Activity Recognition and Uncertain Knowledge in Video Scenes Applied to Health Care Monitoring

Rim Romdhane

UNIVERSITE DE NICE-SOPHIA ANTIPOLIS

ECOLE DOCTORALE STIC

SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

T H E S E

pour l'obtention du grade de

Docteur en Sciences

de l'Université de Nice - Sophia Antipolis

Mention : INFORMATIQUE

présentée et soutenue par

Rim ROMDHANE

Activity Recognition and Uncertain Knowledge in Video Scenes Applied to Health Care Monitoring

Thèse dirigée par Monique THONNAT
soutenue le 30/09/2013

Jury:

Catherine	GARBAY	Directeur de recherche, CNRS - Grenoble	Président
Mounir	MOKHTARI	Professeur, Institut Mines Telecom, HDR UPMC	Rapporteur
Najoua	BEN AMARA	Professeur, Ecole Nationale d'Ingénieurs - Sousse	Rapporteur
Philippe	ROBERT	PU-PH, Université de Nice Sophia Antipolis, CoBTek	Examinateur
François	BREMOND	DR2, INRIA Sophia Antipolis - Méditerranée	Examinateur
Monique	THONNAT	DR1, INRIA Sophia Antipolis - Méditerranée	Directrice de thèse

UNIVERSITE DE NICE-SOPHIA ANTIPOLIS

ECOLE DOCTORALE STIC

SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

T H E S E

pour l'obtention du grade de

Docteur en Sciences

de l'Université de Nice - Sophia Antipolis

Mention : INFORMATIQUE

présentée et soutenue par

Rim ROMDHANE

Reconnaissance d'activités et connaissances incertaines dans les scènes vidéos appliqués à la surveillance de personnes âgées

Thèse dirigée par Monique THONNAT

soutenue le 30/09/2013

Jury:

Catherine	GARBAY	Directeur de recherche, CNRS - Grenoble	Président
Mounir	MOKHTARI	Professeur, Institut Mines Telecom, HDR UPMC	Rapporteur
Najoua	BEN AMARA	Professeur, Ecole Nationale d'Ingénieurs - Sousse	Rapporteur
Philippe	ROBERT	PU-PH, Université de Nice Sophia Antipolis, CoBTek	Examinateur
François	BREMOND	DR2, INRIA Sophia Antipolis - Méditerranée	Examinateur
Monique	THONNAT	DR1, INRIA Sophia Antipolis - Méditerranée	Directrice de thèse

ACKNOWLEDGMENTS

Merci pour Pr. Najoua Essoukri Ben Amara et Pr. Mounir Mokhtari d'avoir accepté de rapporter ma thèse. Je voudrai leur remercier pour leurs conseils et remarques pertinentes.
Je tiens aussi à remercier Pr. Catherine Garbay pour avoir accepter de présider ce jury.

Merci à Monique Thonnat et à François Brémond de m'avoir donné la chance de faire cette thèse dans leur équipe. Merci à tous les deux pour avoir été disponible, patient et m'avoir guidée tout au long de la thèse. Je tiens particulierement à remercier Monique d'avoir été attentionée et compréhensive aprés mon accouchement des jumeaux.

Merci à Pr. Philippe Robert pour avoir été disponible, de son serieux et ambiance amicale. Merci de m'avoir fournit les materiaux et l'environement favorable pour travailler à l'hopital. C'etait un grand plaisir pour moi de travailler avec lui.

Un grand merci à Catherine Martin, Chantal Joncour et Jane Desplanques pour toute l'aide administrative qu'elles m'ont apporté tout au long de ces trois années et pour leur grande gentillesse.

Mes remerciements à tous les membres de l'équipe Pulsar/Stars: Annie, Sabine, Jean-Paul, Salma, Sofia, Carolina, Slawek, Jihed, Nedra, Piotr, Ratnesh, Bernard, Nadia, Etienne, Guillaume, Luis, Phu, Mohammed Becha, Anh-Tuan, Guido, Daniel, Julien Badie, Julien Gueytat, Ikhlef, Leonardo, Chedli, Vasanth, Carlos, Vaibhav, Sahil, Christophe, pour avoir su y faire régner une bonne ambiance amicale et propice au travail.

Un grand merci à toute ma famille, et plus particulièrement à mes parents Mohammed Moncef et Dalila, à mes fréres Riadh, Zhouhair et Mohammed et mes soeurs Leila, Salwa et Jihene et leur familles respectives, pour leurs soutiens et leurs encouragements.

Un grand merci pour ma petite famille, mon mari pour ses encouragements et un trés grand merci à mes amours, mes jumeaux Ahmed et Mariem que dieu les protège pour moi!

Rim Romdhane
Rim.Romdhane@inria.fr
Sophia Antipolis, France

CONTENTS

FIGURES

TABLES

1

INTRODUCTION

Research on activity recognition is receiving an increasing attention from the scientific community today. It is one of the most challenging problem in computer vision and artificial intelligence research. The main goal of the current activity recognition research consists in recognizing and understanding accurate short-term action and long-term complex activities. The advances in low-level processing in video data such as motion detection, object classification and tracking have enabled to focus on higher level analysis of activity recognition [Vu et al., 2003a], [Ryoo and Aggarwal, 2009], [Brendel et al., 2011] and [Kwak et al., 2011]. Recent work focus in the management of the uncertainty to deal with the robustness required by the recent applications [Tran and Davis, 2008], [Ryoo and Aggarwal, 2010], [Lavee et al., 2010a]. There are several interesting application areas for activity recognition systems, mostly video surveillance and health care monitoring. Automatic surveillance systems in public places like airports and subways stations require detection of abnormal and suspicious activities. Health care monitoring consists in monitoring the activities of a person or elderly through cameras to automatically detect early symptoms of certain diseases. It is well known that even subtle changes in the behavior of the elderly can give important signs of progression of certain diseases. Disturbed sleeping patterns could be caused, for example, by heart failure and chronic disease. Changes in gait, on the other hand, can be associated with early signs of neurological abnormalities linked to several types of dementias. These examples highlight the importance of continuous observation of behavioral changes in the elderly in order to detect health deterioration before it becomes critical. Thus a system permitting to analyse the elderly behaviors and looking for changes in their activities is more than needed. Of course, for such a system to be effective, the activity recognition task must provide very accurate results. In fact, if activities are wrongly recognized, the monitoring system may draw erroneous conclusions about the actual adherence

of the patient to the practitioners prescriptions, as well as provide error-prone statistics about the health status of the patient.

The following sections describe the motivations, the objectives of this thesis, the context of the study, my hypotheses, my contributions and the thesis layout.

1.1 Motivations

This work was greatly motivated by research done in understanding human activity. Over the last several years much effort has been put into developing frameworks for activity recognition and employing them in a variety of domains to monitor activities.

One central issue here is the issue of robustness. Robustness is defined as the degree to which a system or a component can function correctly in the presence of invalid inputs or stressful environment conditions. Most systems that have been built to recognize activities have been limited in the variety of activities they recognize and/or in the management of the uncertainty of the recognition. In this work, we propose an approach for video activity recognition that addresses these issues by combining semantic modelling together with a probabilistic reasoning to cope with the errors of low-level (primitive events level) detectors and to handle the uncertainty of the high level event recognition to support demands from health care domains. Our approach aims to provide several services for Alzheimer disease patients in order to help them to retain their independence and to live safely.

1.2 Objectives

The main objective of this thesis is to propose a new activity recognition method able to manage the errors of low level detectors and the uncertainty of high level event recognition to detect interesting activities for health care monitoring. We focus on semantic video event representation and recognition of daily living activities to build an automatic video monitoring system able to help clinician to detect early symptoms of Alzheimer disease.

1.3 Context of the study

Health care technologies for elderly with dementias is a popular area of research. Alzheimer Disease (AD) and related disorders represents a major challenge for health care systems with aging populations. AD is associated with neurodegenerative changes that compromise cognitive and functional abilities and may result in behavioral and neuropsychiatric symptoms (NPS). Many efforts are currently undertaken to investigate AD pathology and develop appropriate treatment strategies. These strategies focus on preserving cognitive and functional abilities and

maintaining quality of life in the AD sufferer. Rating scales are essential tools for the diagnosis of AD, the assessment and careful monitoring of symptoms, as well as the evaluation of treatment effects. However these standard rating scales do not fully capture the complexity of a disease. In fact, AD includes deterioration in cognitive, behavioral and functional domains that do not always progress in parallel and may change idiosyncratically according to the individuality of a given patient [Romdhane et al., 2011]. For this reason, Automatic monitoring of Activities of Daily Living such as eating, using the telephone and managing medications has been a popular focus in gerontechnology.

Detection of these activities would enable systems to monitor and recognize changes in patterns of behavior that might be indicators of developing physical or mental medical conditions. Similarly, it could help to determine the level of independence of elderly. If it is possible to develop systems that recognize such activities, the medical experts may be able to automatically detect changes in patterns of behavior of people that indicate declines in health. This is a challenging domain for multiple reasons. Firstly, for clinicians it is important to establish the exact type of indicators that are clinically relevant and can provide information that can be used in daily practice. Secondly, in the computer vision domain, the challenge is to adapt the technical constraints with the needs of the clinician. For patients and caregivers, participating in an active way and giving an opinion on the feasibility and tolerability of the study is important.

The main impetus of this study comes from the societal objective of assisting and keeping elderly people in their familiar home surroundings, or to enable them to 'age in place'. More specifically, the overall aim is to develop an automatic activity recognition approach for behavioural assessment and preventative care in early and moderate stage AD.

This PhD work has been conducted in the Stars team at INRIA Sophia Antipolis in France. Stars is a multi-disciplinary team at the frontier of computer vision, artificial intelligence and software engineering. Stars work focuses on two main application areas: safety/security and health care. This work takes place in this context and aims at recognizing human activities for health care applications. In this study, we collaborate with clinicians from Nice hospital to determine scenarios to recognize Alzheimer activities that are most important to monitor.

1.4 Thesis Hypothesis

This thesis assume the following hypothesis:

- **Available Detected and Tracked Objects**: In this work, we assume that tracked objects are available. We also assume that the tracking algorithm provides a measurement (i.e. trajectory probability) which is used during the event recognition process.

- **Fixed Video Camera**: In this work, we assume the use of fixed video cameras. The video

cameras are fixed on a wall and without pan, tilt or zoom and without any restriction on the camera orientation. We assume availability of calibration to compute the transformation of the 2D image referential point to a 3D scene referential point.

- **Scene Context**: In this work, we assume the use of scene context. Scene context contains both geometric and semantic description of the specific zones, walls and equipments located in the observed scene.

- **Event models**: We assume the use of a set of event models needed to detect the activities of interest.

1.5 Thesis Contributions

The main contributions of this work are the following:

- **The first main contribution consists in a new activity recognition approach combining semantic event representation and probabilistic event inference to cope with uncertainty.** We use a generic framework for representing the semantics related to events (e.g. events models and objects, contextual information). Then, we define a two-layer semantic-based method (primitive event layer and composite event layer) to perform probabilistic recognition considering the uncertainty of the low-level analysis. The uncertainty is managed by combining the probabilistic output of both layers. Although our approach is demonstrated in the video health care monitoring domain, it is not restricted to a specific domain or implementation.

- **The second contribution consists in a new dynamic linear model for temporal attributes filtering.** We propose a dynamic linear model for computing and updating the attributes of the mobile objects to deal with low-level errors. We compute the confidence of the current attribute value based on the temporal history of its previous values.

- **The third contribution consist in new event modeling specification.** we improve the event description language developed in Stars team [Vu et al., 2003a] and introduce a new probabilistic description based approach to gain in flexibility for event modeling by adding the notion of utility. Utility expresses the importance of sub-events to the recognition of the whole event.

