and testing setups may change, it has to be

considered that a view may not be available during the recognition. In addition, in homes and environments in general it is very difficult to install specific camera setups. Therefore, relaxed multi-view requirements are needed. Camera calibration, specific viewing angles and a constant number of views cannot be guaranteed.

Real-time execution Since the method will be deployed in a person's home for real-time monitoring, the design of the method needs to take into account that the sum of all processing steps should run above video frequency (25 to 30 fps), so that the recognition does not imply any significant delay in the video processing and events can be detected *on the fly*. Moreover, since the method is proposed as a step towards further analysis of higher HBA levels in a hierarchical fashion, it should perform as fast as possible. Otherwise, this would lead to a significant handicap. Note also that the system could be deployed in an embedded hardware architecture, leading to limited processing power and memory availability.

Continuous recognition It has been seen that action recognition is commonly handled as the classification of previously segmented video sequences that contain human actions. In real-world applications as intelligent monitoring, such sequences are not provided. Therefore, the method also needs to support the continuous recognition of video streams and perform action detection and recognition.

Adaptive learning In certain situations, the learning process has to continue while the system is deployed and executing. Since conditions related to the scenario, the actors and the actions that have to be recognised can change during the execution, a dynamic method should be provided. In this way, the learning does not have to start from scratch and adaptive and incremental learning can be handled.

3.1.2 Assumptions

In order to allow this work to place the main focus on human action recognition, some assumptions are made. Specifically, regarding visual tracking, the presented work builds upon the premises stated below.

As it has been mentioned in chapter 2, tasks like motion tracking and human detection and identification are wide research fields on their own. These tasks are therefore out of the scope of this thesis. Great advances have been made in the last two decades to solve these issues, and this effort still continues (Chen, 2012; Yang et al., 2011). In the following technical proposals, it will be assumed that: 1) a single person is being monitored; 2) the person in the image is the one that has to be monitored; and 3) the given image recorded a sufficient part of the human body to distinguish the posture throughout the whole course of the action. This means that tasks as cross-camera tracking and person re-identification are not dealt with in this work. Notwithstanding, the method needs to able to benefit from these techniques if used as preprocessing stages in order to relax these assumptions.

Please note that other assumptions exist, but these depend on the data acquisition and the employed technical setup. For instance, image quality and resolution, as well as the distance to the subject, are relevant for background subtraction, although specific requirements depend on the used method. Furthermore, several of the related difficulties can be solved with more advanced devices that allow human detection and depth-based segmentation, like RGB-D sensors.

3.2 Architectural design

As it has been mentioned in chapter 1, the present thesis is part of a greater research work that is being carried out in the University of Alicante. In the 'TALISMAN+' research project, a vision-based system for the monitoring of people's daily living at home is being developed. This system relies on a multi-view setup of cameras installed in people's homes in order to apply single- and multi-view human behaviour analysis. At higher processing levels, the visual information is fused with the information provided from environmental sensors (motion detectors, thermometer, gas, fire and flooding sensors, etc.). A reasoning system then decides if the detected activity corresponds to any of the monitored events

and acts accordingly, for instance, firing an alarm or enabling certain actuators. Furthermore, it also considers ethical issues in order to preserve the privacy of the inhabitants by means of image post-processing and permission management, among other techniques. In this way, a caregiver can be informed of the event that occurred if this is necessary. This may be a professional tele-assistance service that is directly connected to the home, or an informal caregiver (*e.g.* a relative) that can be informed with a text message. By means of the privacy layer, the appropriate textual or visual information can be shared depending on the permissions of the observer and the risk involved in the detected event.

The results of the monitoring are logged 24 hours a day, making it therefore possible to perform long-term analysis of the human behaviour in order to apply early risk detection. For instance, early detection of senile dementia may be performed based on long-term analysis of human gait and falls (Nakamura et al., 1996) or by means of sleep monitoring (Weldemichael and Grossberg, 2010). Figure 3.1 shows a diagram of the complete architecture of the system, where the parts that belong to this thesis are highlighted.

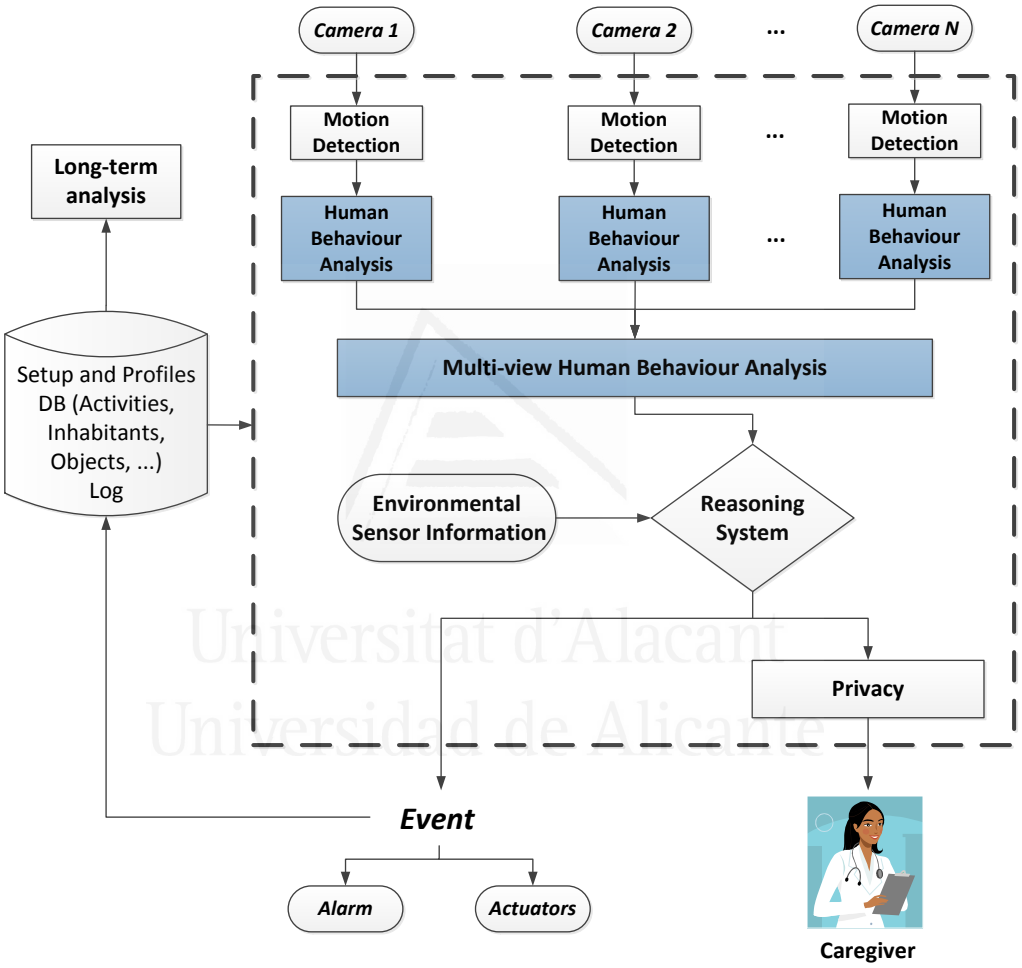


Figure 3.1: Architecture of the intelligent monitoring system to promote independent living at home and support AAL services. The parts that belong to this thesis are shaded in blue.

Part II

Chapter 4

A method based on a bag-of-key-poses model and multiple views

As it has been introduced in the last chapter, our proposal relies on taking advantage of the rich sensor information that cameras provide in order to recognise human actions in the daily living of people at home. For this purpose, a new classification method is presented for the learning and recognition of human actions based on vision. The method is especially conceived to establish a classification framework and support different algorithmic design decisions. In this chapter, the general outline of the method laid out in pose representation, learning and recognition stages is detailed.

4.1 Pose representation

At the pose representation stage, visual cues of the image are extracted, and then employed during the classification in order to detect and distinguish between actions. As it has been seen in chapter 2, *global* (also known as *dense* or *holistic*) and *local* (also known as *sparse*) representations of images can be obtained. The first require a region of interest and therefore the human body needs to be detected in the image, usually with background subtraction and blob extraction techniques. While this additional step of pre-processing is a disadvantage, it is overcome by the significant reduction of both image size and inherent complexity of its content.

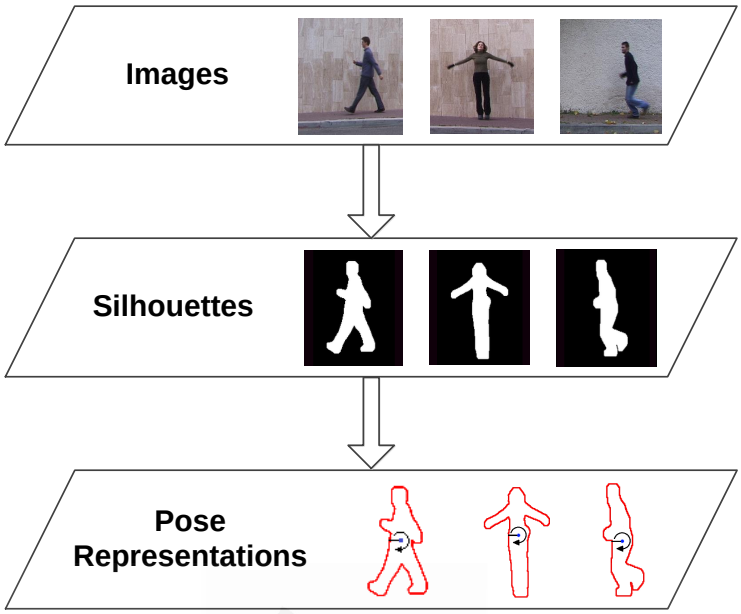


Figure 4.1: Outline of the pose representation process. Based on the recorded video frames, foreground segmentations are obtained. Holistic features can then be extracted relying on the shape of the human silhouettes.

When using local representations, the image is regularly taken as it is and observed as a collection of patches or points. Commonly, different types of salient points are obtained based on shape and gradient changes (like Harris and SUSAN corners, SIFT and SURF points; see Juan and Gwun (2009); Wu et al. (2010) for more details). When considering the temporal evolution of the location or aspect of these points, space-time corners are applied. These encode 3D information of interest points "where the local neighbourhood has a significant variation in both the spatial and the temporal domain" (Poppe, 2010).

In this method, a spatial pose representation is proposed because temporal cues will be considered by the method at a higher learning level (see section 4.2.5). This pose representation can either be given by a holistic feature, which encodes the shape or pose of the human body, or a single compact representation of dense features extracted over the whole image, which can be obtained based on dimensionality reduction or histogram-based techniques.

For instance, as shown in figure 4.1, using RGB colour images, background subtraction can be applied in order to obtain the foreground and extract the blob that corresponds to the human body. Based on this 2D shape, different pose representations may be obtained. Another option is to employ a 3D body pose

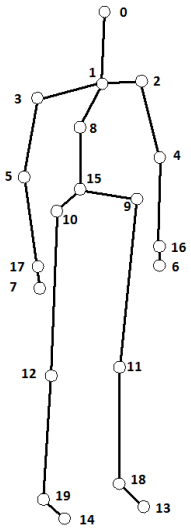


Figure 4.2: Skeleton provided by the Microsoft Kinect SDK relying on the depth data provided by the Microsoft Kinect.

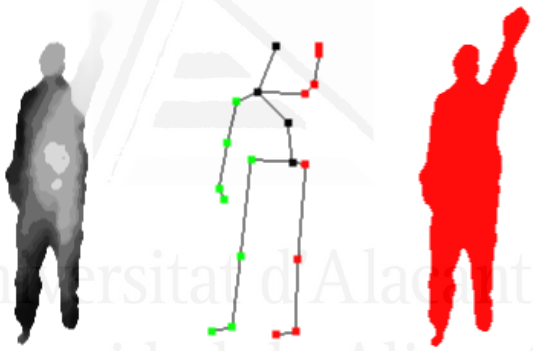


Figure 4.3: Data captured with the Microsoft Kinect device. From left to right, the depth information, the corresponding skeleton model, and the silhouette obtained with depth-based segmentation are shown.

estimation, like the one that can be obtained from depth data. Using a depth sensor, as the Microsoft Kinect device, a marker-less body pose estimation can be computed based on body part recognition (Shotton et al., 2011). Specifically, a skeleton model consisting of a set of joints is generated (see figure 4.2 and 4.3).

4.2 Learning

This section presents the learning stage of the proposed human action recognition method. As is common in machine learning algorithms, training data, *i.e.* previously recorded and labelled video sequences are used as references in order

to learn how human actions look like, and retain the information that is needed to perform a recognition later on.

This learning stage relies on the chosen pose representation to encode or extract important information from a specific video frame of a person which is performing an action. Following the order of processing, first, key poses are proposed so as to learn the most representative poses involved in the motion of human actions. Second, a bag-of-key-poses model is introduced for the learning of key poses from different action classes. Third, fusion of multiple views is performed in order to learn multi-view pose representations or consider multiple views during learning. Fourth, the discriminative value of the key poses is obtained. Fifth and last, temporal cues are learnt by means of sequences of key poses.

4.2.1 Key poses

Lately, several works (Baysal et al., 2010; Cheema et al., 2011; Eweiwi et al., 2011; Thurau and Hlaváč, 2007) have been built upon key poses. Baysal et al. (2010) define key poses as "a set of frames that uniquely distinguishes an action from others". Therefore, the goal of using key poses is to model an action based on its most characteristic poses in time. This makes it possible to significantly reduce the problem scale in exemplar-based recognition methods and, at the same time, to avoid redundant or superfluous learning (Lv and Nevatia, 2007). The underlying idea is that if the human brain is able to recognise what a person is doing seeing a few indicative poses (Giese and Poggio, 2003), action recognition methods may be able to sustain only on pose information.

In this sense, key poses can be seen as the most representative pose representations out of the sample distribution of the feature space. In order to obtain these key poses, different techniques like clustering, vector quantisation and codebook generation algorithms are suitable. For example, in Jurie and Triggs (2005), over-adaptation of clustering algorithms is studied, and a scalable clusterer based on acceptance-radius is proposed for codebook generation. Nonetheless, key poses could also be picked manually by labelling the key frames of the video sequences and extracting their corresponding pose representation. However, the criteria of choice would certainly be subjective and difficult to replicate.

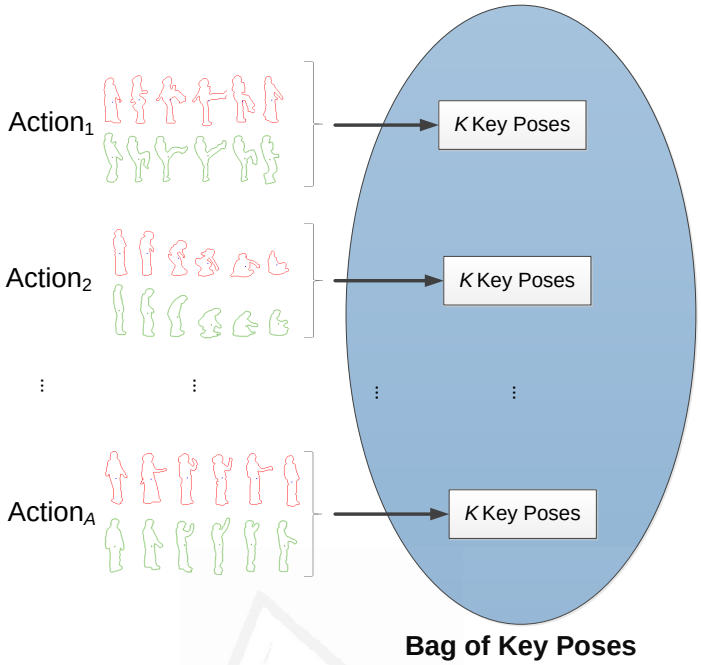


Figure 4.4: Learning scheme of the bag-of-key-poses model. For each action class, K key poses are obtained separately and then joined together. Note that action classes are view independent. Specifically, in this example two views are considered for each class.

4.2.2 Bag-of-key-poses model

The bag-of-words model detailed in section 2.5 relies on a codebook of words, which stand for the most representative sample instances, and a histogram-based representation that encodes the frequency of appearance of these words over space or time. Our proposal is similar in the first step, but instead of words, key poses are used. Furthermore, we do not rely on histograms but on temporal evolution between key poses.

Specifically, the bag-of-key-poses model is generated considering the key poses of each action class. In this way, an equal representation of each of the action classes can be ensured in the bag of key poses. Figure 4.4 shows the learning scheme of the bag-of-key-poses model. For each action class, the available video sequences are processed to generate the corresponding pose representations. Then, K key poses are generated out of each class. This process is repeated for the A action classes the model has to learn, and the obtained key poses are joined together. Hence, a total of $A \times K$ key poses are contained in the bag of key poses.

4.2.3 Multi-view learning

As it has been proposed in chapter 3, the HAR method shall be able to take advantage of multiple cameras focusing on the same field of view. In this way, recognition can be performed despite occlusions and unfavourable viewing angles. In addition, since more characteristic information can be extracted, the classification can be improved. Thus, multiple views can aid the discrimination of ambiguous actions. In order to combine information from multiple sources, which are cameras in this case, information fusion techniques can be applied (Dasarathy, 1994). The 2D data from multiple cameras can be used in order to create a 3D representation (Canton-Ferrer et al., 2006; Yan et al., 2008). This *data fusion* allows to apply a single feature extraction process which minimises information loss. Nevertheless, this requires camera calibration, and 3D representations usually imply a higher computational cost as appropriate 3D features need to be obtained. In *feature fusion*, the fusion process is placed one step further. Single-view features are obtained for each of the camera views, and a common representation is generated for all the features afterwards. The fusion process depends on the type of data. Feature vectors are commonly combined by aggregation functions or concatenation of vectors (Määttä et al., 2010; Wu et al., 2010), or also more sophisticated techniques as canonical correlation analysis (Cilla et al., 2013). The appeal of this type of fusion is the resulting simplicity of transition from single- to multi-view recognition methods, since multi-view data is only handled implicitly. A learning method, which in fact learns and extracts information from actions or poses from multiple views, requires considerations at the learning scheme. Through *model fusion*, multiple views are learnt either as other possible instances of the same class (Wu et al., 2010) or by explicitly modelling each possible view (Cilla et al., 2012). These 2D or 3D models may support a limited or unlimited number of viewpoints. Last but not least, information fusion can be applied at the decision level. In this case, for each of the views a single-view recognition method is used independently, and a decision is taken based on the single-view recognition results. The *best view* is chosen based on one or multiple criteria, like closest distance to the learnt pattern, highest score/probability of feature matching, or metrics which try to estimate the quality of the received input pattern (Iosifidis et al., 2012; Määttä et al., 2010; Wu et al., 2010). Other works analysed the possibility of combining single-view recognition results: Naiel et al. (2011) used a majority voting technique where the camera with minimal

distance is chosen when no majority is reached. In that work, MHI and MEI are reduced with a two-dimensional PCA and classified based on a KNN classifier. The work of Zhu et al. (2013) goes further. The decision is based on accumulation of each view's weighted prediction histogram. These weights are defined as the proportion of sequence segments that share the same maximum voting class labels in their prediction histograms. The prediction histograms are obtained by mapping the local segment features of binary silhouettes using a random forest classifier. In comparison with the former, these approaches have the advantage that all the available viewing angles contribute to the final result. This can potentially improve the recognition of actions whose *best view* may change during its performance (*e.g.* due to occlusions or the type of action itself). However, the main difficulty is to establish this decision rule, because it depends strongly on the type of actions to recognise and on the camera setup. Moreover, decision-level fusion does not combine the multi-view data itself. Hence, the latent characteristic value of fused multi-view data is not exploited. However, its main appeal relies in the distributed execution scheme that is supported by definition.

In this sense, in this HAR method, because of the mentioned associated advantages, fusion of multiple views is considered at the two intermediate levels, *i.e.* feature and model fusion.

Feature fusion

Relying on a scenario with multiple cameras, spatial pose representations are obtained for each of the available video streams as detailed in section 4.1. These pose representations can then be fused into a multi-view pose representation based on feature fusion techniques. The resulting characteristic descriptor is used as the shared representation for all the available views at a given temporal instant t.

By means of this kind of fusion, the remaining stages of the classification method stay unmodified. Using these multi-view pose representations henceforth leads to an implicit learning of multi-view key poses.

As it has been mentioned, the greatest advantage of this type of fusion is its simplicity. Its main weak point is that, unless specifically considered, all the viewpoints are always needed during the recognition.

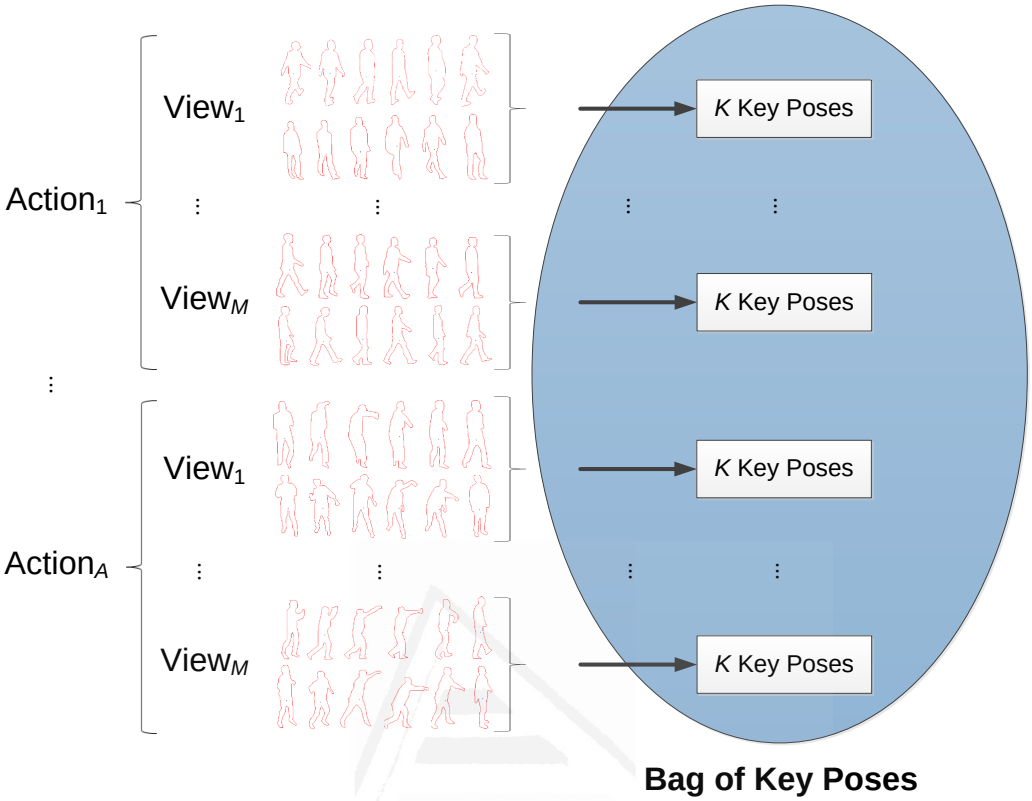


Figure 4.5: Bag-of-key-poses model considering the model fusion of multiple views.

Model fusion

In order to explicitly consider and learn different views of actions, the learning method needs to deal with multiple views. Using the bag-of-key-poses model, multi-view learning can be applied. Let us suppose there are M available viewpoints and A action classes to learn. Considering a single action, all the pose representations are reduced to K key poses per view. This means that the pose representations of each view are considered separately so as to simplify the key pose generation process. Once all the key poses of each view have been generated, this process is repeated for each action class ending up with a bag of key poses of $A \times M \times K$ key poses. Figure 4.5 shows an overview of the learning scheme.

Applying model fusion presents several benefits, especially for the development of AAL services:

1. Once the learning process has been finished, further views and action classes can be learnt without restarting the whole process.

2. The camera setups do not need to match between training and testing stages. More camera views may improve the result of the recognition, though it is not required to have all the cameras available.

3. Each camera view is processed separately and matched with the corresponding view, without requiring to know specifically at which angle it is installed.

4.2.4 Discriminative value of key poses

At this point, the training data has been reduced to a representative model of the key poses involved in each action class. Nevertheless, not all the key poses are equally important. Very common poses such as *standing still* are not able to distinguish between actions, whereas a *bend* pose can most certainly be only found in its own action class. For this reason, a discrimination value w, which indicates the capacity of discrimination between action classes of each key pose kp, is obtained. For this purpose, all available pose representations are matched with their nearest neighbour among the bag of key poses so as to obtain the ratio of within-class matches $w_{kp} = \frac{matches_{kp}}{assignments_{kp}}$. In this manner, *matches* is defined as the number of within-class assignments, *i.e.* the number of cases in which a pose representation is matched with a key pose from the same class; whereas *assignments* denotes the total number of times a key pose got chosen. Please see algorithm 1 for greater detail.

As will be shown in section 4.3.2, the discrimination value of the key poses can be employed to emphasise the matches of relevant poses and ignore those which do not aid in the discrimination of actions.

4.2.5 Sequences of key poses

Key poses have been defined as the most representative instances of the per-class pose representations, which in turn are made up of characteristic information of the spatial properties of the human body. This means that, so far, no temporal cues have been taken into account, and frame-by-frame recognition could be handled without considering any particular motion order. Nevertheless, as video sequences of action performances are used for training, valuable information about the temporal evolution of the shape of the human body and the duration of

Algorithm 1 Pseudocode for obtaining the key pose discrimination value w.

```
——————— Obtain matches and assignments ———————
for each action_class ∈ training_set do
    for each frame ∈ action_class do
        V̄ = pose_representation(frame)
        {kp, kp_class} = nearest_neighbour(V̄, bag_of_key_poses)
        if kp_class = action_class then
            matches_kp = matches_kp + 1
        end if
        assignments_kp = assignments_kp + 1
    end for
end for

——————— Obtain key pose discrimination value ———————
for each kp ∈ bag_of_key_poses do
    if assignments_kp > 0 then
        w_kp = matches_kp / assignments_kp
    else
        w_kp = 0
    end if
end for
```

action sequences is available. Similarly, in an online recognition scenario, silhouettes would be acquired in the particular order of the subject's performance. In consequence, video recognition presents a clear advantage over image recognition that relies on the temporal dimension.

Considering key poses, in Baysal et al. (2010); Cheema et al. (2011) no temporal information is used at all. Thurau and Hlaváč (2007) model the short-term temporal relation between consecutive key poses with n-grams (*trigrams* showed good results at acceptable computational cost). Eweiwi et al. (2011) take into account the temporal context of a small number of frames by means of obtaining temporal key poses based on MHI. On the contrary, our contribution considers long-term temporal relation between key poses and thus takes advantage of the known temporal evolution of key poses over a whole sequence.

To this extent, sequences of key poses are proposed for the learning of the possible transitions between key poses along action performances. Consequently, the long-term temporal evolution of key poses can be modelled by finding the nearest neighbour key pose kp of each video frame's pose representation. The

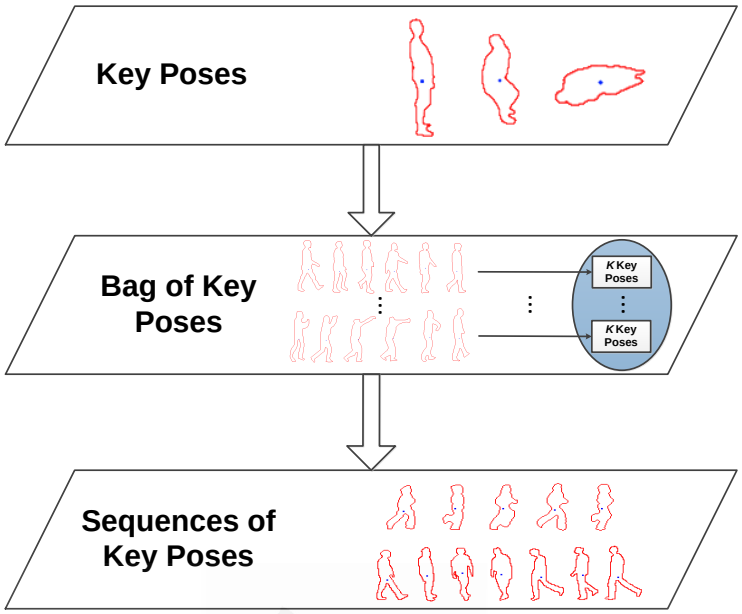


Figure 4.6: Outline of the learning stage. Using the pose representations, key poses are obtained for each action. In this way, a bag-of-key-poses model is learnt. The temporal relation between key poses is modelled using sequences of key poses.

successive nearest neighbour key poses compose a simplified sequence of known key poses and their temporal evolution: $Seq = \{kp_1, kp_2, ..., kp_T\}$. This process is performed for all available training sequences in order to obtain a set of sequences of key poses.

In this step, not only the temporal order of appearance of key poses is modelled, but the training data is also shifted to the shared and defined domain of the bag of key poses. This implies that particular instance- and actor-related noise and outlier values are filtered, and exemplar-based action recognition can be significantly improved.

The diagram of figure 4.6 shows how the mentioned learning stages are connected.

4.3 Recognition

At the recognition stage, classification of unknown video sequences is performed in order to recognise human actions. Since during the learning stage the temporal relation between key poses has been learnt by means of sequences of key poses,

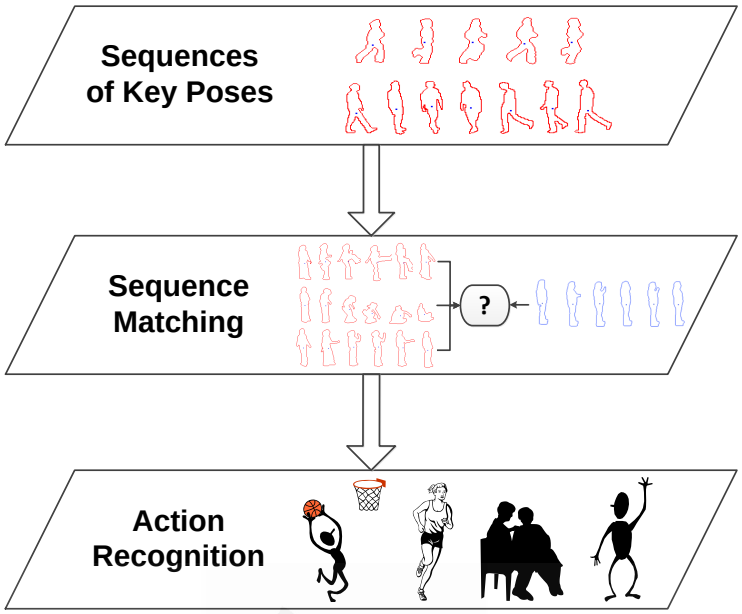


Figure 4.7: Outline of the recognition stage. The unknown sequence of key poses is obtained and compared to the known sequences. Through sequence matching, the action of the video sequence can be recognised.

recognition can be performed by means of sequence matching in order to find the closest training sample.

Given a new video sequence to recognise, the same procedure as in the learning stage is performed so as to obtain a sequence of key poses:

1. The video frames from each view are processed to their single-view pose representation as described in section 4.1.

2. In the case that feature fusion is employed and multiple views are available, the single-view pose representations will be fused to a multi-view pose representation.

3. The equivalent sequence of key poses $Seq' = \{kp'_1, kp'_2, ..., kp'_U\}$ is built by replacing each pose representation with its nearest neighbour key pose among the bag of key poses.

4. If model fusion is employed, these steps are repeated for each view, and one sequence of key poses is obtained per view $(Seq'_1, Seq'_2, ..., Seq'_M)$.

In this sense, the video sequences are transformed into sequences of key poses which can be compared to the learnt set in order to classify the unknown sequence (see figure 4.7).

4.3.1 Sequence matching

Due to the temporal intra-class variance, a suitable distance metric is needed in order to compare sequences of key poses. Different actors can perform the same actions in very different ways and they can do so faster or slower than others. While some motions are indispensable when performing an action, like moving one leg and then the other while walking, these can still appear with a considerable time shift, especially when dealing with elderly people. Therefore, inconsistent time scales have to be supported. This is indispensable, since the pace at which an action is performed depends on the age and condition of the person.

Among suitable distance metrics, temporal alignment algorithms find common patterns between sequences despite temporal shifts. These are very common in research fields like natural language processing, where techniques based on DTW (Sakoe and Chiba, 1978) stand out. This is due to the shared constraint of real-time execution. Also other fields demand efficient and robust comparison techniques, not because of real-time requirements, but because of the massive size of the datasets. Such is the case of genome analysis, where genome sequences are compared with highly optimised tailored techniques. In fact, optimisations of the traditional Needleman-Wunsch (Needleman and Wunsch, 1970) and Smith-Waterman (Smith and Waterman, 1981) algorithms have been studied for more than two decades (Gonnet et al., 1992). These could be used for real-time action recognition if sequences of key poses are seen as genome sequences. Some techniques can also be found in the HBA field. In Zhou and Torre (2009), an extension of canonical correlation analysis is presented for spatio-temporal alignment of human motion between two subjects. In Caspi and Irani (2002), sequence-to-sequence alignment is studied taking into account correspondences in time and space, such as moving objects or illumination changes.

In this manner, using the appropriate distance metric $d(Seq, Seq')$, the nearest neighbour sequence of key poses can be found and its action class supplies the final result of the recognition.

If the model fusion option is employed, the nearest neighbour sequence of key poses is found for each view. The action class of the match with the lowest distance is chosen as the final result of the recognition, *i.e.* the result is based on the *best view*. This means that only a single view is required in order to perform the recognition, even though better viewing angles may improve the result. Note that this process is similar to decision-level fusion, but in this case recognition relies on the same multi-view learning model, *i.e.* the bag of key poses.

4.3.2 Relevance of key pose matches

In order to compare sequences, not only the temporal alignment of elements is needed, but an element-wise distance metric is also required. Therefore, $d(Seq, Seq')$ will have to compare the elements of the sequences, which consist in key poses. The distance between two key poses $d(kp, kp')$ is obtained based on 1) a distance metric considering the similarity of the feature elements; and 2) the relevance of the match of key poses. As seen before, not all the key poses are as relevant for the purpose of identifying the corresponding action class. Hence, it can be determined how relevant a specific match of key poses is based on their discriminative values w_{kp} and $w_{kp'}$.

In this sense, the distance between key poses is obtained as:

$$d(kp, kp') = \hat{d}(kp, kp') + z \ rel(kp, kp') \ , \tag{4.1}$$

$$rel(kp, kp') = |dev(kp, kp') * w_{kp} * w_{kp'}| \ , \tag{4.2}$$

$$dev(kp, kp') = \hat{d}(kp, kp') - average\_distance \ , \tag{4.3}$$

where $average\_distance$ corresponds to the average distance between key poses computed throughout the learning stage. As it can be seen in equation 4.1, the feature distance $\hat{d}(kp, kp')$ and the relevance $rel(kp, kp')$ of the match are taken into account.

The relevance $rel(kp, kp')$ is determined based on the discriminative values of the key poses and the deviation of the feature distance. Consequently, matches of key poses which are very similar or very different are considered more relevant than those that present an average similarity. The value of z depends on the desired behaviour. Table 4.1 shows the chosen value for each case. In pairings of discriminative key poses which are similar to each other, a negative value is chosen

Table 4.1: Value of z based on the pairing of key poses and the signed deviation. *Ambiguous* and *discriminative* stand for respectively low and high discriminative values, which have to be determined during experimentation.

Signed deviation	Pairing	z
$dev(kp, kp') < 0$	*discriminative*	-1
$dev(kp, kp') > 0$	*discriminative*	$+1$
any	*ambiguous*	-1
any	*discriminative and ambiguous*	$+1$
any	*other pairings*	± 1

in order to reduce the feature distance. If the distance among them is higher than average, this indicates that these important key poses do not match well together and therefore the final distance is increased. For ambiguous key poses, *i.e.* key poses with low discriminative value, pairings are not as important for the distance between sequences. On the other hand, a pairing of a discriminative and an ambiguous key pose should be disfavoured as these should match with instances with similar weights. Otherwise, the operator is based on the sign of $dev(kp, kp')$, which means that low feature distances are favoured and high feature distances are penalised. Thus, the shape-based similarity between key poses and the relevance of the specific match are taken into account in sequence matching.

4.4 Remarks

In this chapter, a method based on a bag-of-key-poses model and multiple views has been proposed for human action recognition. Using spatial pose representations, the characteristic information of the pose or shape of the human body is extracted. The most representative instances out of these pose representations, the key poses, are generated. The per-class key poses are learnt by means of a bag-of-key-poses model, which supports feature and model fusion techniques in order to learn from multiple views. Temporal relationship between key poses is then taken into account by modelling sequences of key poses. Recognition can be performed with sequence matching algorithms, where it is proposed to consider both the key pose feature distance and the relevance of the specific match of key poses based on their discriminative values.

As it has been mentioned and will be seen in following chapters, this method serves as a framework in which different algorithmic choices can be made with regard to visual features, specific algorithms for key pose generation or sequence alignment, as well as distance metrics and multi-view fusion techniques. Several of these options will be taken into account in the following chapter.

Chapter 5

Multi-view human action recognition in real time

In chapter 4, a method for HAR based on a bag-of-key-poses model and multiple views has been presented. This method is employed in this chapter in order to apply multi-view HAR in real time using RGB colour images. To this extent, in the following sections, a novel visual feature for real-time recognition is proposed and the required algorithmic choices are detailed. In addition, different proposals for multi-view fusion are presented and compared in the experimentation.

In the first section 5.1, a visual feature for pose extraction on RGB images is proposed, and the algorithmic choices are made to implement the proposed method. In the experimentation that follows, first the proposed approaches for the fusion of multiple views, that have been introduced in section 4.2.3, are considered using an existing visual feature. Next, the proposed visual feature is evaluated and compared with other similar techniques. Finally, in 5.2 an improved fusion of multiple views is presented and compared against previous options. Table 5.1 summarises the development stages and their particular objectives. The corresponding sections and experimentations, as well as the employed visual feature and method are detailed.

Table 5.1: Different development stages of the proposed method for multi-view human action recognition in real time. The different objectives, employed visual features and method names are detailed. Please look up the corresponding sections for greater detail.

Section	Part	Objective	Visual Feature	Method
Section 5.1	Fusion of multiple views	Evaluation of the method - *Feature Fusion* and *Model Fusion* of multiple views	Distance signal of (Dedeoğlu et al., 2006)	*HAR Method*
Section 5.1	Radial summary feature	Evaluation of the proposed feature and comparison with other techniques	*Radial Summary*	*HAR Method with Model Fusion*
Section 5.2	Experimentation	Improvement of the fusion of multiple views	*Radial Summary*	*Weighted Feature Fusion Scheme*

5.1 Implementation of the method

5.1.1 Visual feature extraction for pose representation

In section 4.1, the use of spatial pose representations has been proposed for the extraction of visual characteristics of video data. The feature extraction stage is especially sensitive to recognition accuracy and speed, since later stages depend on the quality and efficiency of this process. In this section, existing visual features based on RGB colour images will be reviewed, and a novel feature called *Radial Summary*, that is especially designed for real-time recognition, will be presented.

Since real-time recognition is an essential requirement for the AAL application of this method, computationally demanding 3D models and reconstructions are discarded in this work. Similarly, dense representations, like salient points, commonly require to process the image block-wise at different scales, and lead to a large set of features. Then, these have to be simplified with dimensionality-reduction- or histogram-based techniques. For these reasons, the usage of holistic features is proposed for pose representation. Hence, a single global pose representation provides the characteristic information of the shape of the person for each given frame. As it has been mentioned earlier, the shape of the person can be obtained by means of extracting the human silhouette. Depending on the scenario setup, this can be done with traditional background subtraction techniques (see section 2.4), using depth-based segmentation, or even with infra-red, thermal or laser sensors.

Among holistic pose representations, we find silhouette-based features which either rely on the whole shape of the silhouette or only on the contour points. In Thurau and Hlaváč (2007), action primitives are extracted reducing the dimensionality of the binary images that contain the foreground segmentation with PCA. Polar coordinates are considered in Hsieh et al. (2011), where three radial histograms are defined for the upper part, the lower part and the whole human body. Each polar coordinate system has several bins with different radii and angles, and the concatenated normalised histograms are used to describe the human posture. Similarly, in Lv and Nevatia (2007) a log-polar histogram is computed choosing the different radii of the bins based on logarithmic scale. Silhouette contours are employed by Dedeoğlu et al. (2006) with the purpose of creating a distance signal based on the pointwise Euclidean distances between each contour point and the centroid of the silhouette. Conversely, Htike et al. (2011) compute the pairwise distances between contour points to build a histogram of distances resulting in a rotation, scale and translation invariant feature. In Boulgouris et al. (2006), the whole silhouette is used for gait recognition. An angular transform based on the average distance between the silhouette points and the centroid is obtained for each circular sector. This shows robustness to segmentation errors. Similarly, in Wang et al. (2007), the shape of the silhouette contour is projected on a line based on the $\Re$ transform, which is then made invariant to translation. Silhouettes can also be used to obtain stick figures, for instance, by means of skeletonisation. Chen et al. (2006) applied star skeletonisation to obtain a five-dimensional vector in star fashion considering the head, the arms and the legs as local maxima. Pointwise distances between contour points and the centroid of the silhouette are used to find the five local maxima. In the work of İkizler and Duygulu (2007) a different approach based on a "bag of rectangles" is presented. In their proposal, oriented rectangular patches are extracted over the human silhouette, and the human pose is represented with a histogram of circular bins of 15° each. Tran and Sorokin (2008) merge both silhouette shape and optical flow in a 286-dimensional feature, which also includes the context of 15 surrounding frames reduced by means of PCA. Recently, this feature has been used successfully in other works as, for instance, in Cilla et al. (2013). Rahman et al. (2013) take an interesting approach proposing a novel feature extraction technique, which relies on the surrounding regions of the subjects. These negative

spaces present advantages related to robustness to boundary variations caused by partial occlusions, shadows and non-rigid deformations.

As has previously been introduced, our goal is to perform HAR in real time, and to do so even in scenarios with multiple cameras. Therefore, the computational cost of feature extraction needs to be minimal. This leads us to silhouette contours. Human silhouettes contain rich shape information and can be extracted relatively easily. In addition, silhouettes and their contours show certain robustness to lighting changes and small viewpoint variations compared with other techniques, like optical flow (Ángeles Mendoza and Pérez de la Blanca, 2007). Using only the contour points of the silhouette results in a significant dimensionality reduction by discarding the interior points.

The following variables are used along this section:

1. the number of contour points n;
2. the number of radial bins B;
3. and the indices i, j, k and l, where $i, k, l \in [1...n]$ and $j \in [1...B]$.

We use the border following algorithm of Suzuki and Abe (1985) to extract the n contour points $P = \{p_1, p_2, ..., p_n\}$, where $p_i = (x_i, y_i)$. Our proposal consists in dividing the silhouette contour in B radial bins of the same angular size. Taking the centroid of the silhouette as the origin, the specific bin of each contour point can be assigned. Then, in difference to Hsieh et al. (2011); Lv and Nevatia (2007), where radial or log-polar histograms are used as spatial descriptors, or Chen et al. (2006), where star skeletonisation is applied, in our approach, a summary representation is obtained for each of the bins. The concatenation of these summary representations returns the final feature, leading to a more efficient representation (figure 5.1 shows an overview of the process).

The motivation behind using a radial scheme is two-fold. On the one hand, it relies on the fact that when using a direct comparison of contours, even after length normalisation as in Dedeoğlu et al. (2006), spatial alignment between feature patterns is still missing. Each silhouette has a distinct shape depending on the actor and the action class, and therefore a specific part of the contour can have more or less points in each sample. Using an element-wise comparison

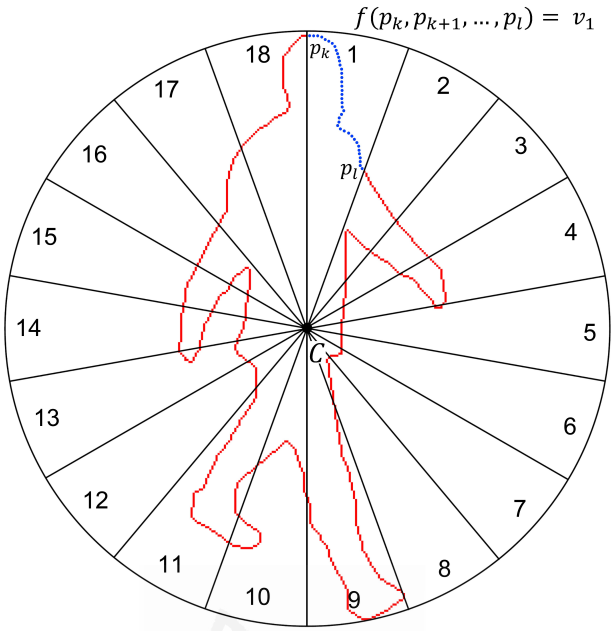


Figure 5.1: Overview of the feature extraction process: 1) All the contour points are assigned to the corresponding radial bin; 2) for each bin, a summary representation is obtained (example with 18 bins).

of the radial bins of different contours, we ignore how many points each sample has in each bin. This avoids an element-wise comparison of the contour points, which would imply the erroneous assumption that these are correlated. On the other hand, this radial scheme allows us to apply an even further dimensionality reduction by obtaining a representative summary value for each radial bin.

The following steps are taken to compute the feature:

1. The centroid of the contour points $C = (x_c, y_c)$ is calculated as:

$$x_c = \frac{\sum_{i=1}^{n} x_i}{n}, y_c = \frac{\sum_{i=1}^{n} y_i}{n} \ . \tag{5.1}$$

2. The pointwise Euclidean distances between each contour point and the centroid $D = \{d_1, d_2, ..., d_n\}$ are obtained as in Dedeoğlu et al. (2006):

$$d_i = \|C - p_i\|, \quad \forall i \in [1...n] \ . \tag{5.2}$$

3. Considering a clockwise order, the corresponding bin b_i of each contour point p_i is assigned as follows (for the sake of simplicity $\alpha_i = 0$ is considered as $\alpha_i = 360$):

$$\alpha_i = \begin{cases} \arccos(\frac{y_i - y_c}{d_i}) \cdot \frac{180}{\pi} & \text{if } x_i \geq 0 \ , \\ 180 + \arccos(\frac{y_i - y_c}{d_i}) \cdot \frac{180}{\pi} & \text{otherwise} \ , \end{cases} \tag{5.3}$$

$$b_i = \left\lceil \frac{B \cdot \alpha_i}{360} \right\rceil , \quad \forall i \in [1...n] \ . \tag{5.4}$$

4. Finally, a summary representation is obtained for the points of each bin. The final feature $\bar{V}$ results of the concatenation of summary representations, which are normalised to unit sum in order to achieve scale invariance:

$$v_j = f(p_k, p_{k+1}, ..., p_l) \ / \ b_k ... b_l = j \wedge k, l \in [1...n] \ , \tag{5.5}$$

$$\bar{v_j} = \frac{v_j}{\sum_{j=1}^{B} v_j} \ , \tag{5.6}$$

$$\bar{V} = \bar{v_1} \parallel \bar{v_2} \parallel ... \parallel \bar{v_B} \ . \tag{5.7}$$

The function f could be any type of function which returns a significant value or property of the input points. We tested three types of summaries (*variance*, *max value* and *range*), based on the previously obtained distances to the centroid, whose results will be analysed in section 5.1.3.

The following definitions of f are used:

$$f_{var}(p_k, p_{k+1}, ..., p_l) = \sum_{i=k}^{l} (d_i - \mu)^2 \ , \tag{5.8}$$

where μ is the average distance of the contour points of each bin.

$$f_{max}(p_k, p_{k+1}, ..., p_l) = \max(d_k, d_{k+1}, ..., d_l) \ , \tag{5.9}$$

$$\begin{aligned} f_{range}(p_k, p_{k+1}, ..., p_l) = \max(d_k, d_{k+1}, ..., d_l) \\ - \min(d_k, d_{k+1}, ..., d_l) \ . \end{aligned} \tag{5.10}$$

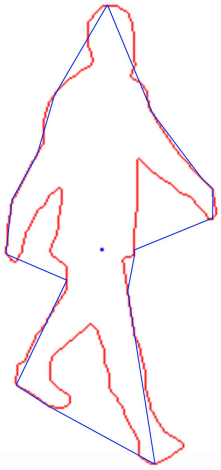


Figure 5.2: Example of the result of applying the f_{max} summary function.

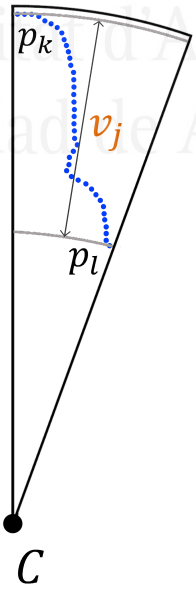


Figure 5.3: Graphical explanation of the statistical range f_{range} of a sample radial bin.

In figure 5.2 an example of the result of the f_{max} summary function can be seen, and figure 5.3 shows a graphical explanation of the f_{range} summary function.

5.1.2 Algorithmic choices for learning and recognition

In order to implement the classification method presented in chapter 4, a few algorithmic choices have to be performed regarding 1) key pose generation, 2) multi-view learning and 3) distance metrics. These are going to be presented below, having the technical requirements of the application scenario in mind.

Key pose generation

The bag-of-key-poses model relies on the learning of the key poses involved in each action class. For the purpose of generating these key poses, an automatic and unsupervised learning approach has been employed. Therefore, key poses can be obtained for any set of action classes, and without having *a priori* knowledge about the key poses that should be involved. First, all the frames of the video sequences are processed in order to obtain their pose representation. Then, the per-class key poses are learnt by means of K-means clustering. Hence, the extracted features of all available images of the same action class $samples = \{sample_1, sample_2, ..., sample_N\}$ are grouped into K clusters. Each cluster centre of $centres = \{centre_1, centre_2, ..., centre_K\}$ represents a key pose kp, since it is a characteristic instance among the training data that represents a dense area in the feature space. The process of clustering is repeated λ times, so as to avoid local minimum, and the best result is taken. For this purpose, clustering results are evaluated with respect to a compactness metric based on the sum of distances. Given that the clustering process returns the corresponding label of each sample, $labels = \{label_1, label_2, ..., label_N\}$, in which $label_g$ stands for the index of the cluster assigned to $sample_g$, the following metric is minimised:

$$\arg\min_{centres} \sum_{g=1}^{N} |sample_g - centre_{label_g}| \ . \tag{5.11}$$

This key pose learning process is repeated individually for the training samples of each action class. In this way, a set of K key poses is obtained for each action and joined together in the bag of key poses. In figure 5.4, the resulting key poses for the *CollapseLeft* action are shown.

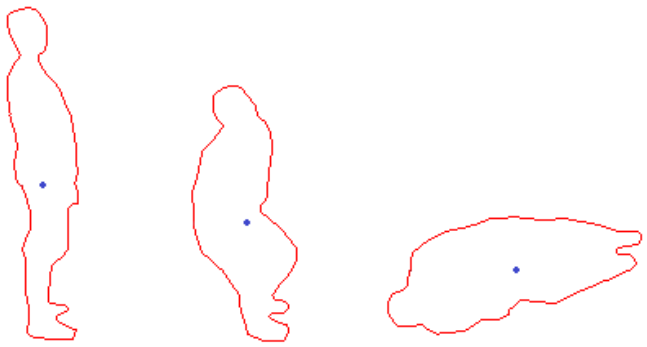


Figure 5.4: Sample key poses obtained with K-means clustering. The *CollapseLeft* action from the MuHAVi dataset is shown.

Multi-view learning

The first of the proposed options for learning from multiple views requires a specific feature fusion operator in order to combine single-view pose representations into a multi-view pose representation. Among the possibilities seen in section 4.2.3, feature concatenation has been chosen. This is motivated by the inherent low dimensionality of the detailed *Radial Summary* feature. It allows to employ feature concatenation and to retain all the characteristic information without prohibitively increasing the final feature size, which would lead to a higher computational cost and make classification more complex. For the same reason, the need of too large training sets caused by the *curse of dimensionality* can be prevented (Ross and Govindarajan, 2005). Furthermore, since the instances of different views of the same feature are combined and normalisation has been applied (see section 4.1), the feature sets are compatible regarding size and scale.

Assuming that M video streams of the same field of view are available, first each frame is individually processed to its pose representation. Then the multi-view pose representation is obtained by frame-by-frame concatenation of single-view pose representations (see figure 5.5):

$$\bar{V}_1 \circ \bar{V}_2 \circ \ldots \circ \bar{V}_M \,. \tag{5.12}$$

This step is identically performed with train and test instances, using multi-view pose representations at the succeeding stages. As a result, when feeding the model with multi-view pose representations, a bag of multi-view key poses (see figure 5.6) is inherently obtained.

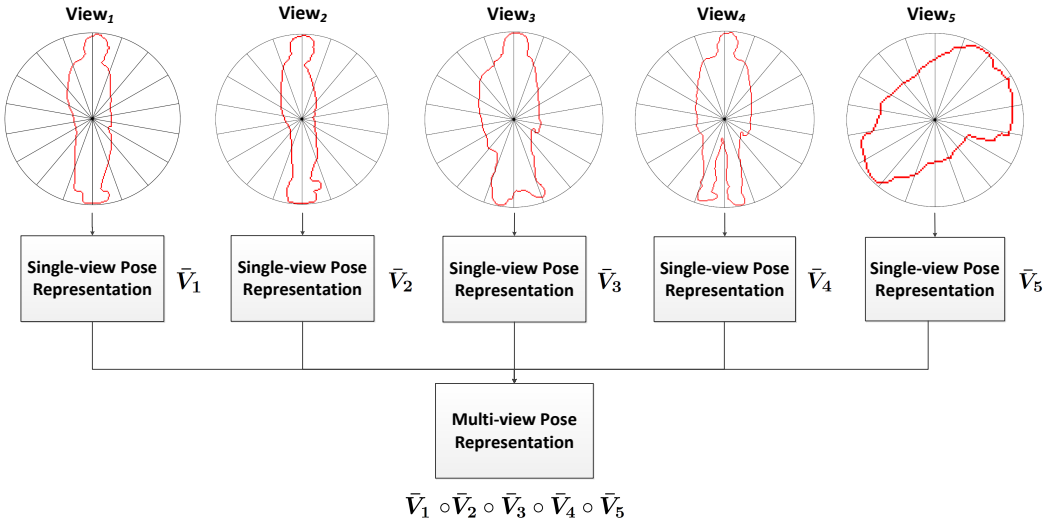


Figure 5.5: Overview of the feature fusion process of the multi-view pose representation. This example shows five different views of a specific pose taken from the *walk* action class from the IXMAS dataset ($View_1$ to $View_4$ correspond to side views and $View_5$ to a top view).

If model fusion is used, the single-view pose representations are directly processed in the bag-of-key-poses model, and no further decisions need to be taken in the implementation.

Distance metrics

As it has been mentioned in the last chapter, once the bag-of-key-poses model is learnt, it is used to translate the training and test sequences to sequences of key poses. In order to compare an unknown sequence of key poses to the labelled ones, sequence matching is proposed. For this purpose, the DTW algorithm (Sakoe and Chiba, 1978) has been chosen because it is able to successfully align sequences with inconsistent time scales, therefore taking into account the actor-related differences among action performances.

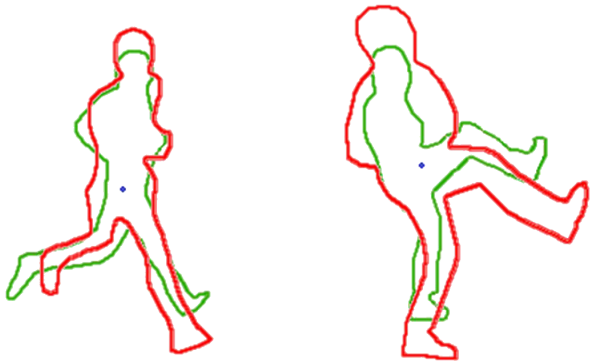


Figure 5.6: Multi-view key poses: *RunLeftToRight* (left) and *KickRight* (right) from the MuHAVi dataset obtained by means of feature concatenation. Silhouettes in red are taken from the first view, and silhouettes in green from the second. A 45° angle exists between them.

The DTW distance $d_{DTW}(Seq, Seq')$ between two sequences of key poses $Seq = \{kp_1, kp_2, ..., kp_T\}$ and $Seq' = \{kp'_1, kp'_2, ..., kp'_U\}$ is defined as:

$$d_{DTW}(Seq, Seq') = dtw\,(T, U)\ , \tag{5.13}$$

$$dtw\,(t,u) = \min \left\{ \begin{array}{l} dtw\,(t-1,u)\,, \\ dtw\,(t,u-1)\,, \\ dtw\,(t-1,u-1) \end{array} \right\} + d(kp_t, kp'_u)\ . \tag{5.14}$$

In such a way, the label of the closest training sequence will be returned as the final result of the classification. Figure 5.7 shows an example for simplified one-dimensional sequence elements. It can be observed how the first and last elements are always matched, and the alignment in between is chosen based on the lowest distance. In this example, the final DTW distance $d_{DTW}(Seq_1, Seq_2) = dtw(5, 6) = 9$.

$$\hat{d}(kp, kp') = \|kp - kp'\|_1\ . \tag{5.15}$$

Last but not least, the alignment of sequences requires to compare the elements. The distance between key poses $d(kp, kp')$ defined in equation (4.1) from section 4.3.2 relies both on the feature distance and the relevance of the match of key poses. The feature distance $\hat{d}(kp, kp')$ has been obtained using the Manhattan distance (see equation 5.15), which means that the individual differences between the corresponding radial summary values are considered. Furthermore,

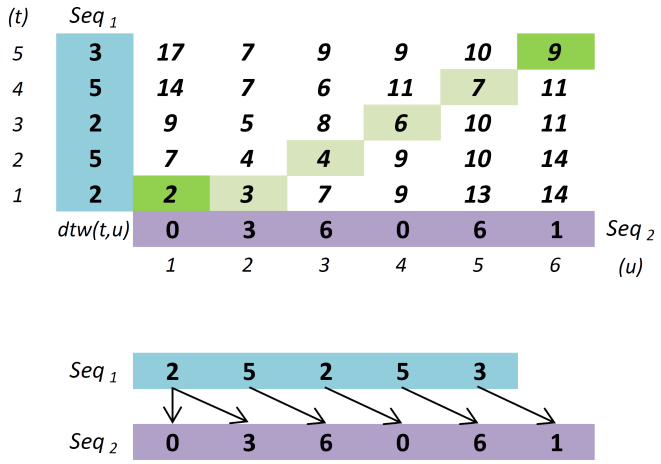


Figure 5.7: An example of the DTW algorithm for a simplified one-dimensional case (top), and the final alignment between elements (bottom) are shown. Note that the matrix elements indicate the accumulated distance between elements for the partial alignment with the lowest distance (following equation 5.14).

this distance can be obtained very efficiently and is also suitable if feature fusion is employed, since no assumption is made regarding the n-dimensional feature space. Note that this feature distance has also been employed to find the nearest neighbour key poses in the learning of sequences of key poses.

5.1.3 Experimentation

In this section, experimental results are shown, that have been obtained on publicly available datasets using the proposed method. First the employed evaluation methodology and technical details of the test environment are described. Then, the presented options for fusion of multiple views are studied. This is performed using a state-of-the-art pose representation feature. This allows us to separate the conclusions that can be drawn regarding this matter from the performance evaluation of our feature extraction proposal. In this way, first, the appropriate fusion of multiple views can be established, and second, the proposed visual feature can be evaluated. Furthermore, an analysis of the behaviour of the proposed method with respect to its parameters is also provided. Finally, an extended validation is performed in order to find possible weaknesses. Discussion regarding the obtained results, as well as comparisons with the state of the art will be given along the obtained results.

Evaluation methodology

Tests have been performed on several human action datasets with different levels of difficulty. These are related to image quality, corresponding binary segmentations, number of action classes and subjects, viewing angles, scenario-specific conditions, etc. Both single- and multi-view datasets have been employed. As input data, binary segmentations, obtained either manually or automatically, have been tested.

The following types of cross validation have been performed:

1. Leave-one-sequence-out (LOSO): In this cross validation test, all the sequences but one are used for training, and the system is tested with the remaining one. This procedure is repeated for each sequence and the final result is determined by the average accuracy score.

2. Leave-one-actor-out (LOAO): Similarly, in this test the sequences of all but one actor are used for training, and the sequences of the unknown one are used for testing. Again, this is done for all the available actors and the results are averaged. This test presents an additional difficulty due to the inherent differences among subjects. Actor variance due to clothes, body build and the personal way in which each subject performs an action commonly leads to worse classification results.

As in these cross validations training and testing sets change in each fold, overfitted learning can be prevented.

Note that based on empirical evidence, the discriminative values to be considered as *ambiguous* have been established as $w < 0.1$ and as *discriminative* if $w > 0.9$. Similarly, the clustering process has been repeated six times ($\lambda = 6$) considering the trade-off between compactness and learning time.

Regarding the performed temporal evaluation, a test environment with the following characteristics has been employed:

1. Tests have been performed on a standard PC with an Intel Core 2 Duo CPU at 3 GHz and 4 GB of RAM running Windows 7 x64. The proposed method has been implemented using the .NET Framework and the OpenCV library (Bradski, 2000).

2. Execution time has been measured using the hardware counter *QueryPerformanceCounter* with a precision of μs.

3. Temporal performance has been measured for the whole classification process starting with the input of binary segmented images, and going through contour extraction, feature extraction and sequence matching. No further optimisations or special purpose hardware have been employed.

Along this section, the Weizmann, MuHAVi and IXMAS datasets are used as benchmarks. These have been preferred over other datasets, like the ones mentioned in section 2.8.1 or others, as the UT-Interaction dataset presented in Ryoo and Aggarwal (2009), because these include ground truth labelling at the HBA level of actions and also provide foreground segmentations. Weizmann and IXMAS include binary segmentations which have been automatically obtained by means of background subtraction techniques. However, the MuHAVi dataset provides manually annotated silhouettes. These allow to separate the difficulties related to the background subtraction task from the ones that correspond to the silhouette-based action recognition. This also means that the silhouettes which can be obtained based on the foreground segmentations from the Weizmann and IXMAS datasets are of a much lower quality, and include noise and incompleteness. Furthermore, whereas in Weizmann only recordings of a single view are available, MuHAVi and IXMAS provide two and five views respectively. Note that the MuHAVi dataset comes in two versions: MuHAVi-14 includes 14 different action classes, and in MuHAVi-8 the same sequences are grouped together in eight action classes, where the direction of the performance is ignored. Although having fewer action classes normally makes discrimination easier, these classes are also harder to learn as intra-class variance is increased. A complete list of the involved actions, technical details and sample images can be seen in appendix A.

Fusion of multiple views

The first experimentation has been performed on the options for fusion of multiple views of the method, since a choice among these has to be performed in order to apply further evaluation. For this purpose, a state-of-the-art feature has been used for pose representation. Similarly to the *Radial Summary* feature, it is based on the silhouette boundary, although it does not provide any spatial alignment nor dimensionality reduction.

Specifically, the distance-signal feature of Dedeoğlu et al. (2006), that has been detailed in section 5.1.1, relies on the Euclidean distances between contour points and centroid (*i.e.* $D = \{d_1, d_2, ..., d_n\}$), which are normalised to a constant

	CollapseLeft	CollapseRight	GuardToKick	GuardToPunch	KickRight	PunchRight	RunLeftToRight	RunRightToLeft	StandupLeft	StandupRight	TurnBackLeft	TurnBackRight	WalkLeftToRight	WalkRightToLeft
CollapseLeft	4/4													
CollapseRight		4/4												
GuardToKick			8/8											
GuardToPunch			1/8	7/8										
KickRight					8/8									
PunchRight						8/8								
RunLeftToRight							3/4	1/4						
RunRightToLeft								4/4						
StandupLeft									1/2	1/2				
StandupRight										4/4				
TurnBackLeft											2/2			
TurnBackRight												3/4	1/4	
WalkLeftToRight													4/4	
WalkRightToLeft														4/4

Figure 5.8: Confusion matrix of the LOSO cross validation on the MuHAVi-14 dataset.

length L and unit sum. The feature size L, as well as the number of key poses per action class (and per view in the case of the model fusion approach) K, are detailed for the best results achieved.

In the following, results for both feature and model fusion of multiples views will be presented and compared with the available state-of-the-art results of the MuHAVi dataset. Lastly, results without any type of data fusion are obtained too. In this case, the system is fed with video sequences ignoring from which view they come, and at testing, each view is considered as an independent test sequence.

Leave-one-sequence-out cross validation The result of the LOSO cross validation applying the model fusion option can be seen in figure 5.8, where the obtained confusion matrix for MuHAVi-14 is shown. It can be observed that only four sequences are misclassified, achieving a final recognition rate of 94.1%. Furthermore, the matrix shows that errors are mostly made between very similar actions as *StandupLeft* and *StandupRight*. In MuHAVi-8 (see figure 5.9) only one sequence is misclassified and a final recognition rate of 98.5% is reached.

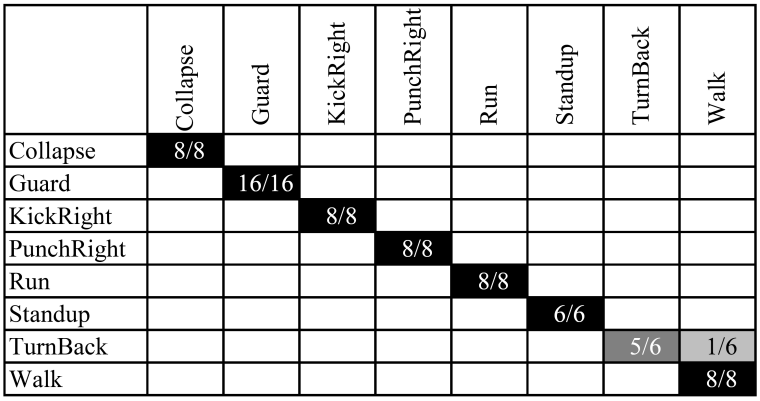


Figure 5.9: Confusion matrix of the LOSO cross validation on the MuHAVi-8 dataset.

Table 5.2: Comparison of our results with similar state-of-the-art approaches on the MuHAVi dataset (all use LOSO cross validation).

Approach		**MuHAVi-14**		**MuHAVi-8**
Singh et al. (2010) (baseline)		82.4%		97.8%
Cheema et al. (2011)		86.0%		95.6%
Martínez-Contreras et al. (2009)		-		98.4%
Eweiwi et al. (2011)		91.9%		98.5%
Without Fusion	$(L = 600, K = 120)$	85.3%	$(L = 350, K = 60)$	95.6%
Feature Fusion	$(L = 200, K = 140)$	92.6%	$(L = 300, K = 100)$	97.1%
Model Fusion	$(L = 450, K = 60)$	94.1%	$(L = 250, K = 75)$	98.5%

Table 5.2 shows the results that have been achieved with the feature fusion approach or without considering any multi-view recognition. It can be seen that model fusion clearly outperforms these methods, and that a significant improvement is obtained when multiple views are learnt explicitly at the model level. Moreover, in comparison with the baseline and the latest and highest recognition rates of the state of the art, our approach presents, to the best of our knowledge, the highest result so far on the MuHAVi-14 dataset.

Leave-one-actor-out cross validation In the LOAO cross validation, robustness to actor variance is evaluated. Table 5.3 shows the results of the LOAO test. Again, model fusion achieves significantly better results than the other methods. Interestingly, feature fusion performs worse than the single-view recognition. This can be attributed to the increased actor variance that results of using the multi-view pose representation. These results outperform the currently available recognition rates by 8.9% and 10.3% respectively.

Table 5.3: Comparison of results of the MuHAVi LOAO test.

Approach		**MuHAVi-14**		**MuHAVi-8**
Singh et al. (2010) (baseline)		61.8%		76.4%
Cheema et al. (2011)		73.5%		83.1%
Eweiwi et al. (2011)		77.9%		85.3%
Without Fusion	$(L = 200, K = 80)$	81.6%	$(L = 300, K = 60)$	92.6%
Feature Fusion	$(L = 200, K = 100)$	80.9%	$(L = 200, K = 100)$	91.2%
Model Fusion	$(L = 450, K = 60)$	86.8%	$(L = 250, K = 75)$	95.6%

The presented method performs especially well in this test, because of the performed shift from sequences of pose representations to sequences of key poses. This moves the test data domain to our domain of key poses and constitutes an essential step in the process, as noise and possible dissimilarities between actors are filtered.

Given these results, the model fusion of multiple views will be employed in the remainder of this section.

Temporal evaluation As stated beforehand, our work has been driven by the ambition of creating a multi-view action recognition method which could deal with both the increased complexity of multi-view learning and the need of an adequate recognition speed in order to perform real-time action recognition. This guided the decisions that have been taken regarding the design of the multi-view action recognition method whose temporal performance has been tested.

Running the MuHAVi-14 benchmark, each sequence is processed in 1.14 s achieving a recognition speed of 51 fps. As MuHAVi-8 presents fewer classes, the recognition speed rises to 66 fps, *i.e.* 0.88 s per sequence. It is worth mentioning that the presented approach also proves to be efficient at the training stage. An average training speed of 39 and 50 fps has been measured for MuHAVi-14 and MuHAVi-8 respectively.

Radial summary feature

Once the fusion of multiple views has been validated, we are now able to evaluate the proposed *Radial Summary* feature using the model fusion approach. In this experimentation, the presented method is tested on two datasets which serve as benchmarks (Weizmann and MuHAVi). On this single- and multi-view data, our learning algorithm is used with the proposed feature, and the results of the

Table 5.4: Comparison of recognition results with different summary values (*variance*, *max value*, *range*) and the features of Boulgouris et al. (2006) and Dedeoğlu et al. (2006). Best results have been obtained with $K \in [5, 130]$ and $B \in [8, 46]$. (Bold indicates highest success rate.)

Dataset	Test	Boulgouris et al. (2006)	Dedeoğlu et al. (2006)	f_{var}	f_{max}	f_{range}
Weizmann	LOSO	65.6%	78.5%	90.3%	**93.5%**	**93.5%**
Weizmann	LOAO	78.5%	80.6%	92.5%	94.6%	**95.7%**
MuHAVi-14	LOSO	61.8%	94.1%	**95.6%**	91.2%	**95.6%**
MuHAVi-14	LOAO	52.9%	86.8%	70.6%	**91.2%**	88.2%
MuHAVi-8	LOSO	69.1%	98.5%	**100%**	**100%**	**100%**
MuHAVi-8	LOAO	67.6%	95.6%	83.8%	**98.5%**	97.1%

three chosen summary representations (*variance*, *max value* and *range*) are compared. In addition, the distance-signal feature of Dedeoğlu et al. (2006) and the silhouette-based feature of Boulgouris et al. (2006), which have been summarised in section 5.1.1, are used as a reference in order to make possible a comparison between features. Lastly, our approach is compared with the state of the art in terms of recognition rates and speed.

The feature of Boulgouris et al. (2006) has been originally designed for gait recognition. It presents advantages regarding, for instance, robustness to segmentation errors, since it relies on the average distance to the centroid of all the silhouette points of each circular sector. Nevertheless, on the tested action recognition datasets it returned low success rates, which are significantly outperformed by the other four contour-based approaches. Both the feature of Dedeoğlu et al. (2006) and ours are based on the pointwise distances between the contour points and the centroid of the silhouette. Our proposal distinguishes itself in that a radial scheme is applied in order to spatially align contour parts. Further dimensionality reduction is also provided by summarising each radial bin in a single characteristic value. Table 5.4 shows the performance we obtained by applying the existing feature of Dedeoğlu et al. (2006) to our HAR method. Whereas on the Weizmann dataset the results are significantly behind the state of the art, the results for the MuHAVi dataset are promising. The difference of performance can be explained with the different qualities of the binary silhouettes, since the sil-

Table 5.5: Comparison of recognition rates and speeds obtained on the Weizmann dataset with other state-of-the-art approaches.

Approach	Actions	Test	Rate	fps
İkizler and Duygulu (2007)	9	LOSO	100%	N/A
Tran and Sorokin (2008)	10	LOSO	100%	N/A
Fathi and Mori (2008)	10	LOSO	100%	N/A
Hernández et al. (2011)[a]	10	LOAO	90.3%	98
Cheema et al. (2011)	9	LOSO	91.6%	56
Sadek et al. (2012)[a]	10	LOAO	97.8%	18
Our approach	10	LOSO	93.5%	188
Our approach	10	LOAO	95.7%	188
Our approach	9	LOAO	97.6%	188
Our approach[a]	10	LOAO	97.8%	188

[a] Using 90 out of 93 sequences (repeated samples are excluded).

houettes from the MuHAVi dataset have been manually annotated. This stands in contrast to the other datasets whose silhouettes have been obtained automatically, presenting therefore segmentation errors. This leads us to the conclusion that the visual feature of Dedeoğlu et al. (2006) is strongly dependant on the quality of the silhouettes.

Table 5.4 also shows the results that have been obtained with the different summary functions of our proposal. The *variance* summary representation, which only encodes the local dispersion and does not reflect the actual distance to the centroid, achieves an improvement in some tests at the cost of obtaining poor results on the MuHAVi actor-invariance tests (LOAO). The *max value* summary representation solves this problem and returns acceptable rates for all tests. Finally, with f_{range}, the *range* summary representation obtains the best overall recognition rates, achieving our highest rates for the Weizmann dataset and the MuHAVi LOSO tests. This leads us to choose this summary function in forthcoming proposals and tests.

In conclusion, the proposed *Radial Summary* feature improves the results obtained with similar features as Boulgouris et al. (2006); Dedeoğlu et al. (2006) substantially. Its low-dimensionality also offers an additional advantage in computational cost (feature size is reduced from $L \approx 300$ points in Dedeoğlu et al. (2006) to $B \approx 20$ radial bins in our approach).

Table 5.6: Comparison of recognition rates and speeds obtained on the MuHAVi-14 dataset with other state-of-the-art approaches.

Approach	LOSO	LOAO	fps
Singh et al. (2010)	82.4%	61.8%	N/A
Eweiwi et al. (2011)	91.9%	77.9%	N/A
Cheema et al. (2011)	86.0%	73.5%	56
Our approach	**95.6%**	**88.2%**	**93**

Table 5.7: Comparison of recognition rates and speeds obtained on the MuHAVi-8 dataset with other state-of-the-art approaches.

Approach	LOSO	LOAO	fps
Singh et al. (2010)	97.8%	76.4%	N/A
Martínez-Contreras et al. (2009)	98.4%	-	N/A
Eweiwi et al. (2011)	98.5%	85.3%	N/A
Cheema et al. (2011)	95.6%	83.1%	56
Our approach	**100%**	**97.1%**	**94**

Comparison with the state of the art Table 5.5 compares our approach with the state of the art regarding the Weizmann dataset. Some authors excluded the *skip* action because it tends to decrease the overall recognition rate, due to its inter-class similarity. Therefore, we indicated the number of actions used in each test. Several works achieve perfect recognition on this dataset, but most of them do not present any temporal evaluation and their suitability for real-time applications is arguable. It can be seen that our method places itself well in terms of both recognition accuracy and recognition speed, when comparing it to methods that target fast human action recognition. More importantly, the measured frame rate shows suitability for real-time applications.

On the MuHAVi-14 and MuHAVi-8 datasets our approach significantly outperforms the known recognition rates of the state of the art (see tables 5.6 and 5.7). To the best of our knowledge, this is the first work to report a perfect recognition on the MuHAVi-8 dataset when performing the LOSO cross validation test. The equivalent test on the MuHAVi-14 dataset returned an improvement of 9.6% in comparison with the work of Cheema et al. (2011), which also shows real-time suitability.

We also want to point out the robustness of our method with respect to the LOAO cross validation. Dissimilarities among action performances from different actors lie in pace, shape and motion. Our approach clearly outperforms latest results on both versions of the MuHAVi dataset. As seen in the results of Singh et al. (2010) and Cheema et al. (2011), this test presents a higher difficulty and the improvements achieved by our proposal constitute a significant benefit.

Regarding temporal performance, our proposal performs at over 180 and 90 fps on the Weizmann and MuHAVi datasets respectively. This difference is mainly due to the resolution of the binary segmentations, since the MuHAVi silhouettes are 16 times larger (see appendix A) and no resizing is applied. Note that with respect to the feature of Dedeoğlu et al. (2006), a performance gain of over 80% is achieved for the MuHAVi-14 dataset using our proposal for visual feature extraction. It is also worth mentioning that the training stage of the presented approach runs at similar rates between 92 and 221 fps.

Parametrisation

The presented method uses two parameters which are not given by the constraints of the dataset and the action classes which have to be recognised, and therefore have to be established by design. So far, it has been observed that the method's performance depends on the choice of these parameters. For this reason, the related behaviour should be further analysed. The first of these parameters is found at the feature extraction stage, *i.e.* the number B of radial bins. A lower value of B leads to a lower dimensionality which decreases the computational cost and may also improve noise filtering, but at the same time it reduces the amount of characteristic data. This data is needed in order to differentiate action classes. The second parameter is the number K of key poses per action class and view. In this case, the appropriate amount of representatives needs to be found to capture the most relevant characteristics of the sample distribution in the feature space, discarding outliers and non-relevant areas. Again, higher values lead to an increase of the computational cost of the classification. Therefore, a compromise needs to be reached between classification time and accuracy.

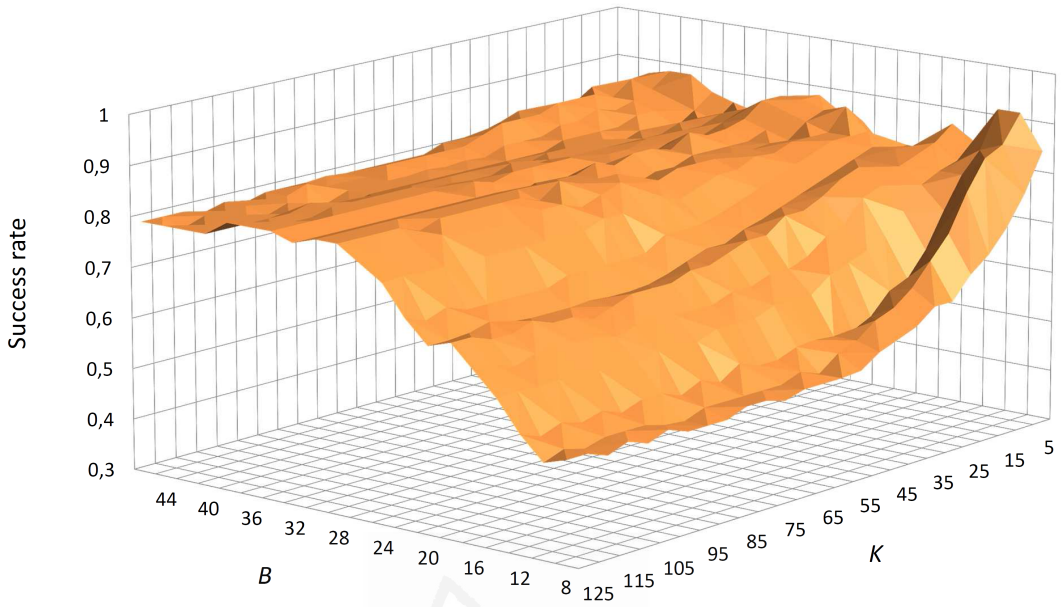


Figure 5.10: First quartile (Q1) values of the obtained success rates for $K \in [5,\ 130]$ and $B \in [8,\ 46]$ (MuHAVi-8 LOAO test). Note that outlier values below $1.5 \times IQR$ are not predominant.

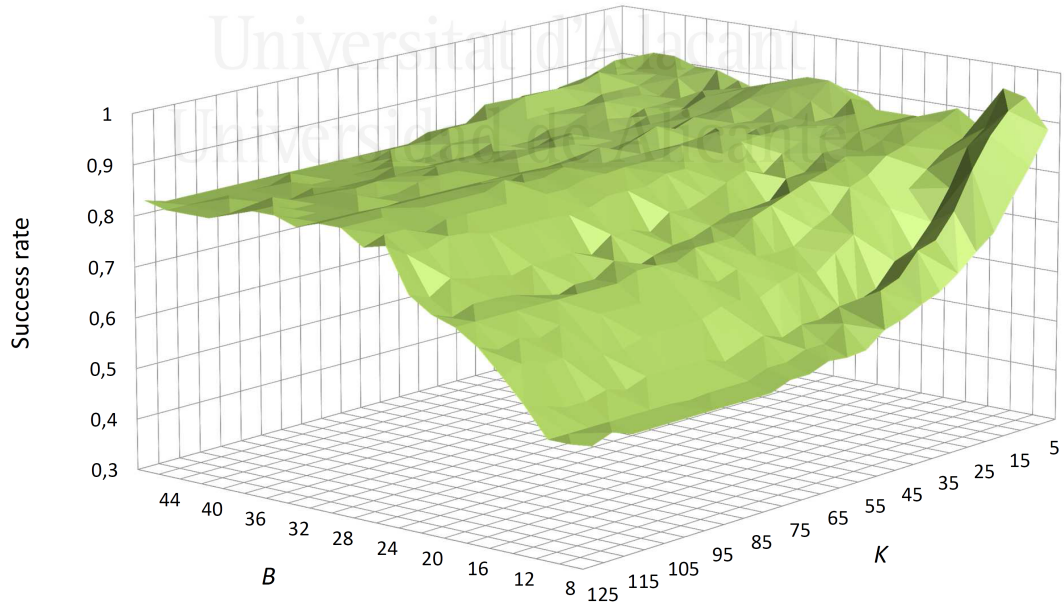


Figure 5.11: Third quartile (Q3) values of the obtained success rates for $K \in [5,\ 130]$ and $B \in [8,\ 46]$ (MuHAVi-8 LOAO test). Note that outlier values above $1.5 \times IQR$ are not predominant.

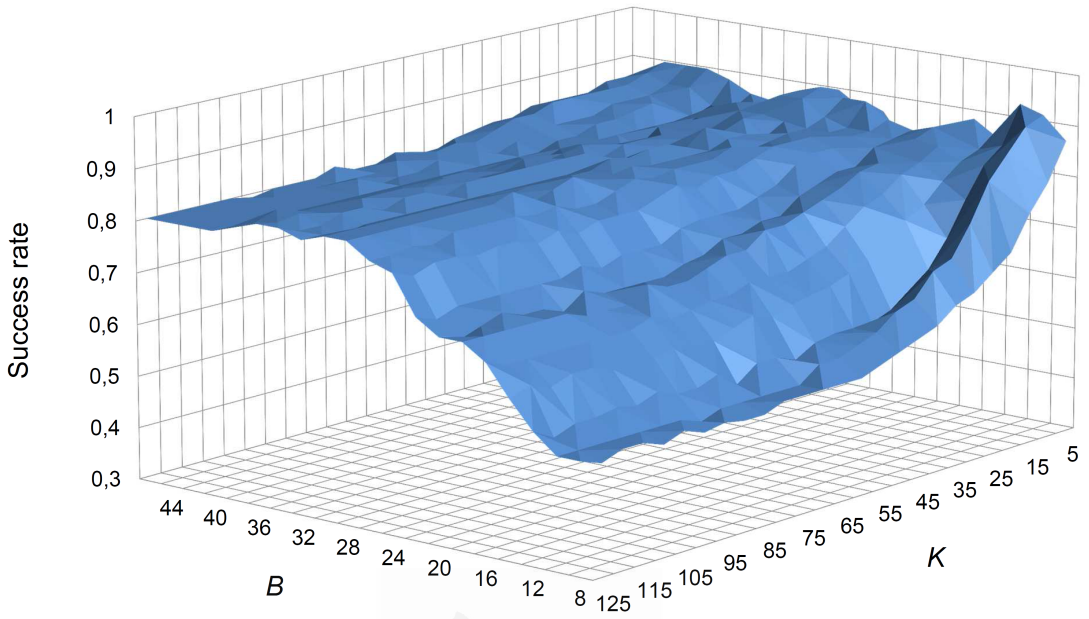


Figure 5.12: Median values of the obtained success rates for $K \in [5, 130]$ and $B \in [8, 46]$ (MuHAVi-8 LOAO test).

In order to examine the behaviour of the proposed algorithm with respect to these two parameters, a statistical analysis has been performed. Due to the non-deterministic behaviour of the K-means algorithm (caused by random initialisation), classification rates vary among executions. We executed fifty repetitions of the MuHAVi-8 LOAO cross validation with the same configuration (26 000 tests in total) and obtained the first quartile, third quartile and median values (see figures 5.10, 5.11 and 5.12). It can be observed that a high value of key poses, *i.e.* feature space representatives, only leads to a good classification rate if the feature dimensionality is not too low; otherwise, a few key poses are enough to capture the relevant areas of the feature space. Note also that a higher feature dimensionality does not necessarily require a higher number of key poses, since it does not imply a broader sample distribution of the feature space. Finally, with the purpose of obtaining high and reproducible results, the parameter values have been chosen based on the highest median success rate (92.6%), which has been obtained with $B = 12$ and $K = 5$ in this case. Since lower values are preferred for both parameters, the lowest parameter values are used if several combinations reach the same median success rate.

In conclusion, in order to setup the method in an intelligent monitoring system, we recommend to follow the proposed selection of parameter values based on

Table 5.8: Comparison of feature and model fusion of multiple views for the IXMAS dataset. The result of the LOAO cross validation is shown.

Approach	**Cam0-1**	**Cam0-2**	**Cam0-3**	**Cam0-4**
Feature Fusion	80.8%	81.6%	83.6%	**89.9%**
Model Fusion	74.2%	75.0%	**76.3%**	75.3%

the highest median success rate obtained on the training set. However, in order to be able to compare our results with the state of the art, we employ maximal recognition rates.

Extended validation

So far, the performed evaluation has returned very satisfactory results confirming the suitability of this implementation of the method for multi-view HAR in real time. Nonetheless, doubts remain regarding the stability of the multi-view learning. An issue that comes to mind is that, although model fusion has shown a superior performance, it has only been tested with two views so far. Data from more viewing angles will lead to a more dense bag of key poses, and the recognition of the *best view* will be more difficult and computationally expensive, since all the views have to be tested. Therefore, further validation is required.

The IXMAS dataset includes 11 action classes which have been performed by 12 different actors, three times each. These 396 performances have been recorded from four side views and one top view (look up appendix A for sample images). These amount of views is considered quite high for our application scenario. In order to verify that the developed approach is able to combine information from five views and take advantage of the additional viewing angles, both the feature fusion and model fusion options have been tested on growing sets of views of the IXMAS dataset.

We assume as baseline for the comparison the recognition rate that is obtained with *Cam0* without any type of fusion. Applying the LOAO cross validation, which is commonly employed for this dataset, a rate of 67.7% is returned. The results obtained with fusion of multiple views, which can be seen in table 5.8, confirm our doubts. Using the feature fusion, the addition of more views continuously provides a significant increase in the recognition rate. In contrast, for the model fusion only a slight improvement is observed until the top-view camera (*Cam4*) is added. Furthermore, for this dataset feature fusion performs better.

From two to five views a difference from 6.6 upto 14.6% can be observed. This can be attributed to the conditions in which the dataset was recorded. The actors were asked to behave natural and to freely choose their position and orientation. As a result, and in contrast to the MuHAVi dataset, cameras do not match with viewing angles across samples. This leads us to the conclusion that a more robust technique for combination of multiple views has to be developed in order to obtain successful and stable results across datasets.