- **The fourth main contribution consist in a knowledge base for health care monitoring.** An ontology for health care monitoring was proposed, in this ontology, we presented several concepts useful for health care monitoring domain. Clinical scenario composed of

a set of activity models for Alzheimer monitoring have been elaborated with the help of clinicians. The overall aim of the clinical scenario was to enable the participants to undertake a set of daily tasks that could realistically be achieved in the setting of an observation room. the defined activity models have been evaluated on a set of videos collected from hospital. Several criteria was elaborated in close collaboration with clinicians to obtain a quantifiable assessment of instrumental activities of daily living (IADLs) of Alzheimer people compared to healthy older people.

1.6 Thesis Layout

This thesis is organized as follows:

1. In **chapter 1**, we introduce the motivation, the objective and the context of the study.

2. In **chapter 2**, we first discuss related work in the area of activity recognition and then we describe the different technologies for monitoring human activities for the health care domain.

3. In **chapter 3**, we present an overview of the proposed activity recognition approach. We give a general architecture of the proposed approach. We describe the inputs, the outputs and the major sources of knowledge of our approach. We define the activity recognition problem as a key component of automatic health care monitoring.

4. In **chapter 4**, we present the proposed video event ontology for representing the knowledge about activities of interest for experts from different domains.

5. In **chapter 5**, we present the proposed approach for the recognition of primitive events and temporal composite events. This is a novel approach combining logic and probabilistic reasoning for event detection.

6. In **chapter 6**, we evaluate the proposed approach and we test the proposed activity models in a set of scenarios performed in a realistic experimental site in Nice hospital.
   We present the obtained results of the performance of the vision algorithm and the medical results.

7. Finally, in **chapter 7**, we conclude this work, by summarizing the contributions of this thesis, and by presenting short-term and long-term perspectives.

2

State Of the Art

One of the most challenging issue in computer vision research is the activity recognition problem for several reasons including noise and uncertainty of low-level output and the ambiguity for translating the semantic (high level) definition of events into a formalism for representation and recognition of events. The main issues of activity recognition are:

- The first issue is to proper represent and model the activities of interest, according to the objectives of real world applications.
- The second issue is to proper choose the appropriate event inference formalism for the detection of high level events.
- The third issue for an accurate detection of meaningful events is to deal with the low-level data processing noise.

In this chapter, we first introduce related work on the approaches for activity representation and recognition. We will discuss the approaches which manage the uncertainty of recognition. Second, we discuss previous work on activity recognition frameworks applied to the health care domain.

2.1 Activity Recognition Approaches

Automatic activity recognition is a very important and active area of research. Research on human activity recognition has emerged as an application domain of computer vision research. Technological advancements have enabled the development of systems able to monitor people behavior and provide information that is filtered and processed in order to infer people activities. Activity recognition approaches can be divided into two main approaches: (i) Probabilistic approaches and (ii) Description-based approaches. The two kind of approaches are described below in this section. We discuss also the approaches which combine logic and probabilistic reasoning.

2.1.1 Probabilistic Approaches

There is a vast amount of literature in the area of computer vision, where the aim is to recognize different types of human activities using probabilistic approaches. Probabilistic methods are designed to search event that most likely fit a model of behavior. These approaches include Neural Networks, Bayesian Networks and Hidden Markov Models. The main characteristic of these techniques is to model explicitly the uncertainty. Neural Networks and Bayesian Networks are well adapted to model the uncertainty in the recognition of events depending on the visual features at a given time. However, Hidden Markov Models deal with sequential events.

2.1.1.1 Neural Networks

The Basic artificial neuron is a model inspired by the biological neuron in the brain. We briefly introduce how biological neuron works in the following way [Wells, 2001]:

'Biological neuron receives inputs along the "dendrites" and add's them up. The neuron shall then produce an output if the sum is greater than the "threshold value". The dendrites are connected to the output's of other neurons via special junctions known as "synapses"'.

Thus the basic artificial neuron model performs a weighted sum on its inputs and compares the resulting value to its internal threshold level. If the the threshold is exceeded, the neuron is activated (fig.2.1). A Neural Network consists of an interconnected group of artificial neurons (fig.2.2). Each input is sent to every neuron in the hidden layer and then each hidden layer's neuron's output is connected to every neuron in the next layer. There can be any number of hidden layers within a Neural Network. Neural Network is an adaptive system that changes its structure during a learning phase. Commonly Neural Networks are adjusted, or trained, so that a particular input leads to a specific target output. Such a situation is shown in fig.2.3

Neural Network techniques have been one of the first approach to be used for activity recognition [Howell and Buxton, 2001], [Chen and Zhang, 2006], [Chen and Zhang, 2007]. Howell

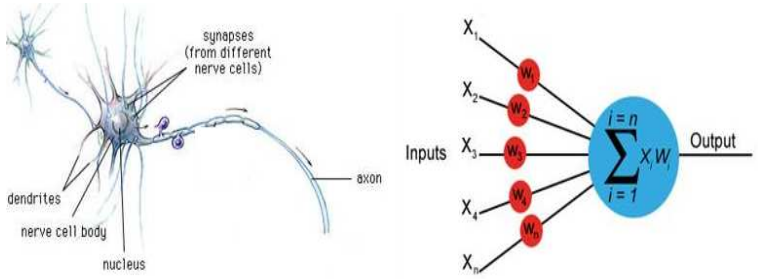


Figure 2.1: Representation of the biological (left) and artificial neuron (right): each input into the artificial neuron has its own weight associated with it, illustrated by the red circle. The neuron (blue circle) sums all the input values multiplied by them weight and compares to a threshold [Wells, 2001].

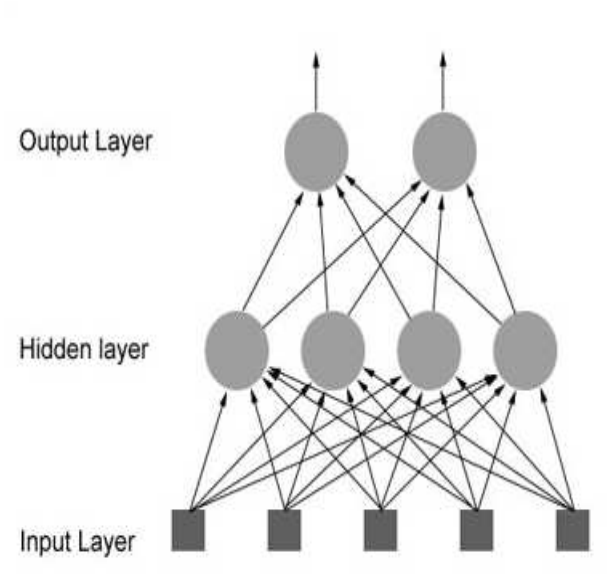


Figure 2.2: Representation of a Neural Network.

and Buxton [Howell and Buxton, 2001] have used Neural Network approaches for human behaviour recognition. More precisely, Howel and Buxton used a time-delay variant of the Radial Basis Function (RBF) network to recognize simple pointing and waving hand gestures in image sequences. However, they do not deal with complex behavior involving a large number of physical objects and complex temporal constraints.

In [Chen and Zhang, 2006], [Chen and Zhang, 2007] the authors have used Neural Networks to recognize events from traffic surveillance videos using the trajectory and the interaction of the observed object (i.e. vehicles) (fig.2.4). A series of vectors is used to represent the

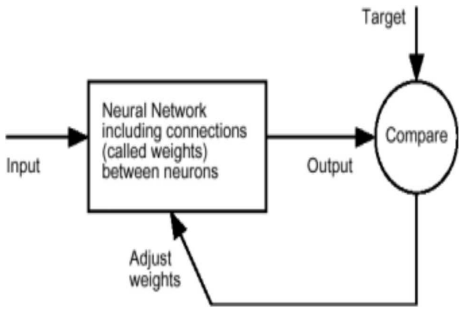


Figure 2.3: The Neural Network is adjusted, based on a comparison of the output and the target, until the network output matches the target.

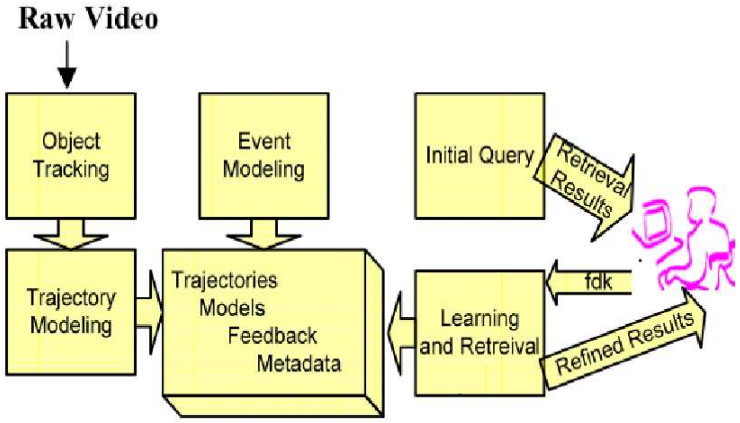


Figure 2.4: Architecture of the approach proposed in [Chen and Zhang, 2006] to detect accidents in traffic surveillance videos.

trajectory model of a vehicle at each point i (equation 2.1), v_i is the velocity, $vdiff_i$ and θ_i are respectively the difference of velocity and the difference of angle of two consecutive points.

$$\alpha = [\alpha_1, ..., \alpha_n]; \quad \alpha_i = [v_i, vdiff_i, \theta_i] \tag{2.1}$$

The trajectory model of two vehicles that appear together is represented as follow (2.2), where $v1diff_i$ and $v2diff_i$ are the velocity changes of the two vehicles at the sampling point i, cat_i is the distance between the two vehicles.

$$\alpha = [\alpha_1, ..., \alpha_n]; \quad \alpha_i = [v1diff_i, v2diff_i, cat_i] \tag{2.2}$$

The authors extended the Neural Network model for temporal processing to integrate the trajectory, they use a temporal window of size m, to predict the value of a vector x_i based on

the m preceding observed data (2.3).

$$x_i = f(x_{i-m}, ..., x_{i-1}) \tag{2.3}$$

The problem becomes a classification task, mapping a temporal sequence onto a class label $c_k \in C$, C is the set of all class labels (2.4).

$$f_c : (x_{i-m}, ..., x_{i-1}) \longrightarrow c_k \in C \tag{2.4}$$

The Neural Network approach tries to recognise activities taking advantage of learning techniques which can adapt the recognition algorithm to uncertain environments. However, Neural Network is mostly applied to the recognition of atomic events. It is not efficient to cope with complex activities involving a large number of physical objects and complex temporal constraints because it leads to a combinatorial explosion of possible activities corresponding to all combinations of physical objects detected in the scene.

2.1.1.2 Bayesian Networks

Another popular approach for activity recognition is the use of Bayesian networks. Bayesian networks (BNs) are well adapted to cope with the problem of uncertain environments. The main advantage of Bayesian networks is that they are capable of modeling the uncertainty of the recognition by using probabilities based on the Bayes' theorem of probability theory

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \tag{2.5}$$

Where,

- $P(B|A)$ is the conditional probability of an event B, given the event A. It is also called the likelihood.
- $P(A)$ is the prior probability (or 'unconditional' or 'marginal' probability) of A. It is "prior" in the sense that it does not take into account any information about B.
- $P(B)$ is the prior or marginal probability of B, and acts as a normalizing constant.
- $P(A|B)$ is the conditional probability of A, given B. It is also called the posterior probability because it is derived from or depends upon the specified value of B.

A Bayesian network consists of a directed acyclic graph in which nodes represent random variables which may be discrete or continuous (i.e. parametric distribution), and arcs represent the causal relationships (i.e. conditional dependencies or links) between variables. For a link

between two variables, $A \rightarrow B$, the overall joint distribution is specified by the product of the prior probability $P(A)$ and the conditional probability $P(B|A)$. The dependencies are specified a priori and used to create the network structure. The distributions $P(A)$ and $P(B|A)$ must be specified beforehand to form the network from domain knowledge.