5.1.4 Discussion and conclusion

In this section, the implementation of the method proposed in chapter 4 has been broken down into two parts: 1) visual feature extraction for pose representation and 2) algorithmic choices for learning and recognition. In the first part, the *Radial Summary* feature has been proposed. This silhouette-based feature allows to spatially align the shape of the silhouette of the human body, and to obtain a very low dimensional feature vector by reducing each radial bin to a single value. Three different summaries have been proposed for the experimentation. In the second part, the algorithmic choices have been detailed regarding the appropriate algorithm for key pose generation, a feature fusion operator and suitable distance metrics. Once all the parts of the method were defined, experimental results have been obtained especially regarding the learning from multiple views and feature extraction stages. Regarding the parameters of the method, the influence of these has been studied and insight has been given on how they should be established. Although good results have been obtained during the experimentation, extended validation has suggested that a more sophisticated approach is needed for the learning from multiple views in order to achieve consistently good results and make the most out of the available data. For this reason, this goal will be pursued in the following section.

5.2 A weighted feature fusion scheme for improving multi-view recognition

In the last section, two approaches for the fusion of multiple views have been considered. Both presented advantages and disadvantages. The proposed model fusion learning scheme allows to recognise actions from any learnt viewpoint, without having to know the angle of the camera, and it only requires a single view in order to perform recognition. Although model fusion has shown a superior performance, this could only be achieved when the cameras consistently matched with specific viewing angles, which is not a realistic scenario. On the other hand, the feature fusion approach handles the recognition of all views at once, which lowers the computational cost. The experimental results obtained suggest that this approach is more stable, and is also able to continuously improve its result if more views are added to the recognition. A disadvantage with respect to the model fusion approach is that the *best view* is not determined and the results relies equally on all the available viewing angles.

This thought is further considered in this section. A proposal is given in order to obtain an intelligent fusion of multiple views and combine the advantages of both feature and model fusion. Experimental results are presented in order to verify the proposal. Two possible data acquisition options for the extraction of silhouettes are considered. These are based on background subtraction and depth-based segmentation. Finally, discussion and conclusions are presented.

5.2.1 A weighted feature fusion scheme

Some viewing angles may be more or less useful than others depending on the captured images. Intuitively, the front camera should provide information with a greater characteristic value than the one that is recording the rear, since actions are normally performed towards the front. On the other hand, they may be equally useful if actions are performed sidewise. In order to confirm this idea, single-view recognition results have been studied. It has been observed that, in general, lateral views perform similarly when several action classes are considered, and only the top view returned steadily worse results (see table 5.9). This behaviour can be explained regarding the different types of actions. Each one of them may be recognised more easily from different views depending on the in-

Table 5.9: Comparison of single-view recognition rates obtained for the IXMAS dataset. The result of the LOAO cross validation is shown (bold indicates highest).

Cam0	**Cam1**	**Cam2**	**Cam3**	**Cam4**
67.7%	**78.8%**	69.9%	75.0%	55.1%

volved motion, and on the position and orientation of the subject. Therefore two conclusions can be stated: First, the usefulness of each view must be obtained automatically in the learning process, since it depends on the specific camera setup. Second, the specific weights to be assigned to each view depend on the action to recognise, since different action classes may be recognised best from different viewing angles.

In this sense, a weighted feature fusion scheme is proposed, that uses specific weights for each view and action class, and takes them into account in the comparison of features. Thus, intelligent information fusion can be applied using *a priori* knowledge about the input data. For this purpose, the feature fusion approach based on feature concatenation presented in section 5.1.2 is further developed below.

Let us recall that M stands for the available camera views and A for the number of action classes that have to be learnt. The camera weights are obtained as follows:

$$r_{m,a} = \text{Test}(m)_a, \quad \forall m \in [1...M] \quad \wedge \quad \forall a \in [1...A] \ , \tag{5.16}$$

where $\text{Test}(m)$ evaluates the recognition of the A action classes in a single-view test using only view m, and returns an array of the recognition results of each action class. Therefore, the per-class success rates of each camera are used as weights in order to determine how useful a camera is when recognising each of the action classes. Finally, these weights are normalised to unit-sum.

In order to apply a weighted distance between multi-view key poses, we redefine $d(kp, kp')$, where the previously obtained camera weights $r_{m,a}$ are now employed in order to consider feature distance and viewpoint relevance instead of the relevance of the match of key poses. Since each action class has a different weight, we take the known class a of the training sequence Seq so as to choose the appropriate weight. In other words, we always suppose that the current comparison is a right match and therefore apply the corresponding weights,

which indicate to which degree each view should be taken into account for this particular class. The distance $d(kp, kp')$ is obtained as follows:

$$d(kp, kp') = \sum_{m=1}^{M} r_{m,a}(\|\bar{V}_m - \bar{V}'_m\|_1) \ , \tag{5.17}$$

where $kp_i = \bar{V}_1 \circ \bar{V}_2 \circ \ldots \circ \bar{V}_M$ and $kp'_j = \bar{V}'_1 \circ \bar{V}'_2 \circ \ldots \circ \bar{V'}_M$.

Using this weighted feature fusion scheme, recognition is performed as detailed in 4.3.1 by returning the label of the best match, *i.e.* the instance with the lowest DTW distance.

Note that camera setups are subject to change. This method does not rely on camera calibration, and small viewpoint variations are inherently supported by the contour-based feature (Ángeles Mendoza and Pérez de la Blanca, 2007). Each camera view is explicitly considered in the feature fusion by preserving the origin of the individual feature part. In the case that a fewer number of cameras were available in the test scenario, only the matching camera views would be compared. Thus, as happened for the model fusion case, recognition can be performed if at least one of the camera views used during training is available.

Figure 5.13 shows a diagram that details the different stages of the method and summarises the specific decisions that have been made for the development.

5.2.2 Experimentation

Most state-of-the-art research works, as those mentioned in section 2, include experimental results on publicly available datasets. This is useful and necessary in order to compare the different proposed approaches in terms of recognition accuracy. Nonetheless, most works only detail the results obtained on one or two specific datasets. We have found that further evaluations need to be performed in order to test the required robustness and generality, since commonly results vary significantly with respect to the type of data. This circumstance has also been observed during the experimentation of section 5.1. In this sense, and as stated in Nebel et al. (2011), suitability of current methods for HAR in real-world applications is still arguable. For this reason, so as to confirm the suitability of the present approach, in this section a wide experimentation is made, and performance results are included for all the tests. Please note that the same evaluation methodology used in the last section has been employed. In addition,

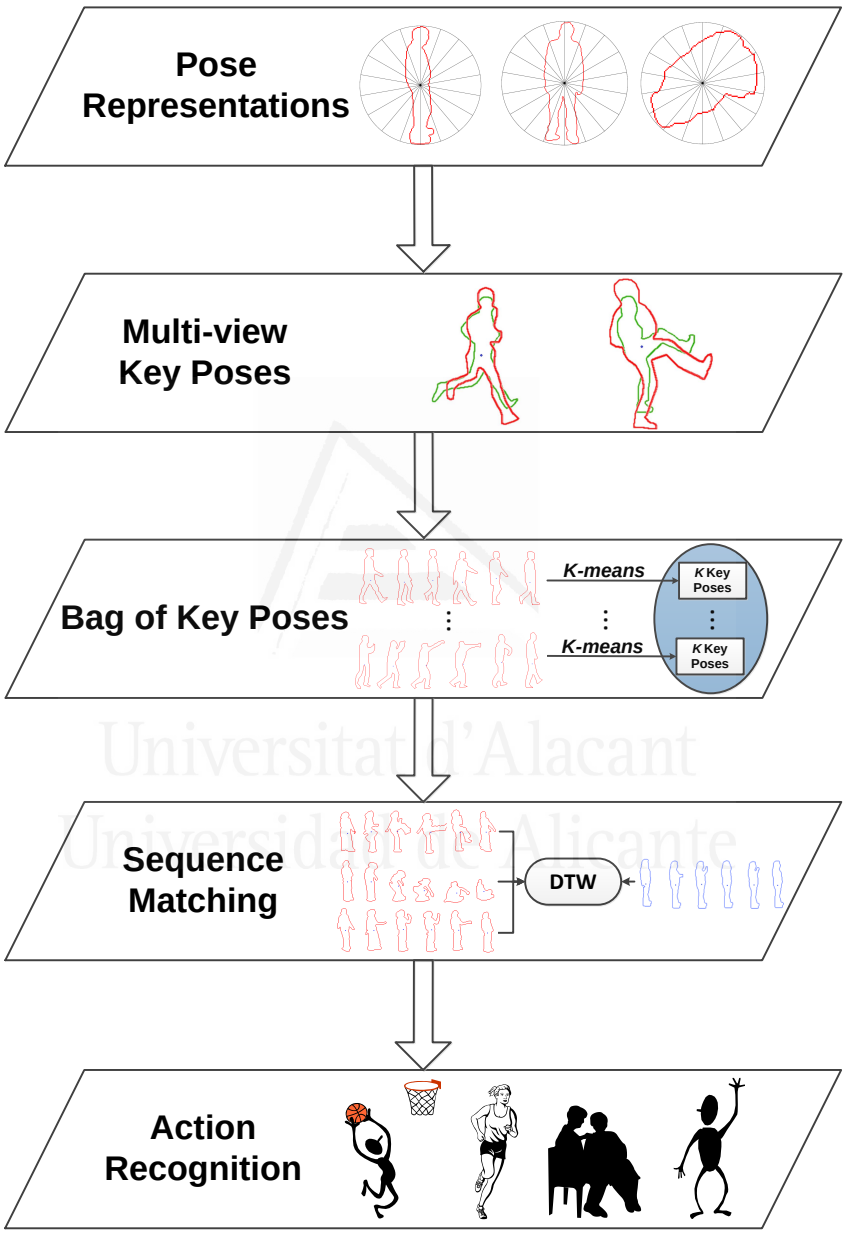


Figure 5.13: Outline of the stages of the method and the applied techniques.

some implementation-specific optimisations have been developed: 1) we have introduced thread-level parallelism to the K-means algorithm, and 2) we used an upper bound when searching for the *nearest neighbour* key poses. The new recognition rates are compared with those that have been obtained in the last section and with the corresponding state-of-the-art works.

Finally, the constant parameters of the algorithm (the number of radial bins B and key poses K) have been chosen based on exhaustive experimentation. The presented results have been obtained with $B \in [10, 34]$ and $K \in [4, 130]$. Note that in order to apply the proposed weighted feature fusion scheme, the camera weights are learnt using the same type of cross validation test but with single-view data.

Using RGB data

In this first part of the experimentation, silhouettes retrieved from traditional RGB colour images are employed. As the foreground masks are provided by the corresponding datasets, comparison between silhouette-based approaches can be performed.

MuHAVi dataset Since the proposed method targets multi-view human action recognition, the MuHAVi dataset has also been included in this experimentation. It is one of the multi-view datasets with the highest image quality. The image resolution is 720×576 px, and manually annotated silhouettes are provided for a subset of two camera views (front-side and 45°). We tested both versions of the dataset with either 14 (MuHAVi-14) or the simplified 8 (MuHAVi-8) action classes.

In order to provide a complete comparative analysis, we followed the authors guidelines and applied, in addition to the LOSO and LOAO cross validations, a leave-one-view-out cross validation (LOVO). Similarly to the other tests, in this one, view invariance is tested by using one view for training and the other for testing, and vice versa. Accordingly, single-view pose representations are employed in this test.

As table 5.10 shows, in MuHAVi-8 perfect recognition has been obtained for the LOSO and LOAO cross validation tests, and, at the same time, a superior recognition speed is achieved. In the case of MuHAVi-14, the currently available recognition rates have also been significantly outperformed (see table 5.11). Our

Table 5.10: Comparison of recognition rates and speeds obtained on the MuHAVi-8 dataset with other state-of-the-art approaches.

Approach	**LOSO**	**LOAO**	**LOVO**	**fps**
Singh et al. (2010)	97.8%	76.4%	50.0%	N/A
Martínez-Contreras et al. (2009)	98.4%	-	-	N/A
Eweiwi et al. (2011)	98.5%	85.3%	38.2%	N/A
Cheema et al. (2011)	95.6%	83.1%	57.4%	56
Model Fusion	100%	97.1%	-	94
Weighted Feature Fusion Scheme	**100%**	**100%**	**82.4%**	**98**

Table 5.11: Comparison of recognition rates and speeds obtained on the MuHAVi-14 dataset with other state-of-the-art approaches.

Approach	**LOSO**	**LOAO**	**LOVO**	**fps**
Singh et al. (2010)	82.4%	61.8%	42.6%	N/A
Eweiwi et al. (2011)	91.9%	77.9%	55.8%	N/A
Cheema et al. (2011)	86.0%	73.5%	50.0%	56
Model Fusion	95.6%	88.2%	-	93
Weighted Feature Fusion Scheme	**98.5%**	**94.1%**	**59.6%**	**99**

method is not intended for cross-view recognition and no fusion of multiple views can be applied in this case. Despite of this fact, the behaviour of the proposed radial silhouette-based feature considering a 45° shift is promising. To the best of our knowledge, these are the highest results reported so far on this dataset.

IXMAS dataset The INRIA Xmas Dataset (Weinland et al., 2006) is the most popular multi-view human action recognition dataset. This benchmark presents an above average difficulty due to multiple reasons: 1) the important number of different actions (11) and actors (12) results in a higher inter-class similarity, which explains why the available methods commonly achieve lower recognition rates on this dataset (*e.g.* Liu et al. (2013); Razzaghi et al. (2013)); 2) as it has been seen earlier, cameras do not match with viewing angles, so this dataset is implicitly testing view invariance; and 3) a few actors performed actions differently as, for instance, with the opposite body part, which is why several authors decided to exclude some samples or actors.

In our tests, we used the configuration that has been proposed by the authors of the dataset. Therefore, the *point* and *throw* actions have been excluded.

Table 5.12: Comparison with other multi-view human action recognition approaches of the state of the art. The rates obtained in the LOAO cross validation performed on the IXMAS dataset are shown (except for (Cherla et al., 2008) where the type of test is not stated).

Approach	Actions	Actors	Views	Rate	fps
Yan et al. (2008)	11	12	4	78%	N/A
Wu et al. (2011)	12	12	4	89.4%	N/A
Cilla et al. (2012)	11	12	5	91.3%	N/A
Weinland et al. (2006)	11	10	5	93.3%	N/A
Cilla et al. (2013)	11	10	5	94.0%	N/A
Holte et al. (2012)	13	12	5	100%	N/A
Cherla et al. (2008)	13	N/A	4	80.1%	20
Weinland et al. (2010)	11	10	5	83.5%	∼500
Model Fusion	11	12	4	76.3%	137
Feature Fusion	11	12	5	89.9%	207
Weighted Feature Fusion Scheme	11	12	5	91.4%	207

Table 5.12 shows the comparison of our results with the state of the art. The number of action classes, actors and views is indicated, since test configuration vary. It can be observed that both Yan et al. (2008) and Cherla et al. (2008) excluded the top view, and Wu et al. (2011) obtained their best result excluding one of the side views. Recently, Holte et al. (2012) achieved perfect recognition using a sophisticated method based on 4D spatio-temporal interest points and optical flow histograms. However, these recognition rates decrease when we look at approaches that target real-time applications. When comparing both the recognition rate and speed, our method stands out achieving 91.4% recognition accuracy at 207 fps with the proposed weighted feature fusion scheme.

In order to give further insight about how the weighted feature fusion scheme considers the multiple views, in table 5.13 the specific normalised weights of each view and action class are detailed. It can be observed that on average *cam1* presents a slightly higher weight. This is related to the camera setup, *cam1* mostly recorded the front view. Its greater discriminative value, which also has been observed in the single-view recognition rates detailed in table 5.9, leads to higher weights in general. Similarly, *cam4* recorded a top view which makes it difficult to distinguish some actions (as *sit down* and *get up*). However, when taking a closer look at the action class-specific weights that have been assigned, considerable differences among the action classes show up. The fine-

Table 5.13: Camera weights that have been obtained for the five views of the IXMAS dataset using the proposed weighted feature fusion scheme.

Action	**cam0**	**cam1**	**cam2**	**cam3**	**cam4**
check watch	0.18	0.24	0.19	0.18	0.21
cross arms	0.16	0.23	0.21	0.20	0.20
scratch head	0.17	0.21	0.20	0.23	0.19
sit down	0.23	0.23	0.21	0.23	0.10
get up	0.22	0.23	0.21	0.23	0.11
turn around	0.20	0.21	0.20	0.21	0.18
walk	0.22	0.21	0.20	0.21	0.16
wave	0.20	0.25	0.16	0.16	0.23
punch	0.20	0.20	0.23	0.18	0.19
kick	0.21	0.26	0.17	0.25	0.11
pick up	0.21	0.21	0.21	0.22	0.15
Average	0.20	0.23	0.20	0.21	0.17

grained camera-action weights allow us to capture this condition, and apply an appropriate feature fusion.

As it has been mentioned before, the IXMAS dataset requires the classification algorithm to support the recognition from arbitrary views, *i.e.* it cannot be assumed that a camera view only recorded a specific orientation of the subject. This is unavoidable in a real scenario, since subjects should be able to perform an action regardless of their orientation or location in the field of view. Regarding this matter, Cherla et al. (2008) reorganised the available video images in six out of eight possible 45° orientations so as to achieve view consistency. In order to evaluate how much our result was influenced by subject orientation, we applied a similar configuration with seven side views and one top view. Interestingly, the obtained recognition rates did not change significantly ($\pm$ $\sim$1%). This can be made clear looking closer at the learning stage of the algorithm. Even if a single camera view contains multiple orientations, the obtained key poses do represent these differences as long as samples are included in the training data. Therefore, inconsistent viewing angles can be recognised and, although our silhouette-based feature is not view-invariant, the learning algorithm itself handles this situation well and does not necessarily depend on fixed orientations.

Using depth data

Once the desired recognition robustness and speed has been reached, the next question to address is the data acquisition problem. The presented method relies on previously obtained human silhouettes, and certainly depends on their quality and completeness. Commonly, human silhouettes are extracted based on traditional background subtraction techniques, which, as shown in Horprasert et al. (1999); Kim et al. (2005), can perform in real time. Although an acceptable result can be obtained, difficulties related to shadows and lighting remain (Cristani et al., 2010). We have shown in previous sections that the presented visual feature achieves to reduce this noise by grouping the contour points into radial bins. Nevertheless, recent advances related to low-priced depth sensors, as the Microsoft Kinect, allow to obtain reliable depth information in real time. Furthermore, as these depth sensors rely on structured infra-red light, good results can also be obtained in darkness. This is convenient for the recognition of ADL at home, where no or few lighting may be available during the night. Therefore, we chose to validate the data acquisition stage of our proposal using the human silhouettes provided by these kind of RGB-D sensors. Since our target application is related to providing AAL services at home, and the Microsoft Kinect shows to be proficient in indoor scenarios, this option suits us particularly well both as a low-cost and real-time applicable solution. In addition, depth, infra-red or laser sensors allow to preserve privacy, as RGB information is not essential for silhouette-based human action recognition.

In this sense, the presented method has been tested on two datasets that contain depth data. The first one, *DAI RGBD Dataset* has been recorded by ourselves in order to acquire multi-view depth data. In the case of the second one, a much larger publicly available dataset has been used.

DAI RGBD Dataset In our setup we used two Microsoft Kinect devices recording a front and a 135° backside view. Twelve actions have been performed by three male actors with noteworthy differences related to their body build. The performed actions include *bending, standing still, sitting down* and *standing up*, among others. A few sample silhouettes are shown in figure 5.14. These have been obtained using depth-based segmentation.

Table 5.14 shows the results that have been achieved using the depth-based silhouettes for both types of cross validations. So as to provide further insight,

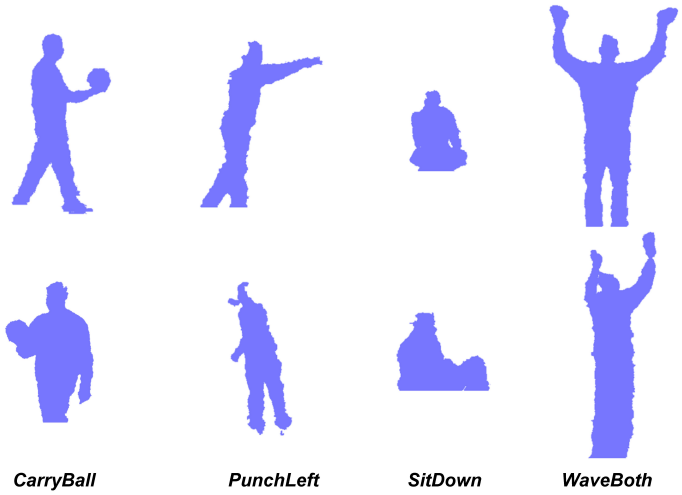


Figure 5.14: Sample silhouettes of our RGB-D dataset: front view at the top and backside view at the bottom.

the rates with feature fusion of multiple views are provided for reference. As it can be seen, the a priori knowledge about the proficiency of each camera in recognising specific action classes leads to a substantial increase in recognition accuracy.

DHA dataset The depth-included human action video dataset (DHA) (Lin et al., 2012) is aimed at providing a large dataset for human action recognition relying on single-view depth data. It contains 23 action classes performed by 21 different actors (12 male and 9 female). Both colour and depth data are provided.

Specifically, we used the configuration which includes the ten actions of the Weizmann dataset and applied the LOSO cross validation (see table 5.15). The obtained recognition rate outperforms the baseline by 4.4%. Tests have also been performed on the complete dataset with all the 23 action classes. In this case, our method reaches a success rate of 72.1%. The larger number of classes leads to lower inter-class distance and consequently to more misclassifications, since our approach is similarity-based.

Table 5.14: Cross validation results obtained on our multi-view depth dataset.

Approach	**LOSO**	**LOAO**	**fps**
Feature Fusion	83.3%	94.4%	80
Weighted Feature Fusion Scheme	**94.4%**	**100%**	80

Table 5.15: LOSO cross validation results obtained on the DHA dataset (10 Weizmann actions).

Approach	**LOSO**	**fps**
Lin et al. (2012)	90.8%	N/A[a]
Weighted Feature Fusion Scheme	**95.2%**	99

[a] An average descriptor extraction time of 3.9 s per video sequence is stated.

In conclusion, the performed experimentation shows that the proposed approach reaches high recognition rates steadily. Furthermore, an outstanding proficiency to actor variance has been unveiled by the LOAO cross validation tests. Regarding the temporal performance, a recognition speed that is significantly above video frequency has been measured. Note that the varying recognition speeds are caused by the specific parametric configuration of the tests as well as the different image resolutions. It is also worth mentioning that the contour extraction stage took on average ∼84% of the total processing time, which means that after the contour points have been obtained, the processing runs at over 1000 fps, leading to an almost insignificant recognition time.

5.2.3 Discussion and conclusion

In this section, an improved implementation of the proposed human action recognition method based on multi-view fusion and key pose sequence classification has been presented. A multi-view pose representation is generated by means of feature fusion based on feature vector concatenation. Instead of a *blind* concatenation, the aptitude of each viewing angle is evaluated performing single-view tests, and a weighted feature fusion scheme is applied in the feature comparison. This approach takes into account the usefulness of each camera in recognising a specific action class, because actions may be recognised best from different viewing angles. An extensive experimentation has been performed on publicly available datasets, which include human silhouettes obtained either manually or automatically. The method's data acquisition has also been validated by using human silhouettes that have been obtained with depth cameras performing in real time. Results show the sought-after behaviour, not only outperforming (both in terms of recognition accuracy and speed) other state-of-the-art methods that target real-time suitability, but also demonstrating the desired robustness among a wide variety of action classes, actors and other test scenario-related conditions. This leads us

to consider that this approach can be used for real-world applications. Due to its temporal efficiency, it could also be employed as a part of a long-term human behaviour recognition system, which relies on the result of HAR to recognise behaviours.

The proposed weighted feature fusion scheme has been tested with up to five viewing angles. Although it is improbable in our application scenario, more camera views could be available. A much higher number of cameras would lead to a temporal performance decay, since the feature dimensionality would increase with each new view. Therefore another feature fusion technique, as an aggregation function, should be considered in order to reduce the dimensionality of the resulting feature vector. Finally, our method does not target view invariance explicitly, and even if some view shifts are successfully supported, a sufficient subset of the possible subject orientations needs to be included in the training data.

5.3 Remarks

In this chapter, the implementation of the method for human action recognition presented in chapter 4 has been studied. Particular attention has been given to the pose representation and the learning from multiple views, as well as to the behaviour of the method with respect to its parameters and the data acquisition stage. Regarding the first, the *Radial Summary* feature has been proposed, which achieves to outperform other silhouette-based representations relying on spatial alignment and dimensionality reduction. With respect to the learning from multiple views, it has been observed that proposals at the feature and model level bring along different advantages and disadvantages. The required robustness across different datasets with varying scenarios has been achieved with a weighted feature fusion scheme. Experimental results confirm that the designed method is proficient in the recognition of a variety of human actions performed in different conditions. The method stands out in its ability of fusing multiple views and reaching a superior recognition speed. Due to the need of human silhouettes in order to be able to apply the proposed approach, data acquisition has been validated using both RGB and depth data.

In conclusion, in view of the obtained results, the presented method is considered to provide an efficient and robust technique for the recognition of human

actions, especially for our application scenario. However, due to the method's simplicity of implementation and reduced temporal cost, it can be used in a wide range of applications as HCI, gaming and video surveillance. The method can relatively easily be developed for a low-cost embedded hardware architecture, and more interestingly, it can be employed for the recognition of higher HBA levels.

Chapter 6

Optimising human action recognition

Chapters 4 and 5 have described how human action recognition can be performed based on vision in order to ease human behaviour analysis in people's homes. The desired robustness of the method towards different subject- and scenario-related conditions has been verified based on benchmarks of varying characteristics. However, once the conditions of the application scenario are known, it is self-evident that the recognition could be improved if the method were able to adjust to its specific characteristics. Among others, these are the actions that have to be recognised, the subject's way of performing them and the environment which is being monitored. In this sense, different optimisations of the presented method for human action recognition are introduced in this chapter. In these proposals, the best performing feature subset, training set and parameters are sought in order to improve the final recognition. Furthermore, an adaptive learning scheme is proposed to support learning in execution time by means of an evolving bag of key poses.

Exhaustive search algorithms are not suitable for this purpose because of the excessively large number of possible configurations. Since we do not know the solution landscape of the optimisation problem, and there may be multiple optimal solutions and local optima, we are also not able to determine an analytic method to solve them. For this reason, evolutionary algorithms (EA) may be employed to find 'good solutions', as EA support the guided search of multi-dimensional spaces. These nature-inspired algorithms rely on a population of individuals, where each individual encodes a potential solution of the optimisa-

tion problem. A fitness value, which is directly related to the given optimality criterion, is defined for the individuals. In this way, a fitness-biased selection is then applied through generations in order to evolve to increasingly better solution candidates. Once convergence is reached, the solution is commonly provided by the fittest individual. For more details about EA techniques and their application to optimisation problems, we refer to De Jong (2006).

In the following sections, three different proposals are studied for the optimisation of the presented HAR method for different scenarios and from lower to higher complexity of optimisation objectives. Please note that, due to the simultaneous development of classification and optimisation methods during this thesis, the first proposal is applied on the first implementation of the method presented in section 5.1, whereas the second and third ones use the improved fusion of multiple views proposed in section 5.2.

6.1 Feature subset selection

6.1.1 Introduction

As it has been seen in chapter 2, current efforts in human action recognition rely on achieving the required robustness to instance, actor and view variance (Poppe, 2010), as well as real-time suitability (Shotton et al., 2011), which is essential in most HBA applications. Therefore, the current goal is to improve existing approaches in order to reach high and stable recognition rates, and, at the same time, simplify the classification algorithm and feature extraction process, thus minimising the involved computational cost. In this sense, we propose and evaluate novel visual feature-related improvements which especially address these goals.

The feature subset problem arises when using feature vectors, since these might contain irrelevant or redundant information that could negatively affect the accuracy of a classifier (Cantú-Paz, 2004). In fact, irrelevant and redundant features interfere with useful ones, and most classifiers are unable to discriminate them (Kira and Rendell, 1992; Lanzi, 1997). Also, identifying relevant features is of great importance for classification tasks, as it improves accuracy and reduces the computational costs involved (Casado Yusta, 2009; Yang and Honavar, 1998).

In the following, a human action recognition optimisation based on evolutionary feature subset selection is proposed. The *Radial Summary* feature presented in section 5.1.1 is especially suitable for feature subset selection, because its spatial definition is inherently related to the parts of the human body. Body parts as arms, legs, hip and head might not be necessary, or they may even introduce noise to the feature descriptor. With this in mind, an evolutionary feature subset selection is introduced to further improve the feature extraction stage. This proposal is evaluated using the human action recognition approach presented in section 5.1, relying on model fusion of multiple views in this case, and performing an experimentation on the publicly available MuHAVi dataset.

Evolutionary feature subset selection

Evolutionary feature subset selection has been used for decades (Siedlecki and Sklansky, 1989). The basic approach is to consider a binary vector where each gene represents whether or not a specific feature element is considered. Two main models are presented to implement this (Cantú-Paz, 2004; Casado Yusta, 2009): the *wrapper* model, and the *filter* model. The latter selects the features based on *a priori* decisions on feature relevance according to some measures (information, distance, dependence or consistency) (Liu and Motoda, 1998), but it ignores the learning algorithm underneath. In the former, and on the contrary, the feature selection algorithm exists as a *wrapper* around the learning algorithm. In the search of a feature subset, the learning algorithm itself is used as a part of the evaluation function (John et al., 1994). The main disadvantage of the wrapper approach is the time needed for the evaluation of each feature subset (Lanzi, 1997). On the other hand, filter-based approaches, although faster, find worse subsets in general (Cantú-Paz, 2004). A third possibility is to embed the feature selection in the construction of the classifier. During training, the classifier selects the appropriate features to improve the results (Guyon and Elisseeff, 2003). Besides, as stated by Espejo et al. (2010), the application of genetic programming (GP) for inducing classifiers usually implies a feature selection process which is inherent to the evolution of classifiers.

Evolutionary feature subset selection has been used broadly in computer vision applications, including: gender recognition from facial images (Sun et al., 2002), object detection (faces and vehicles) (Sun et al., 2004), target detection in synthetic aperture radar (SAR) images (Bhanu and Lin, 2003), image annotation

(Li et al., 2010; Lu et al., 2008), and seed discrimination (Chtioui et al., 1998). For a complete review of feature subset selection see the work of Casado Yusta (2009).

Feature selection for human action recognition

Regarding the specific application of feature selection to HAR, a few works can be found in the literature. The approach of Jhuang et al. (2007) uses a zero-norm SVM for feature selection of position-invariant spatio-temporal features. They concluded that feature density could be reduced up to 24 times without sacrificing accuracy. A multi-class delta latent Dirichlet allocation model for feature selection is introduced in Bregonzio et al. (2010), where collaborative selection is performed by taking into account the correlation between samples. Feature selection can also be performed based on a different criteria instead of the highest classification accuracy. In this sense, in Kovashka and Grauman (2010), multiple kernel learning (MKL) is applied in order to reach the most discriminative neighbourhoods among interest points. Similarly, feature selection based on entropy is used in İkizler et al. (2008). In this work, a new shape descriptor is defined based on the distribution of border lines, which are histogrammed over 15° orientations. Selecting the bins with the highest entropy, the regions in which most of the motion occurs are found. In this way, feature size could be reduced by a factor of three.

In conclusion, we have seen that evolutionary algorithms are proficient for feature subset selection and, on the other hand, feature selection has been used successfully before for similar purposes. This encourages us to apply an evolutionary approach for human action recognition in order to find the best performing subset for our feature.

6.1.2 Evolutionary algorithm

In this work, an evolutionary wrapper approach with a steady-state reproduction scheme is proposed for feature subset selection. Therefore, a binary selection approach is employed in which, through evolution of individuals, the optimal choice of feature elements is sought.

Specifically, an individual of the population is encoded as a binary array Q whose elements $q_j, \forall j \in [1...B]$ represent whether or not a radial bin is selected (respectively $q_j = 1$ or $q_j = 0$). Having fixed the population size N_I, the following steps are performed:

1. The population is initialised with one individual with $q_j = 1, \forall j \in [1...B]$, so as to consider the default feature from the first iteration on. The values of the remaining individuals are set randomly.

2. The fitness value of each individual is evaluated. The aptitude of each of the generated feature subset selections is tested using the human action recognition approach, detailed in section 5.1, with model fusion of multiple views. The fitness value of the individual is set to the returned success rate.

3. The population is ordered by descending fitness value. In the case that two individuals present the same fitness value, the one with the lowest number of selected elements ($q_j = 1$) comes first. Hence, among the fittest individuals those are preferred that result in a higher temporal and spatial efficiency.

4. A new individual is created using one-point crossover. Parents are chosen using the ranking selection method. The new individual is built up by the elements 1 to o from the first parent and $o+1$ to B from the second parent (o denotes the randomly selected crossover point).

5. Each gene of the new individual is mutated according to a probability mut_F. This mutation consists in changing the binary value of the gene.

6. The fitness value for the new individual is obtained. If it does not exist in the current population, it is added to the population. Otherwise, the fitness is updated with the new value in case it is improved. Recall that the non-deterministic behaviour of the classification process, which is caused by the random initialisation of the K-means algorithm, leads to obtain different fitness values for the same feature vectors.

7. The individual with the lowest fitness value is removed from the population in order to apply elitism.

8. The algorithm returns to step 3 until the best-performing individual does not increase its fitness value for a specified number of generations max_{gen}.

In this way, through continuous evolution and selection of the fittest individuals, the sought feature selection is found. Each time a new individual is created, its fitness value is obtained. This value is defined as the success rate of the human action recognition method using that specific selection of feature elements during the whole classification (learning and recognition).

6.1.3 Experimentation

The proposed approach has been evaluated on the MuHAVi dataset. This allows us to compare the obtained performance with the results previously shown in section 5.1. These correspond to:

1. The proposed HAR method with model fusion of multiple views, where the distance-signal feature of Dedeoğlu et al. (2006) was employed (named *HAR method* from now on).

2. The use of the better performing *Radial Summary* feature, but without the proposed feature selection (named *HAR method + Radial Summary* from now on).

For this purpose, the same evaluation methodology (see section 5.1.3) has been employed. With respect to the constant parameters of the method, the present results have been obtained relying on the best performing setup without optimisation ($B \in [12, 30]$ and $K \in [9, 100]$). In the evolutionary algorithm, a population of $N_I = 10$ individuals and a single offspring per generation have been used. The remaining parameters have been set empirically as follows: $mut_F = 0.1$ and $max_{gen} = 250$. Note that using the specified data, convergence has been reached in a limited number of generations (under ∼900), and therefore no alternative stopping criteria has been applied.

In table 6.1, the recognition rates are shown for both types of cross validation tests. The second last and last column represent the accuracy before and after optimisation, respectively. We already know that when using the radial summary feature (individual with $q_j = 1, \forall j \in [1...B]$), the originally obtained recognition rates are improved. It seems obvious that starting with an adequate accuracy level is essential if we want to perform feature subset selection successfully. Furthermore, besides the low dimensionality of the feature, its definition, which is directly related to the parts of the body, leads to an inherent proficiency for feature subset selection. This is confirmed in the returned fitness values for the best

Table 6.1: Benchmark results obtained with the original method, the *Radial Summary* feature and the proposed binary feature subset selection. Both LOSO and LOAO cross validations are performed.

Dataset	Test	***HAR method***	***HAR Method + Radial Summary***	***Feature Subset Selection***
MuHAVi-14	LOSO	94.1%	95.6%	**98.5%**
MuHAVi-14	LOAO	86.8%	88.2%	**94.1%**
MuHAVi-8	LOSO	98.5%	**100%**	**100%**
MuHAVi-8	LOAO	95.6%	97.1%	**100%**

Table 6.2: Analysis of variance test for the LOAO cross validations. The results of 20 replicates have been compared between the *HAR Method* and the proposed feature subset selection of the *Radial Summary* feature. A confidence level of 99% ($\alpha = 0.01$) is considered.

Output measure	**Source of variation**	**Sum of squares**	**DoF**	**Mean squares**	***F***	***p*-value**
MuHAVi-14 LOAO	Between groups	0.229537	1	0.229537	947.862653	1.78102×10^{-28}
	Within groups	0.009202	38	0.000242		
MuHAVi-8 LOAO	Between groups	0.165585	1	0.165585	1630.778024	8.02151×10^{-33}
	Within groups	0.003858	38	0.000102		

evolved individuals. In both type of tests, significant performance improvements are shown.

In order to analyse the statistical significance of the obtained performance gain of this enhanced feature extraction over the original method, an analysis of variance (ANOVA) test has been performed for the results of the MuHAVi LOAO tests. In table 6.2, it can be seen that for both datasets highly statistically significant improvements have been obtained (.01 level). In both cases, a *large* effect size has been measured indicating that these improvements are also practically significant (for MuHAVi-14 and MuHAVi-8, Cohen's $d = 9.86$ and 12.94 respectively).

Further details about the generated feature subset selections are specified in table 6.3, where the values of the best individuals are shown. It can be observed that several radial bins may be ignored, and that by doing so, the feature size is reduced removing unnecessary elements. Moreover, the recognition accuracy is increased since these bins introduced noise. Interestingly, although perfect

Table 6.3: The values $u_j, \forall j \in [1...B]$ of the final individuals of each of the run tests.

Dataset	Test	B	Individual
MuHAVi-14	LOSO	16	*0100110100001110*
MuHAVi-14	LOAO	12	*011110101111*
MuHAVi-8	LOSO	12	*000111000011*
MuHAVi-8	LOAO	30	*000010111011111010110001001011*

Table 6.4: Comparison of recognition rates obtained on the MuHAVi dataset with other state-of-the-art approaches.

	MuHAVi-14		MuHAVi-8	
Approach	**LOSO**	**LOAO**	**LOSO**	**LOAO**
Singh et al. (2010) (baseline)	82.4%	61.8%	97.8%	76.4%
Martínez-Contreras et al. (2009)	-	-	98.4%	-
Cheema et al. (2011)	86.0%	73.5%	95.6%	83.1%
Eweiwi et al. (2011)	91.9%	77.9%	98.5%	85.3%
Feature Subset Selection	**98.5%**	**94.1%**	**100%**	**100%**

recognition has been obtained in the MuHAVi-8 LOSO cross validation test using the default feature, the applied evolutionary feature subset selection maintains this rate with less than half of the feature vector elements. On average, in our tests we achieved to improve or preserve the highest recognition rate reducing ∼47% of the feature elements. Figure 6.1 shows one of the selections on a sample silhouette.

Finally, our results are compared with other state-of-the-art recognition rates. As table 6.4 shows, very considerable improvements have been attained, especially regarding the actor-invariance tests. To the best of our knowledge, these are the highest rates reported so far for this dataset.

Temporal Evaluation

So as to verify the computational cost of the proposed approach, we tested our method with the same setup as in previous performance tests.

Regarding the feature extraction stage, most of the time, *i.e.* 76.36s, is spent processing the contour points of the image. The remaining feature extraction and classification stages are performed in only 1.50 and 5.86s respectively. Applying feature subset selection, this time is reduced by 14%.

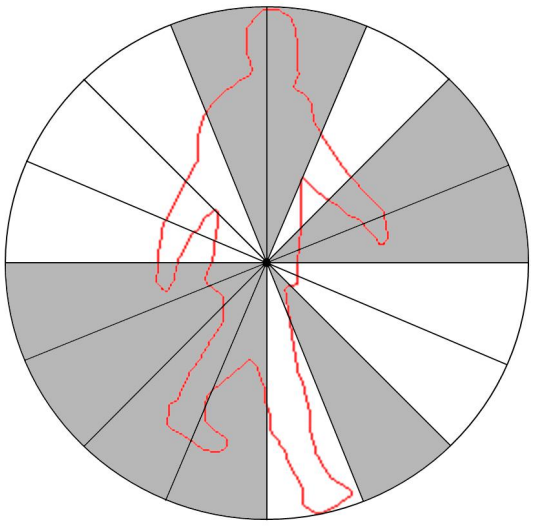


Figure 6.1: Resulting feature subset selection of the MuHAVi-14 LOSO cross validation test (dismissed radial bins are shaded in grey). Through evolutionary selection, 7 out of 16 radial bins have been selected, as these provided the best result.

6.1.4 Discussion and conclusion

In this section, a feature-level optimisation for human action recognition has been introduced. The previously presented *Radial Summary* feature has been further improved. The applied spatial alignment leads to a feature which is very proficient for feature subset selection. In order to find the best selection of radial bins, an evolutionary feature subset selection has been proposed for the binary selection of feature elements. An evolutionary algorithm is used as search heuristic so as to provide the best possible solution in a limited amount of time. The obtained optimisation achieved perfect recognition on the MuHAVi-8 LOSO and LOAO cross validation tests, reducing the number of necessary feature elements by ∼47% (on average). In result, the recognition rates of the original human action recognition approach have been increased, and new reference rates have been established for the MuHAVi dataset.

Regarding future works, it would be interesting to consider a feature selection based on real-valued weights, so that a feature element could be taken into account more or less than others without completely discarding it. This would open a wider search space with potential solutions for more fine-grained optimisations. Furthermore, the optimal selection of features may change for the different types of actions to recognise. *Walking*, *running* and *jogging* may affect mostly the legs and the arms, whereas in bending, the upper part of the body may be the

most relevant. Therefore, an action class-specific feature subset selection can be obtained. Similarly, since a multi-view setup is employed, a feature selection should be computed for each of the views, because the relevant body parts may change depending on the viewpoint. Multiple optimisation objectives, as both the highest recognition rate and the least number of selected feature elements, could also be considered.

6.2 Coevolutionary instance, feature and parameter selection

6.2.1 Introduction

So far, we have seen that feature selection represents a great advantage if the actions and the actors' ways of performing them are known. Nevertheless, the feature set is not the only configuration that can be subject to optimisation. For instance, the training set and the parameters of the method can also be optimised. In this section, multiple optimisation targets are addressed. First, we seek the best possible set of instances. Due to different kinds of recording errors, noise and instance- or subject-related peculiarities (as clothes, body build, etc.), not all the available instances of a training set are equally useful. Whereas having more samples is usually valuable to learn the intra-class variance of an action class, this is not the case for random noise appearances and outlier values, which tend to spoil and overfit the learning model (Cano et al., 2005). Furthermore, optimising the set of instances to the smallest one which maintains or improves the initial recognition rate leads to a significant spatial and temporal improvement. Second, since feature subset selection has shown to be successful, it is also applied in order to obtain the optimal selection of elements out of the feature vector. In this case, a feature selection is applied for the feature vector which results from the feature fusion of multiple views. This means that a feature selection is obtained for each view in order to take into account that different body parts may be relevant depending on the viewpoint. Finally, the optimal values of the method's parameters are determined in order to achieve the best empirical configuration, *i.e.* the one that leads to the highest recognition rate. The use of evolutionary algorithms for parameter selection is the most common, since it has been applied for decades (De Jong, 1975).

So as to compute these optimisations in an acceptable amount of time, a cooperative coevolutionary algorithm is proposed. Unsupervised selection of instances, features and parameter values is simultaneously performed by using a coevolutionary algorithm with a cooperative behaviour. This choice is motivated by the fact that coevolutionary algorithms make it possible to split the domain of the problem by relying on the *divide and conquer* strategy, and tackle each part of the optimisation problem with respect to their different solution spaces and data types. Furthermore, the cooperative coevolution allows us to consider the intrinsic dependencies which may exist between optimisation goals by using a global fitness function and evaluating the cooperation between populations (Derrac et al., 2012). As experimentation in section 6.2.3 will detail, this proposal achieves a significant increase in recognition accuracy and a considerable decrease in spatial and temporal complexity on publicly available datasets.

Data reduction based on evolutionary algorithms

As it has been introduced briefly, two of our optimisation targets address data reduction. These are the best performing selection of instances and feature subset that have to be sought. As stated in Liu and Motoda (2002), this can be seen as selecting rows (training instances) and columns (features) out of the training data. In this sense, a two-fold objective is pursued. First, the recognition rate can be improved by filtering noisy and outlier data (this data could lead to overfitting Wilson and Martinez (2000)). Thus, a more consistent learning model can be obtained. Second, execution time can be reduced without compromising the success rate if the redundant training data is ignored.

Whereas a solid state of the art exists regarding instance selection (Grochowski and Jankowski, 2004; Jankowski and Grochowski, 2004; Wilson and Martinez, 2000), EA are still sparingly being used for this purpose. Cano et al. (2003) elaborated a comparison between evolutionary and non-evolutionary instance and feature selection methods, and concluded that the former consistently performed better in both terms of recognition accuracy and spatial and temporal performance. A generational genetic algorithm (GA), a steady-state GA, a heterogeneous recombination and cataclysmic mutation (CHC) adaptive search algorithm, and a population-based incremental leaning specific EA have been included in the comparison (Cano et al., 2003). In García et al. (2008), a memetic algorithm is proposed for instance selection, tackling the problem of selection in large scale

databases. A cooperative coevolutionary algorithm is used for instance selection in García-Pedrajas et al. (2010), where the obtained results compared favourably with standard and also recently published state-of-the-art algorithms (see Garcia et al. (2012); Olvera-López et al. (2010) for more details).

Finally, some works apply both instance and feature subset selection simultaneously. Kuncheva and Jain (1999) used a GA to overcome the disadvantages of a consecutive approach. A similar method is presented in Ros et al. (2008), where additional heuristics are considered in order to promote diversity and elitism in the population. A cooperative coevolutionary algorithm is successfully employed in Derrac et al. (2010) on datasets of different data nature. Doucette et al. (2012); McIntyre and Heywood (2011) combined competitive and symbiotic (cooperative) coevolutionary multi-objective optimisation and GP classifiers. Competition provides a mechanism for scaling to potentially large unbalanced datasets, while cooperation allows the decomposition of the training set to improve the results. Feature subset selection is embedded in the GP classifiers.

Cooperative coevolutionary algorithms

A coevolutionary algorithm (CEA) can be defined as one or more EA, in which the fitness value of an individual of one of the populations depends on its relationships to the individuals from the other populations (Wiegand, 2004). In other words, the search problem is divided into sub-problems, where each population handles one of them separately. Nonetheless, the proficiency of the individuals is evaluated in a correlative way (Derrac et al., 2010). Coevolutionary algorithms can be categorised by means of the type of relation between individuals. Whereas *cooperative* CEA reward individuals that work well together, *competitive* CEA follow a predator-prey relationship rewarding those individuals whose opponents perform poorly against them (Wiegand, 2004).

Coevolutionary algorithms are being used successfully in different domains as process planning and scheduling (Kim et al., 2003), multiobjective optimisation (Tan et al., 2006) and clustering (Potter and Couldrey, 2010), among others.

6.2.2 Optimisation of multiple sets

In this section, the proposed cooperative coevolutionary algorithm for the simultaneous selection of instances, features and parameter values is detailed. For this

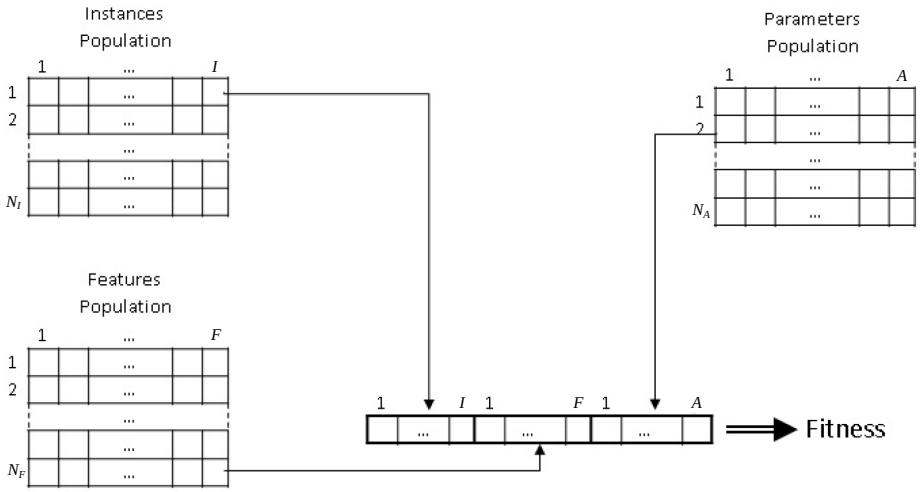


Figure 6.2: The fitness value is obtained by evaluating the human action recognition method with the configuration indicated by one individual from each population.

purpose, we will present how the populations and the individuals are defined, and which steps are executed in the coevolution.

Population structure

Since our problem of finding an optimal classification configuration is divided in the search of a selection of instances, a feature subset and parameter values, we use these sub-problems to define three populations. Individuals of each population are combined to build a possible solution that can be evaluated using the fitness function. In our case, the fitness function relies on the success rate of the human action recognition algorithm using the configuration determined by the individuals (see figure 6.2).

Note that in this case, our objective of optimisation is the final HAR method which relies on the improved fusion of multiple views based on the presented weighted feature fusion scheme (see section 5.2).

Individuals' representation

Instance and feature subset selections are regarded as binary selections. This means that a choice of the training instances that will be used during learning is established, and that the specific feature elements of the feature vector that will

be used during the classification are selected. Naturally, instance selection does not apply for the sequences that have to be recognised in a test.

The individuals of these two populations of instances and features are encoded as Boolean arrays, in which each gene indicates whether or not this element is selected. The individuals of the instance population have I elements, one for each training instance. And the individuals of the feature population have one element for each of the F elements of the feature vector, where B radial bins are included for each of the M views.

Considering the parameters of the HAR method, the number of radial bins B has to be established beforehand in order to apply feature selection and optimise the set and size of the feature vector. The number of key poses K remains subject to optimisation. Nonetheless, since multiple values can be set and optimised relying on individuals encoded as arrays, a more fine-grained parametrisation can be applied if a class-specific number of key poses K_a is learnt for each action class a. This would allow the method to consider the appropriate amount of representative poses for each action. Therefore, for A action classes, the same amount of parameters $K_1, K_2, ..., K_A$ are indicated, leading to an individual of A integer values.

Coevolutionary algorithm

Algorithm 2 details the specific steps that are executed in the proposed coevolutionary algorithm. Following details should be considered:

1. The individuals that are used for recombination and mutation are selected using the ranking method (De Jong, 2006), which means that individuals with better fitness present a higher probability to be chosen.

2. A one-point crossover recombination operator is employed.

3. In the case of the instance and feature populations, the mutation operator switches the Boolean value of each gene according to a probability mut_I and mut_F respectively. Whereas in the parameter populations, two possible mutations are employed: 1) applying a random increase or decrease to the current integer value, or 2) resetting the parameter to a normal random value. These mutations are applied with a probability of mut_P.

Algorithm 2 Cooperative coevolutionary algorithm

Initialise randomly three populations for instances, features and parameters with respectively N_I, N_F and N_P individuals

Order each population by descending fitness

repeat
  for each population **do**
    for number of new individuals to be created **do**

      ——— Generate a new individual ———
      Create one new individual ind_1 by recombination
      Mutate ind_1

      ——— Build a solution ———
      Select by ranking two individuals ind_2 and ind_3 each one of them from one of the other populations
      Build a solution ind as combination of ind_1, ind_2 and ind_3

      ——— Calculate and update fitness ———
      Calculate $fitness(ind)$ as the classification rate
      $fitness(ind_1) = fitness(ind)$
      if $fitness(ind_2) < fitness(ind)$ **then**
        $fitness(ind_2) = fitness(ind)$
      end if
      if $fitness(ind_3) < fitness(ind)$ **then**
        $fitness(ind_3) = fitness(ind)$
      end if
    end for

    ——— Generate next generation's population ———
    Order each population by descending fitness
    Select next generation's population with elitism
  end for
until $generations\_without\_changes > max_{gen}$

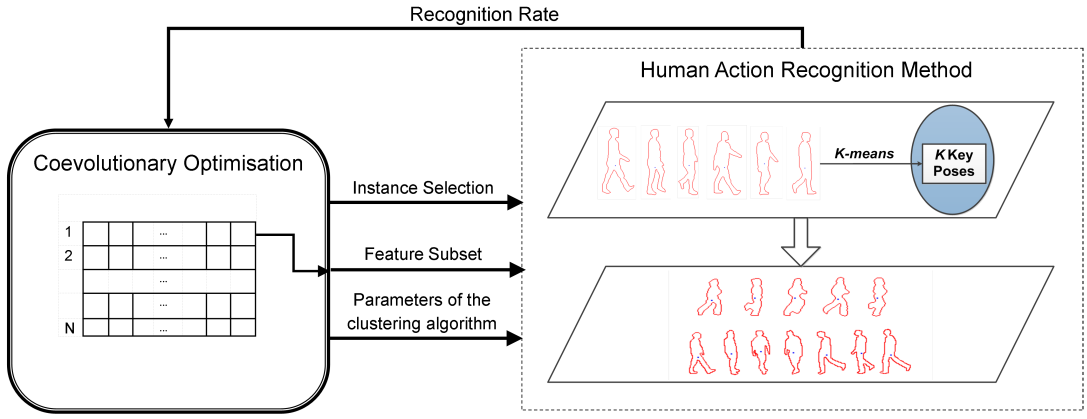


Figure 6.3: This figure shows how the CEA optimisation interacts with the human action recognition method: The individuals are tested using their encoded instance selection, feature subset and clustering parameter values. The employed human action recognition method obtains the multi-view pose representations and learns the bag of key poses out of the training data. Classification is performed matching sequences of key poses. Then, the obtained recognition rate is used as fitness value so as to order the populations and apply elitism.

4. The fitness value is obtained by combining one individual from each population (ind_1, ind_2 and ind_3) and evaluating the success rate of the human action recognition method of section 5.2 with the configuration encoded in the individuals.

   As algorithm 2 shows, whereas ind_1 constitutes the new individual, ind_2 and ind_3 are selected out of their populations based on ranking. As stated in Coello et al. (2006), choosing the best individuals would cause undersampling and excessive greediness in dependant populations.

5. The obtained fitness value is adopted by the new individual ind_1, and by ind_2 and ind_3 if it improves their current fitness value.

6. When populations are ordered by descending fitness value, we also apply an optimisation of spatial and temporal constraints. If two individuals present the same fitness value, the most efficient one is given priority. This means that in the case of the instance and feature populations, the individual with fewer selected values is favoured. In the parameter population, the individual with lower aggregate sum is preferred, since a higher value indicates a greater amount of key poses and this results in a more costly classification.

Figure 6.3 shows how the proposed coevolutionary algorithm employs the same wrapper model that has been seen in the evolutionary algorithm from section 6.1 in order to optimise the recognition of human actions.

6.2.3 Experimentation

In this section, the performed experimentation on the two publicly available Weizmann and MuHAVi datasets is detailed. Again, the same evaluation methodology and test environment as in previous tests has been employed. The specific configuration details about the performed tests are indicated as follows:

1. The size of the instance and parameter individuals is given by the specific dataset, *i.e.* I equals the number of training instances, and A equals the number of action classes to recognise.

2. The number of elements of the feature vector $F = B \times M$, because in the multi-view recognition B feature vector elements are employed for each of the M views.

3. The indicated results have been obtained with populations of ten individuals ($N_I = N_F = N_P = 10$) and a single offspring per generation.

4. Regarding the mutation operator, random probability values mut_I, mut_F and mut_P are employed. The two possible mutations of the parameter individuals which have been previously detailed are chosen with a 50-50 chance. A range between 5 and 130 key poses is considered for the parameters $K_1, K_2, ..., K_A$.

5. We set $max_{gen} = 250$, *i.e.* the evolution is considered to be stable after 250 generations without changes in its best performing individual.

6. The number of radial bins B is detailed for each test. Its value has been obtained based on a statistical analysis of the classification results of the HAR method, following the same procedure as in section 5.1.3. Likewise, in order to obtain reproducible results, we chose the value of B based on the highest median success rate, which has been obtained with $B = 14$ in the case of the Weizmann dataset (94.6%). If multiple configurations return the same result, the lowest value will be chosen, as this reduces the feature size and the computational cost of the classification.

Table 6.5: Comparison of recognition rates obtained on the Weizmann dataset with other state-of-the-art approaches. In this test, 14 radial bins have been used ($B = 14$).

Approach	**#Actions**	**Rate**
Tran and Sorokin (2008)	10	100%
Weinland et al. (2010)	9	100%
Naiel et al. (2011)	10	98.9%
Sadek et al. (2012)	10	97.8%
Hernández et al. (2011)	10	90.3%
Our approach before CEA optimisation	10	95.7%
Our approach after CEA optimisation	10	100%

Weizmann dataset

The Weizmann dataset includes 93 sequences of ten different action classes which have been performed by nine actors. Even if it is only intended for single-view human action recognition, it is commonly used as a baseline benchmark. Table 6.5 shows a comparison of our results, before and after applying the proposed CEA optimisation, and the ones that can be found among the state of the art. All of them report results for the LOAO cross validation.

It can be observed that by means of the applied CEA optimisation of both learning and classification stages, perfect recognition is reached for this dataset.

MuHAVi dataset

The MuHAVi dataset includes a total of 136 sequences, which respectively correspond to 14 or 8 action classes in the MuHAVi-14 and MuHAVi-8 benchmark. Tables 6.6 and 6.7 show the obtained results applying the LOSO and LOAO cross validations once again. Our method not only outperforms all the available recognition rates on both versions of the dataset, but it also reaches perfect recognition on all four tests with the proposed optimisation based on a coevolutionary algorithm. To the best of our knowledge, this is the first work reporting perfect recognition.

Optimisation results

In table 6.8, the results of the proposed optimisation are shown. For each of the applied tests, the values of the best performing individuals from the three populations are detailed. The *instances* column shows the number of training

Table 6.6: Comparison of recognition rates obtained on the MuHAVi-14 dataset with other state-of-the-art approaches. In this test, 12 and 10 radial bins have been employed respectively in the LOSO and LOAO cross validations.

Approach	**LOSO**	**LOAO**
Singh et al. (2010)	82.4%	61.8%
Cheema et al. (2011)	86.0%	73.5%
Eweiwi et al. (2011)	91.9%	77.9%
Our approach before CEA optimisation	**98.5%**	**94.1%**
Our approach after CEA optimisation	**100%**	**100%**

Table 6.7: Comparison of recognition rates obtained on the MuHAVi-8 dataset with other state-of-the-art approaches. In this test, 12 and 18 radial bins have been employed respectively in the LOSO and LOAO cross validations.

Approach	**LOSO**	**LOAO**
Cheema et al. (2011)	95.6%	83.1%
Singh et al. (2010)	97.8%	76.4%
Martínez-Contreras et al. (2009)	98.4%	-
Eweiwi et al. (2011)	98.5%	85.3%
Our approach before/after CEA optimisation	**100%**	**100%**

instances that have been selected out of the available ones. In the case of the *features* column, the feature subset selection of the multi-view pose representation is detailed. Finally, the number of key poses used to represent the action classes is indicated in the *parameters* column.

On average, the coevolutionary algorithm reduced the required training instances to ∼66% of the total number of video sequences. The feature subset used during learning and classification has been reduced by ∼36%. Best results have been obtained representing each action class with an average of ∼26 key poses.

With the purpose of providing further insight about how the obtained results correspond to the data, we analysed the Weizmann LOAO cross validation test. In table 6.9, the resulting selection of instances is given in terms of both actors and action classes. It can be observed that the sequences of actors as *Moshe* and *Shahar* have been selected less often, which could be related to the way they performed some of the actions or the noise of the recording. For instance, in the *skip* sequence from *Shahar*, the arm motion is much more pronounced than in the other samples. Regarding the action-wise selection, we can observe that a fairly similar distribution of actions has been selected, since samples from all actions

Table 6.8: In this table the values of the individuals with the highest fitness value of each of the run tests are shown.

Dataset	Test	Instances	Features	Parameters
Weizmann	LOAO	71/93	11110110110111	72, 37, 25, 22, 19, 78, 25, 47, 55, 106
MuHAVi-14	LOSO	81/136	11101110010111 0111011111	6, 5, 22, 31, 5, 79, 5, 5, 5, 5, 5, 5, 5, 5
MuHAVi-14	LOAO	99/136	10101010100111 111010	4, 4, 10, 112, 14, 29, 6, 54, 5, 95, 23, 27, 6, 20
MuHAVi-8	LOSO	87/136	00010001101101 0110110010	6, 6, 6, 6, 6, 6, 6, 6
MuHAVi-8	LOAO	77/136	10100011010101 10111011101111 00101100	22, 38, 21, 27, 11, 40, 32, 75

are necessary in order to recognise them. In the case of *wave1*, fewer samples are necessary because all of them are performed with the right hand and mostly only the noise of the background segmentation differs between the samples.

Applying a similar analysis to the obtained feature subset, in figure 6.4 it can be seen which feature elements of the radial scheme have been discarded for the actions of the Weizmann dataset due to redundant or noisy components.

Finally, regarding the selection of parameter values, note that the data given in table 6.9 follows the alphabetical order of the action classes. In this case, significant differences can be observed regarding the different number of key poses that were necessary to model the representative poses of action classes as *run* ($K = 19$) and *wave2* ($K = 106$) and obtain the best result.

Table 6.9: Selection of instances in terms of action classes and actors obtained for the LOAO cross validation test on the Weizmann dataset. Note that normally each actor performed each action once, but *Lena* performed twice *run, skip* and *walk*.

Actor	*Bend*	*Jack*	*Jump*	*Pjump*	*Run*	*Side*	*Skip*	*Walk*	*Wave1*	*Wave2*	Total
Daria	1	1	1	0	1	1	0	1	0	1	**7/10**
Denis	1	1	1	1	1	0	1	1	1	1	**9/10**
Eli	1	1	1	1	1	0	0	1	1	1	**8/10**
Ido	0	1	1	0	1	1	1	1	1	1	**8/10**
Ira	1	1	0	1	1	1	1	1	0	1	**8/10**
Lena	1	1	1	1	1/2	0	2/2	2/2	0	1	**10/13**
Lyova	1	1	1	0	1	1	1	1	1	1	**9/10**
Moshe	1	1	0	1	0	1	1	0	0	1	**6/10**
Shahar	1	0	1	1	0	1	0	1	1	0	**6/10**
Total	**8/9**	**8/9**	**7/9**	**6/9**	**7/10**	**6/9**	**7/10**	**9/10**	**5/9**	**8/9**	**71/93**

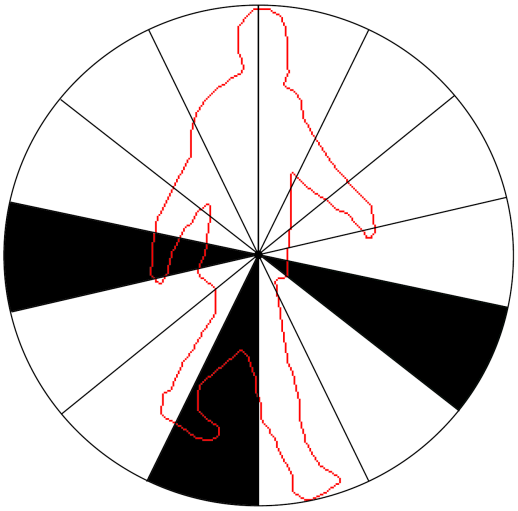


Figure 6.4: Feature selection that has been obtained for the Weizmann dataset. Discarded elements are shaded in black.

To illustrate the performance gain achieved with this optimisation, we performed a temporal evaluation on the Weizmann dataset. The learning stage of the algorithm is executed in 0.72s, whereas the testing stage requires 3.29s for the whole dataset. Applying the obtained configuration, an improvement of respectively ∼41% and ∼32% can be observed. The final processing rate considering the binary silhouette images as input, and applying contour extraction and classification is 210 fps.

6.2.4 Discussion and conclusion

In this section, a human action recognition optimisation based on a cooperative coevolutionary algorithm has been presented. By means of evolutionary search, the redundant or noisy training instances that confuse or unnecessarily extend the learning process are filtered. Similarly, the feature subset which includes the relevant parts of the human body in order to recognise human actions is selected considering the available views. Last but not least, the appropriate number of key poses is sought in order to represent the different poses involved in each action class. This configuration is employed in the proposed HAR method, which relies on a multi-view silhouette-based pose representation and a weighted feature fusion scheme.

Results obtained during the experimentation have shown that our proposal improves the initial recognition rates, reaching perfect recognition for all the applied benchmarks, and considerably outperforming the available rates of the MuHAVi dataset. Furthermore, the execution time and memory needs of the learning algorithm are reduced, since the evolutionary algorithm implicitly prioritises the more efficient configurations.

In conclusion, the presented work successfully optimises both the recognition and the temporal and spatial performance of human action recognition. Since it only has to be applied as a pre-classification stage, it is also suitable for online human action recognition, where the obtained configuration can be applied.

For future works, so as to find the *optimal* configuration of a HAR method intended to perform in a real world application, a compound of benchmarks could be used assigning the appropriate weights to consider the specific needs of the system. In this way, a robust optimisation can be obtained if no scenario-specific training set is available.

6.3 Adaptive human action recognition with an evolving bag of key poses

6.3.1 Introduction

In intelligent environments, and especially at home, robots can be very useful as assistive and proactive agents supporting several safety and health care scenarios as, for instance, monitoring or mobility assistance (Dubowsky et al., 2000). These care services, among others, can potentially improve the independent living of the elderly, or serve senior assisted living facilities. It can be seen that several research fields are involved in this robotics application area due to the related multidisciplinary challenges. From the acceptance of robotic technology at home and what tasks a robot should be able to perform (Broadbent et al. (2009), Broadbent et al. (2009)), to the need of socially interactive robots (Fong et al., 2003), across many others. A key requirement in order to successfully support the mentioned care services is advanced human-robot interaction (HRI). Reliable support can only be assured if robots are intelligent enough to analyse and understand the scenario they perform in and the events that occur, in order to be able to interact appropriately. Furthermore, this scene analysis and human

behaviour understanding abilities are essential with regard to reaching the desired *level of autonomy* of the robot (Goodrich and Schultz, 2007). Specifically, computer vision has shown to be a powerful tool that provides rich sensor data from the environment to deal with these tasks (Cipolla and Pentland, 1998). This way, in this section we study to apply and optimise the recognition of human actions in order to enrich the capacities of HRI of autonomous robots. The application of HAR to autonomous robots comes along with several additional hurdles. Since human behaviours are subject to change depending on the specific scenario and actor, and moreover, the behaviours can vary over time, an adaptive learning process is needed. The robot requires autonomous mental development capacities in order to dynamically adapt its knowledge to recognise new scenarios as, for instance, new actions. This process needs to happen incrementally, as the robot should be able to learn continuously over time without requiring to start from scratch, but evolving its recognition capabilities towards the new data that needs to be discriminated. Furthermore, a robot presents several limitations related to the sensor data that can be collected due to space and weight constraints, and also related to the computational capacity. Therefore, a relatively simple camera setup should be employed and efficient real-time recognition algorithms are required. These constraints can be fulfilled by the previously proposed real-time HAR method.

So far EA have been used only for optimisation purposes in this thesis in order to improve the performance of the classification method modifying its training and feature sets, and parameters. Nevertheless, this does not mean that EA are no more than just a set of optimisation techniques (Cagnoni, 2008), since there are several applications in which the EA and the learning algorithm work actively together, as it is also the case of the present contribution. These EA techniques as online solution components are also analysed in Cagnoni (2008), where it can be seen that they are applied for both interactive and incremental learning.

In this sense, the developed static HAR method is extended in this section in order to support a dynamic behaviour, *i.e.* incremental and adaptive learning. Given that the learning memory of the method is constituted by the bag-of-key-poses model, this model is evolved in our dynamic proposal in order to continuously consider more and more action classes and learning data over time. In other words, the learnt model is incrementally fed with new data which leads to the evolution of the bag-of-key-poses model. Not only the learning model is adapted,

but also the configuration of the classification is continuously optimised for the current scenario. For this purpose, an evolutionary optimisation method with the same three optimisation targets from last section (training instances, feature subset and parameter values) is proposed. The best performing configuration is learnt through evolution using the current training data.

In the performed experimentation, it is analysed how the method develops and adapts itself to the increasingly difficult recognition task. It can be seen that consistently high results are obtained over the different datasets and input data types, and a substantial performance increase is obtained with respect to the static method.

Incremental and adaptive human action recognition

Incremental and adaptive learning techniques have been applied rather sparingly to the field of human action recognition (Kapp et al., 2011; Minhas et al., 2012; Ryoo et al., 2010; Wang et al., 2012) since they are more frequent in related fields as, for instance, visual tracking (Lim et al., 2005; Yang et al., 2011). In incremental learning the goal is to improve the learnt model or exemplar-based data combining the previous experience with the knowledge extracted from the new example(s) in order to both successfully recognise the new samples and also improve the recognition of existing ones. Therefore, it is also known as *iterative* learning (Jain et al., 2006). Adaptive learning is closely related, but it focuses on the capacity of adaptation of the learning towards the new data. Specifically, in incremental learning approaches it is difficult to set the appropriate parameter values of the algorithm if the data is initially unknown. For this reason, the algorithm's configuration should dynamically adapt itself to the new requirements by tuning its configuration (Kapp et al., 2011), for instance, by means of evolutionary algorithms. Ryoo et al. (2010) proposed a method to learn novel human activities incrementally. Based on an incremental codebook, mining of visual words is performed. Local spatio-temporal features are clusterised to obtain the bag-of-words model. When a novel activity is added to the system, new visual words are generated and the existing ones are adapted or merged. Recognition is performed based on visual words histograms, which are also sequentially updated. The method achieves similar activity classification rates as other non-incremental approaches. In Minhas et al. (2012), snippet-level action recognition is targeted using a recursively trained classifier based on a single-hidden layer feed forward

neural network which is extended to present an incremental behaviour. Nonetheless, it is also adaptive, as the input weights are set randomly initially and then adjusted by means of a generalised inverse operation of the hidden layer weight matrices. In this work, the shape of an actor is approximated by adaptively changing intensity histograms to extract pyramid histograms of oriented gradient features. The performance is analysed based on the length of the snippets. It can be seen that with only two frames, a recognition rate of over 80% is achieved employing from 10 to 50% of the training data of the Weizmann dataset. Wang et al. (2012) rely on wearable sensor data instead of vision. Probabilistic neural networks and an adjustable fuzzy clustering algorithm are employed so as to support incremental learning by means of addition of new information and new activities, but also removing existing ones. The possible noise present in the training dataset is explicitly considered by differentiating the importance of pattern neurons.

Application to autonomous robotics

In autonomous robotics, all the aforementioned machine learning and artificial intelligence techniques come together. In Siciliano and Khatib (2007), evolutionary robotics is defined as "a method for automatically generating artificial brains and morphologies of autonomous robots", where the goal is to achieve that robots develop their own control systems without human intervention. A key element towards this long-term goal is that the robot should be aware of its environment. Visual perception is pursued in robot vision. For instance, in Yorita and Kubota (2009) people tracking in dynamic environments is performed based on face detection and fuzzy evaluation. A local genetic algorithm based on clustering is used as search method. Robot control behaviours are targeted in Meeden (1996). An EA is employed for reinforcement learning in order to provide guidance to the robot for its actions. Finally, feature selection is applied to develop the sensitivity to relevant features in the scene. Feature selection is therefore evolving along with active vision with the purpose of simplifying the computational complexity of vision-based behaviour analysis (Siciliano and Khatib, 2007). Similarly, in Roy et al. (2013), both instance and class-specific feature selection is applied for automated learning. Patterns are selected based on sampling, and seeking to represent the class regions with a minimal amount of instances. Regarding the features, those that best separate a class from the others are selected based on

a compactness measurement. This is applied to a hypersphere net classification algorithm which is similar to radial basis function (RBF) nets.

6.3.2 Evolving bag of key poses

As it has previously been introduced, the HAR method presented in chapter 4 relies on a bag-of-key-poses model, which can be seen as a learning memory. In a static environment, the complete training set is established at the beginning of the learning process. In this section, we extend the model in order to allow an adaptive and incremental learning. The bag-of-key-poses model will continuously evolve in order to adapt to the new data and, at the same time, the method will be optimised trying to find the best training set, features and parameters to improve the recognition rate by adapting the configuration to the current scenario.

The evolutionary optimisation system consists of a single population of individuals representing combinations of training set, feature subset and parameters. The fitness value of these individuals is given by the global recognition rate that is obtained using their configuration with the HAR method. The optimisation process follows the evolutionary algorithm 3. Next, information about this evolutionary process is explained in detail.

Individuals' representation

Individuals are encoded as the concatenation of three different vectors representing the training set and feature selections, and the parameters of the clustering algorithm.

The characteristics of each of the vectors are (see also figure 6.5):

- Training set: This vector is made up of I elements corresponding to the number of instances available for training. This number is variable as new sequences increase the size of the vector. It is encoded as a binary vector, where each gene represents the selection of that instance during the learning.

- Features: This vector contains F elements, one for each feature element of the feature vector. This number is constant. Each gene of this binary vector indicates whether or not the corresponding feature element should be used during the classification process.

Algorithm 3 Evolutionary algorithm

Initialise the populations with N_I individuals generated randomly

Rank the population by fitness using the recognition rate obtained with the configuration coded by each individual

repeat
  for number of new individuals to be created **do**

    ——— Generate a new individual ———
    Create one new individual *ind* by crossover
    Mutate *ind*

    ——— Calculate fitness ———
    Calculate $fitness(ind)$
  end for

  ——— Generate next generation's population ———
  Rank the population by fitness
  Select next generation's population with elitism

until a termination condition is achieved or forever

- Parameters: This vector defines A parameter values for the A action classes. Therefore, it is increased if a new action must be learnt. Each gene takes an integer value in a range of allowed number of key poses.

Each individual of the population stores the bag of key poses obtained with its configuration in order to be used if new instances are included in the learning. This will allow the online updating of the bag of key poses when new instances are added to the training set.

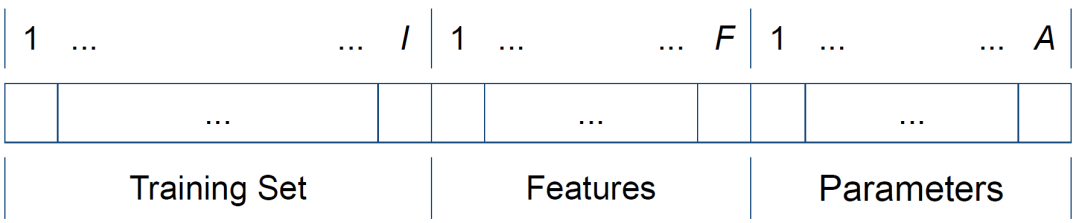


Figure 6.5: Structure of the individuals' representation.

Crossover

The usual method in evolutionary computation is to perform a single crossover operation to the individual. However, since the individual considered in this work has three different parts, one crossover is applied to each one of the vectors (training set, features and parameters) as if these were independent individuals, similarly to the behaviour of the CEA employed in section 6.2. Nevertheless, the use of a CEA with different populations for each of the vectors is not used in this case. A CEA would hinder the management of the key poses associated with each of the individuals, because these would be shared among individuals of different populations. Therefore, a one-point crossover operator is applied to each part of the individual. As before, the parents are selected by ranking among the individuals of the population.

Mutation

Similarly to the crossover operation, a mutation is performed over each part of the individual with different probabilities. Instance and feature vectors use standard mutation, *i.e.* each gene changes its value according to probabilities mut_I and mut_F. Each gene in the parameter population is mutated with a mut_P probability. This mutation can be done in two different ways (with equal chance): slightly modifying its value applying a Gaussian noise, or setting it to a random value in an interval.

Adaptive learning of new actions

Different learning trajectories could be considered. As it has been mentioned, in this work, the learning of new data which belongs to previously unknown classes is considered. The inclusion of more and more action classes reduces the inter-class distance and makes the discrimination more complex. Figure 6.6 shows a graphical explanation on how this affects the evolving bag-of-key-poses model. Learning a new action involves the generation of the specific K key poses for that action in the bag of key poses. This is achieved by:

1. Updating the length of the training set vector of each individual increasing it by the number of training instances of the new action class. The new binary genes are set randomly following the same distribution than when the population was initialised.

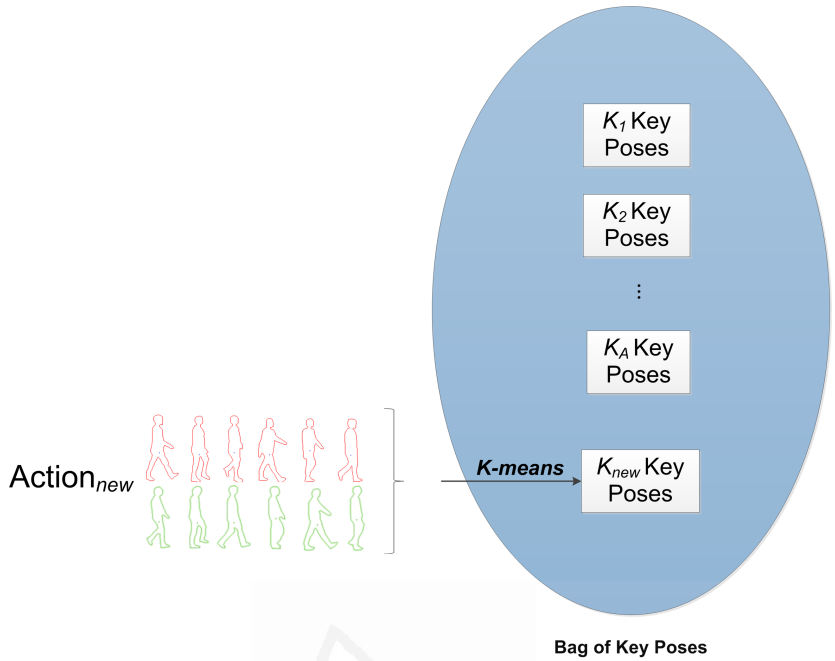


Figure 6.6: Learning of new actions. It can be seen that for a new action class the bag of key poses is updated with the corresponding K_{new} key poses.

2. Increasing the length of the parameter vector with a gene associated to the number of key poses of the new action $Action_{new}$. This gene is set to a random value in the interval of possible K.

3. Learning the K_{new} key poses for the new action executing the K-means clustering algorithm for each individual of the population and recalculating its fitness. The key poses for the actions that have already been learnt do not need to be obtained, since they are stored along with each individual.

Once the current population is updated with the new action, new individuals can be created and the evolution continues.

6.3.3 Experimentation

For the experimentation, three state-of-the-art publicly available benchmarks have been tested, and both single- and multi-view datasets have been chosen. These are the single-view Weizmann dataset and two multi-view datasets, the MuHAVi and the challenging IXMAS dataset. Regarding the employed implementation of the HAR method, although the final version from section 5.2 is employed in this case, note that since the learning data changes over time no

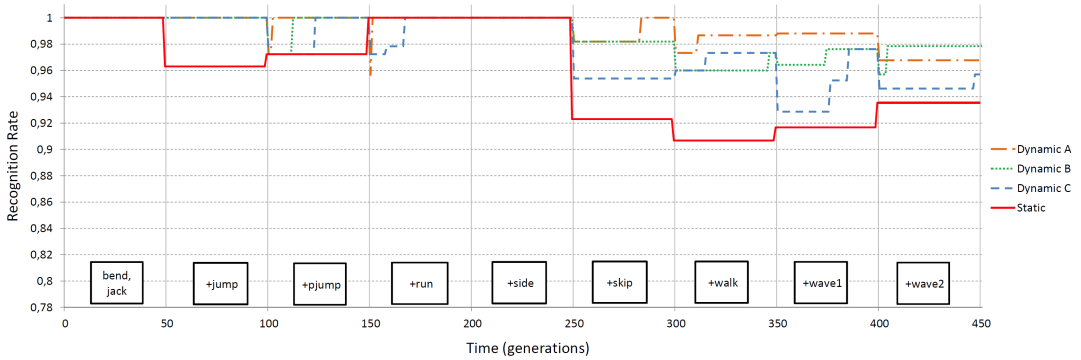


Figure 6.7: Results of the dynamic learning of the Weizmann dataset. Recognition rates for the static learning and three different dynamic runs are shown. The number of employed radial bins $B = 14$ and the default value for $K_1, K_2, ..., K_A$ is 10 (with a range from 4 to 30).

a priori knowledge can be obtained about the proficiency of each view, and therefore, equally weighted views are used leading to a simple feature fusion of multiple views.

The carried out performance analysis is shown in the graphs of figures 6.7 to 6.9 and 6.11. These include a static line that represents the best solution considering that the system is using all the training instances, all the features and the default value for the parameters. This line varies during a stage because of the non-deterministic behaviour of the K-means clustering. Obviously, the system always uses the configuration that has obtained the best recognition results until that moment. An equivalent number of generations (*i.e.* iterations in this case) per stage is considered. Therefore, the recognition rate can increase during a learning stage if a better performing clustering is obtained for the same data. The dynamic lines represent three different runs of our evolving proposal with different random initialisations.

The graphs show that when a new action is included, the recognition rate usually suffers a dramatic change. Sometimes the recognition rate decreases. This is because the new action is difficult to recognise, or because there are other similar actions that have already been learnt by the system and misclassifications are produced. At other times, the global recognition rate improves since the new action is easy to recognise by the system and this increases the overall recognition rate.

As it will be seen for each particular dataset, the recognition rate is globally better in any of the dynamic runs. The inclusion of new actions affects less than

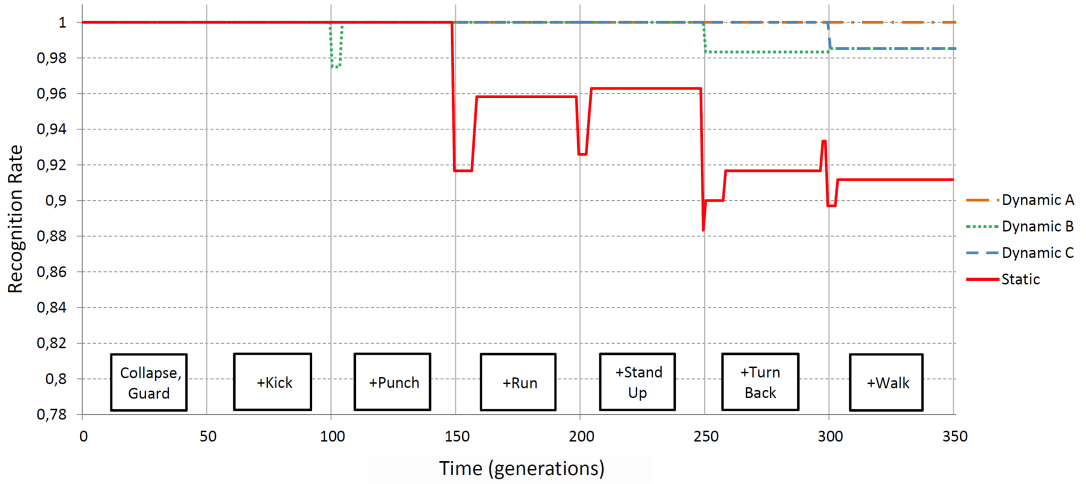


Figure 6.8: Results of the dynamic learning of the MuHAVi-8 dataset. Recognition rates for the static learning and three different dynamic runs are shown. The number of employed radial bins $B = 18$ and the default value for $K_1, K_2, ..., K_A$ is 7 (with a range from 4 to 30).

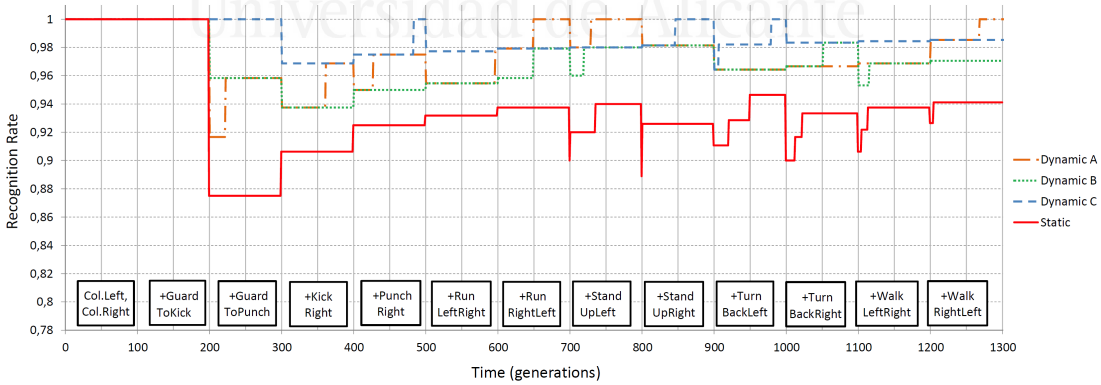


Figure 6.9: Results of the dynamic learning of the MuHAVi-14 dataset. Recognition rates for the static learning and three different dynamic runs are shown. The number of employed radial bins $B = 10$ and the default value for $K_1, K_2, ..., K_A$ is 4 (with a range from 4 to 30).

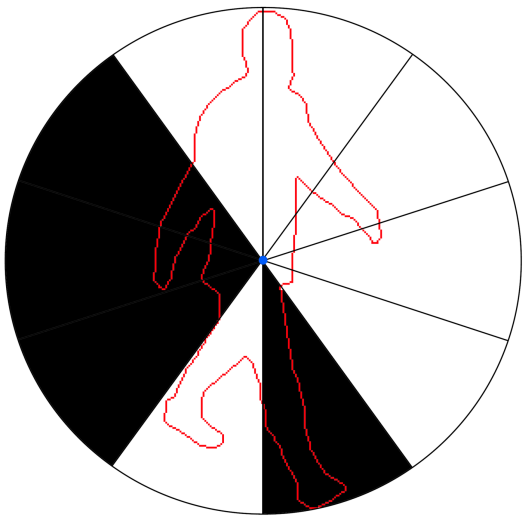


Figure 6.10: Final feature subset selection that has been obtained for the MuHAVi-14 dataset. White bins are selected, black ones are discarded.

in the static version. But if the classification is affected, the learning memory rapidly evolves, updating the appropriate selection of training set, features and parameter values, and thus improving the recognition rate.

Figure 6.7 shows the results that have been obtained on the Weizmann dataset with three different runs of the dynamic human action recognition method (with incremental and adaptive learning) and the static approach. Each 50 generations, a new action class is incorporated to the evolutionary process. As it can be seen, depending on the newly introduced action (and on the already learnt ones) the static result decreases or increases. Especially, the *skip* action, which is commonly excluded by other authors due to its intra-class similarity (Saghafi and Rajan, 2012; Shao and Chen, 2010), causes a dramatic performance drop from which the static approach does not achieve to recover. In contrast, the dynamic approach handles this drop and also others very well. In few iterations the system adapts to the new data and achieves to return high results steadily. Furthermore, the drops are not as relevant as with the static approach (particularly in *Dynamic A* and *B*), which means that the dynamic approach leads to a more stable and general learning model.