Bayesian networks are mainly used for calculating conditional probabilities of variables related to each other by relations of cause and effect. This computation is called inference. A particular type of inference is the updating of probabilities that occurs following a set of observations. However, depending on the network structure, these calculations are more or less complex. While the general inference problem in BNs is NP-hard, efficient algorithms for inference exist under certain BN structure assumptions.

A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. In naive Bayes classifier each data instance X is assigned to one of m classes $C1, C2, ..., Cm$ which it has the highest posterior probability conditioned on X (i.e. the class which is most probable given the prior probabilities of the classes and the data X [Squire, 2004]. That is to say, X is assigned to class C_i if and only if:

$$P(C_i|X) > P(C_j|X) \quad for\ all\ j\ such\ that\ 1 \leq j \leq m. \tag{2.6}$$

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \tag{2.7}$$

Naive Bayes classifier have been often used to recognize elementary actions at the numerical level [Nevatia et al., 2004] with only one physical object. Buxton and Gong [Buxton and Gong, 1995] proposed an interpretation of the scene based on Bayesian networks characterizing the positions and interactions between objects observed in a traffic surveillance zone.

Hongeng et al. [Hongeng and Nevatia, 2001], [Nevatia et al., 2004] have used Bayesian networks and multi-agent architecture in the recognition of complex events (fig. 2.5). These events are represented in four hierarchical layers:
- The first layer corresponds to the detection and monitoring of observed objects (or agents).
- The second corresponds to the updating of properties associated with agents (distance between objects, changing speeds and directions of observed objects ...)
- The third defines the simple events (approaching another object, to slowdown,...)
- The latter defines the scenarios, using the simple events to model these scenarios, using logical rules of sequences and durations.

However only the simple events were recognized by BNs, the temporal aspect was considered by HMMs.

BNs have been successfully applied to person interaction [Park and Aggarwal, 2004] such as 'shake hands', parking lot surveillance [Nevatia et al., 2004], traffic monitoring [Kumar et al., 2005]

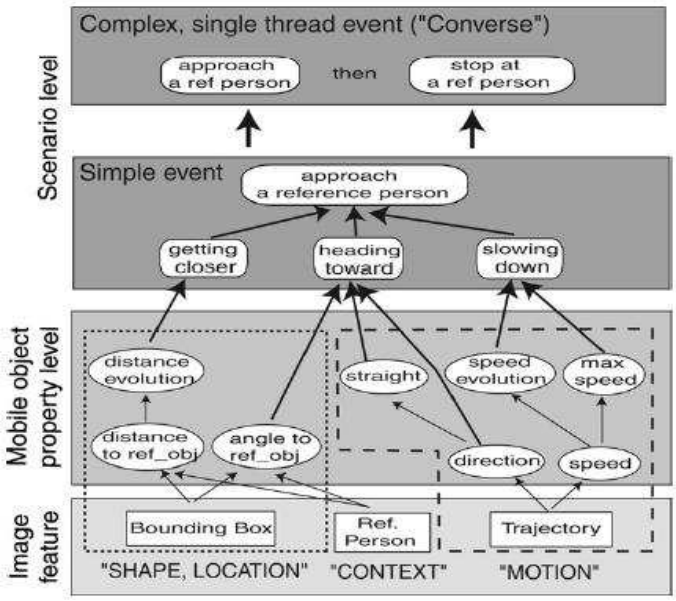


Figure 2.5: Representation of the complex event 'converse' as described in [Nevatia et al., 2004].

and detection of left luggage [Lv et al., 2006]. However, BNs are not adapted to model temporal compositions of events like sequential events. In fact, a major limitation of Bayesian networks is their poor representation of temporal dynamics. Time is not straightforward modeled in Bayesian Network formalism, temporal representation in Bayesian networks is often done using a static representation, where time points or time slices are represented as static processes. Dynamic Bayesian Networks (DBNs, BNs for dynamic processes) have been used to represent temporal sequencing of BNs [Town, 2006]. Dynamic Bayesian Networks have been used successfully to recognize short temporal actions [Reddy et al., 2003], but the recognition process depends on time segmentation: when the frame-rate or the activity duration changes, the DBN has to be re-trained. DBNs pose specific computational challenges. In fact, DBNs are able to model temporal relationships between variables, they do this by breaking up time into relevant discrete time-steps (or time slice), and placing a structurally similar copy of the network within each time-step. A causal relationship between nodes in time-step k and nodes in time-step $k+1$ are then inserted (fig. 2.6). For complex networks (i.e. network with a large number of nodes, 20 or more nodes) this would not be feasible as it makes the network gets very large, very quickly and increases in complexity very quickly, in turn it makes greatly increasing the amount of computational power required to run it ([Nicholson and Korb, 2006]).

Other DBNs have been studied to recognize more complicated composite events with sampling-based inference algorithms [Laxton et al., 2007]. However, DBNs are still insufficient to describe various events because their scenarios are bounded by their conditional dependency

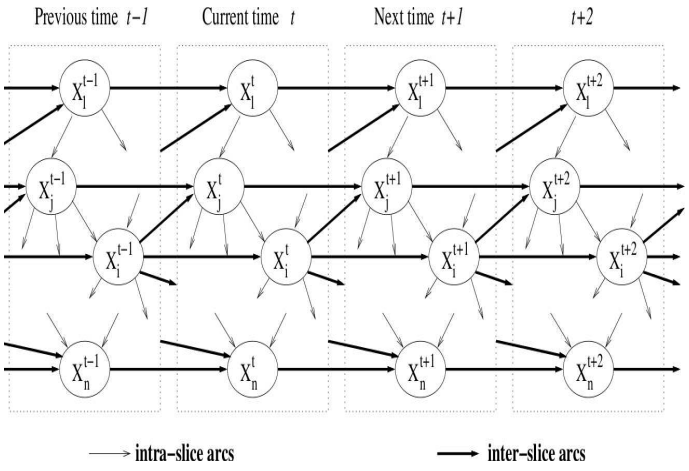


Figure 2.6: Example of a DBN as described in [Nicholson and Korb, 2006]. Each variable in a DBN is associated with a time slice t and denoted Xt. Intra-slice arcs: $X_i^t -> X_j^t$. Inter-slice (temporal) arcs: $X_i^t -> X_i^{t+1}$ and $X_i^t -> X_j^{t+1}$.

structures: to represent all the conditional dependencies in time-series data, the network can become excessively complex. Also, the inference algorithms for DBNs involve heuristics such as beam search [Shi et al., 2004], [Shi et al., 2006] or greedy pruning of candidates [Laxton et al., 2007] to explore the large search space: inferring the most likely observed sequence of events by computing the probability of future events at time t given the past observation at time $t - i$. The span time $t - (t - i)$ is typically large ([Darwiche, 2000]).

The main advantage of Bayesian Networks is that they are well adapted to model the uncertainty of the recognition by using probabilities. However, the main drawback is that they are not adapted to model temporal relations because time needs to be explicitly indicated.

One among the most popular approach able to model the uncertainty and recognize activities with temporal relations and in particular sequential events is the Hidden Markov Models described in the section below.

2.1.1.3 Hidden Markov Models

Hidden Markov Models (HMMs) have been used to model uncertainty of the observed environment and in particular, the uncertainty of temporal relations of events. HMMs and their extensions have been most widely used among probabilistic approach for activity recognition, motivated primarily by their successful use in speech recognition.

The formal definition of a HMM model λ is specified by the tuple (Q, O, A, B, π) as follows ([Blunsom, 2004]):

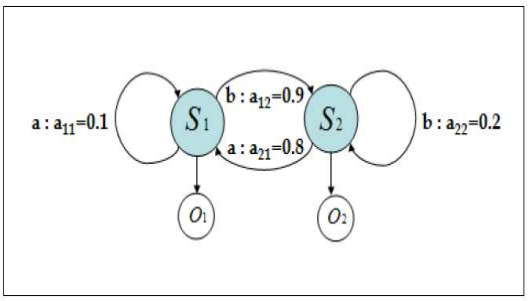


Figure 2.7: A HMM model λ is specified by the tuple (Q, O, A, B, π) where, $Q = \{S_1, S_2\}$ is the set of possible states, O is the set of observation symbols, A is the state transition probability matrix $(a_{ij} = P(q_{t+1} = j|q_t = i))$, B is the observation probability distribution $(b_j(k) = P(o_t = k|q_t = j))$ and π is the initial state distribution.

$$\lambda = (Q, O, A, B, \Pi) \tag{2.8}$$

Q is defined to be a fixed state sequence of length T, and corresponding observations O:

$$Q = q_1, q_2, ..., q_T \tag{2.9}$$

$$O = o_1, o_2, ..., o_T \tag{2.10}$$

S is the state alphabet set, and V is the observation alphabet set:

$$S = (s_1, s_2, ..., s_N) \tag{2.11}$$

$$V = (v_1, v_2, ..., v_N) \tag{2.12}$$

A is a transition array, storing the probability of state j following state i . Note the state transition probabilities are independent of time:

$$A = [a_{ij}]; a_{ij} = P(q_t = s_j|q_{t-1} = s_i) \tag{2.13}$$

B is the observation array, storing the probability of observation k being produced from the state i, independent of t:

$$B = [b_i(k)]; b_i(k) = P(o_t = v_k|q_t = s_i) \tag{2.14}$$

Π is the initial probability array:

$$\Pi = [\Pi_i]; \Pi_i = P(q_1 = s_i) \tag{2.15}$$

Two assumptions are made by the model. The first, called the Markov assumption, states that the current state is only dependent on the previous state, this represents the memory of the model:

$$P(q_t|q_{t-1}, q_{t-2}, ..., q_1) = P(q_t|q_{t-1}) \tag{2.16}$$

The independence assumption states that the output observation at time t is only dependent on the current state, it is independent of previous observations and states:

$$P(o_t|o_{t-1}, ..., o_1, q_t, ..., q_1) = P(o_t|q_t) \tag{2.17}$$

Given a HMM λ and a sequence of observations O, the probability of the observations O for a specific state sequence Q of length T is:

$$P(O|Q, \lambda) = \prod_{T}^{t=1} P(o_t|q_t, \lambda) = b_{q_1}(o_1) \times b_{q_2}(o_2)... \times b_{q_T}(o_T) \tag{2.18}$$

and the probability of the state sequence is:

$$P(Q|\lambda) = \Pi_{q_1} \times a_{q_1 q_2} \times a_{q_2 q_3} \times ... \times a_{q_{T-1} q_T} \tag{2.19}$$

Thus we can calculate the probability of the observations given the model as:

$$P(O|\lambda) = \sum_{Q} P(O|Q, \lambda).P(Q|\lambda) = \sum_{q_1 ... q_T} \Pi_{q_1} \times b_{q_1}(o_1) \times a_{q_1 q_2} ... a_{q_{T-1} q_T} \times b_{q_T}(o_T) \tag{2.20}$$

HMMs is among the most popular probabilistic approach for activity recognition[Oliver et al., 2002], [Duong et al., 2005]. HMMs have been widely used for the recognition of activities based on trajectories [Cuntoor et al., 2005] and motion [Achard et al., 2008].