Figures 6.8 and 6.9 show the results on the MuHAVi dataset. Whereas in MuHAVi-8 the static approach starts to suffer an important performance decrease when the fifth action class (*Run*) is introduced, the dynamic approach shows to support the incremental learning well. The run corresponding to *Dynamic A*

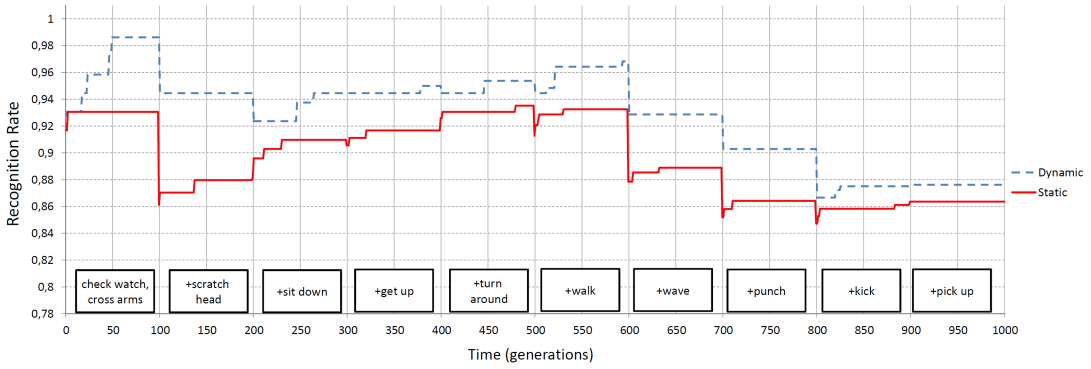


Figure 6.11: Results of the dynamic learning of the IXMAS dataset. Recognition rates for the static and dynamic learning are shown. The number of employed radial bins $B = 27$ and the default value for $K_1, K_2, ..., K_A$ is 130 (with a range from 4 to 130).

presents a constant perfect recognition throughout the learning from two to eight different action classes. With MuHAVi-14, 100 generations have been employed for each stage, since due to the higher number of action classes (14), the system requires more time to adapt to the new data. Again, our proposal shows that it successfully adapts to the new action classes to be recognised, and it is less affected by the new data. The obtained recognition rates are globally better and more stable. As an example, the final feature subset selection of the best performing individual is shown in figure 6.10.

Finally, the method has been tested also on the much larger IXMAS dataset. Figure 6.11 compares the dynamic and the static behaviour. Since the dataset includes nearly 2 000 video sequences, only one full run could be executed. Nonetheless, also on this challenging data, the proposed method presents an importantly improved behaviour with lower performance drops and a continuously higher recognition rate over the different learning stages. It can be seen that in this case, although the learnt selection of features, training instances and parameter values successfully supports the inclusion of new actions to recognise, less improvement is observed during the learning stages. This is related to the limited amount of generations used for each learning stage, as well as to the data variance.

As it has been seen, the adaptive learning algorithm shows steadily promising results despite the singularities of different action classes, actors and scenario-related conditions.

6.3.4 Discussion and conclusion

In this section, an adaptive human action recognition method that can be applied to intelligent environments and autonomous robots has been presented. Based on an evolutionary algorithm, an incremental and adaptive learning of human actions is supported with an evolving bag-of-key-poses model. At the same time, through evolution, the best performing selection of training instances, features and parameters is sought. Therefore, the contribution of the evolutionary approach is two-fold. On the one hand, it serves as an optimisation method in order to improve the adaptation to the new data to be recognised and increase the recognition rate. On the other hand, it guides the dynamic behaviour in which new data is incrementally learnt and the bag-of-key-poses model is evolved. In this way, the method is able to successfully support the inclusion of new data with small performance changes, that are overcome in few generations in which the method adapts to the new data. The approach has been validated on three different datasets. On multiple publicly available datasets of different difficulty and image quality, good to outstanding results have been obtained throughout the whole learning. The analysis shows that, although the obtained results vary due to the random initialisation and non-deterministic behaviour, the proposed dynamic approach achieves superior results steadily in comparison to the static incremental learning. This leads us to consider that this method is suitable for its application to autonomous robots in order to enrich the essential task of human-robot interaction.

In future works, the proposed adaptive learning should be compared with other state-of-the-art incremental learning and continuous adaptation methods. In order to make this comparison possible, a consensus must be reached on how this performance can be measured in terms of continuous recognition rates and adaptation times. Other learning trajectories should be studied too. The intuitive idea of starting with the most simple actions and continue to learn more and more difficult ones requires a measurement of difficulty. This could be based on a binary classification rate, since a multi-class one would depend on the other classes.

6.4 Remarks

This chapter has provided three different approaches for the optimisation of human action recognition. Using EA as support techniques, the search of the best performing feature subset, training instances and parameter values has been addressed. Whereas in the first proposal feature selection led to a significant improvement of performance, in the second proposal a coevolutionary optimisation of the three sets achieved the best results so far. Finally, in the third proposal, an EA is not only employed as a search algorithm, but also as a guide to support adaptive and incremental learning. In this regard, it has been seen that the presented HAR method can easily be tailored for a specific scenario in order to exploit the implied constraints and improve the final recognition. The proposed optimisation techniques have produced substantial improvements, considering both the classification accuracy and the computational cost of the recognition.

In conclusion, this chapter has provided a two-fold contribution. On the one hand, the presented method and the proposed implementation has proven its flexibility towards optimisation for specific setups and application scenarios. On the other hand, the proposed optimisation techniques based on EA and the wrapper model may be used with other classification methods that not necessarily have to deal with HAR. Therefore, its application can be considered for a wide range of applications.

Chapter 7

Continuous human action recognition

Until now, human action recognition has been addressed by classifying short video sequences that contain single actions. Therefore, two strong assumptions have been made: 1) segmented video sequences which only contain a single action each are provided, and 2) all the video sequences necessarily match with one of the learnt action classes. Whereas these assumptions are commonly made in the state of the art as it has been seen in section 2.5, and most of the datasets provide such a data, these do not hold true in practical situations.

In this chapter, two new objectives are introduced to handle continuous recognition of human actions without relying on the aforementioned assumptions. Thus, a mechanism has to be introduced to handle the continuous video stream provided by the cameras and, at the same time, a *null class* has to be considered in order to discard unknown actions and avoid false positives. This class corresponds to all the behaviours that may be observed and have not been modelled during the learning.

7.1 Introduction

As it has been mentioned before, human action recognition is in demand in a wide range of application areas. In video annotation, one could assume that a video segmentation has been applied beforehand (manually or automatically) and that therefore, only sequence recognition has to be performed. Nonetheless, in scenarios where recognition has to be applied in real time, as HCI, gaming, or

video surveillance, this process has to happen *on the fly*. This is also the case of AAL. In people's homes, cameras will provide a continuous video stream which can contain actions at any moment. This leads us to continuous (also known as online) human action recognition. In other words, an unsegmented video stream has to be analysed in order to detect actions at any point.

Another restriction, which comes along with dealing with the raw video stream of the cameras, is that actually these may not record the expected actions. The person could be performing an unknown action, or nothing at all[1]. Therefore, the proposed system needs to be robust enough in order to discard unknown actions that otherwise would result in misclassifications.

In the following, works which specifically address continuous human action recognition (CHAR) will be reviewed. Next, a proposal is introduced to extend the method from chapter 4 to CHAR. The approach is verified in the experimentation, introducing an evaluation scheme especially adjusted for the AAL application. Finally, some remarks are given.

7.2 Related work

Determining the relevant segments of a continuous video stream may be trivial for a human, but it certainly involves a great difficulty for an automated CV system. This explains why few works deal with the related additional constraints.

Some works try to find the boundaries of the actions in order to apply temporal segmentation. These boundaries can be detected based on discontinuities or extremes in acceleration, velocity or curvature (Kellokumpu, 2011). Once the resulting video segments are obtained, sequence-based action recognition can be applied. Such a temporal segmentation is performed in Vitaladevuni et al. (2008), where atomic movements are localised in the video stream based on so-called 'ballistic movements'. These are defined as impulsive movements, which involve a sudden propulsion of the limbs, and rely on the acceleration and deceleration of start and end of the ballistic segments. A trajectory-based motion feature (*i.e.* MHI) is employed along velocity magnitude features based on silhouette transformation, frame differences and optical flow. Two approaches are tested for the temporal segmentation. The first proposal handles alignment of

[1]Note that actions where the person is still can be learnt too (*e.g. standing still* or *sitting*). Such actions are included in our self-recorded DAI RGBD dataset.

the optical flow direction by means of dynamic programming. Whereas in the second, assuming that boundaries are characterised by zero velocity, movement begin-end detection is performed with a boosting based classifier. The first option performed better, since it does not classify specific temporal moments, but aligns a globally optimal segmentation taking into account movement direction. In Guo et al. (2012), start and end key frames of actions are identified. Segmentation is performed based on the posterior probability of model matching considering recognition rounds. Depending on the accumulated probability, rounds are ended if a threshold is reached. Adjacent rounds classified as the same action classes are connected into a single segment. Lu et al. (2013) deal with temporal segmentation of successive actions. During the learning, only a few characteristic frames are selected based on change, which leads to an outstanding temporal performance of the recognition. Likelihood of action segments is computed considering pair-wise representations of characteristic frames. Although good results are obtained, no further instructions are provided on how an actual continuous video stream would be handled.

A very popular technique in video and audio processing is the sliding window approach. Sliding windows allow to analyse different overlapped segments of the stream in order to isolate a region of interest and then perform classification comparing the window to a set of training templates. If a variable size is also considered, both window position and size dynamically change so that all necessary locations and scales are taken into account. Some works have applied the sliding window technique to CHAR (Bobick and Davis, 2001; Hu et al., 2009; Kavi and Kulathumani, 2013; Zelnik-Manor and Irani, 2006). In Bloom et al. (2013), a sliding window is employed to accumulate and smooth the frame-wise predictions of a frame-based low-latency recognition. Low-latency CHAR is also considered in Nowozin and Shotton (2012), where so-called *action points* are proposed as natural temporal anchors. These are especially useful for gaming. Two approaches are proposed. The first relies on a continuous observation HMM with firing states that detect action points. And the second employs a direct classification based on random forests classifiers and sliding window.

In conclusion, by means of sliding window techniques, the temporal segmentation is simplified, since no specific boundaries have to be detected. However, due to its computational cost it may only be used if the applied segment analysis can be performed very efficiently.

7.3 Continuous human action recognition based on action zones

In order to extend the proposed method for CHAR, proposals have been made for both the learning and the recognition stages. Specifically, *action zones* are introduced for the detection and recognition of temporal segments that contain relevant parts of the actions.

Definition 7.1. *Action zones correspond to the most discriminative segments with respect to the other action classes in the course of an action.*

Based on definition 7.1, for instance, the *punch* action contains an action zone corresponding to the segment from where the arm is partially outstretched, until it is completely outstretched and returns to the initial position. In other words, the part where the person is standing still is ignored, since it is not discriminative with respect to other actions. Similarly, in a *bend* action, the segment is taken in which the person is at least partially bent down. In this way, the most relevant segments can be identified in order to ease the differentiation between actions. Furthermore, action zones are shorter than the original sequences. For this reason, the matching time will be significantly reduced. Since the underlying HAR method also presents a very low computational cost, a sliding window approach may be employed without prohibitively reducing the temporal performance. These action zones are thus learnt in the training stage. During the recognition, the mentioned sliding window approach is employed to detect the equivalent segments and match them with the previously learnt action zones.

7.3.1 Learning of action zones

Initially, the same learning is performed as detailed in section 4.2, relying on the *Radial Summary* feature and the weighted feature fusion scheme presented in chapter 5. Since segmented sequences are still needed for the learning process, these can easily be obtained relying on the frame-wise ground truth and discarding the segments where no action is performed. In this way, the bag-of-key-poses model and the sequences of key poses are obtained and used for the learning of action zones.

Action zones may be located at different parts of the actions depending on the type of action and how the action ground truth has been labelled. However,

based on the provided definition, action zones can be detected automatically by analysing the transition of key poses. For this purpose, we can take advantage of the discrimination value of each key pose w_{kp}, which has been defined in section 4.2.4 as the ratio of within-class matches. It therefore indicates how valuable a key pose is for distinguishing action classes. In this way, by analysing the matched key poses of the previously obtained sequences of key poses, the most discriminative parts can be detected.

Specifically, for each training sequence of action class a and a specific temporal instant t, the following steps are taken for the corresponding frame:

1. The *Radial Summary* feature $\bar{V}(t)$ of the current frame is matched with the key poses of the bag-of-key-poses model. For each action class a, the nearest neighbour key pose $kp_a(t)$ is obtained.

2. The raw class evidence values $H_{raw_1}(t), H_{raw_2}(t), ..., H_{raw_A}(t)$ are computed based on the ratio between the discrimination value $w_{kp_a(t)}$ and the distance $dist_{kp_a(t)}$, where $dist_{kp_a(t)}$ denotes the distance between the pose representation and the matched key pose $kp_a(t)$. Hence, the discrimination value will be taken into account depending on how well the key pose defines the current pose.
$$H_{raw_a}(t) = \frac{w_{kp_a(t)}}{dist_{kp_a(t)}}, \quad \forall a \in [1...A] \ . \tag{7.1}$$

3. Normalisation is applied with respect to he highest value observed:
$$H_{norm_a}(t) = \frac{H_{raw_a}(t)}{H_{raw_{max}}(t)}, \quad \forall a \in [1...A] \ . \tag{7.2}$$

4. Gaussian smoothing is performed centred in the current frame, considering only the frames from a temporal instant $u \leq t$. In this way, we do not take into account future frames, as this would require to delay the recognition for a constant time interval. Convolution is applied to the history $H_{norm}(u)$ values with a Gaussian filter kernel in order to generate $H_{smooth}(t)$. Discrete kernel values are processed based on approximating the continuous values (see Russ (2006)):
$$G(u) = \frac{1}{\sigma\sqrt{2\pi}} \, e^{\frac{-(u-\mu)^2}{2\sigma^2}}, \quad / \quad u \leq t \ . \tag{7.3}$$

5. Attenuating the resulting value, the final class evidence $H(t)$ is obtained:

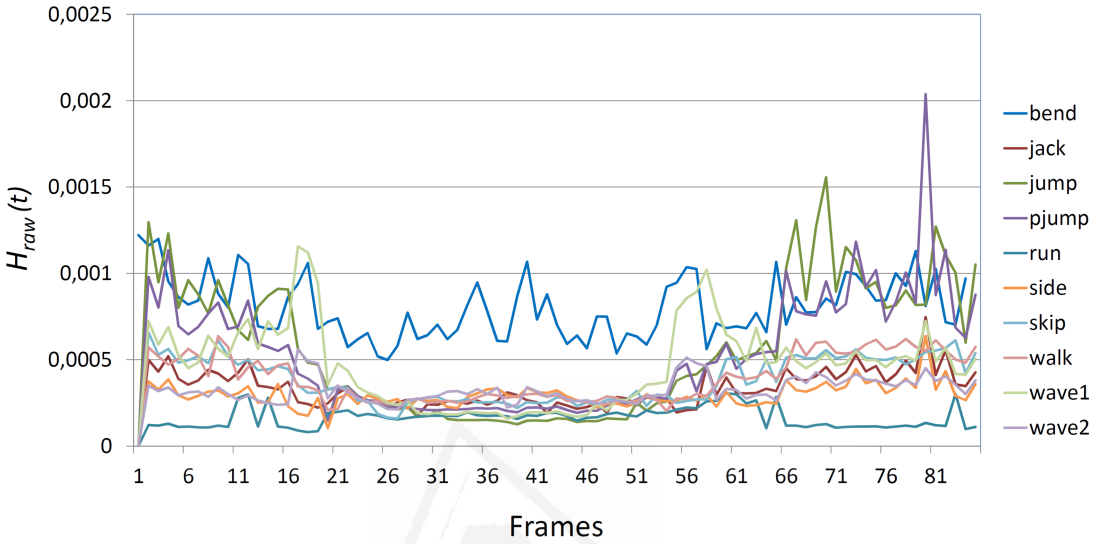
$$H_a(t) = e^{10H_{smooth_a}(t)}, \quad \forall a \in [1...A] \ . \tag{7.4}$$

Figure 7.1 shows the $H(t)$ evidence values that have been obtained over the course of a *bend* action. In comparison to the raw values, here outliers have been filtered and the differences between classes have become more pronounced. As it can be seen, the evidence of the *bend* class is significantly higher than the others in the central part of the sequence. This is due to the fact that the person is initially standing still. He or she then bends down and, finally, returns to the initial position. The segment that corresponds to the poses in which the person is bent down is the most discriminative one. The poses of this segment match with the most discriminative key poses of the *bend* action class, whereas the ratio between discrimination value and distance is lower for the other classes. For this reason, action zones can be detected by defining the thresholds $HT_1(t), HT_2(t), ..., HT_A(t)$ that have to be reached by the class evidence values of these segments. Specifically, an action zone will be collected from the frame on where:

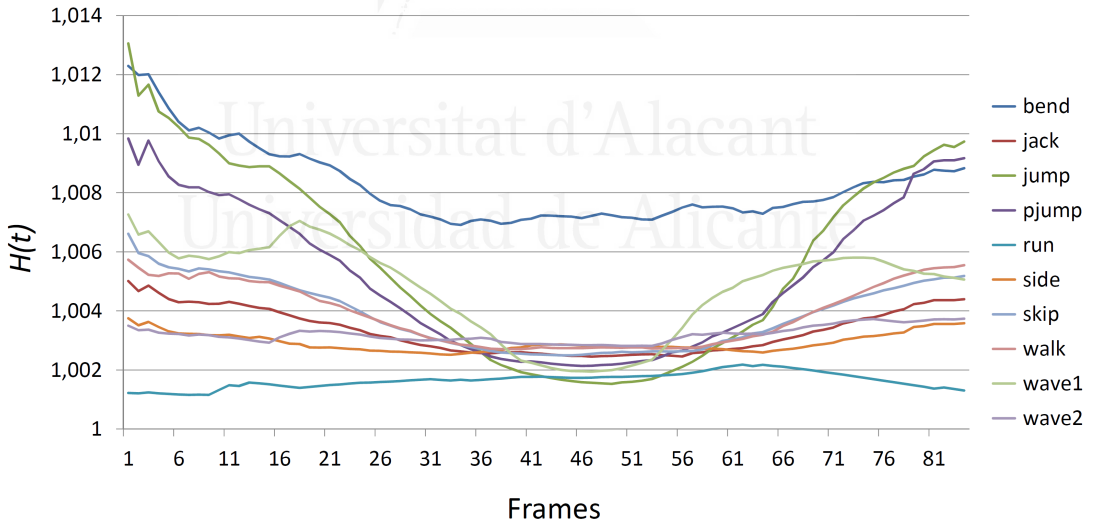
$$H_{action}(t) > H_{median}(t) + HT_{action} \ , \tag{7.5}$$

where *action* corresponds to the action class of the current sequence and $H_{median}(t)$ indicates the median value out of $H_1(t), H_2(t), ..., H_A(t)$. An action zone will end if this condition ceases to be met. The median value is employed because the expected peak of $H_{action}(t)$ would influence the average. Moreover, this approach also works if action segments present a high evidence value for more than one action class, which may happen for very similar actions.

Last but not least, the class evidence can behave differently in specific actions. In figures 7.2(a) and 7.3, the obtained graphs for two cyclic actions (*jumping jack* and *walk*) are shown. For *jumping jack*, several short action zones could be found choosing the appropriate threshold HT_a. Observing the silhouettes of the sequence in figure 7.2(b), it can be seen that peaks correspond to the frames in which the limbs are outstretched. However, for the *walk* action, no clear peaks can be distinguished. This is obvious regarding that all the postures are as discriminative, since the person walks into the image and out of it without stopping. For this special cases, the action zone is modelled as the central third part of the labelled sequence. This results in sufficient frames to capture the

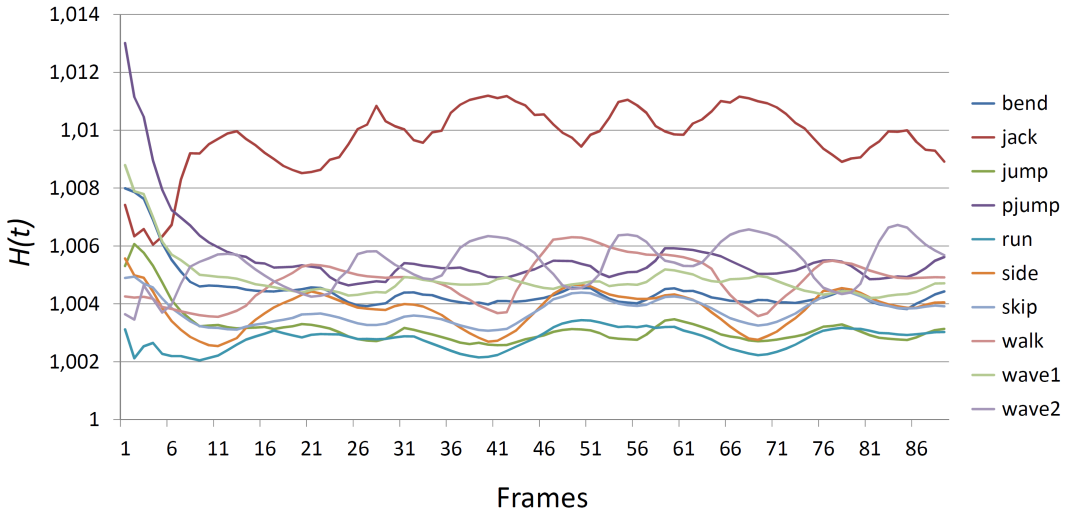


(a) Raw values

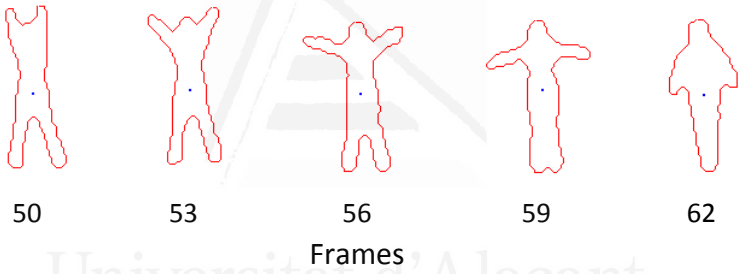


(b) Processed values

Figure 7.1: Evidence values of each action class before and after processing are shown for a *bend* sequence of the Weizmann dataset.



(a) Evidence values



(b) Corresponding silhouettes

Figure 7.2: Evidence values $H(t)$ of each action class and the corresponding silhouettes of one of the peaks of evidence are shown for a *jumping jack* sequence of the Weizmann dataset.

pattern of the repetitive motion of the action, and reduces the computational cost of the sequence matching that has to be applied afterwards.

In this way, the action zones are learnt and their segment of the sequence of key poses is employed for classification.

7.3.2 Continuous recognition

As it has earlier been mentioned, in this proposal the classification is performed by recognising action zones. For the continuous evaluation of the incoming video stream, a sliding window technique is employed. More specifically, a sliding and

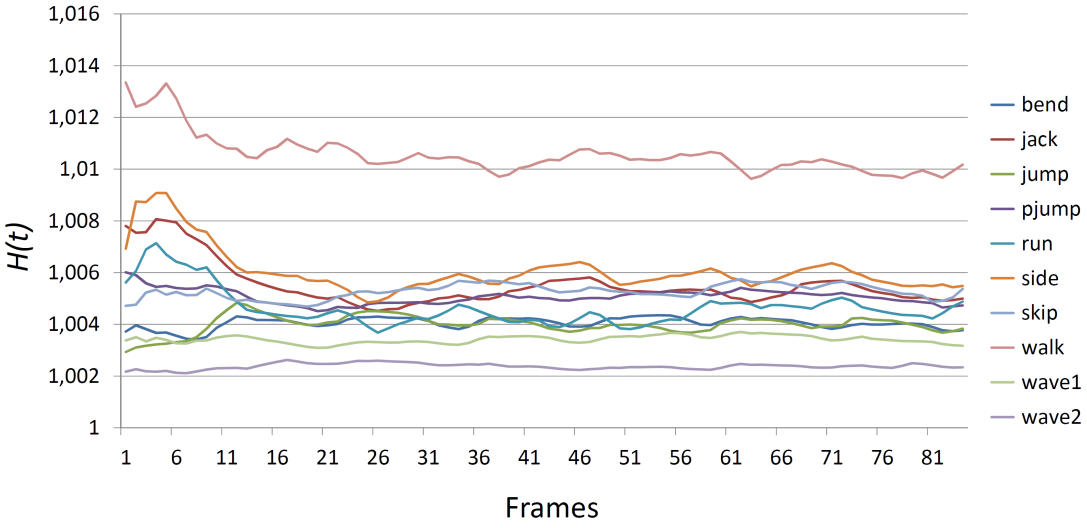


Figure 7.3: Evidence values $H(t)$ of each action class are shown for a *walk* sequence of the Weizmann dataset.

growing window is used to process the continuous stream at different overlapping locations and scales.

Algorithm 4 details the process: The sliding and growing window grows δ frames in each iteration and slides γ frames if the window has reached its maximal length $length_{max}$. If at least $length_{min}$ frames have been collected, the segment of the video stream (or video streams if available) S that corresponds to the window is compared to the known action zones. The best match is obtained as before based on matching the segments of key poses using DTW. Then, a threshold value DT_a is taken into account in order to trigger the recognition. This value DT_a indicates the highest allowed distance $dist_{min}$ in a per-frame basis. In this way, only segments which match well enough with an action zone are classified. Eventually, the unrecognised frames will be discarded and considered to belong to the *null class*.

Algorithm 4 Continuous recognition: sliding and growing window

Let δ denote the number of frames the window grows in each step.
Let γ denote the number of frames the windows moves when slid.
Let S denote the video stream.

$start = 0$
$length = 0$

repeat
——————— Sliding and growing window ———————
$length = length + \delta$

if $length > length_{max}$ **then**
Discard γ frames considered to belong to the *null class*
$start = start + \gamma$
$length = length - \gamma$
end if

——————— Compare to action zones ———————
if $length \geq length_{min}$ **then**
$dist_{min} = max\_value$
for each $action\_class \in training\_set$ **do**
for each $action\_zone \in action\_class$ **do**
$dist = d_{DTW}(action\_zone, S[start : start + length])$
if $dist < dist_{min}$ **then**
$dist_{min} = dist$
$a = action\_class$
end if
end for
end for

——————— Recognise or continue ———————
if $dist_{min} \leq length \times DT_a$ **then**
Recognise segment $S[start : start + length]$ as action class a
$start = start + length$
$length = 0$
end if
end if

until end of stream or forever

7.4 Experimentation

In this section, the performed experimentation will be detailed. Since several new parameters have been introduced for this CHAR approach, insights about how to establish them are given. Furthermore, in order to test the approach, an appropriate evaluation scheme has to be defined. Finally, the achieved results are presented.

7.4.1 Parametrisation

Special consideration has been given to the parameters $HT_1, HT_2, ..., HT_A$ and $DT_1, DT2, ..., DT_A$. The first ones define the threshold that has to be surpassed by the class evidence $H_{action}(t)$ in comparison to the $H_{median}(t)$ value. Different values are admitted for each action class, since the class evidence behaves differently for each type of action. This has been shown in figures 7.1 to 7.3. In the case of the second set of parameters, each action class is considered to require a specific similarity between sequence segments and action zones in order to confirm the match as a recognition and avoid false positives for 'poor matches'. Some actions may be naturally performed in many different ways, whereas others are more similar in general between samples and subjects. This leads us to two sets of A parameters that are difficult to establish empirically, as exhausting tests are unaffordable. Although a manual approach could have been employed, this would be a strongly scenario-specific and slow process. For this reason, an appropriate search heuristic is required.

As it has been seen in chapter 6, EA are proficient in guided search. Relying on the previously used wrapper approach and the cooperative CEA from section 6.2, the best performing combination of HT and DT values can be found, taking into consideration their intrinsic dependencies. Therefore, two populations have been defined for the two arrays of A values. The scale of the real numbers is changed (to 10^2 and 10^4 respectively for HT and DT) so that integer values can be employed and the best performing value can be found in the corresponding ranges: $HT \in [5,\ 150]$ and $DT \in [10,\ 250]$. Then, the same crossover and mutation operators are applied as before to obtain the fittest individuals of each population by means of coevolution (see integer values population of the number of key poses $K_1, K_2, ..., K_A$ in section 6.2.2). The obtained parameter values are detailed further below.

7.4.2 Continuous evaluation

So far, for action recognition based on sequences, the evaluation scheme has been straightforward. Since the ground truth label of each sequence is known, the ratio of correctly classified sequences in the test has been used as accuracy score. Nevertheless, in the continuous evaluation, several new constraints appear. Depending on the application scenario, one might be interested in the number of repetitions of each action. This happens in gaming (*e.g.* three *punches*), whereas in video surveillance the fact that the action happened is more relevant (*e.g. punching*). In AAL, it is especially important not to miss any actions, because this could result in safety issues (*e.g. falling down*). A delay of a few seconds may be acceptable if this improves the recognition avoiding false negatives. As a result, the applied evaluation scheme varies between authors.

A common option is to apply frame-by-frame evaluation as in Lu et al. (2013); Wang et al. (2012), but the reliability of this approach is arguable. This is due to the lack of correlation between frames and actions. It could happen that only a few last frames of an action are not recognised correctly. This would result in a high frame-by-frame recognition rate (*e.g.* 90%), although only one correct class label and one or more incorrect predictions have been returned by the system. This means that 50% or more of the returned labels were erroneous. For this reason, other proposals have been made. For instance, in Guo et al. (2012), classification of segments is considered and a certain overlap with the ground truth is required. Moreover, in the work of Minnen et al. (2006); Ward et al. (2011), further thought is given to this issue. As it can be seen in those contributions, evaluation schemes can be categorised into three levels of analysis:

1. The already mentioned *frame analysis*, where frames or similar atomic units are evaluated. Although the difficulty to relate this temporal units to the actions themselves has been mentioned, this approach comes along with the advantage of not requiring any temporal alignment.

2. In *event analysis*, the basic unit of comparison is the individual activity occurrence. Approaches may consider only the order of the events and compare them with the recognition (using, for instance, the Levenshtein distance), or take into account as well the occurrence time.

3. A hybrid approach called *segment analysis* combines the previous two relying on segments as the basic unit of comparison. A segment is defined

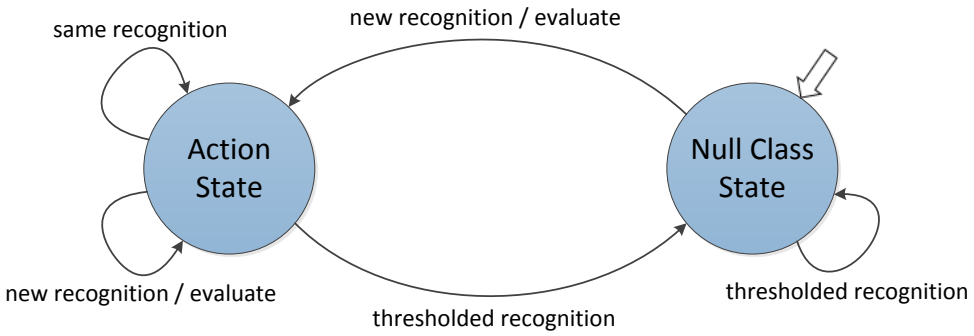


Figure 7.4: This finite-state machine details the logic behaviour of the applied segment analysis.

as "an interval of maximal duration in which both the ground truth and the predicted activities are constant" (Minnen et al., 2006). In this way, despite the fact that segments may have different durations, alignment is given since each ground truth or prediction change leads to a new unit of evaluation.

This last level of analysis has been employed in this work, as it provides a clear way to align the recognitions with the ground truth and avoids the disadvantages of the frame-based analysis. Figure 7.4 shows how the *null class* has been considered in the segment analysis. As it can be seen, only new recognitions (*i.e.* different from the last predicted action class) are taken into account for evaluation. The thresholded recognitions are retained and their segments are considered to belong to the *null class*. In addition, recognitions are accepted for a delay of τ frames after the ground truth indicated the end of the action. Note that this is only allowed if no prediction was given until that moment, *i.e.* the *null class* state was active since the action started and until the delayed recognition has occurred. Otherwise, the action would have already been classified (correctly or wrongly).

In view of the multi-class classification that is performed and that now a *null class* has to be contemplated, results are measured in terms of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN).

Specifically, these results are computed as follows:

1. **True positives**
   (a) If the current action is correctly classified.
   (b) If the last action is correctly classified in the allowed delay of τ frames and no prediction was given for this action until that moment.
2. **False positives**
   (a) If the current action is wrongly classified.
3. **False negatives**
   (a) If the last action has not been recognised in the allowed delay of τ frames.
   (b) If the current action has been classified (correctly or wrongly) and the last action has been ignored as a result.
   (c) If the current action and the last action have been ignored (null class state) and a third action has began. In other words, it is not allowed to recognise actions which are not immediately preceding.
   (d) If the current or the last action has not been classified on time and the video stream has ended. In this case, the allowed delay cannot be completely used and false negatives are computed for both actions.

   Figure 7.5 shows an example of the last three types of false negatives.
4. **True negatives**
   (a) If a segment finishes and it was correctly considered to belong to the *null class*.

These values are accumulated along a cross validation test. Similarly to the previously employed tests, a LOAO cross validation is proposed in which each sequence now includes several continuously performed actions of an actor. Typically, two metrics are employed to evaluate how the method behaves, these are *precision* and *recall*. Whereas the first computes the ratio of correct predictions with respect to the total number of action predictions ($precision = \frac{TP}{TP+FP}$), the second does it with respect to the total number of segments labelled as actions

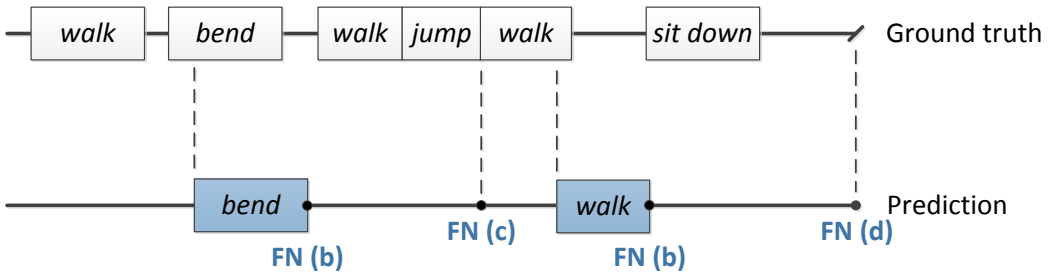


Figure 7.5: Example of how the different types of false negatives are computed.

in the ground truth ($recall = \frac{TP}{TP+FN}$). To combine these metrics into a single measure of performance, the F_1-measure is used:

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{7.6}$$

7.4.3 Results

Event though the IXMAS dataset is mainly used for action recognition based on sequences, it actually provides a continuous ground truth labelling. Actors performed the actions one after the other, and the frames which do not correspond to any action have been labelled with the *null class* (in these frames, the actor is waiting to perform the next action, walking in and out of the image, etc.). Therefore, the IXMAS dataset can be used for the evaluation of CHAR.

Some works provide results on the Weizmann dataset. In this case, the sequences of the same actor are concatenated into a single continuous sequence. Consequently, unnatural transitions are created due to the gaps of information. Nevertheless, tests have been performed on this dataset for illustrative purposes so that a comparison with other approaches can be made.

With regard to the introduced parameters, the following values have been used during the experimentation (these have been chosen based on experimentation):

1. The Gaussian smoothing applied to the $H(t)$ class evidence is of $\sigma = 10.486$ frames. Since approximate discrete values are applied for the convolution, a total of 22 history frames are taken into account and the rest is considered zero.

Table 7.1: Dataset-specific parameter values of the applied tests.

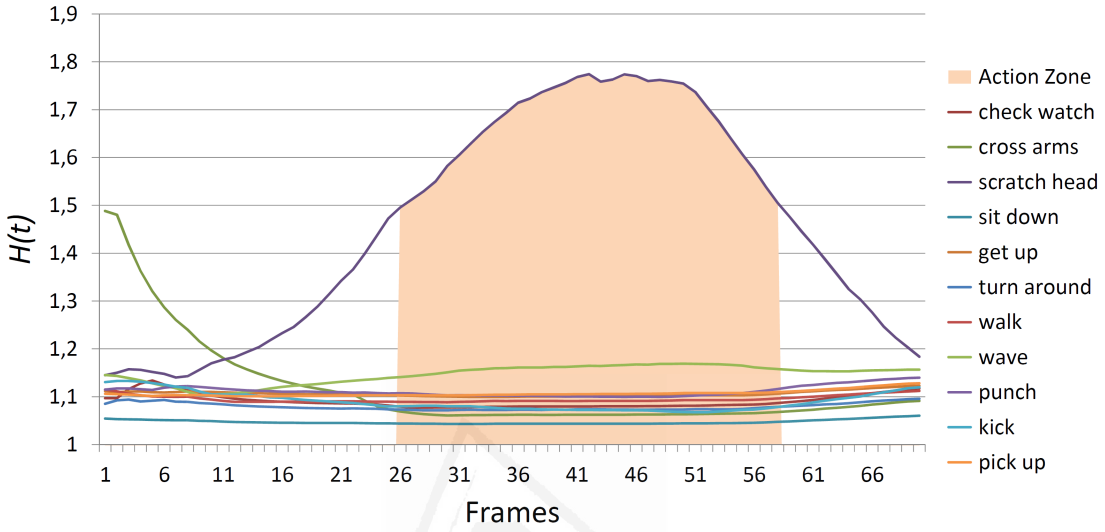
Dataset	**Test**	***K***	***B***	**Actions**	**HT** (10^2)	**DT** (10^4)
IXMAS	LOAO	130	27	*check watch*	80	123
				cross arms	94	105
				scratch head	97	131
				sit down	138	187
				get up	21	152
				turn around	53	119
				walk	53	110
				wave	31	96
				punch	72	118
				kick	5	70
				pick up	46	20
Weizmann	LOAO	10	14	*bend*	145	232
				jack	116	240
				jump	140	150
				pjump	14	121
				run	113	149
				side	35	188
				skip	74	174
				walk	81	150
				wave1	111	129
				wave2	68	121

2. Regarding the sliding and growing window, in each iteration the window grows 5 frames ($\delta = 5$), and when the maximal length $length_{max}$ is reached, the window slides 10 frames ($\gamma = 10$).

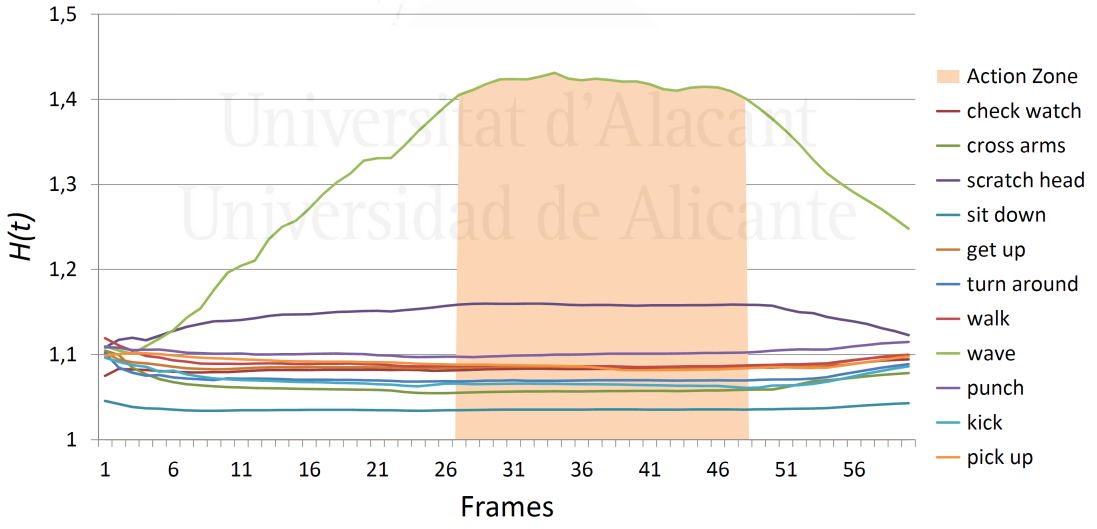
3. A delayed recognition is accepted within a period of 60 frames, corresponding to approximately 2 seconds ($\tau = 60$).

4. The remaining parameter values have been chosen for each dataset. Tables 7.1 and 7.2 show the corresponding values. The best performing combination of HT and DT thresholds obtained by means of the CEA are also detailed. In figure 7.6, two sample sequences can be seen where the action zones that have been obtained using these HT class evidence thresholds are highlighted.

Comparison with other state-of-the-art works is difficult in CHAR. Depending on the application scenario, authors applied different evaluation schemes, as it has been seen in section 7.4.2. Wang et al. (2012) and Lu et al. (2013) both employed frame analysis and reported respectively 76.5% and 81.0% accuracy on



(a) *Scratch head* action zone



(b) *Wave* action zone

Figure 7.6: Evidence values $H(t)$ of each action class and the detected action zones are shown for a *scratch head* and a *wave* sequence of the IXMAS dataset.

the IXMAS dataset. However, the first relies on per-frame accuracy, whereas the second measured the average per-class frame accuracy. In the case of the Weizmann dataset, for example, Guo et al. (2012) performed CHAR and reached a score of 97.8%. Segment analysis is employed in this case, although the rate of correctly classified segments is computed requiring a 60% overlap with the ground truth.

Table 7.2 shows the scores that have been achieved by our approach over the ideal value F_1-measure of 1.0. As it has been seen along this work, the IXMAS dataset presents several additional difficulties and these also had an impact on this result. Furthermore, the segments labelled as *null class* in which 'other actions' are performed can easily lead to an increase of false positives. In order to show the benefit gained from the action zones approach, tests have also been performed using the entire segmented sequences instead. In this way, larger segments are considered by the sliding and growing window and these are compared to the original action sequences provided by the ground truth. It can be seen that the proposed continuous recognition based on action zones provides a substantial performance increase and leads to higher scores.

The temporal performance has also been evaluated for this continuous approach. While the sliding and growing window technique is computationally demanding, this is offset by the proposed action zones. Due to the short lengths of both action zones and temporal windows, the comparisons between them is very efficient. Using the same test environment as in previous tests and computing contour extraction, key pose matching and continuous recognition, a rate of 196 fps has been measured on the Weizmann dataset.

Table 7.2: Obtained results applying CHAR and segment analysis evaluation (LOAO cross validation test). Results are detailed using the segmented sequences or the proposed action zones.

Dataset	**Approach**	$length_{min}$	$length_{max}$	F_1**-measure**
IXMAS	Segmented sequences	10	120	0.504
IXMAS	Action zones	3	30	0.705
Weizmann	Segmented sequences	10	120	0.693
Weizmann	Action zones	3	20	0.928

7.5 Remarks

In this chapter, the proposed method for human action recognition has been extended to support continuous human action recognition. Proposals have been made at the learning and recognition stage. The concept of action zones has been introduced to define and automatically learn the most discriminative segments of action performances. Relying on these action zones, recognition can be carried out by finding the equivalent segments that clearly define the action that is being performed. For this purpose, a sliding and growing window approach has been employed. Finally, after reviewing the existing evaluation schemes, segment analysis is used introducing special considerations for the specific AAL application of our work. Tests have been performed relying on the whole segmented sequences or only on the action zones, and significant differences can be seen. By means of action zones, higher accuracy scores are obtained. Real-time suitability of this continuous approach has also been verified. This is indispensable for most of the possible applications, and a necessary premise for online recognition.