Layered HMMs [Oliver et al., 2002] have been proposed to model events such as interaction between multiple mobile objects. In [Rosario et al., 2000] the authors have constructed a variant of the basic HMM, the coupled HMM (CHMM), to model human-human interactions. More specifically, they have coupled the hidden states of two different HMMs by specifying their dependencies. As a result, their system was able to recognize complex interactions between two persons, such as the concatenation of two persons meeting and walking together. However, they deal only with sequential events. Several studies used HMM for sport activity recognition. In [Chiang et al., 2007] the authors have studied baseball in order to determine the successful strikes. Their approach is to extract key frames from the baseball video. They

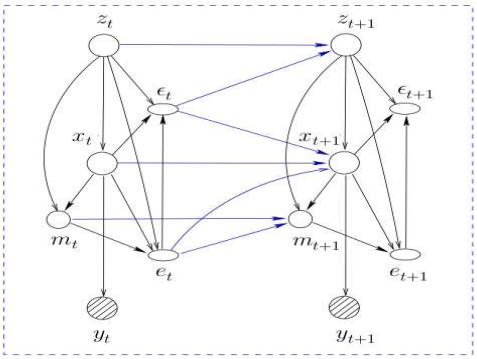


Figure 2.8: DBN representation of S-HSMM for two time slices [Duong et al., 2005].

have used an estimate of the global movement, size and position of the objects calculated from key images. The created Hidden Markov Model seeks to recognize events. However, they can only recognize four types of event (e.g. Base hit, Ground outs, Air outs, Strikeout) which are simple and domain specific. In [Duong et al., 2005] the authors introduce the switching Hidden Semi-Markov Model (S-HSMM) to deal with time duration modeling. Activities are modeled in the S-HSMM in two ways: the bottom layer represents atomic activities and their duration using HSMMs; the top layer represents a sequence of high-level activities where each high-level activity is made of a sequence of atomic activities. This extension attempts to introduce more semantic in the formalism at the cost of tractability: the extension complicates the structure of the graph model (i.e. additional relations between nodes are inserted) which can quickly leads computation in difficulties especially for graph with large number of nodes and time slices (fig.2.8). Figure 2.8 shows a DBN representation of the S-HSMM for two time slices. At each time slice t, a set of variables $Vt = \{z_t, t, x_t, e_t, m_t, y_t\}$ is maintained. At the top level, z_t is the current top-level state acting as a switching variable; t is a boolean-valued variable set to 1 when the z_t-initiated semi-Markov sequence ends at the current time slice. At the bottom level, x_t is the current child state in the z_t-initiated semi-Markov sequence; e_t is a boolean-valued variable set to 1 when x_t reaches the end of its duration. m_t represents the current phase of x_t . y_t is the observed alphabet.

The advantage of HMMs compared to Neural Network and Bayesian Network is the ability to recognize sequence of events. However, they are limited in recognising sequence of events which involves several mobile objects because the probability of being in a state for a mobile object has to be combined with the probability of being in another state for all other mobile objects. These combination leads the recognition process to a combinatorial explosion.

2.1.2 Description-based Approaches

Description-based approaches have also been largely used to recognize activities for few decades. A description-based approach recognizes activities with spatio-temporal structures. The main trend consists in designing symbolic network whose nodes correspond to boolean recognition of simpler events. Nodes represent a high-level activity describing the temporal, spatial, and logical relationships of simpler activities (i.e. sub-events). That is, description-based approaches model an activity as an occurrence of its sub-events that satisfy certain relations. Therefore, the recognition of the activity is performed by searching the sub-events satisfying the relations specified in its representation. All description-based approaches are inherently hierarchical (since they use sub-events to represent activities), and they are able to handle activities with concurrent structures. In description-based approaches, a time interval is usually associated with an occurring sub-event to specify necessary temporal relationships among sub-events. Allen's temporal predicates [Allen, 1983], [Allen and Ferguson, 1994] have been widely adopted for these approaches to specify relationships between time intervals [Pinhanez and Bobick, 1998], [Siskind, 2001], [Nevatia et al., 2003]. 13 basic predicates that Allen has defined for instance: before, meet, during, starts, overlaps and finishes. Note that the predicates before and meets describe sequential relationships while the other predicates are used to specify concurrent relationships (Figure. 2.9).

The first description-based approaches have been developed in the 70s and include plan recognition and event calculus. However, these approaches have not been applied to scene understanding. Other approaches include grammars that have been proposed to parse simple actions recognized by vision modules [Ivanov et al., 2005]. Logic and Prolog programming have also been used to recognize activities defined as predicates [Davis and Shet, 2005]. Chronicle recognition [Dousson and Ghallab, 1993] and Constraint resolution [Thonnat and Rota, 1999], [Rota and Thonnat, 2000] have also been used. Constraint Satisfaction Problem (CSP) has been applied to model activities as constraint networks [Reddy et al., 2009].

2.1.2.1 Constraint-based approaches

Early work in constraint recognition introduces the notion of chronicles, undirected constraint graphs describing the temporal constraints of atomic sub-events [Ghallab, 1996]. Techniques which are based on constraint resolution are among the most sophisticated event recognition techniques to date. Rota and Thonnat [Rota and Thonnat, 2000] represented an activity (i.e. scenario) by a set of positive (+) and/or negative (-) variables corresponding (at each instant t) to the detection of individuals, equipment, instantaneous recognised events. A positive variable ($x_i : +$) corresponds to an expected object/event, whereas, a negative variable ($x_i : -$) corresponds to an object/event that is not allowed to occur during the recognition of a

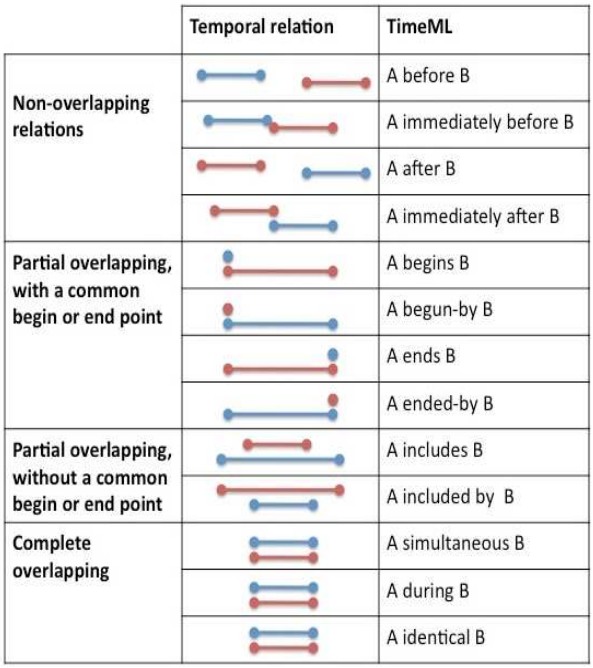


Figure 2.9: Allen's Temporal relations.

given activity. These variables are linked by a set of conditions (c_j) corresponding to temporal constraints and/or non-temporal constraints. Each constraint is a boolean predicate involving these variables. A constraint is called negative constraint if it involves at least one negative variable, otherwise, it is called a positive constraint. Figure 2.10 show an example of 'a person is far from an equipment' scenario represented by Rota and Thonnat [Rota and Thonnat, 2000]. The authors define the recognition problem as a boolean problem:

$$P_0(M, A, F) \tag{2.21}$$

where F is a set of facts, A is an ordered subset $\{f_1, ...f_k\}$ of F and M is an event model. A fact is a structured object defined by seven sets of attributes: name, type, date, geometry, velocity, properties and reference (see fig.2.10 and fig.2.11).

$A = \{f_1, ...f_k\}$ is a solution of P_0 if and only if:

$$\exists x_{k+1} \in F, ..., \exists x_n \in F \;\; c_j(f_1, ...f_k) = true \;\; \forall j \in \{1, ...p\} \tag{2.22}$$

A scenario is recognized if all its positive constraints are satisfied and its negative constraints are not satisfied.

The authors in [Dechter et al., 1991] have presented the activity recognition as a temporal constraint satisfaction problem. Their framework involves a set of variables $X = \{X_1, ...X_n\}$ and

Variables: $x_1 : +,\ x_2 : +$
Conditions:

$$\begin{cases} type(x_1) = person \\ type(x_2) = equipment \\ distance(x_1, x_2) \geq \alpha_{is\ far\ from} \end{cases}$$

Production: x_3

$$\begin{cases} name(x_3) = is\ far\ from \\ type(x_3) = state \\ date(x_3) = date(x_1) \\ reference(x_3) = (name(x_1), name(x_2)) \end{cases}$$

Figure 2.10: An example of 'a person is far from an equipment' scenario represented by Rota and Thonnat [Rota and Thonnat, 2000].

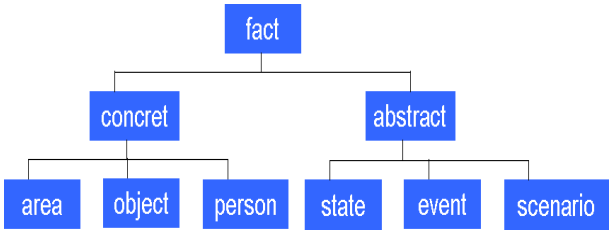


Figure 2.11: Hierarchy of facts[Rota and Thonnat, 2000].

a set of constraints. Each variables presents a time point and each constraint is presented as a set of intervals:

$$\{I_1, ..., I_n\} = [a_1, b_1], ..., [a_n, b_n] \tag{2.23}$$

A unary constraint on a variable X_i restricts the value of X_i to a set of intervall; namely, it respresents the disjunction:

$$(a_1 \leq X_i \leq b_1) \vee ... \vee (a_n \leq X_i \leq b_n) \tag{2.24}$$

A binary constraint on two variables X_i and X_j respresents the permissive value of the distance $X_j - X_i$; it represent the disjunction:

$$(a_1 \leq X_j - X_i \leq b_1) \vee ... \vee (a_n \leq X_j - X_i \leq b_n) \tag{2.25}$$

A network of binary constraints consists on a set of variables, $X_1, ..., X_n$ and a set of unary and binary constraints. It is represented by a 'temporal constraint graph', where nodes respresent

variables and edges correspond to constraints. The edges are labelled by sets of intervalls. A tuple $X = (x_1, ...x_n)$ is called a solution if the assignment $\{X_1 = x_1, ...X_n = x_n\}$ satisfies all the constraints. A value v is a feasible value for variable X_i if there is a solution in which $X_i = v$. Figure 2.12 is an ilustration of a constraint satisfaction graph.

Example . John goes to work either by car (30–40 minutes), or by bus (at least 60 minutes). Fred goes to work either by car (20–30 minutes), or in a carpool (40–50 minutes). Today John left home between 7:10 and 7:20, and Fred arrived at work between 8:00 and 8:10. We also know that John arrived at work about 10–20 minutes after Fred left home. We wish to answer queries such as: "Is the information in the story consistent?", "Is it possible that John took the bus, and Fred used the carpool?", "What are the possible times at which Fred left home?", and so on.

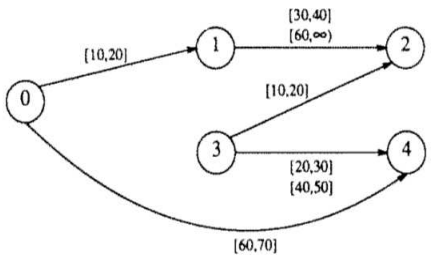


Figure 2.12: A constraint satisfaction problem model and the corresponding graph[Dechter et al., 1991]. This graph involves five variables: x_0, the starting time of the problem, the choosen value is 7:00 am; x_1, x_2:are respectively the time when John left home and arrived at work and x_3, x_4 are respectively the time when Fred left home and arrived at work. There are five constraints involving to the five variables corresponding to the time duration that each person has to take for going to work.

Chleq and Thonnat [Chleq and Thonnat, 1996] have represented a scenario as a set of independant positive/negative instantaneous events. The events composing a scenario are relating by temporal constraints. The algorithm recognize incrementally pre-defined scenarios representing human behaviors in the observed scene. For each scenario, a graph is built: the vertices respresent the time point variables and the edges correspond to the temporal relations.

In [Vu et al., 2002], [Vu et al., 2003a], the authors first propose a language to describe scenario models and second a temporal constraint resolution approach to recognize in real-time scenario occurrences. They represent a scenario model with the list of the actors involved in the scenario and a set of constraints on these actors. An actor can be a person detected as a mobile object by the recognition process or a static object of the observed environment. A person is represented by his/her characteristics: his/her position in the observed environment,

width, velocity, etc. A static object of the environment is defined as a priori knowledge and can be either a zone of interest (a 2D polygonal as the entrance zone) or a piece of equipment (a 3D object as a desk) (fig.2.13, fig.2.14).