In future works, further evaluation should be applied to ease the comparison to other approaches. It could be useful to implement other state-of-the-art techniques and test them in the same conditions as our proposal. Furthermore, a consensus should be reached about the appropriate evaluation schemes. It has also been observed that regarding CHAR, there is a lack of suitable benchmarks including foreground segmentations or depth data. Therefore, new datasets should be created along the corresponding evaluation schemes.

Chapter 8

Concluding remarks

This chapter will conclude this thesis including discussion, conclusions and future directions of the developed work in the following sections.

8.1 Discussion

In this thesis, 2D silhouettes have been employed to extract spatial information and learn its evolution over time. As it has been seen in the analysis of the state of the art, this type of pose representation is view dependant, which means that the recognition is limited to specific viewpoints. However, experimental results have shown that the proposed *Radial Summary* feature supports up to 45° shifts. Furthermore, since a multi-view setup is employed, other viewpoints are considered to overcome this shortage. In this sense, as it has been shown, the proposed method improves its accuracy if further views are added to the classification.

Silhouettes can also be ambiguous. For instance, distinguishing *eating* from *shaving* can be very difficult if we look only at the human silhouette. Nonetheless, this would be solved easily knowing, for example, the person's location at home. This means that the proposed method could benefit enormously from multi-modal data. Therefore, data fusion of environmental sensor information can provide a significant improvement and, together with action recognition, lead to accurate recognition of more complex activities and long-term behaviour.

Another concern regarding silhouettes is that these may be difficult to obtain under certain circumstances. Research on segmentation methods that are robust to diverse lighting conditions and include shadow removal is still being carried

out (Gallego and Pardàs, 2010). To this extent, an alternative data acquisition method based on RGB-D sensors has been verified. This technique is suitable for indoor environments, as homes, and allows to obtain accurate human silhouettes even without illumination, which means that certain activities can be monitored by night. In this sense, it has also been seen that proposed method is able to benefit from preprocessing techniques as tracking and human body detection.

Considering the type of actions that have been tested, the method has shown to support a wide variety of actions performed by different types of subjects. However, it would be interesting to establish an action class hierarchy. For instance, it has been seen that *walking* can be difficult to distinguish from *running*. If these two actions were grouped by a superclass, the monitoring system would probably still be able to act accordingly even if only the superclass is correctly recognised.

The deployment of video surveillance at home raises obvious concerns about privacy issues. As it has been seen in the architecture of the intelligent monitoring system that is being developed, besides human behaviour analysis, a privacy protection system based on image modification or redaction methods is being designed. This system takes into account the personal needs and preferences of the inhabitants. For instance, once the human posture has been recognised, the person in the image could be replaced with an avatar, *i.e.* a 3D human body model. In this way, the identity and appearance of the person can be protected. With the resulting data, HBA can still be performed. Therefore, this is an open issue that is being researched in the 'TALISMAN+' project. Moreover, since the used pose representation relies only on the boundary of the human silhouette, the actual colour image can be discarded in an early processing stage.

Furthermore, it has been confirmed that the presented HAR method is also suitable for other applications besides human action recognition for AAL scenarios. In related publications Chaaraoui et al. (2014); Climent-Pérez et al. (2013), gesture and action recognition related to gaming and NUI is handled. In these works, skeletal 3D pose representations, which have been obtained using the depth data from RGB-D sensors, are successfully classified based on the presented classification method. The obtained results outperform the state of the art. This confirms that the introduced method is proficient also with other pose representations.

Although this work has assumed that a single person is being monitored, in combination with visual tracking techniques, the silhouettes of multiple persons can be provided and their behaviour can be monitored. This should be verified with the appropriate tests.

Finally, the output of the performed human behaviour analysis consists of the recognised action classes along time. The general architecture of the intelligent monitoring system includes a reasoning system, which is responsible of considering this information and acting accordingly depending on the events that are being monitored. It would certainly benefit from a degree-of-belief probability in order to weight each result of the human behaviour analysis and decide, for instance, based on the risk involved in the recognised action. In this sense, a belief measure can be provided based on the final distance of the match of the classification.

8.2 Conclusions

In this work, a 2D template-based non-parametric method has been presented for human action recognition. The requirements of the thesis proposal have been satisfied by means of a very robust technique that has shown to be versatile and efficient. The HAR method based on a bag-of-key-poses model handles single- and multi-view recognition proficiently. State-of-the-art recognition rates have been achieved, outperforming the best known rates in several benchmarks.

The provided contributions can be summarised as follows:

(a) A review of the state of the art of vision-based human behaviour analysis has been provided. A taxonomy is proposed in order to join existing definitions and establish different HBA levels to categorise the works of the research field. (This contribution has been published in Chaaraoui et al. (2012b); Climent-Pérez et al. (2012).)

(b) A HAR method based on a bag-of-key-poses model and multiple views has been proposed to handle multi-view human action recognition in real-time. In contrast to the bag-of-words model, in this case, the temporal relation between representative instances is learnt over time, and specific support of fusion of multiple views is considered.

This method can be used with different pose representations and specific algorithms in its processing stages. Its adaptability to other applications has been verified in related works, where the method has been successfully employed to recognise gaming actions and gestures. (This contribution has been published in Chaaraoui et al. (2012a); Chaaraoui, Climent-Pérez, and Flórez-Revuelta (2013); Chaaraoui, Padilla-López, and Flórez-Revuelta (2013); Climent-Pérez et al. (2013); Chaaraoui et al. (2014).)

(c) In the proposed implementation of the method, the *Radial Summary* feature has been introduced. Based on the boundary of human silhouettes, a very characteristic and low dimensional feature is obtained. Regarding the learning from multiple views, several proposals have been made leading to a final result based on a weighted feature fusion scheme. Stable and accurate results have been obtained on several publicly available benchmarks, performing better and faster than existing real-time methods.

The data acquisition stage of the method has been validated by means of an alternative method using RGB-D sensors. Based on these devices, reliable human silhouettes can be obtained in indoor environments. Experimental results show that this type of data can be classified successfully. (This contribution has been published in Chaaraoui and Flórez-Revuelta (2013a,b).)

(d) In order to learn scenario-specific constraints and customise the system to the recognition of a specific set of actions and subjects, optimisation methods are proposed. Based on evolutionary algorithms, feature subset, training instances and parameter values have been selected. Furthermore, an adaptive method is proposed to handle scenarios in which the learning has to continue during the execution of the system. Experimental results indicate that accurate recognition can be obtained based on these proposals. (This contribution has been published in Chaaraoui and Flórez-Revuelta (2013a,b).)

(e) The method has been extended to support the recognition of continuous video streams. For this purpose, action zones that allow to detect the most discriminative parts of an action have been proposed. Consequently, continuous action detection and recognition can be performed with a sliding window approach.

(f) The proposed techniques have been designed to satisfy specific demands of AAL services, like relaxed camera setup requirements, adaptive learning, continuous recognition and real-time execution. During the carried out experimentations, temporal evaluation has been performed to verify that the proposals are in fact suitable for online recognition. The measured performance has consistently been above video frequency, enabling the recognition in real time.

As a result, the introduced method is conceived as part of an intelligent monitoring system that aims to provide AAL services and extend the independent life at home of elderly and impaired people. This system will have to consider specific requirements of AAL applications, like the needed infrastructure, constraints of the scenario to be monitored, minimum acceptable accuracy, etc. (This contribution has been published in Chaaraoui, Padilla-López, Ferrández-Pastor, García-Chamizo, Nieto-Hidalgo, Romacho-Agud, and Flórez-Revuelta (2013).)

In conclusion, this thesis provides valuable advances for human action recognition. Given its temporal efficiency and relaxed constraints regarding data acquisition and camera setups, this work may fill the gap between motion and activity recognition, easing the development of further long-term analysis of human behaviour.

8.3 Future work

Regarding future lines of research, several proposal-specific ones have already been introduced in the corresponding chapters of this thesis. These future directions are listed below from lower to higher level of processing:

Pose representations During the review of the state of the art, it has been seen that local and global pose representation can be extracted out of the video data. Furthermore, 3D models and spatio-temporal (x, y, t or x, y, z, t) information can be extracted. Therefore, a wide variety of options exist which could be considered for real-time recognition. This would allow, for instance, to apply more fine-grained motion recognition.

Key poses In this work, key poses have been obtained by means of a clustering algorithm. As has already been mentioned, a great number of these ex-

ists and more advanced methods have been presented in chapter 2. Other codebook generation techniques could be employed too. For this reason, potential for improvement could be found in this area.

Information fusion Regarding information fusion, multiple options have been seen for the fusion of multiple views. For instance, if a 3D pose representation were employed, it may benefit from data fusion, since this would allow to obtain a more accurate pose information by combining the multi-view data. Not only different levels, but also different data types can be subject of information fusion. In other words, multiple pose representations, or non-visual data from other sensors can be combined in order to improve the quality of the recognition.

Bag-of-key-poses model It has been seen that the bag-of-key-poses model allows to learn representative instances of the feature data, whose temporal evolution can then be employed for sequence modelling. Although, this method has been designed for the purpose of this work, the learning method is not necessarily limited to visual features nor HAR. As a matter of fact, it can be applied to diverse kinds of pattern recognition problems where features are extracted along an ordered sequence of instances. This leads to a wide area of possible applications and future research.

Distance metrics Also a choice of distance metrics regarding both the comparison of pose representations and key poses, and the matching of sequences has been made. In chapter 4, several options have been proposed for this part of the method. Further testing and consideration of techniques from other research fields should be carried out in this direction.

ADL and long-term behaviour In order to move forward to the recognition of more complex activities and long-term behaviour, the proposals of this work can be applied. For example, sequences of actions can be tracked along time. By means of probabilistic graphical models, the order and repetitions of the actions can be learnt and activities could be recognised. Furthermore, long-term HBA can be applied exploiting the knowledge base of the performed actions along the day.

Evaluation In this work, several different datasets have been used as benchmarks for evaluation. Nonetheless, no appropriate real-world data of elderly or impaired people has been found. In association with care giving organisations and volunteers, this data could be obtained. Since it will not always be possible to employ a scenario-specific dataset to train the system, the available datasets should be merged with a shared ground truth basis, so that a general learning can be applied and a single training process may serve for multiple testing scenarios.

Deployment Real-time recognition has been a main requirement of this thesis, and, as a result, the proposals present a low computational demand. Despite of this fact, in future work it is necessary to test the method in the technical setup conditions of intelligent environments, and handle the deployment in the available hardware and infrastructure. This opens several issues regarding video networks and the required management software, among others.

In general, however, two main future lines of work define how this thesis can be continued. On the one hand, as it has been stated before, based on sequences of actions and other multi-modal data from environmental sensors, more complex activities can be recognised. This would allow to increase the degree of semantics and the time frame of the analysis. It would also require object recognition and modelling, since objects provide essential information about the activities that are carried out. On the other hand, the presented method is able to recognise motion-based actions, *i.e.* the actions need to involve significant body motion. But, for instance, gesture recognition requires tracking the movement of the hands or the face. In this sense, the spatial information could be enhanced with other features, as dense spatio-temporal interest points or 3D body pose estimation. A preliminary approach that exploits this option has already been tested with successful results (Chaaraoui, Padilla-López, and Flórez-Revuelta, 2013).

Appendix A

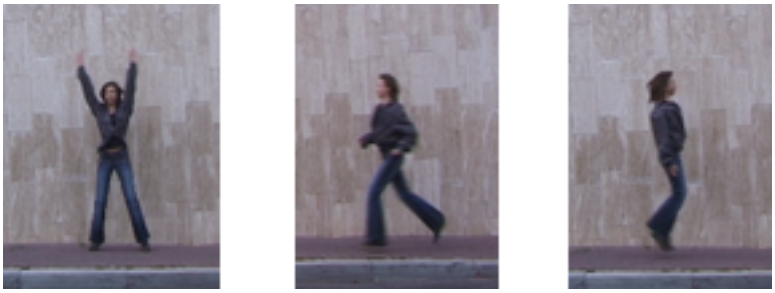
Datasets

In this appendix, technical details and sample images are provided for the RGB and RGB-D datasets, that have been employed in the experimentation of this thesis. For each dataset, a few sample images are shown, as well as the corresponding silhouette contours and centroids which have been used for feature extraction. For greater detail about the datasets, we refer to the corresponding references.

A.1 RGB images

A.1.1 Weizmann

The Weizmann dataset presented in Gorelick et al. (2007) is a single-view (static front-side camera) outdoor dataset. It provides 180 × 144 px resolution images of ten different actions performed by nine actors. This results in a total of 93 sequences, since the actress *Lena* performed *run*, *skip* and *walk* twice. It has a relatively simple background, provides automatically extracted silhouettes (we use the version without post-alignment) and has become a reference in human action recognition. Actions include *bending (bend), jumping jack (jack), jumping forward (jump), jumping in place (pjump), running (run), galloping sideways (side), skipping (skip), walking (walk), waving one hand (wave1)* and *waving two hands (wave2).* It is worth mentioning that several works exclude the *skip* action, as it commonly shows higher error rates and also weakens the recognition of other actions.



(a) RGB images

(b) Silhouette contours

Figure A.1: Sample images and silhouette contours from the Weizmann dataset. From left to right the *jumping jack*, *running* and *jumping forward* actions are shown for the actress *Daria*.

The silhouettes have been obtained through background subtraction techniques. Therefore, they present noise and incompleteness.

A.1.2 MuHAVi

The MuHAVi dataset (Singh et al., 2010) is a more recent and complex benchmark with multi-view images. It provides 720 × 576 px resolution images on a complex background with street light illumination. Its full version includes 17 different actions performed by seven actors and has been recorded indoors with eight CCTV cameras, each one at 45° to its neighbours. A manually annotated subset (*MuHAVi-MAS*) provides silhouettes for two of these views (front-side and 45°) and two actors, labelling 14 (MuHAVi-14: *CollapseLeft, CollapseRight, GuardToKick, GuardToPunch, KickRight, PunchRight, RunLeftToRight, RunRightToLeft, StandupLeft, StandupRight, TurnBackLeft, TurnBackRight, WalkLeftToRight* and *WalkRightToLeft*) or eight (MuHAVi-8: *Collapse, Guard, KickRight, PunchRight, Run, Standup, TurnBack* and *Walk*) actions in its merged version. Actors performed up to four different samples of each action leading to a

set of 136 sequences, or 68 multi-view sequences. In difference to MuHAVi-14, in MuHAVi-8 the direction in which an action is performed is ignored. This means that although fewer action classes need to be recognised, these are more difficult to learn since they present more differences.

(a) RGB images

(b) Silhouette contours

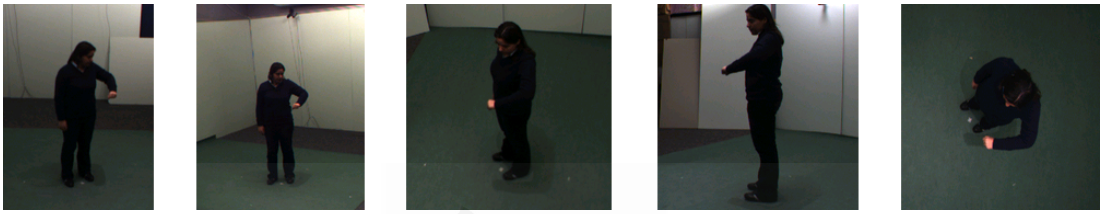
Figure A.2: Sample images and silhouette contours from the MuHAVi-MAS dataset. From left to right the *KickRight*, *CollapseLeft* and *PunchRight* actions are shown.

As it can be seen in fig A.2, the manually annotated silhouettes are of greater quality. The purpose of providing these is to separate the difficulties which arise from the background subtraction task from the silhouette-based action recognition.

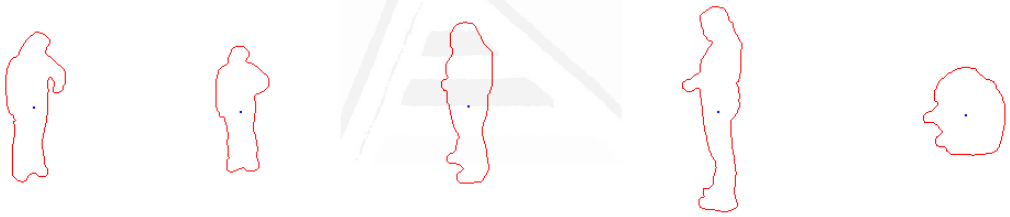
A.1.3 IXMAS

The INRIA Xmas motion acquisition sequences (IXMAS) dataset (Weinland et al., 2006) is popular among human action recognition methods that are specifically designed for multi-view recognition. The IXMAS dataset includes multi-view data and is especially aimed at view-invariance testing. It provides 390 $\times$ 291 px resolution images from five different angles including four sides and one top-view camera. A set of 12 actors have been recorded performing 14 different actions (*check watch, cross arms, scratch head, sit down, get up, turn*

around, walk, wave, punch, kick, point, pick up, throw over head and *throw from bottom up*) three times each, resulting in a dataset with over 2 000 sequences. This benchmark presents an increased difficulty because subjects were asked to freely choose their position and orientation. Therefore, each camera has captured different viewing angles, which makes methods which rely on fixed camera views (front, side, etc.) unsuitable. The authors of the dataset suggested to exclude the *point* action and the two throwing actions, leading to a benchmark of 1 980 sequences, or 396 multi-view sequences.



(a) RGB images



(b) Silhouette contours

Figure A.3: Sample images and silhouette contours from the IXMAS dataset. From left to right camera views 1 to 5 of the *check watch* action are shown.

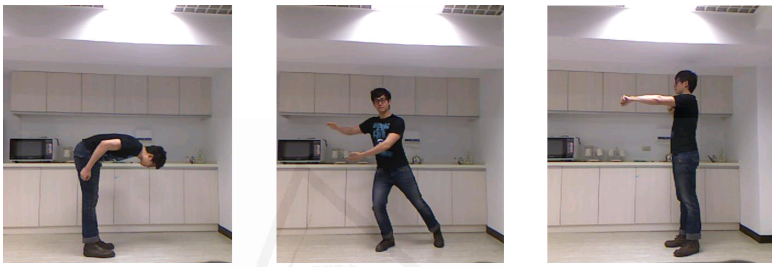
Similarly to the Weizmann dataset, automatically obtained foreground masks are provided, leading to noisy and unreliable silhouette data.

A.2 RGB-D data

A.2.1 DHA

The depth-included human action video dataset (DHA) (Lin et al., 2012) has a two-fold objective. On the one hand, it provides RGB-D data, *i.e.* along the RGB colour information of each pixel, the real-world distance to the object is detailed. This makes it possible to apply 3D or depth-based techniques, as well as to obtain a reliable silhouette. On the other hand, it provides a greater number

of classes and samples, which is essential for learning. The dataset relies on single-view 640 × 480 px RGB-D data of action classes, which are consistent with the traditional motion-based human action datasets like Weizmann (Gorelick et al., 2007) and KTH (Schuldt et al., 2004). In this sense, the same ten actions of the Weizmann dataset have been recorded. In addition, the following actions are available: *front clap*, *arm swing*, *leg kick*, *rod swing*, *side box*, *side clap*, *arm curl*, *leg curl*, *golf swing*, *front box*, *taichi*, *pitch* and *kick*. The usage of different subsets of respectively 10, 17 or all the 23 actions is proposed. Up to 21 actors were recorded reaching a total of 483 sequences.



(a) RGB images

(b) Silhouette contours

Figure A.4: Sample images and silhouette contours from the DHA dataset. From left to right the *bend*, *taichi* and *side box* actions are shown.

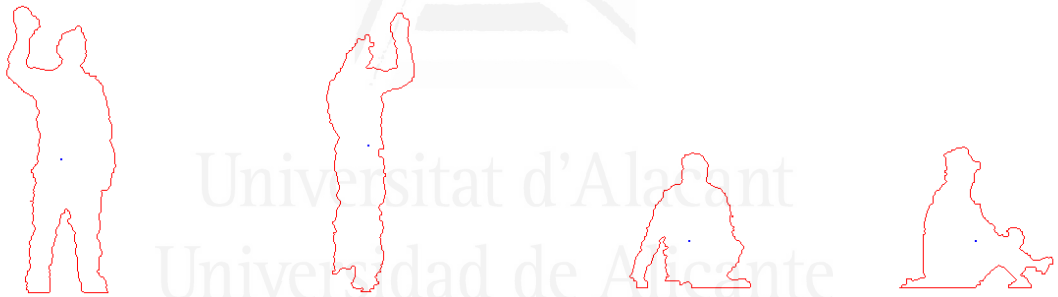
A.2.2 DAI RGBD

The DAI RGBD dataset has been recorded by the *Domotics and Ambient Intelligence Research Group* of the University of Alicante. The goal of this dataset is to provide multi-view RGB-D data for human action recognition. In our setup, we used two Microsoft Kinect devices recording a front and a 135° back-side view. Cameras were located at 2.5 and 3.5 m respectively, recording an indoor scenario with halogen lighting. This dataset includes 640 × 480 px RGB-D

data of the following 12 actions classes: *Bend, CarryBall, CheckWatch, Jump, PunchLeft, PunchRight, SitDown, StandingStill, Standup, WaveBoth, WaveLeft* and *WaveRight*. These actions have been performed by three male actors with noteworthy differences regarding their body build. Therefore, 72 sequences or 36 multi-view sequences are available. Silhouettes can be obtained using depth-based segmentation.

In future works, we intend to publish an expanded version of this dataset with more types of human motion (gestures, actions and ADL), subjects and samples.

(a) RGB images



(b) Silhouette contours

Figure A.5: Sample images and silhouette contours from the DAI RGBD dataset. From left to right the *WaveRight* and *SitDown* actions are shown for the front and backside view.

Appendix B

Publications

In this appendix, the resulting publications of the contributions presented in this thesis are listed. These include both the detailed proposals as well as applications.

B.1 Journals

I Chaaraoui, A.A., Climent-Pérez, P., Flórez-Revuelta, F., 2012. A Review on Vision Techniques Applied to Human Behaviour Analysis for Ambient-Assisted Living. Expert Systems with Applications 39 (12), 10873 - 10888.

Journal: Expert Systems with Applications
Impact factor: 2.203 (JCR 2011), 1.854 (JCR 2012)
Category: Computer Science - Artificial Intelligence
Position of journal: 22 of 111 (JCR 2011), 31 of 115 (JCR 2012)

II Chaaraoui, A.A., Climent-Pérez, P., Flórez-Revuelta, F., 2013. Silhouette-based Human Action Recognition Using Sequences of Key Poses. Pattern Recognition Letters 34 (15), 1799 - 1807. Smart Approaches for Human Action Recognition.

Journal: Pattern Recognition Letters
Impact factor: 1.034 (JCR 2011), 1.266 (JCR 2012)
Category: Computer Science - Artificial Intelligence
Position of journal: 63 of 111 (JCR 2011), 56 of 115 (JCR 2012)

III Chaaraoui, A.A., Flórez-Revuelta, F., 2013. Optimizing Human Action Recognition Based on a Cooperative Coevolutionary Algorithm. Engineering Applications of Artificial Intelligence. Advances in Evolutionary Optimization Based Image Processing. DOI 10.1016/j.engappai.2013.10.003

Journal: Engineering Applications of Artificial Intelligence

Impact factor: 1.665 (JCR 2011), 1.625 (JCR 2012)

Category: Computer Science - Artificial Intelligence

Position of journal: 34 of 111 (JCR 2011), 38 of 115 (JCR 2012)

IV Chaaraoui, A.A., Padilla-López, J.R., Climent-Pérez, P., Flórez-Revuelta, F., 2014. Evolutionary Joint Selection to Improve Human Action Recognition with RGB-D Devices. Expert Systems with Applications 41 (3), 786 - 794. Methods and Applications of Artificial and Computational Intelligence.

Journal: Expert Systems with Applications

Impact factor: 2.203 (JCR 2011), 1.854 (JCR 2012)

Category: Computer Science - Artificial Intelligence

Position of journal: 22 of 111 (JCR 2011), 31 of 115 (JCR 2012)

B.2 Conferences and workshops

I Climent-Pérez, P., Chaaraoui, A.A., Flórez-Revuelta, F., 2012. Useful Research Tools for Human Behaviour Understanding in the Context of Ambient-Assisted Living, in: Novais, P., Hallenborg, K., Tapia, D.I., Rodríguez, J.M.C. (Eds.), Ambient Intelligence - Software and Applications. Springer Berlin / Heidelberg. Vol. 153 of Advances in Intelligent and Soft Computing, pp. 201 - 205.

Conference: 3rd Intl. Symposium on Ambient Intelligence (ISAmI 2012)

Publication type: Short conference paper

II Chaaraoui, A.A., Climent-Pérez, P., Flórez-Revuelta, F., 2012. An Efficient Approach for Multi-view Human Action Recognition based on Bag-of-Key-Poses, in: Salah, A.A., Ruiz-del Solar, J., Meriçli, C., Oudeyer, P.Y. (Eds.), Human Behavior Understanding. Springer Berlin / Heidelberg. Vol. 7559 of Lecture Notes in Computer Science, pp. 29 - 40.

Conference: IEEE/RSJ Intl. Conference on Intelligent Robots and Systems (IROS 2012)

Publication type: Full workshop paper

III Climent-Pérez, P., Chaaraoui, A.A., Padilla-López, J.R., Flórez-Revuelta, F., 2013. Optimal Joint Selection for Skeletal Data from RGB-D Devices using a Genetic Algorithm, in: Batyrshin, I., Mendoza, M. (Eds.), Advances in Computational Intelligence. Springer Berlin / Heidelberg. Vol. 7630 of Lecture Notes in Computer Science, pp. 163 - 174.

Conference: 11th Mexican Intl. Conference on Artificial Intelligence (MICAI 2012)

Publication type: Full conference paper

IV Chaaraoui, A.A., Flórez-Revuelta, F., 2013. Human Action Recognition Optimization based on Evolutionary Feature Subset Selection, in: Proceedings of the Fifteenth Annual Conference on Genetic and Evolutionary Computation Conference (GECCO'13), ACM, New York, NY, USA, pp. 1229 - 1236.

Conference: Genetic and Evolutionary Computation Conference (GECCO 2013)

Publication type: Full conference paper

CORE Ranking: A (2013)

V Chaaraoui, A.A., Padilla-López, J.R., Ferrández-Pastor, F.J., García-Chamizo, J.M., Nieto-Hidalgo, M., Romacho-Agud, V., Flórez-Revuelta, F., 2013. A Vision System for Intelligent Monitoring of Activities of Daily Living at Home, in: Nugent, C., Coronato, A., Bravo, J. (Eds.), Ambient Assisted Living and Active Aging. Springer International Publishing. Vol. 8277 of Lecture Notes in Computer Science, pp. 96 - 99.

Conference: 5th Intl. Work-conference on Ambient Assisted Living (IWAAL 2013)

Publication type: Short conference paper

VI Chaaraoui, A.A., Padilla-López, J.R., Flórez-Revuelta, F., 2013. Fusion of Skeletal and Silhouette-based Features for Human Action Recognition with RGB-D devices, in: 2013 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). To be presented in the 3rd Workshop on Consumer Depth Cameras for Computer Vision (CDC4CV13).

Conference: IEEE Intl. Conference on Computer Vision (ICCV 2013)

Publication type: Full workshop paper

Appendix C

Resumen

En este apéndice se detalla un resumen global de los resultados alcanzados en esta tesis, así como una discusión y conclusión sobre los mismos. Para mayor detalle, se ruega que se consulte la versión completa en inglés de este trabajo.

C.1 Introducción

C.1.1 Motivación

En la actualidad, se están experimentando enormes avances en las tecnologías de la información, comunicación y control (TICC). Esto se debe principalmente a la reducción de coste y tamaño, y al aumento simultáneo de capacidad computacional. Pero también ha sido influenciado por la implantación de Internet y la reciente extensión de las tecnologías móviles, que han permitido desarrollar una infinidad de servicios y aplicaciones. En concreto, en el campo de la inteligencia artificial se han podido observar éxitos considerables en las áreas de la minería de datos, el reconocimiento de patrones y la visión por computador. Estos avances están comenzando a implantarse también en el área de la automatización en el hogar y los entornos inteligentes. En estos son necesarias las TICC para posibilitar una interacción natural e intuitiva entre la persona y su entorno. Para ello, el entorno también debe ser capaz de detectar y comprender qué está sucediendo. El paradigma que surge de esta problemática se denomina inteligencia ambiental (AmI). La gran demanda de este tipo de tecnología se debe a su multitud de aplicaciones, entre las que no solo encontramos servicios de confort y entretenimiento, sino también de asistencia a las personas. En el campo de la vida asistida por

el entorno (*ambient-assisted living*, AAL), la inteligencia ambiental se aplica a la promoción y extensión de la vida independiente de las personas de edad avanzada o con diversidad funcional. Mediante AAL se pueden proporcionar una gran variedad de servicios de salud y seguridad en el hogar para aumentar la autonomía y el bienestar de las personas, como, por ejemplo, la supervisión de la medicación y la monitorización inteligente (Kleinberger et al., 2007; Sun et al., 2009). Existe una gran demanda de este tipo de servicios debido al actual envejecimiento de la población. La Oficina Estadística de la Comisión Europea prevé que para 2060 la relación entre población trabajadora y jubilada habrá pasado de cuatro a dos trabajadores por jubilado. Teniendo en cuenta que los miembros de la UE gastan aproximadamente un cuarto de su PIB en protección social, preocupa que estos logros se puedan mantener en el actual contexto económico y demográfico (EC, 2012). Por este motivo, es clave poder ayudar a las personas a mantenerse activas e independientes con el fin de combatir el envejecimiento demográfico.

Mediante el uso de las TICC es posible dar soporte a los servicios AAL y dotarlos de la infraestructura necesaria. En concreto, los sensores proporcionan conocimiento sobre el entorno en forma de datos de distinta naturaleza (binarios, medidas medioambientales, detección del movimiento, datos de audio y vídeo...). Mediante el aprendizaje automático se pueden detectar patrones en estos datos que permiten inferir la actividad humana. Este es el objeto de estudio del análisis del comportamiento humano (*human behaviour analysis*, HBA). Entre otros, los datos de tipo visual son muy valiosos para el HBA, ya que aportan una información sensorial amplia y su obtención resulta menos invasiva que el uso de sensores corporales (Nakashima et al., 2009).

Por este motivo, este campo ha adquirido financiación pública recientemente para una gran cantidad de proyectos de investigación. Esto se ha visto reflejado en el *AAL Joint Programme*, dotado de un presupuesto de 600 millones de euros para los años comprendidos entre 2008 y 2013. Debido a su éxito, ya se ha propuesto una continuación en el Programa Marco de la UE 'Horizonte 2020' (EC, 2013). En concreto, el proyecto de investigación TALISMAN+ — SisTemA inteLIgente para Seguimiento y proMoción de la autonomíA persoNal — está financiado por el Minsterio de Ciencia e Innovación (TIN2010-20510-C04-02). Este proyecto pretende dar soporte a servicios AAL mediante avances centrados en las siguientes áreas: 1) sensorización, 2) monitorización, 3) razonamiento, 4) orquestación de servicios y 5) seguridad. Su principal obje-

tivo es ofrecer la base científica y la infraestructura tecnológica necesaria para dar soporte a la autonomía personal de las personas de edad avanzada o con diversidad funcional, promoviendo el desarrollo de la industria española en este campo. El grupo de Domótica y Ambientes Inteligentes de la Universidad de Alicante es responsable del subproyecto vision@home, cuyo objetivo es desarrollar un sistema de monitorización y reconocimiento de la actividad de las personas en su hogar utilizando tecnología de visión, teniendo en cuenta cuestiones de privacidad. Por tanto, el presente trabajo se desarrolla en el marco de este proyecto.

Esta tesis está dirigida a prestar apoyo a escenarios de vida asistida por el entorno aportando avances en el análisis del comportamiento. La visión se considera la principal fuente de información para analizar y comprender la actividad de las personas en el hogar. Por tanto, el objetivo es dotar a los entornos inteligentes de la capacidad de detectar y comprender la actividad de las personas mediante técnicas basadas en visión por computador.

C.1.2 Objetivos de la tesis

El trabajo que se presenta en esta tesis contempla los siguientes objetivos:

(a) Estudio de métodos de comprensión de la actividad de las personas basados en visión, con el objetivo de clasificar los trabajos existentes y establecer un marco teórico, así como de identificar el objeto especifico de estudio que debe ser abordado para dar soporte al análisis del comportamiento.

(b) Propuesta de un método para la monitorización de la actividad de las personas en espacios interiores que sea capaz de inferir el conocimiento a partir de la observación visual. Se debe contemplar tanto la detección de eventos relevantes como la recopilación de datos estadísticos sobre el comportamiento humano (por ejemplo, para la monitorización de la salud o la detección de anomalías).

(c) Satisfacer requisitos específicos de los servicios AAL, como la ejecución continua en tiempo real y el manejo de varias cámaras.

(d) Alcanzar y verificar la robustez necesaria para reconocer actividades humanas en una gran variedad de circunstancias, considerando las diferencias entre los sujetos, las condiciones de los sujetos y los distintos entornos.

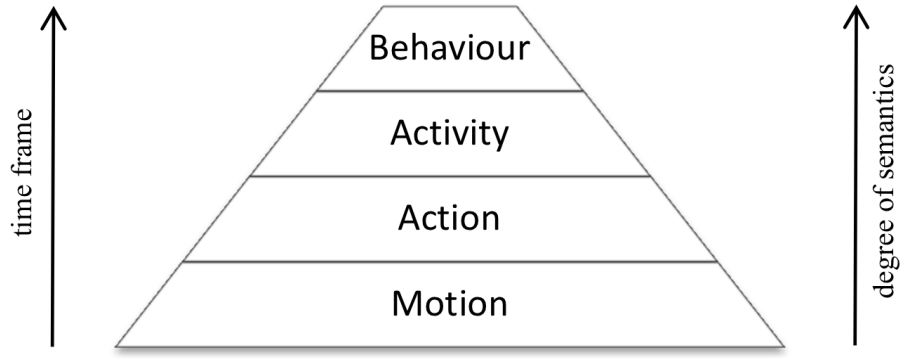


Figura A.1: Clasificación de los niveles de análisis del comportamiento humano.

C.1.3 Propuesta de solución

Antes de poder diseñar una propuesta de solución para el objetivo de este trabajo, se ha de revisar el trabajo existente en este campo. Esta tarea se ha abordado en el capítulo 2 de este documento, donde se ha establecido el marco teórico de la presente tesis doctoral. En concreto, se ha definido una taxonomía que establece los niveles de análisis del comportamiento de las personas (véase la figura A.1). Como se puede observar, se detallan cuatro niveles que abordan el análisis de menor a mayor complejidad temporal y semántica: movimiento, acciones, actividades y comportamientos. Mientras que en el análisis del movimiento se detectan la postura y la dirección de la mirada, en el nivel de acciones cabe asignar un significado a una secuencia de movimientos comprendidos en un intervalo temporal de unos segundos. Por tanto, se pueden reconocer acciones como caminar, sentarse, o caerse al suelo. En cuanto a las actividades, estas comprenden una mayor complejidad temporal y semántica, y pueden contener interacciones con otras personas u objetos. En un intervalo temporal de unos segundos a unos minutos, se pueden reconocer actividades cotidianas como cocinar, ducharse o hacer la cama. Estas actividades pueden definirse como secuencias de acciones que tienen como objetivo o resultado una actividad concreta. Finalmente, el nivel de comportamientos comprende entre días y semanas de análisis y permite aprender rutinas y hábitos, así como detectar desviaciones en estos.

Una de las conclusiones extraídas del estudio realizado es que el nivel que mayor interés científico e industrial está recibiendo es el reconocimiento de acciones. Existe una multitud de aplicaciones, desde la interacción persona-com-

putador y los videojuegos hasta la videovigilancia. El reconocimiento de acciones se aborda mediante el modelado del movimiento, clasificando el significado semántico de una secuencia de posturas. Sin embargo, en niveles superiores, como el reconocimiento de actividades más complejas y el estudio de comportamientos, se utilizan directamente características de bajo nivel; de este modo, no se abordan los niveles inferiores previamente, lo cual resulta en una fuerte limitación a problemas y entornos muy concretos. Esto nos lleva a proponer una solución basada en el reconocimiento de acciones, con el objetivo de desarrollar un método lo suficientemente fiable y eficiente para poder facilitar el análisis de niveles superiores del comportamiento humano a largo plazo.

También se ha podido observar que la inmensa mayoría de propuestas se basan en el uso de una única cámara y no suelen contemplar restricciones de ejecución en tiempo real. Además, los métodos existentes parecen funcionar bien únicamente para el reconocimiento de un conjunto reducido y específico de acciones. En este sentido, pretendemos proponer los avances necesarios para mejorar la estabilidad del reconocimiento de acciones. Asimismo, dado que hoy en día muchos de los escenarios de aplicación disponen de más de una cámara cubriendo el mismo campo de visión, se propone combinar la información proporcionada por todas las cámaras para evitar posibles oclusiones o ángulos ambiguos y mejorar el reconocimiento final. En concreto, se establecen los siguientes requisitos para abordar el reconocimiento de acciones de múltiples vistas teniendo en cuenta escenarios AAL:

Reconocimiento de una gran variedad de acciones El método propuesto debe ser capaz de reconocer acciones basadas en movimiento, como andar, correr, caerse, agacharse, levantarse, sentarse, etc. Por otra parte, se debe aportar una técnica para configurar el sistema al reconocimiento de un conjunto especifico de acciones en función de las necesidades de la aplicación.

Soporte de requisitos de escenario Dado que la forma en la que se realizan las acciones depende de los sujetos y su estado, así como del entorno, el método deberá ser estable ante una gran variedad de escenarios y sujetos de diferente edad y género. Dichas condiciones deben ser aprendidas para que el sistema sea capaz de mejorar si estas son conocidas y estables.

Aprendizaje de múltiples vistas El sistema debe aprender de múltiples vistas si estas están disponibles. Por tanto, se debe soportar el reconocimiento con una y varias cámaras, y mejorar el resultado de la clasificación conforme se van añadiendo vistas.

Ejecución en tiempo real Dado que el sistema se instalará en hogares para la monitorización en tiempo real, el diseño del método deberá contemplar que la suma de todos las fases de procesamiento pueda ejecutarse por encima de la frecuencia de vídeo (entre 25 y 30 fps), de modo que el análisis no suponga un retraso significativo y los eventos puedan detectarse en el momento.

Reconocimiento continuo Aunque comúnmente se aborda el reconocimiento de acciones por secuencias de vídeo previamente segmentadas, en la monitorización inteligente en el hogar se tiene que realizar un reconocimiento continuo. Por tanto, el método debe soportar la detección de acciones en flujos continuos de vídeo.

Aprendizaje adaptativo En algunas situaciones, el proceso de aprendizaje debe continuar cuando el sistema se encuentra en ejecución. En estos casos es preciso un método dinámico. Mediante el aprendizaje adaptativo e incremental se pueden tener en cuenta los cambios relacionados con el escenario, los sujetos o las acciones a reconocer.

Finalmente, cabe tener en cuenta que esta propuesta forma parte del proyecto de investigación TALISMAN+ anteriormente mencionado. En este proyecto se está desarrollando un sistema de monitorización de las actividades cotidianas de las personas en el hogar. En la figura A.2 se muestra la arquitectura de este sistema. Este se basa en una configuración de múltiples cámaras instaladas en el hogar para aplicar un análisis del comportamiento basado en información visual. En niveles más altos de procesamiento, esta información se combina con sensores del entorno. El sistema de razonamiento es el encargado de detectar si los datos reconocidos corresponden a algún evento monitorizado y actuar en consecuencia. Además, también se tienen en cuenta cuestiones éticas de privacidad mediante el procesamiento posterior de la imagen y la gestión de permisos, entre otros. De esta forma, se puede informar a un cuidador compartiendo la información visual o textual adecuada, dependiendo de los permisos del observador y de la gravedad del evento detectado.

Figura A.2: Arquitectura del sistema de monitorización inteligente propuesto para promover la vida independiente en el hogar y dar soporte a servicios AAL. Se indican en azul los bloques lógicos que pertenecen a esta tesis.

C.2 Reconocimiento de acciones

C.2.1 Método propuesto

En esta sección se va a presentar un resumen del método propuesto para el reconocimiento de acciones en secuencias de vídeo. Para mayor detalle, se hace referencia a los capítulos 4 y 5 de este documento.

Características visuales

Con el objetivo de extraer un vector característico de la información que contiene una imagen con respecto a la postura de la persona, se propone utilizar una representación global. En concreto, se utiliza únicamente el borde de la silueta de la persona para codificar su forma. Estas siluetas pueden obtenerse mediante la extracción del fondo en imágenes RGB convencionales, pero también mediante la segmentación basada en profundidad si se dispone de datos RGB-D, como los que ofrece el dispositivo Microsoft Kinect. En concreto, se propone una característica visual nueva cuyas principales ventajas son su dimensionalidad y coste computacional reducidos. Utilizando un esquema radial, se obtiene para cada sector circular un valor resumen que representa los puntos del contorno que contiene. De este modo, se obtiene una alineación espacial. Entre las diferentes propuestas, el rango estadístico de las distancias entre los puntos de cada sector y el centroide es el que mejor resultado ha dado. En la figura A.3 se muestra el esquema utilizado. Esta característica se obtiene para cada fotograma de las secuencias de vídeo.

Aprendizaje

Una vez que se ha extraído la característica visual correspondiente, se propone realizar un aprendizaje basado en secuencias de posturas clave. En concreto, se define el modelo *bag of key poses*. En este se aprenden las posturas clave (*key poses*) más representativas de cada tipo de acción mediante un algoritmo de agrupamiento. Posteriormente, se aprende la relación temporal entre las posturas clave modelando las secuencias de posturas clave correspondientes a las secuencias de aprendizaje. De esta forma, se obtiene un conjunto de secuencias de posturas clave que pueden ser utilizadas en el reconocimiento.

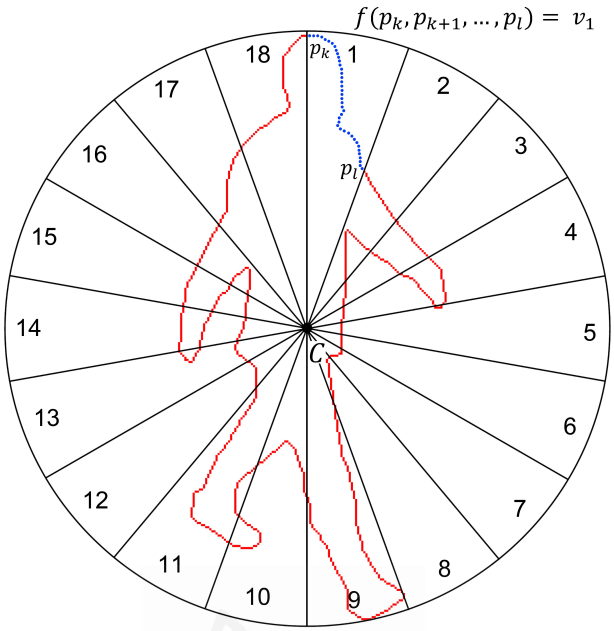


Figura A.3: Esquema utilizado para la extracción de características: En primer lugar, se asignan los puntos del contorno a sus correspondientes sectores; después, para cada sector se obtiene un valor resumen (ejemplo con 18 sectores).

Para poder aprender de múltiples vistas se utiliza una fusión de características. Dado que, dependiendo del tipo de acción, algunos ángulos pueden aportar mayor información que otros, se obtiene un peso concreto para cada clase y cada ángulo. De esta forma se utiliza una fusión inteligente de características en la que se asigna un valor específico a cada uno de los componentes.

Reconocimiento

En el reconocimiento se realiza inicialmente el mismo procesamiento. La secuencia de vídeo se convierte en una secuencia de posturas clave, y dicha secuencia se empareja con las secuencias del conjunto de entrenamiento. Para ello se utiliza el algoritmo *dynamic time warping* (DTW) que permite el alineamiento temporal de secuencias. Esto es importante, ya que personas de diferente edad y estado pueden realizar las acciones a ritmos muy dispares. En este emparejamiento se tiene en cuenta el esquema de pesos previamente aprendido para fusionar todas las vistas disponibles en el reconocimiento. Además, se permite que la configuración de cámaras no coincida entre aprendizaje y reconocimiento.

En el diagrama de la figura A.4 se muestran las distintas fases de procesamiento implicadas en el método de clasificación.

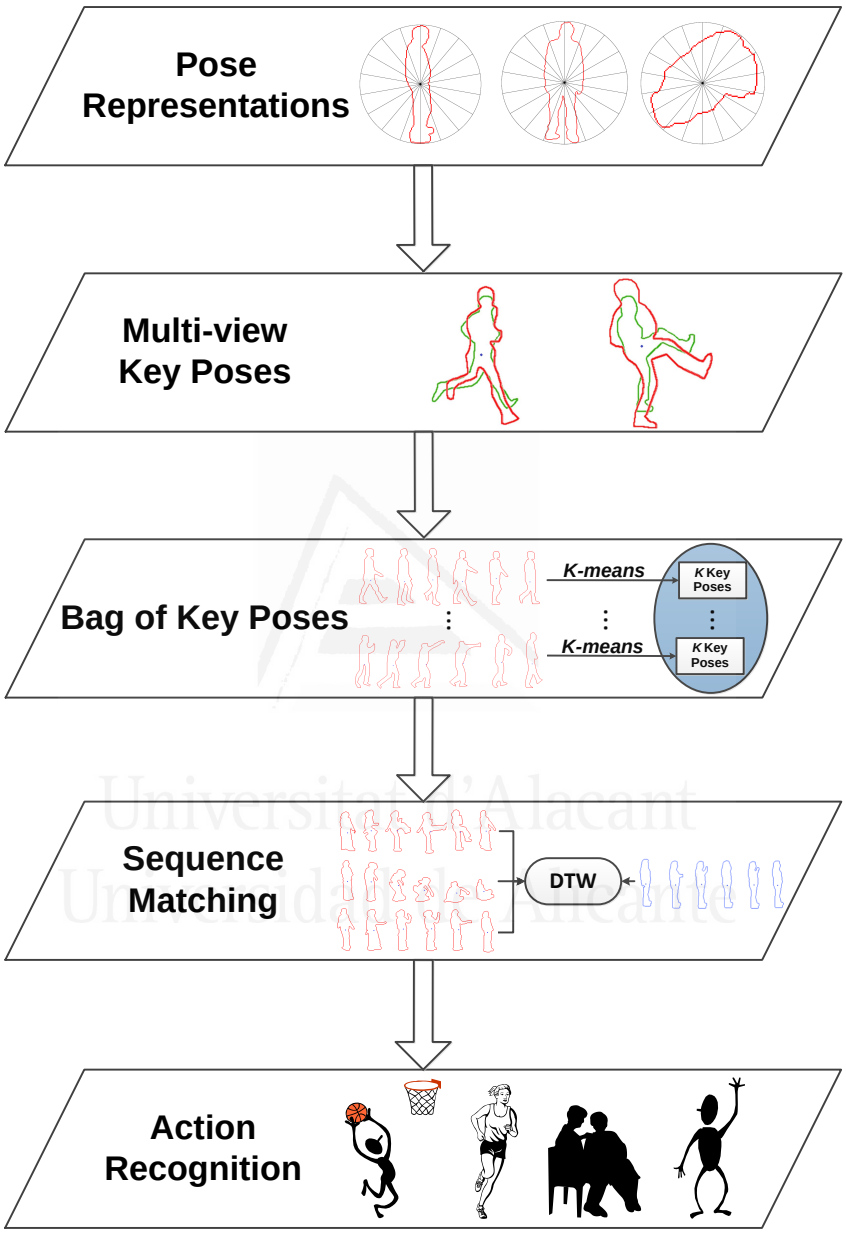


Figura A.4: Visión esquemática de las fases de procesamiento y las técnicas aplicadas en el método de clasificación propuesto.

C.2.2 Optimizaciones y extensiones

Además del método de reconocimiento de acciones, se proponen varias técnicas de optimización y extensiones. En estas se contempla la posibilidad de ajustar la configuración del método a un escenario concreto para mejorar su rendimiento.

Selección de instancias, características y parámetros

Una vez que el sistema se encuentra instalado en un entorno específico, se conocen los sujetos que van a ser monitorizados y cómo realizan las acciones. Por tanto, existe un enorme potencial de mejora relacionado con la optimización del conjunto de aprendizaje y la parametrización del método. Este concepto se estudia en las secciones 6.1 y 6.2, donde se propone el uso de algoritmos evolutivos para la búsqueda guiada del conjunto óptimo de instancias, características y parámetros del método. Los métodos propuestos, entre los que se encuentran una selección evolutiva de características y una selección coevolutiva de instancias, características y parámetros, permiten obtener mejoras muy significativas, sobrepasando ampliamente las tasas obtenidas hasta el momento en el estado de la cuestión.

Reconocimiento adaptativo

También se propone un método de aprendizaje incremental y adaptativo que habilita el aprendizaje en tiempo de ejecución. De este modo, se da soporte a escenarios de aplicación en los que, en tiempo de ejecución, pueden verse alteradas las acciones a reconocer o las características del entorno y de los sujetos. El método diseñado es capaz de aprender nuevos datos de aprendizaje de forma incremental, sin tener que iniciar un aprendizaje nuevo, a la vez que se adapta a los nuevos datos para mejorar el reconocimiento del conjunto de acciones.

Reconocimiento continuo

Dado que en nuestra aplicación AAL cabe procesar uno o varios flujos continuos de vídeo para detectar acciones, se ha extendido el método propuesto para soportar el reconocimiento continuo. Para ello, se ha presentado el concepto de *action zones*. Estas se definen como las partes de una acción que resultan más discriminativas en comparación con otras clases. La detección de estas zonas permite reducir

significativamente la longitud de las secuencias de aprendizaje, excluyendo las partes que no se diferencian lo suficiente entre acciones.

En el reconocimiento, estas *action zones* son emparejadas con el flujo continuo de vídeo mediante una técnica basada en una ventana deslizante y creciente. De esta forma se detectan y reconocen las acciones analizando el vídeo en diferentes posiciones y escalas temporales. Además, se propone una metodología de evaluación apropiada para el reconocimiento de acciones en aplicaciones AAL. Basándonos en el análisis por segmentos, se define el uso de la clase nula para evitar falsos positivos. Esta clase corresponde a todos los tipos de comportamiento que pueden darse y que no han sido aprendidos por el sistema.

C.2.3 Resultados

En este apartado se exponen los resultados alcanzados durante las distintas fases de experimentación de este trabajo. Por tanto, se incluyen tanto las tasas de reconocimiento obtenidas con el método propuesto de reconocimiento de acciones, como las correspondiente a las técnicas de optimización aplicadas. Esta experimentación se ha realizado sobre los siguientes bancos de datos públicos: Weizmann (Gorelick et al., 2007), MuHAVi (Singh et al., 2010), IXMAS (Weinland et al., 2006) y DHA (Lin et al., 2012). Entre estos se encuentran dos conjuntos de datos con múltiples vistas de la escena (MuHAVi con dos vistas laterales e IXMAS con cuatro vistas laterales y una superior). En el caso del DHA, se ha realizado la grabación mediante sensores RGB-D, que aportan datos de profundidad para cada píxel de la imagen. Estos permiten obtener una segmentación de la silueta de alta calidad. Además, se ha realizado la grabación de un banco de datos propio denominado DAI RGBD con el fin de obtener datos de múltiples vistas de tipo RGB-D. Mediante estos bancos de datos se ha podido verificar la estabilidad del método ante una gran variedad de acciones, sujetos y escenarios. Las características concretas de cada uno de ellos se encuentran detalladas en el apéndice A.

Utilizando estos datos, se han realizado diferentes pruebas de validación cruzada. En concreto, se han utilizado las pruebas de validación cruzada *leave one actor out* (LOAO), *leave one sequence out* (LOSO) y *leave one view out* (LOVO), en las que se utilizan subconjuntos basados en los datos de un actor, una secuencia o una cámara específica respectivamente. Sobre estos subconjuntos se realiza una prueba por cada uno de ellos. En el entrenamiento, el método utiliza todos

Tabla A.1: Comparación de tasas de reconocimiento y velocidades obtenidas con el banco de datos Weizmann.

Método	**Nº Acciones**	**Prueba**	**Tasa**	**fps**
İkizler and Duygulu (2007)	9	LOSO	100%	N/A
Tran and Sorokin (2008)	10	LOSO	100%	N/A
Fathi and Mori (2008)	10	LOSO	100%	N/A
Hernández et al. (2011)	10	LOAO	90,3%	98
Cheema et al. (2011)	9	LOSO	91,6%	56
Sadek et al. (2012)	10	LOAO	97,8%	18
Método propuesto	10	LOSO	93,5%	188
Método propuesto	10	LOAO	97,8%	188
Método optimizado	10	LOAO	100%	210

Tabla A.2: Comparación de tasas de reconocimiento y velocidades obtenidas con el banco de datos MuHAVi-14.

Approach	**LOSO**	**LOAO**	**LOVO**	**fps**
Singh et al. (2010)	82,4%	61,8%	42,6%	N/A
Eweiwi et al. (2011)	91,9%	77,9%	55,8%	N/A
Cheema et al. (2011)	86,0%	73,5%	50,0%	56
Método propuesto	**98,5%**	**94,1%**	**59,6%**	**99**
Método optimizado	**100%**	**100%**	-	-

los subconjuntos excepto uno, mientras que en la fase de prueba se utiliza el subconjunto desconocido para el sistema. Dado que esto se realiza para todos los subconjuntos existentes y se obtiene el resultado medio, se puede evitar un sobreajuste del aprendizaje.

Como se puede observar en las tablas A.1, A.2 y A.3, el método propuesto alcanza tasas de reconocimiento altas de forma estable en las diferentes pruebas. Si comparamos estas tasas con las obtenidas por métodos que también intentan abordar el reconocimiento en tiempo real, observamos que el método propuesto dispone de claras ventajas de rendimiento. En el caso del banco de datos MuHAVi, a nuestro leal saber y entender, las tasas obtenidas son las más altas publicadas hasta el momento. Cabe destacar la amplia mejora en la prueba LOAO, que demuestra la gran estabilidad del método ante las diferencias entre actores. Además, también se han obtenido las mejores tasas paras las pruebas LOVO, donde se realiza un entrenamiento con las grabaciones de una cámara y

Tabla A.3: Comparación de tasas de reconocimiento y velocidades obtenidas con el banco de datos MuHAVi-8.

Método	**LOSO**	**LOAO**	**LOVO**	**fps**
Singh et al. (2010)	97,8%	76,4%	50,0%	N/A
Martínez-Contreras et al. (2009)	98,4%	-	-	N/A
Eweiwi et al. (2011)	98,5%	85,3%	38,2%	N/A
Cheema et al. (2011)	95,6%	83,1%	57,4%	56
Método propuesto	**100%**	**100%**	**82,4%**	**98**

Tabla A.4: Comparación de tasas de reconocimiento y velocidades obtenidas con el banco de datos IXMAS en la prueba LOAO.

Método	**Nº Acciones**	**Nº Actores**	**Nº Vistas**	**Tasa**	**fps**
Yan et al. (2008)	11	12	4	78%	N/A
Wu et al. (2011)	12	12	4	89,4%	N/A
Cilla et al. (2012)	11	12	5	91,3%	N/A
Weinland et al. (2006)	11	10	5	93,3%	N/A
Cilla et al. (2013)	11	10	5	94,0%	N/A
Holte et al. (2012)	13	12	5	100%	N/A
Cherla et al. (2008)	13	N/A	4	80,1%	20
Weinland et al. (2010)	11	10	5	83,5%	∼500
Método propuesto	11	12	5	91,4%	207

se realiza la prueba con las grabaciones de otra distinta situada a un ángulo de 45°. Las optimizaciones propuestas suponen una mejora sustancial de las tasas de reconocimiento.

En el caso del banco de datos IXMAS (véase la tabla A.4), se observa que en el estado de la cuestión varía la configuración con la que se ha utilizado. En concreto, se muestran trabajos que han abordado el aprendizaje de múltiples vistas. Aunque en Holte et al. (2012) se ha conseguido obtener un reconocimiento perfecto, observamos que las tasas de reconocimiento bajan de forma drástica si buscamos métodos que aseguren la ejecución en tiempo real. De este modo, destacan las tasas de clasificación y velocidad de ejecución obtenidas por nuestra propuesta.

Las tablas A.5 y A.6 muestran los resultados que se han obtenido con los datos de tipo RGB-D. Mediante sensores de profundidad es posible obtener siluetas de gran calidad en espacios interiores, como las habitaciones de una vivienda. Además, el sistema utilizado permite detectar a la persona en la imagen y realizar

un seguimiento en tiempo real. Utilizando estos datos, de nuevo el método propuesto alcanza tasas de reconocimiento y de velocidad de ejecución excepcionales.

Finalmente, en la tabla A.7 se muestran los resultados obtenidos mediante el reconocimiento continuo propuesto. En el caso del banco de datos IXMAS, existe un etiquetado continuo de las secuencias. Sin embargo, para Weizmann comúnmente se concatenan las secuencias de cada actor, por lo que las secuencias resultantes contienen cortes. Se muestran las tasas de valor-F_1 con las que se evalúa la relación entre precisión y exhaustividad. Estas se han obtenido utilizando tanto las secuencias segmentadas originales como las *action zones*, y se puede observar que la detección automática de *action zones* aporta claramente una ventaja de rendimiento.

Por tanto, se puede concluir que mediante las contribuciones de esta tesis es posible obtener resultados estables en una amplia variedad de escenarios con diferentes características. Asimismo, el método es apropiado para el análisis del comportamiento en tiempo real, lo que nos lleva a considerar que esta propuesta es compatible con aplicaciones del mundo real y, en concreto, con los servicios AAL.

C.3 Observaciones finales

C.3.1 Discusión

En la presente tesis se han utilizado siluetas 2D para extraer características espaciales y aprender su evolución a lo largo del tiempo. Como se ha visto en el análisis del estado de la cuestión, este tipo de representaciones no es independiente del ángulo, lo cual limita el reconocimiento a ángulos específicos. Sin embargo, en los resultados experimentales se ha observado que la característica visual propuesta basada en el contorno de la silueta soporta rotaciones de hasta 45°. Además, dado que se emplea una configuración con múltiples vistas, el reconocimiento contempla varios ángulos para compensar esta desventaja.

Tabla A.5: Comparación de tasas de reconocimiento y velocidades obtenidas con el banco de datos propio DAI RGBD.

Método	**LOSO**	**LOAO**	**fps**
Método propuesto	**94,4%**	**100%**	80

Tabla A.6: Comparación de tasas de reconocimiento y velocidades obtenidas con el banco de datos DHA.

Método	**LOSO**	**fps**
Lin et al. (2012)	90,8%	N/A
Método propuesto	**95,2%**	99

Tabla A.7: Resultados obtenidos mediante el reconocimiento continuo y la evaluación por segmentos en la prueba LOAO. Se incluyen las tasas obtenidas con y sin el uso de las *action zones*.

Banco de datos	**Método**	**valor-F_1**
IXMAS	Secuencias segmentadas	0,504
IXMAS	Action zones	0,705
Weizmann	Secuencias segmentadas	0,693
Weizmann	Action zones	0,928

En cuanto a la obtención de este tipo de datos, la extracción de siluetas no es trivial. Siguen existiendo dificultades relacionadas con las sombras y los cambios de iluminación (Gallego and Pardàs, 2010), por lo que no siempre se pueden obtener siluetas de calidad. Por este motivo, se ha propuesto un método alternativo para la adquisición de datos. Mediante sensores RGB-D se pueden obtener siluetas de gran calidad gracias a la detección y el seguimiento de la persona en la imagen. Por otra parte, es posible obtener estos datos en la oscuridad, lo que permite que el análisis pueda realizarse también durante las actividades nocturnas, por ejemplo, para la monitorización del sueño.

El desarrollo de tecnologías de monitorización visual en el hogar implica cuestiones éticas de privacidad. Como se ha podido observar en el diagrama arquitectural del sistema de monitorización inteligente que se está desarrollando, además del análisis del comportamiento, se consideran técnicas de privacidad basadas en procesamientos posteriores de la imagen. Por ejemplo, es posible reconocer la postura de la persona en la imagen y sustituirla por un avatar, es decir, un modelo 3D del cuerpo humano. De esta manera se puede proteger la identidad y la apariencia de la persona. Utilizando este modelo, sigue siendo posible ejecutar el análisis del comportamiento. Asimismo, dado que la característica visual empleada se basa únicamente en el contorno de la silueta, es posible descartar los datos de color en una fase temprana de procesamiento.

Cabe destacar que el método propuesto de clasificación del comportamiento humano también ha sido utilizado con éxito para otros tipos de aplicaciones, como el reconocimiento de acciones relacionadas con los videojuegos (Chaaraoui et al., 2014; Climent-Pérez et al., 2013) donde se ha utilizado una representación en esqueleto 3D de la persona.

C.3.2 Conclusiones

En este trabajo se ha presentado un método no paramétrico basado en plantillas 2D para el reconocimiento de acciones. En las contribuciones aportadas, se han cumplido los requisitos planteados en la propuesta de solución mediante una técnica muy estable que ha demostrado ser versátil y eficiente. El método de reconocimiento de acciones basado en el modelo *bag of key poses* maneja de forma idónea el reconocimiento con una o varias cámaras. De este modo, se han alcanzado tasas de reconocimiento superiores a las del estado de la cuestión para varias de las pruebas realizadas.

Las contribuciones aportadas por esta tesis doctoral se puede resumir como sigue:

- Se ha realizado un estudio del estado de la cuestión del análisis del comportamiento humano basado en visión. En este se ha propuesto una taxonomía con el fin de unificar las definiciones existentes y establecer los diferentes niveles de análisis para clasificar los trabajos del área de investigación.

- Se ha presentado un método para el reconocimiento de acciones basado en el modelo *bag of key poses* y múltiples vistas, que permite el procesamiento en tiempo real. Asimismo, se ha presentado una característica visual nueva que utiliza el contorno de la silueta humana, y se han obtenido resultados elevados y estables en varios bancos de datos públicos, superando a otros métodos del estado de la técnica en la tasa de reconocimiento y la velocidad del procesamiento.

- Con el fin de aprender las restricciones específicas de los escenarios de aplicación, se han propuesto métodos de optimización. Utilizando algoritmos evolutivos se han definido técnicas para abordar la selección de instancias de aprendizaje, características y parámetros. Además, se ha presentado un

método adaptativo para manejar situaciones en las que el aprendizaje debe continuar durante el tiempo de ejecución. Los resultados obtenidos durante las respectivas fases de experimentación indican que se pueden obtener tasas de reconocimiento excepcionales mediante estas propuestas.

- El método también ha sido extendido para soportar el reconocimiento de flujos continuos de vídeo. Mediante la detección de *action zones* se aborda la detección y la clasificación de acciones utilizando una ventana deslizante.

- Las técnicas propuestas han sido diseñadas para cumplir con las exigencias de los servicios AAL. En consecuencia, el método concebido es parte de un sistema de monitorización inteligente que da soporte a servicios AAL con el fin de extender la autonomía personal de las personas en su hogar.

En conclusión, esta tesis doctoral aporta avances valiosos para el reconocimiento de acciones. Dada su eficiencia temporal y compatibilidad en relación a la adquisición de los datos y las configuración de las cámaras, este trabajo puede acortar las distancias entre el reconocimiento del movimiento y el reconocimiento de actividades y comportamientos complejos, facilitando el desarrollo de técnicas para el análisis del comportamiento humano a largo plazo.

C.3.3 Trabajo futuro

Las líneas de trabajo futuro que ha abierto esta investigación se han resumido en en la sección 8.3. Sin embargo, en general se distinguen dos lineas de trabajo futuro principales. Por un lado, es posible abordar el reconocimiento de actividades de mayor complejidad en base a las secuencias de acciones reconocidas, así como otros datos multimodales del entorno. Esto permitiría aumentar el valor semántico y el intervalo temporal del análisis. Por otro lado, el presente método está limitado al reconocimiento de acciones del cuerpo completo. Sin embargo, otras aplicaciones requieren, por ejemplo, el reconocimiento de gestos. En este sentido, es posible enriquecer la información espacial empleada mediante características locales, como puntos salientes en el espacio y/o en el tiempo o estimaciones 3D del cuerpo humano. En efecto, una propuesta preliminar que se beneficia de esta combinación ya ha sido validada con éxito (Chaaraoui et al., 2013).

List of Acronyms

AAL Ambient-assisted living

ADL Activities of daily living

BoW Bag of words

CEA Coevolutionary algorithm

CV Computer vision

DoS Degree of semantics

DTW Dynamic time warping

EA Evolutionary algorithm

HAR Human action recognition

HBA Human behaviour analysis

HBU Human behaviour understanding

HCI Human-computer interaction

HMM Hidden Markov model

LOAO Leave one actor out

LOSO Leave one sequence out

MEI Motion-energy image

MHI Motion-history image

SVM Support vector machine

List of Symbols

A	Number of action classes
B	Number of radial bins
C	Centroid $C = (x_c, y_c)$, where $x_c = \frac{\sum_{i=1}^{n} x_i}{n}, y_c = \frac{\sum_{i=1}^{n} y_i}{n}$
D	Distances $D = \{d_1, d_2, ..., d_n\}$, where $d_i = \|C - p_i\|$
H	Processed class evidence values
K	Number of key poses per class
M	Number of views
N	Number of training samples per class
P	Contour points $P = \{p_1, p_2, ..., p_n\}$, where $p_i = (x_i, y_i)$
$\bar{V}$	*Radial Summary* pose representation
kp	Nearest neighbour key pose
n	Number of contour points
r	Camera weights
w_{kp}	Discrimination value of the key pose kp
λ	Number of clustering attempts
τ	Allowed recognition delay

Bibliography

Agarwal, A. and B. Triggs (2004). 3D human pose from silhouettes by relevance vector regression. In *IEEE Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004*, Volume 2, pp. II–882 – II–888.

Aggarwal, J. and M. Ryoo (2011). Human activity analysis: A review. *ACM Computing Surveys 43*(3), 16:1 – 16:43.

Altun, K. and B. Barshan (2010). Human activity recognition using inertial/magnetic sensor units. In A. Salah, T. Gevers, N. Sebe, and A. Vinciarelli (Eds.), *Human Behavior Understanding*, Volume 6219 of *Lecture Notes in Computer Science*, pp. 38 – 51. Springer Berlin / Heidelberg.

Aminian, K., P. Robert, E. Buchser, B. Rutschmann, D. Hayoz, and M. Depairon (1999). Physical activity monitoring based on accelerometry: validation and comparison with video observation. *Medical and Biological Engineering and Computing 37*, 304 – 308.

Anderson, D., R. Luke, J. Keller, M. Skubic, M. Rantz, and M. Aud (2009). Modeling human activity from voxel person using fuzzy logic. *IEEE Transactions on Fuzzy Systems 17*(1), 39 –49.

Andriluka, M., S. Roth, and B. Schiele (2009). Pictorial structures revisited: People detection and articulated pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009*, pp. 1014 – 1021.

Ángeles Mendoza, M. and N. Pérez de la Blanca (2007). HMM-based action recognition using contour histograms. In J. Martí, J. Benedí, A. Mendonça, and J. Serrat (Eds.), *Pattern Recognition and Image Analysis*, Volume 4477 of *Lecture Notes in Computer Science*, pp. 394 – 401. Springer Berlin / Heidelberg.

Ashraf, N., Y. Shen, X. Cao, and H. Foroosh (2013). View invariant action recognition using weighted fundamental ratios. *Computer Vision and Image Understanding 117*(6), 587 – 602.

Avgerinakis, K., A. Briassouli, and I. Kompatsiaris (2013). Activity detection and recognition of daily living events. ACM Press. 1st ACM MM Workshop on Multimedia Indexing and Information Retrieval for Healthcare (MIIRH).

Bandouch, J., F. Engstler, and M. Beetz (2008). Accurate human motion capture using an ergonomics-based anthropometric human model. In F. Perales and R. Fisher (Eds.), *Articulated Motion and Deformable Objects*, Volume 5098 of *Lecture Notes in Computer Science*, pp. 248 – 258. Springer Berlin / Heidelberg.

Bao, L. and S. Intille (2004). Activity Recognition from User-Annotated Acceleration Data. In A. Ferscha and F. Mattern (Eds.), *Pervasive Computing*, Volume 3001 of *Lecture Notes in Computer Science*, pp. 1 – 17. Springer Berlin / Heidelberg.

Baysal, S., M. Kurt, and P. Duygulu (2010). Recognizing human actions using key poses. In *20th International Conference on Pattern Recognition, 2010. ICPR 2010*, pp. 1727 –1730.

Beetz, M., J. Bandouch, D. Jain, and M. Tenorth (2010). Towards automated models of activities of daily life. *Technology and Disability 4*, 1 – 11.

Belongie, S., J. Malik, and J. Puzicha (2000). Shape context: A new descriptor for shape matching and object recognition. In *Proceedings of the Neural Information Processing Systems 2000*, pp. 831 – 837.

Ben-Arie, J., Z. Wang, and P. Pandit (2002). Human activity recognition using multidimensional indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24*(8), 1091 – 1104.

Bhanu, B. and Y. Lin (2003). Genetic algorithm based feature selection for target detection in SAR images. *Image and Vision Computing 21*(7), 591 – 608. Computer Vision beyond the visible spectrum.

Bloom, V., V. Argyriou, and D. Makris (2013). Dynamic feature selection for online action recognition. In A. Salah, H. Hung, O. Aran, and H. Gunes (Eds.),

Human Behavior Understanding, Volume 8212 of *Lecture Notes in Computer Science*, pp. 64 – 76. Springer International Publishing.

Bobick, A. and J. Davis (2001). The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence 23*(3), 257 – 267.

Borges, P., N. Conci, and A. Cavallaro (2013). Video-based human behavior understanding: a survey. *IEEE Transactions on Circuits and Systems for Video Technology*. DOI 10.1109/TCSVT.2013.2270402.

Boulay, B., F. Bremond, and M. Thonnat (2006). Applying 3D human model in a posture recognition system. *Pattern Recognition Letters 27*(15), 1788 – 1796.

Boulgouris, N. V., K. N. Plataniotis, and D. Hatzinakos (2006). Gait recognition using linear time normalization. *Pattern Recognition 39*(5), 969 – 979.

Bourdev, L. (2011). *Poselets and their applications in high-level computer vision*. Ph. D. thesis, Berkeley, CA, USA.

Bourdev, L., S. Maji, T. Brox, and J. Malik (2010). Detecting people using mutually consistent poselet activations. In *Proceedings of the European Conference on Computer Vision: Part VI. ECCV 2010*, ECCV'10, pp. 168 – 181. Springer Berlin / Heidelberg.

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Brand, M. (1996). Coupled hidden Markov models for modeling interacting processes. *Neural Computation 405*(405), 1 – 28.

Brdiczka, O., J. Crowley, and P. Reignier (2009). Learning situation models in a smart home. *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics 39*(1), 56 –63.

Bregonzio, M., J. Li, S. Gong, and T. Xiang (2010). Discriminative topics modelling for action feature selection and recognition. In *Proceedings of the British Machine Vision Conference*, pp. 8.1 – 8.11. BMVA Press. DOI 10.5244/C.24.8.

Broadbent, E., R. Stafford, and B. MacDonald (2009). Acceptance of healthcare robots for the older population: Review and future directions. *International Journal of Social Robotics 1*(4), 319 – 330.

Broadbent, E., R. Tamagawa, N. Kerse, B. Knock, A. Patience, and B. MacDonald (2009). Retirement home staff and residents' preferences for healthcare robots. In *The 18th IEEE International Symposium on Robot and Human Interactive Communication, 2009. RO-MAN 2009*, pp. 645 – 650.

Cagnoni, S. (2008). Evolutionary computer vision: A taxonomic tutorial. In *Eighth International Conference on Hybrid Intelligent Systems, 2008. HIS '08*, pp. 1 – 6.

Cano, J., F. Herrera, and M. Lozano (2003). Using evolutionary algorithms as instance selection for data reduction in KDD: an experimental study. *IEEE Transactions on Evolutionary Computation 7*(6), 561 – 575.

Cano, J., F. Herrera, and M. Lozano (2005). A study on the combination of evolutionary algorithms and stratified strategies for training set selection in data mining. In F. Hoffmann, M. Köppen, F. Klawonn, and R. Roy (Eds.), *Soft Computing: Methodologies and Applications*, Volume 32 of *Advances in Soft Computing*, pp. 271 – 284. Springer Berlin / Heidelberg.

Cantú-Paz, E. (2004). Feature subset selection, class separability, and genetic algorithms. In K. Deb (Ed.), *Genetic and Evolutionary Computation - GECCO 2004*, Volume 3102 of *Lecture Notes in Computer Science*, pp. 959 – 970. Springer Berlin / Heidelberg.

Canton-Ferrer, C., J. Casas, and M. Pardas (2006). Human model and motion based 3D action recognition in multiple view scenarios. In *14th European Conference on Signal Processing*, Italy, pp. 1 – 5.

Canton-Ferrer, C., C. Segura, M. Pardas, J. Casas, and J. Hernando (2008). Multimodal real-time focus of attention estimation in smartrooms. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2008. CVPRW 2008*, pp. 1 – 8.

Cardinaux, F., S. Brownsell, M. Hawley, and D. Bradley (2008). Modelling of behavioural patterns for abnormality detection in the context of lifestyle reassurance. In J. Ruiz-Shulcloper and W. Kropatsch (Eds.), *Progress in Pattern Recognition, Image Analysis and Applications*, Volume 5197 of *Lecture Notes in Computer Science*, pp. 243 – 251. Springer Berlin / Heidelberg.

Carranza, J., C. Theobalt, M. A. Magnor, and H.-P. Seidel (2003). Free-viewpoint video of human actors. In *ACM SIGGRAPH 2003 Papers*, SIGGRAPH '03, New York, NY, USA, pp. 569 – 577. ACM.

Casado Yusta, S. (2009). Different metaheuristic strategies to solve the feature selection problem. *Pattern Recognition Letters 30*(5), 525 – 534.

Caspi, Y. and M. Irani (2002). Spatio-temporal alignment of sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24*(11), 1409 – 1424.

CAVIAR Project (2004). Caviar test case scenarios. http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1. Last access on 12/11/2013.

Chaaraoui, A. A., P. Climent-Pérez, and F. Flórez-Revuelta (2012a). An efficient approach for multi-view human action recognition based on bag-of-key-poses. In A. A. Salah, J. Ruiz-del Solar, C. Meriçli, and P.-Y. Oudeyer (Eds.), *Human Behavior Understanding*, Volume 7559 of *Lecture Notes in Computer Science*, pp. 29 – 40. Springer Berlin / Heidelberg.

Chaaraoui, A. A., P. Climent-Pérez, and F. Flórez-Revuelta (2012b). A review on vision techniques applied to human behaviour analysis for ambient-assisted living. *Expert Systems with Applications 39*(12), 10873 – 10888.

Chaaraoui, A. A., P. Climent-Pérez, and F. Flórez-Revuelta (2013). Silhouette-based human action recognition using sequences of key poses. *Pattern Recognition Letters 34*(15), 1799 – 1807. Smart Approaches for Human Action Recognition.

Chaaraoui, A. A. and F. Flórez-Revuelta (2013a). Human action recognition optimization based on evolutionary feature subset selection. In *Proceeding of the fifteenth annual conference on Genetic and evolutionary computation conference*, GECCO '13, New York, NY, USA, pp. 1229 – 1236. ACM.

Chaaraoui, A. A. and F. Flórez-Revuelta (2013b). Optimizing human action recognition based on a cooperative coevolutionary algorithm. *Engineering Applications of Artificial Intelligence*. Advances in Evolutionary Optimization Based Image Processing DOI 10.1016/j.engappai.2013.10.003.

Chaaraoui, A. A., J. R. Padilla-López, P. Climent-Pérez, and F. Flórez-Revuelta (2014). Evolutionary joint selection to improve human action recognition with RGB-D devices. *Expert Systems with Applications 41*(3), 786 – 794. Methods and Applications of Artificial and Computational Intelligence.

Chaaraoui, A. A., J. R. Padilla-López, F. J. Ferrández-Pastor, J. M. García-Chamizo, M. Nieto-Hidalgo, V. Romacho-Agud, and F. Flórez-Revuelta (2013). A vision system for intelligent monitoring of activities of daily living at home. In C. Nugent, A. Coronato, and J. Bravo (Eds.), *Ambient Assisted Living and Active Aging*, Volume 8277 of *Lecture Notes in Computer Science*, pp. 96 – 99. Springer International Publishing.

Chaaraoui, A. A., J. R. Padilla-López, and F. Flórez-Revuelta (2013). Fusion of skeletal and silhouette-based features for human action recognition with RGB-D devices. In *IEEE 14th International Conference on Computer Vision Workshops, 2013. ICCV Workshops 2013*. To be presented in 3rd Workshop on Consumer Depth Cameras for Computer Vision (CDC4CV13).

Chaquet, J. M., E. J. Carmona, and A. Fernández-Caballero (2013). A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding 117*(6), 633 – 659.

Cheema, S., A. Eweiwi, C. Thurau, and C. Bauckhage (2011). Action recognition by learning discriminative key poses. In *IEEE 13th International Conference on Computer Vision Workshops, 2011. ICCV Workshops 2011*, pp. 1302 –1309.

Chen, H.-S., H.-T. Chen, Y.-W. Chen, and S.-Y. Lee (2006). Human action recognition using star skeleton. In *Proceedings of the 4th ACM International Workshop on Video Surveillance and Sensor Networks*, VSSN '06, New York, NY, USA, pp. 171 – 178. ACM.

Chen, L., H. Wei, and J. Ferryman (2013). A survey of human motion analysis using depth imagery. *Pattern Recognition Letters 34*(15), 1995 – 2006. Smart Approaches for Human Action Recognition.

Chen, S. (2012). Kalman filter for robot vision: A survey. *IEEE Transactions on Industrial Electronics 59*(11), 4409 – 4420.

Cheng, H., C. Yang, F. Han, and H. Sawhney (2008). HO2: A new feature for multi-agent event detection and recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2008. CVPRW 2008*, pp. 1 – 8.

Cherla, S., K. Kulkarni, A. Kale, and V. Ramasubramanian (2008). Towards fast, view-invariant human action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2008. CVPRW 2008*, pp. 1 – 8.

Cheung, K., S. Baker, and T. Kanade (2003). Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *IEEE Conference on Computer Vision and Pattern Recognition, 2003. CVPR 2003*, Volume 1, pp. I–77 – I–84.

Chtioui, Y., D. Bertrand, and D. Barba (1998). Feature selection by a genetic algorithm. application to seed discrimination by artificial vision. *Journal of the Science of Food and Agriculture 76*(1), 77 – 86.

Chung, P.-C. and C.-D. Liu (2008). A daily behavior enabled hidden Markov model for human behavior understanding. *Pattern Recognition 41*(5), 1572 – 1580.

Cilla, R., M. A. Patricio, A. Berlanga, and J. M. Molina (2012). A probabilistic, discriminative and distributed system for the recognition of human actions from multiple views. *Neurocomputing 75*(1), 78 – 87. Brazilian Symposium on Neural Networks (SBRN 2010), International Conference on Hybrid Artificial Intelligence Systems (HAIS 2010).

Cilla, R., M. A. Patricio, A. Berlanga, and J. M. Molina (2013). Human action recognition with sparse classification and multiple-view learning. *Expert Systems*. DOI 10.1111/exsy.12040.

Cipolla, R. and A. Pentland (1998). *Computer Vision for Human-Machine Interaction*. Cambridge University Press.

Climent-Pérez, P., A. A. Chaaraoui, and F. Flórez-Revuelta (2012). Useful research tools for human behaviour understanding in the context of ambient assisted living. In P. Novais, K. Hallenborg, D. I. Tapia, and J. M. C. RodrÃguez (Eds.), *Ambient Intelligence - Software and Applications*, Volume 153 of *Advances in Intelligent and Soft Computing*, pp. 201 – 205. Springer Berlin / Heidelberg.

Climent-Pérez, P., A. A. Chaaraoui, J. R. Padilla-López, and F. Flórez-Revuelta (2013). Optimal joint selection for skeletal data from RGB-D devices using a genetic algorithm. In I. Batyrshin and M. Mendoza (Eds.), *Advances in Computational Intelligence*, Volume 7630 of *Lecture Notes in Computer Science*, pp. 163 – 174. Springer Berlin / Heidelberg.

Coello, C. A. C., G. B. Lamont, and D. A. V. Veldhuizen (2006). *Evolutionary Algorithms for Solving Multi-Objective Problems (Genetic and Evolutionary Computation)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Cristani, M., M. Farenzena, D. Bloisi, and V. Murino (2010). Background subtraction for automated multisensor surveillance: a comprehensive review. *EURASIP Journal on Advances in Signal Processing 2010*, 43:1 – 43:24.

Dalal, N. and B. Triggs (2005). Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005*, CVPR '05, Washington, DC, USA, pp. 886 – 893. IEEE Computer Society.

Dasarathy, B. V. (1994). *Decision fusion*, Volume 1994. IEEE Computer Society Press.

De Jong, K. A. (1975). *An Analysis of the behavior of a class of genetic adaptive systems*. Doctoral dissertation, University of Michigan.

De Jong, K. A. (2006). *Evolutionary computation - A unified approach*. MIT Press.

De la Torre Frade, F., J. K. Hodgins, A. W. Bargteil, X. Martin Artal, J. C. Macey, A. Collado I Castells, and J. Beltran (2008). Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) database. Technical Report CMU-RI-TR-08-22, Robotics Institute, Pittsburgh, PA.

Dedeoğlu, Y., B. Töreyin, U. Güdükbay, and A. Çetin (2006). Silhouette-based method for object classification and human action recognition in video. In T. Huang, N. Sebe, M. Lew, V. Pavlovic, M. Kölsch, A. Galata, and B. Kisacanin (Eds.), *Computer Vision in Human-Computer Interaction*, Volume 3979 of *Lecture Notes in Computer Science*, pp. 64 – 77. Springer Berlin / Heidelberg.

Dempster, A. P. (1968). A generalization of bayesian inference. *Journal of the Royal Statistical Society. Series B 30*, 205 – 247.

Derrac, J., S. García, and F. Herrera (2010). IFS-CoCo: Instance and feature selection based on cooperative coevolution with nearest neighbor rule. *Pattern Recognition 43*(6), 2082 – 2105.

Derrac, J., I. Triguero, S. García, and F. Herrera (2012). A co-evolutionary framework for nearest neighbor enhancement: Combining instance and feature weighting with instance selection. In E. Corchado, V. Snášel, A. Abraham, M. Woźniak, M. Graña, and S.-B. Cho (Eds.), *Hybrid Artificial Intelligent Systems*, Volume 7209 of *Lecture Notes in Computer Science*, pp. 176 – 187. Springer Berlin / Heidelberg.

Doshi, A. and M. Trivedi (2010a). Attention estimation by simultaneous observation of viewer and view. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2010. CVPRW 2010*, pp. 21 – 27. IEEE.

Doshi, A. and M. Trivedi (2010b). Examining the impact of driving style on the predictability and responsiveness of the driver: Real-world and simulator analysis. In *2010 IEEE Intelligent Vehicles Symposium (IV)*, pp. 232 – 237. IEEE.

Doucette, J., A. McIntyre, P. Lichodzijewski, and M. Heywood (2012). Symbiotic coevolutionary genetic programming: a benchmarking study under large attribute spaces. *Genetic Programming and Evolvable Machines 13*(1), 71 – 101.

Dubowsky, S., F. Genot, S. Godding, H. Kozono, A. Skwersky, H. Yu, and L. S. Yu (2000). Pamm - a robotic aid to the elderly for mobility assistance and monitoring: a "helping-hand" for the elderly. In *Proceedings of the IEEE International Conference on Robotics and Automation, 2000. ICRA '00*, Volume 1, pp. 570 – 576.

Duong, T., H. Bui, D. Phung, and S. Venkatesh (2005). Activity recognition and abnormality detection with the switching hidden semi-Markov model. In *IEEE Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005*, Volume 1, pp. 838 – 845.

EC (2012). Active ageing special eurobarometer 378. Technical report, DG COMM "Research and Speechwriting" Unit, European Comission. Conducted by TNS Opinion & Social at the request of Directorate-General for Employment, Social Affairs and Inclusion.

EC (2013). Tackling the demographic challenge to create growth and jobs - commission proposal for a renewed Ambient Assisted Living Joint Programme (AAL JP), Digital Agenda for Europe - European Commission. http://ec.europa.eu/information_society/newsroom/cf/dae/itemdetail.cfm?item_id=11393. Last access on 12/11/2013.

Eichner, M. and V. Ferrari (2009). Better appearance models for pictorial structures. In *Proceedings of the British Machine Vision Conference*, pp. 1 – 11.

Espejo, P., S. Ventura, and F. Herrera (2010). A survey on the application of genetic programming to classification. *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews) 40*(2), 121 – 144.

Eweiwi, A., S. Cheema, C. Thurau, and C. Bauckhage (2011). Temporal key poses for human action recognition. In *IEEE 13th International Conference on Computer Vision Workshops, 2011. ICCV Workshops 2011*, pp. 1310 –1317.

Fahlman, S. (2010). The scone knowledge-base project. http://www.cs.cmu.edu/~sef/scone/. Last access on 12/11/2013.

Fathi, A. and G. Mori (2008). Action recognition by learning mid-level motion features. In *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008*, pp. 1 – 8.

Felzenszwalb, P., R. Girshick, D. McAllester, and D. Ramanan (2010). Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32*(9), 1627 – 1645.

Ferrari, V., M. Marin-Jimenez, and A. Zisserman (2008). Progressive search space reduction for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008*, pp. 1 – 8.

Fong, T., I. Nourbakhsh, and K. Dautenhahn (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems 42*(3–4), 143 – 166.

Gallego, J. and M. Pardàs (2010). Enhanced bayesian foreground segmentation using brightness and color distortion region-based model for shadow removal. In *IEEE 17th International Conference on Image Processing, 2010. ICIP 2010*, pp. 3449 – 3452.

García, S., J. R. Cano, and F. Herrera (2008). A memetic algorithm for evolutionary prototype selection: A scaling up approach. *Pattern Recognition 41*(8), 2693 – 2709.

Garcia, S., J. Derrac, J. Cano, and F. Herrera (2012). Prototype selection for nearest neighbor classification: Taxonomy and empirical study. *IEEE Transactions on Pattern Analysis and Machine Intelligence 34*(3), 417 – 435.

García-Pedrajas, N., J. A. Romero del Castillo, and D. Ortiz-Boyer (2010). A cooperative coevolutionary algorithm for instance selection for instance-based learning. *Machine Learning 78*, 381 – 420.

Gavrila, D. (1999). The visual analysis of human movement: A survey. *Computer Vision and Image Understanding 73*(1), 82 – 98.

Giese, M. A. and T. Poggio (2003). Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience 4*(3), 179 – 192.

Gonnet, G. H., M. A. Cohen, and S. A. Benner (1992). Exhaustive matching of the entire protein sequence database. *Science 256*(5062), 1443 – 1445.

Goodrich, M. A. and A. C. Schultz (2007). Human-robot interaction: a survey. *Foundations and Trends in Human-Computer Interaction 1*(3), 203 – 275.

Gorelick, L., M. Blank, E. Shechtman, M. Irani, and R. Basri (2007). Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence 29*(12), 2247 – 2253.