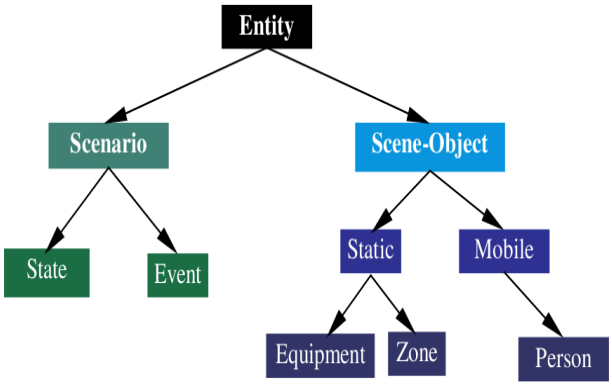


Figure 2.13: Five types of entities are classified into 'scenario' and 'scene-object' in [Vu et al., 2002].

```
State(close_to,
   Actors((p : Person), (eq : Equipment) )
   Constraints((distance(p, eq) ≤ Close_Distance))
   Production((s : State)(Name = "close_to")) )
```

Figure 2.14: An example of the model close to: a person is close to an equipment[Vu et al., 2002].

The event recognition process consists in mapping the set of constraints to a temporal constraints Network and determining whether the video sequence satisfies these constraints (Figures. 2.15, 2.16). The first step of the event recognition process is to compute which event (i.e. scenarios) can be recognized at the current time. They call 'trigger' such a scenario which can be recognized. Once they have recognized a scenario they add to the list of triggers all the more complex scenarios which are ended with the given recognized scenario. For example, once they have recognized the scenario '$close - to$', it is possible that the scenario '$moves - close - to$' would be recognized. To recognize a given scenario the algorithm selects an actor for each actor variable and check whether the selected actors satisfy the constraints defined within the scenario model.

The algorithms proposed in the literature to solve the recognition problem are generally computationally intractable (NP-hard). To overcome this problem, Vu et al. [Vu et al., 2003a], [Vu et al., 2003b] have proposed a scenario model pre-compilation step and achieve a speed up of the algorithm that allows it to be used in real time surveillance applications. This step consists in decomposing the event model into a set of simple event models containing at most two sub-events. However, their approach of recognition is deterministic.

Figure 2.15: Four step of 'Bank attack' scenario [Vu et al., 2004]: (1) at time t_1, the employee is at his position behind the counter, (2) at time t_2, the robber enters by the entrance and the employee is still at his position, (3) at time t_3, the robber moves to the front of the counter and the emplyee is still at his position and (4), at time t_4, both of them arrive to the safe door.

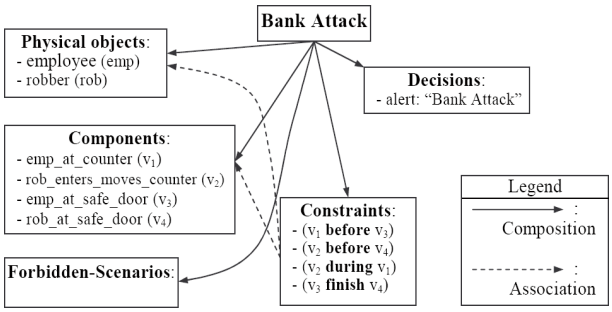


Figure 2.16: Hierarchical model of the 'Bank attack'[Vu et al., 2004]. This scenario is composed of five parts: a set of physical object corresponding to the employee and the robber, a set of temporal variable (components), an empty set of forbidden scenarios, a set of constraints ('before', 'during' and 'finish' [Allen, 1983]) and an alert.

The advantage of constraint-based approaches is that knowledge can be formulated as an ontology for a particular activity domain and easily reused for different application domains [Thonnat and Rota, 1999]. They allow the description of events in a natural language, declarative and easily to be defined and modified by human experts. One drawback is that the event modeling step often needs a strong effort to describe all the events of interest. One solution to solve this problem is to re-use existent ontology and also use learning technics to learn automatically event models. Another limitation is that these approaches do not cope with the uncertainty of recognition.

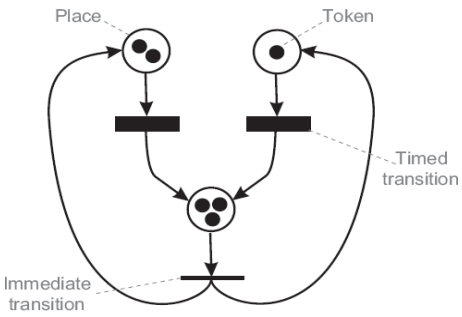


Figure 2.17: An example of a Petri Net [Haas, 2002].

2.1.2.2 Petri Nets

Petri Nets can be graphically represented by a bipartite graph (figure. 2.17) which includes two types of nodes: the places P, which are drawn as circles and the transitions T, which are drawn as bars. Places can contain tokens that are drawn as black dots within places. The state of a Petri-Net is called marking, and is defined by the number of tokens in each place. A transition node is enabled if all input place nodes to that transition have tokens. The possible relationships between temporal intervals (sequential and non-sequential relationships), defined by Allen, are used to describe the relationships between sub-events within the same event.

Petri nets specify the temporal ordering of sub-events in terms of a graph representation. Several researchers have utilized Petri nets to represent and recognize human activities [Lavee et al., 2007], [Perse et al., 2010]. Zaidi [Zaidi, 1999] has showed that Petri nets are able to fully represent temporal relationships described by Allen's temporal predicates. Ghanem et al. [Ghanem et al., 2004] have taken advantage of Petri nets to represent and recognize interactions between humans and vehicles. The authors in [Perse et al., 2010] have developed a method for automatic trajectory-based analysis of multi-agent activities in the sport domain. These researches however, do not take into account the uncertainty inherent in both low level observations and semantic definitions of events.

The advantage of Petri Nets is their capacity to express sequencing and parallel events occurring. A disadvantage of Petri Nets is their deterministic nature, this formalism relies on perfect input data which is unrealistic for real world applications. Recent work aiming at extending this formalism with a convenient probabilistic mechanism to handle the uncertainties will be discussed in section (2.1.4)

2.1.3 Discussion: Description-based approaches vs. Probabilistic approaches

Description-based approaches are suitable for recognizing high-level activities which can be hierarchically described using simpler sub-events. They can more easily incorporate human

knowledge into the systems and require less training data as pointed out by many researchers [Oliver et al., 2002], [Nevatia et al., 2003], [Ryoo and Aggarwal, 2006].

While Description-based approaches allow complex event modeling with complex spatio-temporal constraints, the formalism of the event model is largely deterministic and convenient mechanism to handle uncertainty (both observation and interpretation) is generally unavailable.

Probabilistic approaches provide a probabilistic framework for reliable recognition with noisy inputs. However, they have difficulties representing and recognizing activities with concurrently organized sub-events especially for HMMs which were mostly used to recognize sequential events. But, Petri Nets are well suited for modeling the concurrent composition of events. Description-based approaches are able to represent and recognize human activities with complex temporal constraints. Not only sequentially occurring, but also concurrent organized sub-events using concurrent Allen's temporal predicates. The major drawback of description-based approaches are their inability to compensate for the failures of low-level components (e.g. detection failure).

The probabilistic approaches have the advantage of modeling the uncertainty. They accurately learn event models from training data achieving a high precision within a domain and allowing an intrinsic handling of low level analysis uncertainty. However, most of these methods mainly focus on a specific human activity and their descriptions are not declarative and it is relatively difficult to modify them or add a priori knowledge. The a priori probability needs to be learned and this learning stage is often tiresome due to the construction of the learning set. Additionally, the amount of training data increases with the number of states and observation. For instance, Dynamic Bayesian Networks (DBN) have been used successfully to recognize short temporal actions [Reddy et al., 2003], but the recognition process depends on time segmentation: when the frame-rate or the activity duration changes, the DBN has to be re-trained. The advantage of the HMMs compared to NNs and Bayesian classifier is the ability to recognize sequences of events, however, they are restricted to simple and sequential temporal patterns. Additionally, they are limited when the recognition involves several mobile objects. The probability of being in a state for a mobile object has to be combined with the probability of being in another state for all other mobile objects. This combination leads to combinatorial explosion in the recognition process. Consider the problem of modeling the movement of several objects in a sequence of images. There are M objects, each of which can occupy K positions and orientations in the image, there are K^M possible states of the system underlying an image. A HMM would require K^M distinct states to model this system. This representation is not only inefficient but difficult to interpret. A summary of the advantages of each approach is provided in the table 2.1:

Approaches	Advantages and Drawbacks	References
Probabilistic Approaches		
NNs	+ uncertainty handling - difficulties to cope with temporal relations	[Howell and Buxton, 2001], [Chen and Zhang, 2007]
BNs	+ uncertainty handling - difficulties to process temporal relations	[Nevatia et al., 2004]
HMMs	+ uncertainty handling + can process sequences of events - tedious learning phase - can not cope with complex temporal relations	[Cuntoor et al., 2005], [Achard et al., 2008]
Description-based Approaches		
Constraint-based app	+ Complex event recognition - difficulties to handle the uncertainty	[Ghallab, 1996], [Vu et al., 2003a], [Rota and Thonnat, 2000]
Standard PNs	+ conccurant temporal relationships - deterministic recognition - can not process forbidden events - combinatorial explosion in case of multi-physical objects problem	[Ghanem et al., 2004], [Lavee et al., 2007]

Table 2.1: Advantages and drawbacks of probabilistic and description-based approaches.

2.1.4 Combination of Description-based Approaches and Probabilistic Approaches

The approaches combining logic and probabilistic reasoning have been designed to overcome the limitations of the previous approaches. The logic-probabilistic combination is an interesting field of research in the recent years as a consequence of the above-mentioned limitations ([Ryoo and Aggarwal, 2010], [SanMiguel and Martinez, 2012]). However, it has not been fully explored and many efforts are still needed to provide a complete framework for fully handling the uncertainty of recognition. Ryoo and Aggarwal [Ryoo and Aggarwal, 2006] have first proposed a deterministic description-based approach using a context free grammar (CFG) as a syntax of their representation language to represent composite actions and interactions. Furthermore, Ryoo and Aggarwal [Ryoo and Aggarwal, 2007] , [Ryoo and Aggarwal, 2009], [Ryoo and Aggarwal, 2010] have proposed a probabilistic extension of their recognition framework to deal with uncertainty (fig.2.18). In order to compensate for the complete failure of the atomic-level components, they have taken advantage of the concept of the hallucinated time intervals inserted in the place of missing gestures with an extremely low probability, similar to the one used in [Minnen et al., 2003]. 'Hallucinations' are time intervals which are inserted