Grochowski, M. and N. Jankowski (2004). Comparison of Instance Selection Algorithms II. Results and Comments. In L. Rutkowski, J. Siekmann, R. Tadeusiewicz, and L. Zadeh (Eds.), *Artificial Intelligence and Soft Computing - ICAISC 2004*, Volume 3070 of *Lecture Notes in Computer Science*, pp. 580 – 585. Springer Berlin / Heidelberg.

Guo, P., Z. Miao, Y. Shen, W. Xu, and D. Zhang (2012). Continuous human action recognition in real time. *Multimedia Tools and Applications*, 1 – 18.

Guyon, I. and A. Elisseeff (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research 3*, 1157 – 1182.

Han, J., L. Shao, D. Xu, and J. Shotton (2013). Enhanced computer vision with microsoft kinect sensor: A review. *IEEE Transactions on Cybernetics*. DOI 10.1109/TCYB.2013.2265378.

Hara, K., T. Omori, and R. Ueno (2002). Detection of unusual human behavior in intelligent house. In *Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing, 2002*, pp. 697 – 706.

Hayes, T., M. Pavel, and J. Kaye (2008). An approach for deriving continuous health assessment indicators from in-home sensor data. In *Technology and Aging: Selected Papers from the 2007 International Conference on Technology and Aging*, Volume 21, pp. 130 – 137. IOS Press.

Hernández, J., A. Montemayor, J. José Pantrigo, and A. Sánchez (2011). Human action recognition based on tracking features. In J. Ferrández, J. Álvarez Sánchez, F. de la Paz, and F. Toledo (Eds.), *Foundations on Natural and Artificial Computation*, Volume 6686 of *Lecture Notes in Computer Science*, pp. 471 – 480. Springer Berlin / Heidelberg.

Holte, M., B. Chakraborty, J. Gonzalez, and T. Moeslund (2012). A local 3-D motion descriptor for multi-view human action recognition from 4-D spatio-temporal interest points. *IEEE Journal of Selected Topics in Signal Processing 6*(5), 553 – 565.

Holte, M., C. Tran, M. Trivedi, and T. Moeslund (2012). Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments. *IEEE Journal of Selected Topics in Signal Processing 6*(5), 538 – 552.

Holte, M. B., C. Tran, M. M. Trivedi, and T. B. Moeslund (2011). Human action recognition using multiple views: a comparative perspective on recent developments. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, J-HGBU '11, New York, NY, USA, pp. 47 – 52. ACM.

Hong, X., C. Nugent, M. Mulvenna, S. McClean, B. Scotney, and S. Devlin (2009). Evidential fusion of sensor data for activity recognition in smart homes. *Pervasive and Mobile Computing 5*(3), 236 – 252.

Hongeng, S. and R. Nevatia (2001). Multi-agent event recognition. In *IEEE Eighth International Conference on Computer Vision, 2001. ICCV 2001*, Volume 2, pp. 84 – 91. IEEE.

Hongeng, S. and R. Nevatia (2003). Large-scale event detection using semi-hidden Markov models. In *IEEE Ninth International Conference on Computer Vision, 2003. ICCV 2003*, Volume 2, pp. 1455 – 1462.

Hongeng, S., R. Nevatia, and F. Bremond (2004). Video-based event recognition: activity representation and probabilistic recognition methods. *Computer Vision and Image Understanding 96*(2), 129 – 162. Special Issue on Event Detection in Video.

Horprasert, T., D. Harwood, and L. Davis (1999). A statistical approach for real-time robust background subtraction and shadow detection. In *IEEE Seventh International Conference on Computer Vision Workshops, 1999. ICCV Workshops 1999*, pp. 256 – 261.

Howe, N. (2004). Silhouette lookup for automatic pose tracking. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2004. CVPRW 2004*, pp. 15 – 22.

Hsieh, C., P. Huang, and M. Tang (2011). Human action recognition using silhouette histogram. In *Proceedings of the Thirty-Fourth Australasian Computer Science Conference (ACSC 2011)*, pp. 11 – 15.

Htike, Z., S. Egerton, and K. Chow (2011). Model-free viewpoint invariant human activity recognition. *Lecture Notes in Engineering and Computer Science 2188*(1), 154 – 158.

Hu, W., T. Tan, L. Wang, and S. Maybank (2004). A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews) 34*(3), 334 – 352.

Hu, Y., L. Cao, F. Lv, S. Yan, Y. Gong, and T. Huang (2009). Action detection in complex scenes with spatial and temporal ambiguities. In *IEEE 12th*

International Conference on Computer Vision, 2009. ICCV 2009, pp. 128 – 135.

İkizler, N., R. Cinbis, and P. Duygulu (2008). Human action recognition with line and flow histograms. In *19th International Conference on Pattern Recognition, 2008. ICPR 2008*, pp. 1 – 4.

İkizler, N. and P. Duygulu (2007). Human action recognition using distribution of oriented rectangular patches. In A. Elgammal, B. Rosenhahn, and R. Klette (Eds.), *Human Motion - Understanding, Modeling, Capture and Animation*, Volume 4814 of *Lecture Notes in Computer Science*, pp. 271 – 284. Springer Berlin / Heidelberg.

Intille, S., K. Larson, E. Tapia, J. Beaudin, P. Kaushik, J. Nawyn, and R. Rockinson (2006). Using a live-in laboratory for ubiquitous computing research. In K. Fishkin, B. Schiele, P. Nixon, and A. Quigley (Eds.), *Pervasive Computing*, Volume 3968 of *Lecture Notes in Computer Science*, pp. 349 – 365. Springer Berlin / Heidelberg.

Iosifidis, A., A. Tefas, and I. Pitas (2012). Neural representation and learning for multi-view human action recognition. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pp. 1 – 6.

Jaimes, A. and N. Sebe (2007). Multimodal human-computer interaction: A survey. *Computer Vision and Image Understanding 108*(1-2), 116 – 134.

Jain, G., D. Cook, and V. Jakkula (2006). Monitoring health by detecting drifts and outliers for a smart environment inhabitant. In *Proceedings of the International Conference On Smart Homes and Health Telematics*, pp. 1 – 8.

Jain, M., H. Jégou, and P. Bouthemy (2013). Better exploiting motion for better action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, 2013. CVPR 2013*, Portland, USA.

Jankowski, N. and M. Grochowski (2004). Comparison of instances seletion algorithms I. Algorithms survey. In L. Rutkowski, J. Siekmann, R. Tadeusiewicz, and L. Zadeh (Eds.), *Artificial Intelligence and Soft Computing - ICAISC 2004*, Volume 3070 of *Lecture Notes in Computer Science*, pp. 598 – 603. Springer Berlin / Heidelberg.

Jhuang, H., T. Serre, L. Wolf, and T. Poggio (2007). A biologically inspired system for action recognition. In *IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007*, pp. 1 – 8.

Ji, X., H. Liu, Y. Li, and D. Brown (2008). Visual-based view-invariant human motion analysis: A review. In I. Lovrek, R. Howlett, and L. Jain (Eds.), *Knowledge-Based Intelligent Information and Engineering Systems*, Volume 5177 of *Lecture Notes in Computer Science*, pp. 741 – 748. Springer Berlin / Heidelberg.

Ji, X., S. S. Member, and H. Liu (2010). Advances in view-invariant human motion analysis: A review. *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews) 40*(1), 13 – 24.

John, G., R. Kohavi, and K. Pfleger (1994). Irrelevant features and the subset selection problem. In *Proceedings of the 11th International Conference on Machine Learning*, San Francisco, CA, pp. 121 – 129. Morgan Kaufmann.

Juan, L. and O. Gwun (2009). A Comparison of SIFT , PCA-SIFT and SURF. *International Journal of Image Processing (IJIP) 3*(4), 143 – 152.

Junejo, I., E. Dexter, I. Laptev, and P. Perez (2011). View-independent action recognition from temporal self-similarities. *IEEE Transactions on Pattern Analysis and Machine Intelligence 33*(1), 172 – 185.

Jurie, F. and B. Triggs (2005). Creating efficient codebooks for visual recognition. In *IEEE Tenth International Conference on Computer Vision, 2005. ICCV 2005*, Volume 1, pp. 604 – 610.

Kadir, T. and M. Brady (2003). Scale saliency: A novel approach to salient feature and scale selection. In *International Conference on Visual Information Engineering, 2003. VIE 2003*, pp. 25 – 28. IET.

Kapp, M. N., R. Sabourin, and P. Maupin (2011). A dynamic optimization approach for adaptive incremental learning. *International Journal of Intelligent Systems 26*(11), 1101 – 1124.

Karaman, S., J. Benois-Pineau, R. Mégret, V. Dovgalecs, J. Dartigues, and Y. Gäestel (2010). Human daily activities indexing in videos from wearable

cameras for monitoring of patients with dementia diseases. In *20th International Conference on Pattern Recognition, 2010. ICPR 2010*, pp. 4113 – 4116. IEEE.

Kavi, R. and V. Kulathumani (2013). Real-time recognition of action sequences using a distributed video sensor network. *Journal of Sensor and Actuator Networks 2*(3), 486 – 508.

Kellokumpu, V.-P. (2011). *Vision-based human motion description and recognition*. Ph. D. thesis, University of Oulu, Faculty of Technology, Department of Computer Science and Engineering.

Kim, K., T. H. Chalidabhongse, D. Harwood, and L. Davis (2005). Real-time foreground-background segmentation using codebook model. *Real-Time Imaging 11*(3), 172 – 185. Special Issue on Video Object Processing.

Kim, Y. K., K. Park, and J. Ko (2003). A symbiotic evolutionary algorithm for the integration of process planning and job shop scheduling. *Computers & Operations Research 30*(8), 1151 – 1171.

Kira, K. and L. A. Rendell (1992). The feature selection problem: Traditional methods and a new algorithm. In *Proceedings of the tenth national conference on Artificial intelligence*, AAAI'92, pp. 129 – 134. AAAI Press.

Kleinberger, T., M. Becker, E. Ras, A. Holzinger, and P. Müller (2007). Ambient intelligence in assisted living: Enable elderly people to handle future interfaces. In C. Stephanidis (Ed.), *Universal Access in Human-Computer Interaction. Ambient Interaction*, Volume 4555 of *Lecture Notes in Computer Science*, pp. 103 – 112. Springer Berlin / Heidelberg.

Kovashka, A. and K. Grauman (2010). Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, 2010. CVPR 2010*, pp. 2046 – 2053.

Kuncheva, L. I. and L. C. Jain (1999). Nearest neighbor classifier: Simultaneous editing and feature selection. *Pattern Recognition Letters 20*(11 - 13), 1149 – 1156.

Kwapisz, J., G. Weiss, and S. Moore (2011). Activity recognition using cell phone accelerometers. *ACM SIGKDD Explorations Newsletter 12*(2), 74 – 82.

Lanzi, P. (1997). Fast feature selection with genetic algorithms: a filter approach. In *IEEE International Conference on Evolutionary Computation, 1997*, pp. 537 – 540.

Laptev, I. (2005). On space-time interest points. *International Journal of Computer Vision 64*(2), 107 – 123.

Laptev, I., M. Marszalek, C. Schmid, and B. Rozenfeld (2008). Learning realistic human actions from movies. In *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008*, pp. 1 – 8.

Launila, A. and J. Sullivan (2010). Contextual features for head pose estimation in football games. In *20th International Conference on Pattern Recognition, 2010. ICPR 2010*, pp. 340 – 343.

Lea, C., J. Fackler, G. Hager, and R. Taylor (2012). Towards Automated Activity Recognition in an Intensive Care Unit. In *MICCAI Workshop on Modeling and Monitoring of Computer Assisted Interventions. M2CAI 2012*.

Lester, J., T. Choudhury, and G. Borriello (2006). A practical approach to recognizing physical activities. In K. Fishkin, B. Schiele, P. Nixon, and A. Quigley (Eds.), *Pervasive Computing*, Volume 3968 of *Lecture Notes in Computer Science*, pp. 1 – 16. Springer Berlin / Heidelberg.

Lewandowski, M., D. Makris, and J.-C. Nebel (2010). View and style-independent action manifolds for human activity recognition. In K. Daniilidis, P. Maragos, and N. Paragios (Eds.), *European Conference on Computer Vision. ECCV 2010*, Volume 6316 of *Lecture Notes in Computer Science*, pp. 547 – 560. Springer Berlin / Heidelberg.

Li, W., Z. Zhang, and Z. Liu (2010). Action recognition based on a bag of 3D points. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2010. CVPRW 2010*, pp. 9 – 14.

Li, W., Z. Zhang, Z. Liu, and S. Member (2008). Expandable data-driven graphical modeling of human actions based on salient postures. *IEEE Transactions on Circuits and Systems for Video Technology 18*(11), 1499 – 1510.

Lim, J., D. Ross, R. sung Lin, and M. hsuan Yang (2005). Incremental learning for visual tracking. In *Advances in Neural Information Processing Systems*, pp. 793 – 800. MIT Press.

Lin, Y.-C., M.-C. Hu, W.-H. Cheng, Y.-H. Hsieh, and H.-M. Chen (2012). Human action recognition and retrieval using sole depth information. In *Proceedings of the 20th ACM international conference on Multimedia*, MM '12, New York, NY, USA, pp. 1053 – 1056. ACM.

Liu, C.-D., Y.-N. Chung, and P.-C. Chung (2010). An interaction-embedded HMM framework for human behavior understanding: With nursing environments as examples. *IEEE Transactions on Information Technology in Biomedicine 14*(5), 1236 – 1246.

Liu, H. and H. Motoda (1998). *Feature Selection for Knowledge Discovery and Data Mining*. Boston: Kluwer Academic Publishers.

Liu, H. and H. Motoda (2002). On issues of instance selection. *Data Mining and Knowledge Discovery 6*(2), 115 – 130.

Liu, J., M. Shah, B. Kuipers, and S. Savarese (2011). Cross-view action recognition via view knowledge transfer. In *IEEE Conference on Computer Vision and Pattern Recognition, 2011. CVPR 2011*, pp. 3209 – 3216.

Liu, L., L. Shao, and P. Rockett (2013). Human action recognition based on boosted feature selection and naive bayes nearest-neighbor classification. *Signal Processing 93*(6), 1521 – 1530. Special issue on Machine Learning in Intelligent Image Processing.

Liu, Z. and S. Sarkar (2006). Improved gait recognition by gait dynamics normalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence 28*(6), 863 – 876.

Livingston, M., J. Sebastian, Z. Ai, and J. Decker (2012). Performance measurements for the Microsoft Kinect skeleton. In *IEEE Virtual Reality Short Papers and Posters. VRW 2012*, pp. 119 – 120.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision 60*(2), 91 – 110.

Lu, G., M. Kudo, and J. Toyama (2013). Temporal segmentation and assignment of successive actions in a long-term video. *Pattern Recognition Letters 34*(15), 1936 – 1944. Smart Approaches for Human Action Recognition.

Lu, J., T. Zhao, and Y. Zhang (2008). Feature selection based-on genetic algorithm for image annotation. *Knowledge-Based Systems 21*(8), 887 – 891.

Lucas, B. D. (1981). An iterative image registration technique with an application to stereo vision. *Imaging 130*, 121 – 129.

Lv, F. and R. Nevatia (2007). Single view human action recognition using key pose matching and viterbi path searching. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR 2007*, pp. 1 – 8.

Määttä, T., A. Härmä, and H. Aghajan (2010). On efficient use of multi-view data for activity recognition. In *Proceedings of the Fourth ACM/IEEE International Conference on Distributed Smart Cameras*, ICDSC '10, New York, NY, USA, pp. 158 – 165. ACM.

Mahmoud, S., M. Akhlaghinia, A. Lotfi, and C. Langensiepen (2011). Trend modelling of elderly lifestyle within an occupancy simulator. In *13th International Conference on Computer Modelling and Simulation. 2011 UkSim*, pp. 156 – 161.

Marin-Jimenez, M., A. Zisserman, and V. Ferrari (2011). "Here's looking at you, kid." Detecting people looking at each other in videos. In *Proceedings of the British Machine Vision Conference, 2011. BMVC 2011.*

Martinez, M. and R. Stiefelhagen (2013). Automated multi-camera system for long term behavioral monitoring in intensive care units. In *Proceedings of IAPR Conference on Machine Vision Applications, 2013. MVA 2013.* MVA.

Martínez-Contreras, F., C. Orrite-Urunuela, E. Herrero-Jaraba, H. Ragheb, and S. Velastin (2009). Recognizing human actions using silhouette-based HMM. In *IEEE Int. Conference on Advanced Video and Signal Based Surveillance, 2009. AVSS 2009*, pp. 43 – 48.

Martínez del Rincón, J., M. J. Santofimia, and J.-C. Nebel (2013). Common-sense reasoning for human action recognition. *Pattern Recognition Letters 34*(15), 1849 – 1860. Smart Approaches for Human Action Recognition.

Maurer, U., A. Smailagic, D. Siewiorek, and M. Deisher (2006). Activity recognition and monitoring using multiple sensors on different body positions. In *International Workshop on Wearable and Implantable Body Sensor Networks, 2006. BSN 2006*, pp. 113 – 116.

McIntyre, A. R. and M. I. Heywood (2011). Classification as clustering: A pareto cooperative-competitive GP approach. *Evolutionary Computation 19*(1), 137 – 166.

Mcivor, A. M. (2000). Background Subtraction Techniques. In *Proceedings of Image and Vision Computing, Auckland, New Zealand, 2000*.

Meeden, L. (1996). An incremental approach to developing intelligent neural network controllers for robots. *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics 26*(3), 474 – 485.

Messing, R., C. Pal, and H. Kautz (2009). Activity recognition using the velocity histories of tracked keypoints. In *IEEE 12th International Conference on Computer Vision, 2009. ICCV 2009*, pp. 104 – 111.

Mihailidis, A., J. N. Boger, T. Craig, and J. Hoey (2008). The coach prompting system to assist older adults with dementia through handwashing: an efficacy study. *BMC Geriatrics 8*, 28.

Mihailidis, A., B. Carmichael, and J. Boger (2004). The use of computer vision in an intelligent environment to support aging-in-place, safety, and independence in the home. *IEEE Transactions on Information Technology in Biomedicine 8*(3), 238 – 247.

Mihailidis, A., G. R. Fernie, and W. L. Cleghorn (2000). The development of a computerized cueing device to help people with dementia to be more independent. *Technology and Disability 13*, 23 – 40.

Mikić, I., M. Trivedi, E. Hunter, and P. Cosman (2003). Human body model acquisition and tracking using voxel data. *International Journal of Computer Vision 53*(3), 199 – 223.

Minhas, R., A. Mohammed, and Q. Wu (2012). Incremental learning in human action recognition based on snippets. *IEEE Transactions on Circuits and Systems for Video Technology 22*(11), 1529 – 1541.

Minnen, D., T. Westeyn, T. Starner, J. Ward, and P. Lukowicz (2006). Performance metrics and evaluation issues for continuous activity recognition. In *Proceedings of Performance Metrics for Intelligent Systems Workshop, 2006. PerMIS '06*.

Miranda, L., T. Vieira, D. Martinez, T. Lewiner, A. W. Vieira, and M. F. M. Campos (2012). Real-time gesture recognition from depth data through key poses learning and decision forests. In *XXV Conference on Graphics, Patterns and Images, Sibgrapi 2012.*, Ouro Preto, MG. IEEE.

Moeslund, T. (2001). A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding 81*(3), 231 – 268.

Moeslund, T. B., A. Hilton, and V. Krüger (2006). A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding 104*(2-3), 90 – 126.

Moghaddam, Z. and M. Piccardi (2013). Training initialization of hidden Markov models in human action recognition. *IEEE Transactions on Automation Science and Engineering*. DOI 10.1109/TASE.2013.2262940.

Monekosso, D. N. and P. Remagnino (2010). Behavior analysis for assisted living. *IEEE Transactions on Automation Science and Engineering 7*(4), 879 – 886.

Moutzouris, A., J. Martínez-del Rincón, M. Lewandowski, J. Nebel, and D. Makris (2011). Human pose tracking in low dimensional space enhanced by limb correction. In *IEEE 18th International Conference on Image Processing, 2011. ICIP 2011*, pp. 2301 – 2304.

Munstermann, M., T. Stevens, and W. Luther (2012). A novel human autonomy assessment system. *Sensors 12*(6), 7828 – 7854.

Murphy-Chutorian, E. and M. Trivedi (2009). Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence 31*(4), 607 – 626.

Naiel, M., M. Abdelwahab, and M. El-Saban (2011). Multi-view human action recognition system employing 2DPCA. In *IEEE Workshop on Applications of Computer Vision, 2011. WACV 2011*, pp. 270 – 275.

Nait-Charif, H. and S. McKenna (2004). Activity summarisation and fall detection in a supportive home environment. In *17th International Conference on Pattern Recognition, 2004. ICPR 2004*, Volume 4, pp. 323 – 326.

Najafi, B., K. Aminian, A. Paraschiv-Ionescu, F. Loew, C. J. Büla, and P. Robert (2003). Ambulatory system for human motion analysis using a kinematic sensor: monitoring of daily physical activity in the elderly. *IEEE Transactions on Bio-Medical Engineering 50*(6), 711 – 723.

Nakamura, T., K. Meguro, and H. Sasaki (1996). Relationship between falls and stride length variability in senile dementia of the alzheimer type. *Gerontology 42*(2), 108 – 113.

Nakashima, H., H. Aghajan, and J. C. Augusto (2009). *Handbook of Ambient Intelligence and Smart Environments* (1st ed.). Springer Publishing Company, Incorporated.

Nebel, J.-C., M. Lewandowski, J. Thévenon, F. Martínez, and S. Velastin (2011). Are current monocular computer vision systems for human action recognition suitable for visual surveillance applications? In G. Bebis, R. Boyle, B. Parvin, D. Koracin, S. Wang, K. Kyungnam, B. Benes, K. Moreland, C. Borst, S. DiVerdi, C. Yi-Jen, and J. Ming (Eds.), *Advances in Visual Computing*, Volume 6939 of *Lecture Notes in Computer Science*, pp. 290 – 299. Springer Berlin / Heidelberg.

Needleman, S. B. and C. D. Wunsch (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology 48*(3), 443 – 453.

Nguyen, N., D. Phung, S. Venkatesh, and H. Bui (2005). Learning and detecting activities from movement trajectories using the hierarchical hidden Markov model. In *IEEE Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005*, Volume 2, pp. 955 – 960.

Nicolini, C., B. Lepri, S. Teso, and A. Passerini (2010). From on-going to complete activity recognition exploiting related activities. In A. Salah, T. Gevers, N. Sebe, and A. Vinciarelli (Eds.), *Human Behavior Understanding*, Volume 6219 of *Lecture Notes in Computer Science*, pp. 26 – 37. Springer Berlin / Heidelberg.

Niebles, J. C., H. Wang, and L. Fei-Fei (2008). Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision 79*(3), 299 – 318.

Nowozin, S. and J. Shotton (2012). Action points: A representation for low-latency online human action recognition. Technical report, Microsoft Research Cambridge. Technical Report MSR- TR-2012-68.

Oikonomopoulos, A., M. Pantic, and I. Patras (2008). B-spline polynomial descriptors for human activity recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2008. CVPRW 2008*, pp. 1 – 6.

Oikonomopoulos, A., I. Patras, and M. Pantic (2005). Spatiotemporal salient points for visual recognition of human actions. *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics 36*(3), 710 – 719.

Oikonomopoulos, A., I. Patras, and M. Pantic (2009). An implicit spatiotemporal shape model for human activity localization and recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2009. CVPRW 2009*, pp. 27 – 33.

Oliver, N. M., B. Rosario, and A. P. Pentland (2000). A bayesian computer vision system for modeling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*(8), 831 – 843.

Olvera-López, J., J. Carrasco-Ochoa, J. Martínez-Trinidad, and J. Kittler (2010). A review of instance selection methods. *Artificial Intelligence Review 34*, 133 – 143.

Ozturk, O., T. Yamasaki, and K. Aizawa (2009). Tracking of humans and estimation of body/head orientation from top-view single camera for visual focus of attention analysis. In *IEEE 12th International Conference on Computer Vision Workshops, 2009. ICCV Workshops 2009*, pp. 1020 – 1027.

Park, K., Y. Lin, V. Metsis, Z. Le, and F. Makedon (2010). Abnormal human behavioral pattern detection in assisted living environments. In *Proceedings of the 3rd International Conference on Pervasive Technologies Related to Assistive Environments, 2010*, PETRA '10, New York, NY, USA, pp. 9:1 – 9:8. ACM.

Park, S. and H. Kautz (2008). Hierarchical recognition of activities of daily living using multi-scale, multi-perspective vision and RFID. In *IET 4th International Conference on Intelligent Environments, 2008*, pp. 1 – 4.

Poppe, R. (2007). Vision-based human motion analysis: An overview. *Computer Vision and Image Understanding 108*(1-2), 4 – 18.

Poppe, R. (2010). A survey on vision-based human action recognition. *Image and Vision Computing 28*(6), 976 – 990.

Porle, R. R., A. Chekima, F. Wong, and G. Sainarayanan (2009). Performance of histogram-based skin colour segmentation for arms detection in human motion analysis application. *International Journal of Electronics, Communications and Computer Engineering 1*(3), 403 – 408.

Potter, M. A. and C. Couldrey (2010). A cooperative coevolutionary approach to partitional clustering. In R. Schaefer, C. Cotta, J. Kołodziej, and G. Rudolph (Eds.), *Parallel Problem Solving from Nature, PPSN XI*, Volume 6238 of *Lecture Notes in Computer Science*, pp. 374 – 383. Springer Berlin / Heidelberg.

Quintas, J., K. Khoshhal, H. Aliakbarpour, M. Hofmann, and J. Dias (2011). Using concurrent hidden Markov models to analyse human behaviours in a smart home environment. In *12th International Workshop on Image Analysis for Multimedia Interactive Services, 2011. WIAMIS 2011.*

Rahman, S. A., I. Song, M. Leung, I. Lee, and K. Lee (2013). Fast action recognition using negative space features. *Expert Systems with Applications*. DOI 10.1016/j.eswa.2013.07.082.

Razzaghi, P., M. Palhang, and N. Gheissari (2013). A new invariant descriptor for action recognition based on spherical harmonics. *Pattern Analysis and Applications 16*(4), 507 – 518.

Reale, M., T. Hung, and L. Yin (2010a). Pointing with the eyes: Gaze estimation using a static/active camera system and 3D iris disk model. In *IEEE International Conference on Multimedia and Expo, 2010. ICME 2010*, pp. 280 – 285. IEEE.

Reale, M., T. Hung, and L. Yin (2010b). Viewing direction estimation based on 3D eyeball construction for HRI. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2010. CVPRW 2010*, pp. 24 – 31. IEEE.

Ren, F., J. Huang, R. Jiang, and R. Klette (2009). General traffic sign recognition by feature matching. In *Image and Vision Computing New Zealand, 2009. IVCNZ '09. 24th International Conference*, pp. 409 – 414.

Reyes, M., G. Dominguez, and S. Escalera (2011). Featureweighting in dynamic timewarping for gesture recognition in depth data. In *IEEE 13th International Conference on Computer Vision Workshops, 2011. ICCV Workshops 2011*, pp. 1182 – 1188.

Robertson, N. and I. Reid (2006). A general method for human activity recognition in video. *Computer Vision and Image Understanding 104*(2-3), 232 – 248.

Robertson, N., I. Reid, and M. Brady (2006). Behaviour recognition and explanation for video surveillance. In *The Institution of Engineering and Technology Conference on Crime and Security, 2006*, pp. 458 – 463.

Rodriguez, M., J. Ahmed, and M. Shah (2008). Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008*, pp. 1 – 8.

Roetenberg, D., P. J. Slycke, and P. H. Veltink (2007). Ambulatory position and orientation tracking fusing magnetic and inertial sensing. *IEEE Transactions on Bio-Medical Engineering 54*(5), 883 – 90.

Ros, F., S. Guillaume, M. Pintore, and J. Chrétien (2008). Hybrid genetic algorithm for dual selection. *Pattern Analysis and Applications 11*, 179 – 198.

Rosales, R. and S. Sclaroff (2000). Learning and synthesizing human body motion and posture. In *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, 2000*, pp. 506 – 511.

Ross, A. A. and R. Govindarajan (2005). Feature level fusion of hand and face biometrics. In *Proceedings of SPIE Conference on Biometric Technology for Human Identification II*, Volume 5779, pp. 196 – 204.

Roy, A., P. D. Mackin, and S. Mukhopadhyay (2013). Methods for pattern selection, class-specific feature selection and classification for automated learning. *Neural Networks 41*, 113 – 129. Special Issue on Autonomous Learning.

Russ, J. C. (2006). *The image processing handbook*. CRC Press.

Rybok, L., M. Voit, H. Ekenel, and R. Stiefelhagen (2010). Multi-view based estimation of human upper-body orientation. In *20th International Conference on Pattern Recognition, 2010. ICPR 2010*, pp. 1558 – 1561.

Ryoo, M. and J. Aggarwal (2009). Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *IEEE 12th International Conference on Computer Vision, 2009. ICCV 2009*, pp. 1593 – 1600.

Ryoo, M. S., J. Joung, S. Choi, and W. Yu (2010). Incremental learning of novel activity categories from videos. In *16th International Conference on Virtual Systems and Multimedia, 2010.VSMM 2010*, pp. 21 – 26.

Sadek, S., A. Al-Hamadi, B. Michaelis, and U. Sayed (2012). A fast statistical approach for human activity recognition. *International Journal of Intelligence Science 2*(1), 9 – 15.

Saghafi, B. and D. Rajan (2012). Human action recognition using pose-based discriminant embedding. *Signal Processing: Image Communication 27*(1), 96 – 111.

Sakoe, H. and S. Chiba (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing 26*(1), 43 – 49.

Santofimia, M. J., J. Martinez-del Rincon, and J.-C. Nebel (2012). Common-sense knowledge for a computer vision system for human action recognition. In J. Bravo, R. Hervás, and M. Rodríguez (Eds.), *Ambient Assisted Living and Home Care*, Volume 7657 of *Lecture Notes in Computer Science*, pp. 159 – 166. Springer Berlin / Heidelberg.

Sapp, B., A. Toshev, and B. Taskar (2010). Cascaded models for articulated pose estimation. In *Proceedings of the European Conference on Computer Vision:*

Part II. ECCV 2010, ECCV'10, Berlin, Heidelberg, pp. 406 – 420. Springer-Verlag.

Schuldt, C., I. Laptev, and B. Caputo (2004). Recognizing human actions: a local SVM approach. In *17th International Conference on Pattern Recognition, 2004. ICPR 2004*, Volume 3, pp. 32 – 36.

Sempena, S., N. Maulidevi, and P. Aryan (2011). Human action recognition using dynamic time warping. In *3rd International Conference on Electrical Engineering and Informatics, 2011. ICEEI 2011*, pp. 1 – 5.

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton University Press.

Shakhnarovich, G., P. Viola, and T. Darrell (2003). Fast pose estimation with parameter-sensitive hashing. In *IEEE Ninth International Conference on Computer Vision, 2003. ICCV 2003*, Volume 2, pp. 750 – 757.

Shao, L. and X. Chen (2010). Histogram of body poses and spectral regression discriminant analysis for human action categorization. In *Proceedings of the British Machine Vision Conference, 2010. BMVC 2010*, Volume 4.

Shimizu, H. and T. Poggio (2003). Direction estimation of pedestrian from images. Technical report, Massachusetts Institute of Technology Computer Science and Artificial Intelligence Laboratory.

Shotton, J., A. W. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake (2011). Real-time human pose recognition in parts from single depth images. In *IEEE Conference on Computer Vision and Pattern Recognition, 2011. CVPR 2011*, pp. 1297 – 1304.

Siciliano, B. and O. Khatib (2007). *Springer Handbook of Robotics*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Siedlecki, W. and J. Sklansky (1989). A note on genetic algorithms for large-scale feature selection. *Pattern Recognition Letters 10*(5), 335 – 347.

Singh, S., S. Velastin, and H. Ragheb (2010). MuHAVi: A multicamera human action video dataset for the evaluation of action recognition methods. In *IEEE Int. Conference on Advanced Video and Signal Based Surveillance, 2010. AVSS 2010*, pp. 48 – 55. IEEE.

Sminchisescu, C., A. Kanaujia, Z. Li, and D. Metaxas (2005). Discriminative density propagation for 3D human motion estimation. In *IEEE Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005*, Volume 1, pp. 390 – 397.

Smith, T. F. and M. S. Waterman (1981). Identification of common molecular subsequences. *Journal of Molecular Biology 147*(1), 195 – 197.

Song, Y., L.-P. Morency, and R. Davis (2013). Action recognition by hierarchical sequence summarization. In *IEEE Conference on Computer Vision and Pattern Recognition, 2013. CVPR 2013*.

Spriggs, E., F. De La Torre, and M. Hebert (2009). Temporal segmentation and activity classification from first-person sensing. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2009. CVPRW 2009*, pp. 17 – 24.

Stauffer, C. and W. Grimson (1999). Adaptive background mixture models for real-time tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1999. CVPR 1999*, Volume 2, pp. 246 – 252.

Sun, H., V. De Florio, N. Gui, and C. Blondia (2009). Promises and challenges of ambient assisted living systems. In *Sixth International Conference on Information Technology: New Generations, 2009. ITNG '09*, pp. 1201 – 1207.

Sun, L., U. Klank, and M. Beetz (2009). EYEWATCHME-3D hand and object tracking for inside out activity analysis. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2009. CVPRW 2009*, pp. 9 – 16. IEEE.

Sun, Z., G. Bebis, and R. Miller (2004). Object detection using feature subset selection. *Pattern Recognition 37*(11), 2165 – 2176.

Sun, Z., G. Bebis, X. Yuan, and S. Louis (2002). Genetic feature subset selection for gender classification: a comparison study. In *Proceedings of the Sixth IEEE Workshop on Applications of Computer Vision, 2002. WACV 2002*, pp. 165 – 170.

Sundaram, S. and W. Cuevas (2009). High level activity recognition using low resolution wearable vision. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2009. CVPRW 2009*, pp. 25 – 32.

Suriani, N. S., A. Hussain, and M. A. Zulkifley (2013). Sudden event recognition: A survey. *Sensors 13*(8), 9966 – 9998.

Sutton, C. and A. McCallum (2007). An Introduction to Conditional Random Fields for Relational Learning. In L. Getoor and B. Taskar (Eds.), *Introduction to Statistical Relational Learning*. MIT Press.

Suzuki, S. and K. Abe (1985). Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics, and Image Processing 30*(1), 32 – 46.

Tan, K., Y. Yang, and C. Goh (2006). A distributed cooperative coevolutionary algorithm for multiobjective optimization. *IEEE Transactions on Evolutionary Computation 10*(5), 527 – 549.

Tao, S., M. Kudo, and H. Nonaka (2012). Privacy-preserved behavior analysis and fall detection by an infrared ceiling sensor network. *Sensors 12*(12), 16920 – 16936.

Tao, Y., H. Hu, and H. Zhou (2007). Integration of vision and inertial sensors for 3D arm motion tracking in home-based rehabilitation. *The International Journal of Robotics Research 26*(6), 607 – 624.

Tenorth, M., J. Bandouch, and M. Beetz (2009). The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition. In *IEEE 12th International Conference on Computer Vision Workshops, 2009. ICCV Workshops 2009*, pp. 1089 – 1096.

Thurau, C. and V. Hlaváč (2007). *n*-grams of action primitives for recognizing human behavior. In W. Kropatsch, M. Kampel, and A. Hanbury (Eds.), *Computer Analysis of Images and Patterns*, Volume 4673 of *Lecture Notes in Computer Science*, pp. 93 – 100. Springer Berlin / Heidelberg.

Toyama, K., J. Krumm, B. Brumitt, and B. Meyers (1999). Wallflower: principles and practice of background maintenance. In *IEEE Seventh International Conference on Computer Vision, 1999. ICCV 1999*, Volume 1, pp. 255 – 261.

Tran, D. and A. Sorokin (2008). Human activity recognition with metric learning. In D. Forsyth, P. Torr, and A. Zisserman (Eds.), *European Conference on Computer Vision. ECCV 2008*, Volume 5302 of *Lecture Notes in Computer Science*, pp. 548 – 561. Springer Berlin / Heidelberg.

Turaga, P., R. Chellappa, V. Subrahmanian, and O. Udrea (2008). Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology 18*(11), 1473 – 1488.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind 59*(236), 433 – 460.

Tuytelaars, T. and K. Mikolajczyk (2007). Local invariant feature detectors: A survey. *Foundations and Trends in Computer Graphics and Vision 3*(3), 177 – 280.

van der Meulen, P. and A. Seidl (2007). Ramsis - the leading CAD tool for ergonomic analysis of vehicles. In V. Duffy (Ed.), *Digital Human Modeling*, Volume 4561 of *Lecture Notes in Computer Science*, pp. 1008 – 1017. Springer Berlin / Heidelberg.

Virone, G. and A. Sixsmith (2008). Monitoring activity patterns and trends of older adults. In *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, pp. 2071 – 2074.

Vishwakarma, S. and A. Agrawal (2013). A survey on activity recognition and behavior understanding in video surveillance. *The Visual Computer 29*(10), 983 – 1009.

Vitaladevuni, S., V. Kellokumpu, and L. Davis (2008). Action recognition using ballistic dynamics. In *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008*, pp. 1 – 8.

Wang, L. (2003). Recent developments in human motion analysis. *Pattern Recognition 36*(3), 585 – 601.

Wang, Y., K. Huang, and T. Tan (2007). Human activity recognition based on R transform. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR 2007*, pp. 1 – 8.

Wang, Z., M. Jiang, Y. Hu, and H. Li (2012). An incremental learning method based on probabilistic neural networks and adjustable fuzzy clustering for human activity recognition by using wearable sensors. *IEEE Transactions on Information Technology in Biomedicine 16*(4), 691 – 699.

Wang, Z., J. Wang, J. Xiao, K.-H. Lin, and T. Huang (2012). Substructure and boundary modeling for continuous action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, 2012. CVPR 2012*, pp. 1330 – 1337.

Ward, J. A., P. Lukowicz, and H. W. Gellersen (2011). Performance metrics for activity recognition. *ACM Transactions on Intelligent Systems and Technology 2*(1), 6:1 – 6:23.

Weinland, D., E. Boyer, and R. Ronfard (2007). Action recognition from arbitrary views using 3D exemplars. In *IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007*, pp. 1 – 7.

Weinland, D., M. Özuysal, and P. Fua (2010). Making action recognition robust to occlusions and viewpoint changes. In K. Daniilidis, P. Maragos, and N. Paragios (Eds.), *European Conference on Computer Vision. ECCV 2010*, Volume 6313 of *Lecture Notes in Computer Science*, pp. 635 – 648. Springer Berlin / Heidelberg.

Weinland, D., R. Ronfard, and E. Boyer (2006). Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding 104*(2-3), 249 – 257.

Weldemichael, D. A. and G. T. Grossberg (2010). Circadian rhythm disturbances in patients with alzheimer's disease: a review. *International Journal of Alzheimer's Disease 2010*. Article ID 716453.

Wiegand, R. P. (2004). *An analysis of cooperative coevolutionary algorithms.* . Ph. D. thesis, George Mason University, Fairfax, VA, USA.

Wilson, D. and T. Martinez (2000). Reduction techniques for instance-based learning algorithms. *Machine Learning 38*, 257 – 286.

Wood, A., J. Stankovic, G. Virone, L. Selavo, Z. He, Q. Cao, T. Doan, Y. Wu, L. Fang, and R. Stoleru (2008). Context-aware wireless sensor networks for assisted living and residential monitoring. *IEEE Network 22*(4), 26 – 33.

Wren, C., A. Azarbayejani, T. Darrell, and A. Pentland (1997). Pfinder: real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence 19*(7), 780 – 785.

Wu, B. and R. Nevatia (2005). Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *Proceedings of IEEE Tenth International Conference on Computer Vision, 2005. ICCV 2005*, Volume 1, pp. 90 – 97.

Wu, C., A. H. Khalili, and H. Aghajan (2010). Multiview activity recognition in smart homes with spatio-temporal features. In *Proceedings of the Fourth ACM/IEEE International Conference on Distributed Smart Cameras, 2010. ICDSC '10*, New York, NY, USA, pp. 142 – 149. ACM.

Wu, J., A. Osuntogun, T. Choudhury, M. Philipose, and J. Rehg (2007). A scalable approach to activity recognition based on object use. In *IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007*, pp. 1 – 8. IEEE.

Wu, X., Z. Shi, and Y. Zhong (2010). Detailed analysis and evaluation of keypoint extraction methods. In *International Conference on Computer Application and System Modeling, 2010. ICCASM 2010*, Volume 2, pp. V2–562 – V2–566.

Wu, X., D. Xu, L. Duan, and J. Luo (2011). Action recognition using context and appearance distribution features. In *IEEE Conference on Computer Vision and Pattern Recognition, 2011. CVPR 2011*, pp. 489 – 496.

Xu, C. and J. Prince (1998). Snakes, shapes, and gradient vector flow. *IEEE Transactions on Image Processing 7*(3), 359 – 369.

Yamamoto, M., H. Mitomi, F. Fujiwara, and T. Sato (2006). Bayesian classification of task-oriented actions based on stochastic context-free grammar. In *7th International Conference on Automatic Face and Gesture Recognition, 2006. FGR 2006*, pp. 317 – 322.

Yamato, J., J. Ohya, and K. Ishii (1992). Recognizing human action in time-sequential images using hidden Markov model. In *IEEE Conference on Computer Vision and Pattern Recognition, 1992. CVPR 1992*, pp. 379 – 385.

Yan, P., S. Khan, and M. Shah (2008). Learning 4D action feature models for arbitrary view action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008*, pp. 1 – 7.

Yang, A., S. Iyengar, S. Sastry, R. Bajcsy, P. Kuryloski, and R. Jafari (2008). Distributed segmentation and classification of human actions using a wearable motion sensor network. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2008. CVPRW 2008*, pp. 1 – 8.

Yang, H., L. Shao, F. Zheng, L. Wang, and Z. Song (2011). Recent advances and trends in visual tracking: A review. *Neurocomputing 74*(18), 3823 – 3831.

Yang, J. and V. Honavar (1998). Feature subset selection using a genetic algorithm. *IEEE Intelligent Systems and their Applications 13*(2), 44 – 49.

Yorita, A. and N. Kubota (2009). Fuzzy-based evolutionary robot vision for people tracking. *International Journal of Intelligent Computing in Medical Sciences & Image Processing 3*(2), 119 – 129.

Zelnik-Manor, L. and M. Irani (2006). Statistical analysis of dynamic actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence 28*(9), 1530 – 1535.

Zhou, F. and F. Torre (2009). Canonical time warping for alignment of human behavior. In *Advances in Neural Information Processing Systems*, pp. 2286 – 2294.

Zhou, Z., X. Chen, Y.-c. Chung, Z. He, T. X. Han, and J. M. Keller (2008). Activity analysis, summarization, and visualization for indoor human activity monitoring. *IEEE Transactions on Circuits and Systems for Video Technology 18*(11), 1489 – 1498.

Zhu, F., L. Shao, and M. Lin (2013). Multi-view action recognition using local similarity random forests and sensor fusion. *Pattern Recognition Letters 34*(1), 20 – 24. Extracting Semantics from Multi-Spectrum Video.

Zouba, N., B. Boulay, F. Bremond, and M. Thonnat (2008). Monitoring activities of daily living (ADLs) of elderly based on 3D key human postures. In B. Caputo and M. Vincze (Eds.), *Cognitive Vision*, Volume 5329 of *Lecture Notes in Computer Science*, pp. 37 – 50. Springer Berlin / Heidelberg.