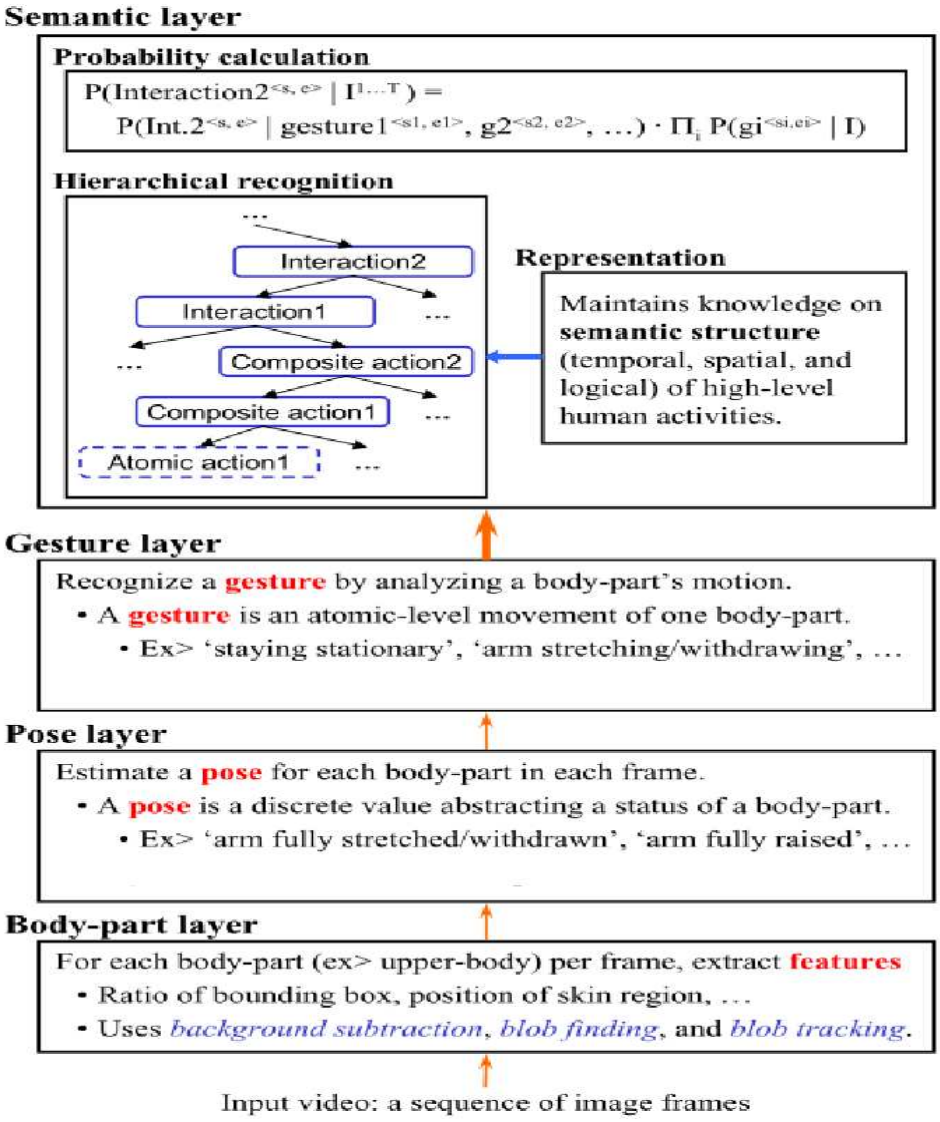


Figure 2.18: Figure illustrating the overall framework of the system proposed in[Ryoo and Aggarwal, 2009].

regardless the gesture recognition results. However, to calculate the probability of a high level activity, the authors have used the dependency information between the activity and its sub-events. The hierarchy tree (fig.2.19) illustrates the dependencies among the activities. They compute the probability of an occurring activity R in one time interval $[s, e]$, given the sequence of images, $P(R^{[s,e]}|I^T)$, I^T represents the images from frame 1 to frame T.

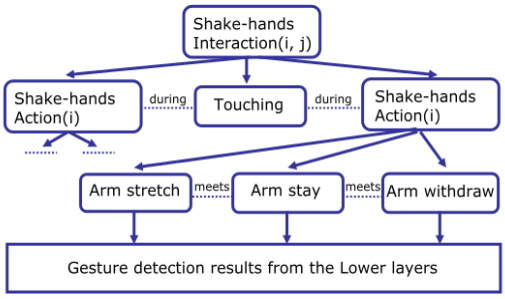


Figure 2.19: Example of recognition process tree of the interaction 'shake-hands' [Ryoo and Aggarwal, 2009].

$$\begin{aligned} P(R^{[s,e]} \mid I^T) &= P(\{R\} \mid sub(\{R\})) \times P(sub(\{R\}) \mid sub(sub(\{R\}))) \\ &\quad \times \ldots \times P(sub^d(\{R\}) \mid I^T) \\ &= \prod_{i=0}^{d-1} P(sub^i(\{R\}) \mid sub^{i+1}(\{R\})) \times P(sub^d(\{R\}) \mid I^T) \\ &= \prod_{i=0}^{d-1} P(sub^i(\{R\}) \mid sub^{i+1}(\{R\})) \times P(a_1, \ldots, a_n \mid I^T) \end{aligned} \tag{2.26}$$

Where, $sub(\{R\})$ is the union of sub-events of the activity R, $a_1, \ldots, a_n$ are leaf nodes (i.e. atomic action) of the tree and d the depth of the tree.

The first limitation lies in the concept of the hallucinated time intervals, this can lead to high rate of false positive recognition, first in the atomic events level and thus false position recognition of more complex events. The seconde limitation of the work relies in the method adopted to compute the probability of an activity which can results to a very low probability value especially if the sub-events of the activity include gestures detected with low probability value (e.g. 'hallucinated' gestures). Finally, another limitation of the work as noticed by the author ([Ryoo and Aggarwal, 2009]) is the time complexity to find optimum recognition solution: it is an NP-hard problem.

Tran and Davis [Tran and Davis, 2008] have combined logic and probabilistic methods to recognize events (fig. 2.20). The authors have represented the knowledge as first-order logic production rules and have successfully applied the probabilistic graphical model, Markov logic networks (MLNs) to probabilistically infer events in a parking lot. Markov logic networks (MLNs) are one type of the unrolled graphical models developed to combine logical and probabilistic reasoning. In MLNs, every logic formula F_i is associated a real-valued weight. Each instantiation of F_i is given the same weight. An undirected network called a Markov Network is constructed such that:

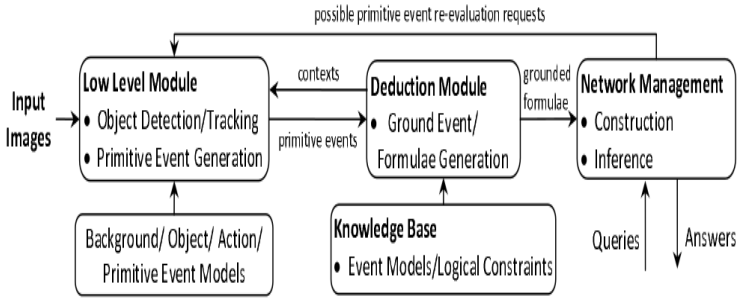


Figure 2.20: Overview of the system proposed in[Tran and Davis, 2008].

- Each of its nodes corresponds to a ground atom (i.e. atomic event),
- If a subset of ground atoms $x_{\{i\}}$ are related to each other by a formula F_i, then a clique C_i is added to the network. C_i is associated with a weight ω_i and a feature f_i defined as follow:

$$\begin{aligned} f_i(x_{\{i\}}) &= 1, \ \ if \ \ F_i(x_{\{i\}}) \ \ is \ \ true \\ &= 0, \ \ otherwise. \end{aligned} \tag{2.27}$$

This network models the joint distribution of the set of all ground atoms. It provides a mean for performing probabilistic inference, the probability of an event given some evidences (observations) is then computed. The probability that a ground atom X_i is equal to x_i given its Markov blanket (neighbours) B_i is [Tran and Davis, 2008]:

$$p(X_i = x_i) = \frac{exp(\sum_{f_j \in F_i} \omega_j f_j(X_i = x_i, B_i = b_i)}{exp(\sum_{f_j \in F_i} \omega_j f_j(X_i = 0, B_i = b_i) + \sum_{f_i \in F_i} \omega_j f_j(X_i = 1, B_i = b_i))} \tag{2.28}$$

Where F_i is the set of all cliques that contain X_i and f_j is computed as in equation (2.27).

However, even though the above-mentioned approach attempts to integrate logical-based inference into a probabilistic framework, it is limited in recognizing complex high level activities because Markov logic networks (MLNs) make an inference using binary predicates without inferring the occurring time of the activity being recognized. MLNs provides the same weight to all the instantiations of a formula (or event model) which it is not realistic for real world applications as the probability of an event instantiation can change from an observation to another. Markov logic network relies on the assumption that an identical sub-event occurs only once during interactions, limiting itself from being applied to dynamically interacting actors (dynamic interaction can contain two or more instantiation of the same sub-event). Working

with Markov logic typically makes three assumptions about the logical representation: different constants refer to different objects (unique names), the only objects in the domain are those representable using the constant and function symbols (domain closure), and the value of each function is always a known constant (known functions). These assumptions ensure that the number of possible worlds is finite and that the Markov logic network will give a well-defined probability distribution.

The probabilistic extensions of Petri Nets have been proposed for uncertain state ordering [Albanese et al., 2008] and duration [Lavee et al., 2010b]. However, they do not consider the uncertainty of the low level analysis. In [Lavee et al., 2010a] authors propose a probabilistic mechanism for recognizing activities modeled as Petri Nets. More formally, an activity Petri Net is a tuple:

$$< P, T, C, events, \delta, S, F > \tag{2.29}$$

Where,
- P is the set of places,
- T is the set of transitions,
- C is the set of connecting arcs,
- $events$ is the set of events that are relevant to the activity.
- $\delta : T \rightarrow events$ is a labeling function mapping transitions to an event label.
- $S \subset P$ is the place node representing the 'start' of the activity and
- $F \subset P$ is the set of place nodes representing the recognition of the activity.

By the mean of Petri Net, the authors calculate the transition and observation probability for each state $a \in A$ (A is the set of all reachable states from the initial state). The normalized transition probability is computed as follow:

$$P(x_t = x'|x_{t-1} = a) = \frac{\hat{P}(x_t = x'|x_{t-1} = a)}{\sum_{x^* \in A} \hat{P}(x_t = x^*|x_{t-1} = a)} \tag{2.30}$$

- The non-normalized probability of staying in the same place $\hat{P}(x_t = a|x_{t-1} = a) = 1$,
- $\hat{P}(x_t = b|x_{t-1} = a) = (\alpha)^\tau$, where τ is the minimum distance between $a \in A$ and $b \in A$; $\alpha \in [0, 1]$ is a parameter.
- Transition to some $c \in A$ that is not reachable from a is given some small non-zero probability ϵ, $\hat{P}(x_t = c|x_{t-1} = a) = \epsilon$.

The observation probability is computed in the same way:

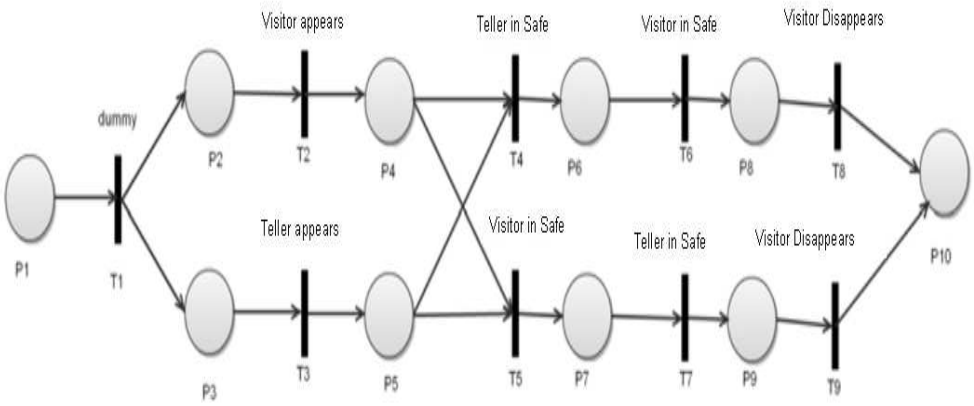


Figure 2.21: Activity Petri Net Describing 'Bank Attack'[Lavee et al., 2010a]. $P = \{P_1, ..., P_8\}$, $T = T_1, ..., T_9$, C, is the set of connecting arcs depicted in the figure, $events$ = {'Visitor Appears', 'Teller Appears', 'Teller in Safe','Visitor in Safe', 'Visitor Disappears'} , $\delta(T_2)$ = 'Visitor Appears' , $\delta(T_3)$ = 'Teller Appears', $\delta(T_4)$= 'Teller in Safe', $\delta(T_5)$= 'Visitor in Safe', $\delta(T_6)$= 'Visitor in Safe', $\delta(T_7)$= 'Teller in Safe', $\delta(T_8)$= 'Visitor Disappears', $\delta(T_9)$= Teller Disappears', $S = \{P1\}$ and $F = \{P10\}$.

$$P(y_t = q|x_t = a) = \frac{\hat{P}(y_t = q|x_t = a)}{\sum_{q' \in Y} \hat{P}(y_t = q'|x_t = a)} \tag{2.31}$$

- $\hat{P}(y_t = q|x_t = a) = 1$, if there is an enabled transition in making a with an event label q.
- $\hat{P}(y_t = q|x_t = a) = \epsilon$, otherwise.

The propagation of uncertainty is processed as follow: at each discrete time slice, there is a set of hypotheses on the state of activity which is called particles. The sum over the weight of all particles in a particular state is the confidence that they are currently in this state.

The main limitation of this work lies in the way that the authors propose to integrate the probability in the Petri Net: the authors have not described how they have calculated the different parameters of the algorithm. For example, to calculate the transition probability $\hat{P}(x_t = b|x_{t-1} = a) = (\alpha)^\tau$, the authors have not specified how the minimum distance τ is calculated and they have not detailed their choice of α, which value can deeply changes the final value of the probability. Another limitation is the non-modular aspect of the Petri Network, which makes difficult any modification of the structure of network.

In [Brendel et al., 2011] the authors have presented a probabilistic event logic (PEL) which uses weighted event-logic formulas to represent probabilistic constraints among events. The knowledge base is representing a set of weighted event-logic formulas:

$$\Sigma = \{(\phi_1, w_1), ...(\phi_n, w_n)\} \tag{2.32}$$

where w_i is a non-negative numeric weight associated with formulas ϕ_n. Σ assign a score S to any interpretation (X, Y). An interpretation is denoted by (X, Y), where X is the set of observable event occurrences, and Y is the set of hidden event occurrences.

$$S((X,Y),\Sigma) = \sum_i w_i.\ |SAT((X,Y),\phi_i)| \tag{2.33}$$

Where $|SAT((X,Y),\phi)|$ is the number of intervals in (X,Y) satisfied by ϕ. The posterior probability of the hidden part of interpretations is :

$$Pr(Y|X,\Sigma) \propto exp(S((X,Y),\Sigma)) \tag{2.34}$$

However, they authors have considered only the recognition of primitive events of basketball game and they have not proposed any approach to deal with low-level uncertainty.

Kwak et al. [Kwak et al., 2011] have adopted constraint flows to summarize the combination of the primitive events composing a complex event to infer the recognition. The constraint flow is traced to search for the feasible interpretation. They propose the use of an objective function to calculate the optimum interpretation by combining two measures:
(i) the '*observation agreement*' formulated by the posterior probability of an interpretation $P(x(v_{1:t})|O_{1:t})$, $v_{1:t}$ and $x(v_{1:t})$ are respectively a tracing result up to the t^{th} time step and a feasible interpretation. O_t is observation at time step t.

$$P(x(v_{1:t})|O_{1:t}) \propto P(O_t|x(v_t)) \times \ P(x(v_t)|x(v_{1:t-1})) \times \ P(x(v_{1:t-1})|O_{1:t-1}) \tag{2.35}$$

Based on the assumption that the primitive events occurs independently of the others, the conditional observation probability in equation (2.35) have been calculated as the following:

$$P(O_t|x(v_t)) = \prod_{i=1}^{n} P(O_t^i|x(v_t)) = \prod_{i=1}^{n} P(O_t^i|x(v_t^i)) \tag{2.36}$$

where n is the number of primitives of the scenario. The state transition probability $P(x(v_t)|x(v_{1:t-1}))$ is assumed to be uniform and equal to 2^{-n}.

(ii) and the '*degree of blankness*' $B(v_{1:t})$ measure cohesion among time intervals: if no primitive event is recognized while the composite event is in progress then the time step is defined as *blank*. A *blank* time interval is a set of *blank* time steps arranged consecutively. the '*degree of blankness*' is considered as a factor of penalty during the computation of the probability.

$$B(v_{1:t}) = \prod_{j=1}^{t} exp(-\beta I_b(v_j)) \tag{2.37}$$

$$\begin{aligned} I_b(v_j) &= 1, \ \text{if } v_j \text{ is blank}, \\ &= 0, otherwise. \end{aligned} \tag{2.38}$$

The limitation of this work during the probability calculation step is, first, the authors have only considered during the computation of $P(x(v_{1:t})|O_{1:t})$ the probability of sub-events occurrences, they do not consider the probability of verification of the constraints (e.g. spatial and temporal constraints). They also have not described the calculation of the parameter β called 'the penalty weight for idle time steps' in equation (2.37). Finally, they authors have not explained why they have assumed a uniform value for the transition probability $P(x(v_t)|x(v_{1:t-1}))$, the value 2^{-n} risks to be very low, when n is high, which can affect the finale value of the probability $P(x(v_{1:t})|O_{1:t})$. Another limitation of this work is that the authors have not considered the low level uncertainty, they have a noisy primitive event detection as they have already mentioned in their work. Finally, the complexity of the recognition algorithm is not described.

Combining logic and probabilistic reasoning is an interesting field of research. The main trend is to integrate a probability reasoning in description-based approaches or to add more logic in probabilistic approaches. The authors in [Tran and Davis, 2008] have attempted to integrate logical-based inference into a probabilistic graphical model, Markov logic networks (MLNs) to probabilistically infer events in a parking lot. In [Ryoo and Aggarwal, 2010], the authors have proposed a probabilistic extension of their description-based recognition framework to deal with uncertainty. The approach we propose is described in chapter 5, fits into this category of approaches that take advantages of the probabilistic methods as well as the representation of event semantics. A summary of the advantages and drawbacks of each approach is provided in the table 2.2.

Reference	Advantages	Drawbacks
[Ryoo and Aggarwal, 2009]	+ High level uncertainty handling. + Low level uncertainty handling. + High level activity recognition.	- High complexity to find optimum recognition solution. - High rate of false positive due to the concept of hallucinated time intervals. - Probability computation value can be very low.
[Tran and Davis, 2008]	+ Recognition uncertainty handling. + Event modeling uncertainty handling.	- Limitation from being applied to dynamic interactions. - Limitated when recognizing high level activity .
[Albanese et al., 2008]	+ Handle the uncertainty of high level activities. + Complex activity detection.	- Complexity of the algorithm not described. - No handling of low-level processing uncertainties.
[Lavee et al., 2010a]	+ Handle the uncertainty of high level activities. + Complex activity detection.	- Probability integration not well defined - Non modular Network structure
[Brendel et al., 2011]	+ Event modelling uncertainty. + Event recognition uncertainty.	- No handling of low-level processing. - Limited to primitive event recognition.
[Kwak et al., 2011]	+ Probabilistic recognition + Complex activity recognition.	- Recognition algorithm complexity not described - Parameter of the probability calculation not well described

Table 2.2: Advantages and drawbacks of the approaches combining probability and logic.

2.2 Health care Monitoring

Healthcare technology for the elderly is a popular area of research. According to the European Union commission's projection, the number of elderly will increase three-fold between 2008 and 2060 [Cardinaux et al., 2011]. In France, the proportion of people aged 75 and over in the population (approximately 7% in 2000) should reach nearly 10 % in 2020. In future years, the difference between the needs of the dependent elderly and the number of places available in hospitals and in specialized centers will become even more important than it is currently. Recently, a number of researchers have developed solutions based on video cameras and computer vision systems with promising results. However, for the domain to reach maturity, several challenges need to be faced, including the development of systems that are robust in the real-world and are accepted by users, carers and society.

2.2.1 Automatic monitoring for health care

Over the last several years much effort has been put into developing and employing a variety of sensors to monitor activities in the health care domain. These sensors include camera networks for people tracking [Sidenbladh and Black, 2001], cameras and microphones for activity recognition [Clarkson et al., 1998], and embedded sensors for activity detection [Wang et al., 2007], [Zouba et al., 2009], [Biswas et al., 2010b].

The change in the manner of doing the activities of daily living is a good indicator of declining health, thus patient monitoring is receiving a big interest from medical experts. Video technology can be used in this context to recognise users' activities. Nait-Charif and McKenna [Nait-Charif and McKenna, 2004] have introduced a method to recognize activity by tracking the user's position using an overhead camera. They tracked a person using an ellipse and they infer the 'unusual inactivity' of a person when the targeted person is detected as inactive outside a normal zone of inactivity like sofa. Diet is also an important factor affecting health. Kim et al. [Kim et al., 2010] have proposed a design concept of an image-based dietary assessment tool implemented in a mobile phone. Food images are recorded using a built-in camera on a mobile phone and sent to a server where the portion size of the meal is estimated (fig.2.22). This information is subsequently used to keep personal dietary records as a means of monitoring energy and nutrient intakes. However many issues are not addressed in this work and the images recording step is very constrained as many recommendations are done to the users:
- full image acquisition: foods should be captured large enough and around the center area in the image so that they can easily be recognized during analysis.
- Highlight and shadow condition: Highlights and shadows hinder the automatic analysis by making the recognition of food areas difficult.
- Spatial consistency: When several foods exist in the images, keeping them at the same po-

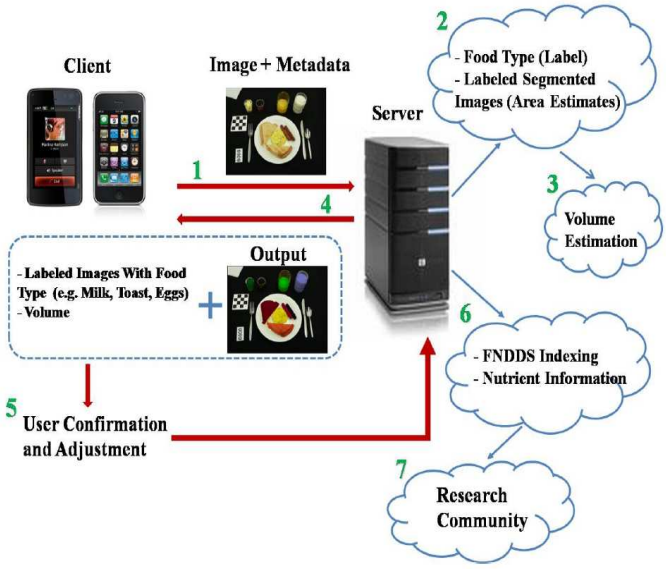


Figure 2.22: Overview of an image-based dietary assessment system in a server-client architecture [Kim et al., 2010].

sition helps the analysis part to automatically recognize foods in different images of the same eating occasion.
Second, users are very involved in the process of recognition as they are engaged to confirm and adjust the food tags in case that the analysis results contain errors.

Currently fall detection is a well researched topic and sensor based solutions are available and commercialized. The majority of such solutions are based on accelerometers, but one of the drawbacks of these systems is that users always need to wear the sensor. The system will not work if the user forgets to wear the device. Alternative solutions are suggested, such as passive fall detection which uses floor vibration sensors (fig.2.23), sound [Alwan et al., 2006] or video based monitoring [Foroughi et al., 2008]. In [Foroughi et al., 2008], the authors uses Neural Network for motion classification but still they recognize short actions (e.g. sit down) and they do not deal with complex behaviors involving temporal relation.

In [Biswas et al., 2010a] the authors discuss an approach toward building a system for assisting people with dementia in their home. They use the concept of micro-context which is information about objects and activities in a smart space, the information are generated through sensors (e.g. RFID, accelerometers) in the ambient environment. In further work [Biswasa et al., 2011], the authors present a design of a prototype to deploy a sensor network for obtaining micro-context information. The sensors incorporated in the network are: pressure sensors, RF Tags, Reed switches, acoustic, motion and inertial sensors. For activity recognition,

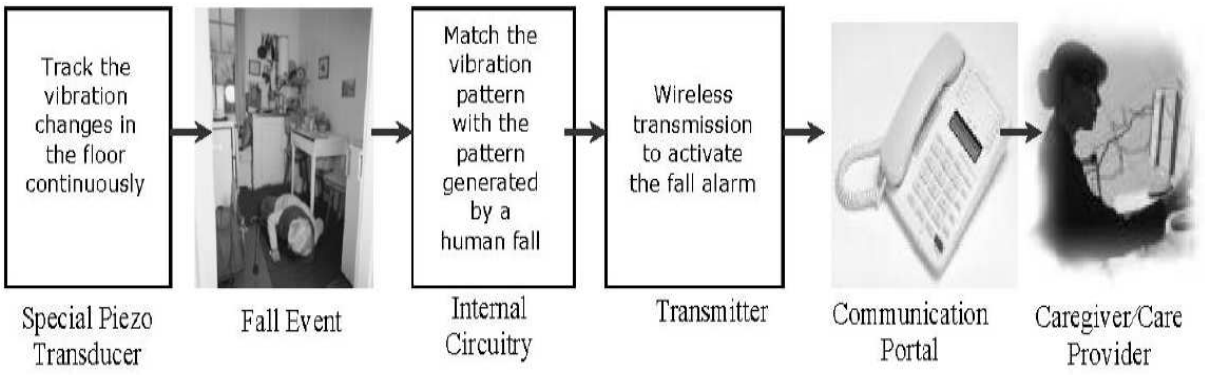


Figure 2.23: Schematic Representation of the Working Principle of the Floor Vibration Based Fall Detector [Alwan et al., 2006].

the authors adopt Dynamic Bayesian Networks to infer high level activities.

In [Tolstikov et al., 2008] the authors have worked on activity recognition and assistance of activities of daily living (ADLs) of elderly. The approach relies on multi-modal information fusion and primitive activity recognition at the low level. They have choosen the model of the Dynamic Bayesian Network (DBN) for the detection of activities. However, they only have been dealing with primitive activities which are very short in duration and they have not addressed high level activity recognition.

The information related to activity, physiological data and diet detected by the systems described above can be subsequently fed back to users to promote health behaviour changes or sent to carers. These technologies can be part of a mechanism to support self-management of people suffering from chronic problems such as cognitive deficiency.

Many health care applications are typically based on solutions employing sensors either embedded in the environment or body worn. The most widely used sensors for health care applications are reported in Table 2.3, along with their domains of application. To reach the level of analysis required for health care applications (e.g recognising activity), a large network of embedded sensors is generally required, therefore systems are usually costly to maintain, are relatively obtrusive (e.g. sensors set up on every cupboard door) and are highly sensitive to the performance of the sensors. Another approach is to use body worn sensors, however, it is recognised that user compliance with wearable systems is poor.

Recently a number of researchers and companies have been looking at developing solutions based on video cameras and computer vision approaches. A monitoring system based on video cameras has potential advantages. In principle a single camera in a room could pick up most of the activities performed in the room and, consequently, could replace a large number of sensors.

Remarkably, the last five years have shown a large interest in video based solutions. One

Sensor	Applications domains	Embedded/Body worn
Motion detector	Fall detection, Activity monitoring, Security	Embedded
Door open	Activity monitoring, Security	Embedded
Electrical appliance	Activity monitoring, Safety	Embedded
Microphone	Fall detection	Embedded
Accelerometer	Fall detection, Activity monitoring	Body worn
RFID	Activity monitoring	Body worn/Embedded
Temperature	Safety	Embedded
Smoke sensor	Safety	Embedded
Camera	**Fall detection, Activity monitoring, Security, Health care**	**Embedded**

Table 2.3: Commonly used sensors and their domains of application.

of the major reasons for this gain of interest is the cost of video cameras. The price of video cameras has drastically lowered within the last decade and good quality video cameras are now available at very low cost. Another important factor in favour of video technology is the maturity of computer vision based technologies. Latest research in computer vision, motivated by needs in domains such as security and surveillance, provides new means of interpreting rich information provided by video cameras.

2.2.2 Discussion and challenges

Many advances have been done for automatic monitoring for the health care domain. However some issues are still to be addressed fully such as dealing with occlusions in environments filled with furniture and various objects. Hardware solutions to this problem include the use of a overhead camera. Alternatively, the solution proposed in [Chen et al., 2007] makes use of multi-views of the scene by combining features from multiple cameras covering the same area.

Another challenge to be addressed is the handling of uncertainty of the recognition of activities.

Another challenge to be addressed is the multi-occupancy of the accommodation. While most health care applications are targeted for people living alone, they are still likely to receive visitors, therefore algorithms for identity recognition (e.g. face recognition) and tracking of multiple persons should be included in health care systems.

2.2.3 Acceptability and Privacy

Little evaluation research exists on user acceptance of automatic monitoring systems. There are only few studies that investigate elderly perception of these technologies or other home-

based technological applications. One of these studies that addresses this concept has been conducted by Vincent et al. [Vincent et al., 2002] who have examined the application of environmental control systems in the homes of users and nurses and concluded that the use of remote control by people with moderate cognitive impairments was difficult, while verbal reminders were greatly appreciated. Demiris et al. [Demiris et al., 2001] have investigated elderly perception of monitoring technology and found that the respondents have an overall positive attitude toward the use of these technologies. The findings from this study indicate that privacy can be a barrier for elderly to adopt these technologies; however their perception of their need for the technology may override their own privacy concerns.

Marquis-Faulkes et al. [Marquis-Faulkes et al., 2005] have presented a qualitative user requirement study for the use of video technology for activity monitoring. Four focus groups were conducted with a total of 37 participants. Three groups were composed of elderly with different levels of dependence and the last group comprised sheltered housing wardens. One outcome of the study was that the participants were generally satisfied to have a video based monitoring device provided that the video footage was analysed only by computer and that no one would actually watch them. The participants in the user group with the highest level of independence expressed concerns about carers having access to too much information about their activities but were in favour of a system that would alert carers in case of fall or detection of abnormal activity.

Turgeon Londei et al. [Londei et al., 2009] have explored the receptivity and requirements of elderly regarding the introduction of an intelligent video monitoring system in Canada. Interviews were conducted with 25 elderly with a history of falls. Both qualitative and quantitative data was reported and analysed. The perception towards video monitoring was generally good with 96% of the participants at least partially favorable to intelligent video monitoring system and 88% accepted the sending of an image to a designated carer.

2.3 Conclusions

Automatic activity recognition is a very important and active area of research. In this chapter, we have discussed many approaches for event detection which are mainly divided in two categories probabilistic approaches and deterministic description-based approaches. The main advantage of the probabilistic approaches is the convenient mechanism of probability reasoning which allow handling the uncertainties for more accurate recognition. But these approaches are limited when addressing high level activity detection. In the other hand, description-based approaches are suitable for recognizing high level activities which can be hierarchically described using sub-event with sequential and/or concurrent temporal relations. However, these approaches suffer from a lack of a convenient mechanism to deal with noisy data.

In this work, our goal is to take advantage of the two approaches for event recognition: building an approach which recognize complex activities and which is able to deal with the uncertainties of recognition. Mainly we aim at :

- explicit representation of the event models.
- handling low level uncertainty for an accurate detection of primitive events.
- handling high level uncertainty by adopting an appropriate high event inference formalism to probabilistically detect the interesting activities.
- apply the proposed approach in health care monitoring to detect daily living activities of elderly suffering of dementia and help clinicians to detect early symptoms of illness.

3

ACTIVITY RECOGNITION FRAMEWORK: OVERVIEW

3.1 Introduction

Activity recognition is an active research domain. The goal of human activity recognition is to provide accurate information about the behavior of tracked objects observed in a scene. As seen in chapter 2, the activity recognition problem has been treated with probabilistic approaches [Oliver et al., 2002],[Nevatia et al., 2004], [Chen and Zhang, 2007] and description-based approaches [Vu et al., 2003a], [Ghanem et al., 2004], [Ryoo and Aggarwal, 2010]. Our goal is to propose a framework that takes the advantages of each approach for a more reliable recognition. The objectives are presented in section 3.2, the terminology is presented in section 3.3. Section 3.4 presents the application domains. Thesis hypotheses are presented in section 3.5. An overview of the proposed cognitive vision approach for activity recognition is described in section 3.6 and finally, the conclusion is presented in section 3.7.

3.2 Objectives

The goal of this work is to propose an approach for activity recognition which is able to manage the uncertainty of the recognition and to handle noisy data. Uncertainty exists in many steps during the activity recognition process and can lead to the mis-detection of the activities of interest or can lead to errors in the activity detection decision thus an approach to characterize the uncertainty and provide the appropriate methodology and the conceptual basis to deal with it is more than needed. The proposed approach takes the advantages of a description-based approach to model and recognize complex events with complex temporal relationships and takes the advantages of probabilistic approaches to handle the uncertainty of the low-level data and propagate the probabilistic reasoning at the high level of recognition. For an application point of view, the proposed work have been mainly applied on health care applications to monitor elderly and people with dementia especially with Alzheimer disease. In order to validate the generality of the proposed approach, it has been also tested on other public and real world video data sets.

3.3 Terminology

In the context of this thesis and before going into details, several concepts must be appropriately defined.

- ***Uncertainty***: There is neither a commonly shared terminology nor agreement on a generic typology of uncertainties in the litterature, as uncertainty can be interpreted differently depending on the discipline and context where it is applied. Typical definitions of uncertainty found in the literature include:
  - In statistics, the estimated amount or percentage by which an observed or calculated value may differ from the true value.
  - *Incomplete information about a particular subject* [Ascough et al., 2008]
  - *Lack of confidence in knowledge related to a specific question* [Sigel et al., 2010]
  - *Any deviation from the unachievable ideal of completely deterministic knowledge of the relevant system* [Walker et al., 2003]
  - *Definitions of uncertainty vary from any departure from the unachievable ideal of complete determinism, to the degree of confidence a person has about the specific outcome of an event or action* [Warmink et al., 2010]
  - *A situation of inadequate information which can be of three sorts: inexactness, unreliability and border with ignorance* [Funtowicz and Ravetz, 1990].

  For activity recognition, uncertainty can be mainly classified in three types :

(1) ***Observation uncertainty*** results for instance from a wrong estimation of object location in the video due to occlusion and poor imaging conditions, uncertainty due to lack of knowledge. In the litterature, it is also called *statistical uncertainty* which is any uncertainty which can be characterized in probabilities [Warmink et al., 2010]. The most obvious example of statistical uncertainty is the *measurement uncertainty* associated to all data. Measurement uncertainty stems from the fact that measurement can practically never precisely represent the 'true' value of that which is being measured. Measurement uncertainty arises from measurement errors (e.g. type of instrument used, quality of calibration, data reading/logging, etc.), type of data recorded and length of record, the type of data processing and the method of data representation.
(2) ***Semantic or logic uncertainty*** uncertainty is introduced through the specification or the modelling of reality. It exists in the interpretation of event occurrences due to the fuzziness inherent in semantic concepts used to specify event models such as 'close to' or 'inside zone', the definition of the context (e.g. the borders of the zones).
(3) ***Decision rule uncertainty*** arises whenever there is ambiguity in the decision. It implies that there is a range of possible outcomes, but the mechanisms leading to these outcomes do not enable the definition of the probability of any particular outcome.

- ***Reliability*** can be defined as the confidence or degree of trust we have on a measurement. In this general sense, reliability measures can be interpreted, modeled and calculated depending on the attributes we want to measure or the detector that we want to evaluate. The detector can be a sensor (e.g. camera) or a video processing task (e.g. classifier, tracker, event detector).

- ***Likelihood*** is the probability or chance that something happens. In statistics, a likelihood function (often simply called the likelihood) is a function of the parameters of a statistical model, defined as follows: The likelihood of a hypothesis (H) after conducting an experiment or gathering data (D) is the probability of the data given the hypothesis, i.e. $\mathcal{L}(\mathsf{H}|\mathsf{D}) = \mathsf{P}(\mathsf{D}|\mathsf{H})$.

- ***Video sequence***: temporal sequence of images which are generated by a video camera.

- ***Scene***: the physical space where a real world event occurs and which can be observed by one or several video cameras. A scene without any physical object of interest is called an empty scene.

- ***Physical object***: a real world object in the scene. There are two types of physical objects: mobile object and contextual object.

- ***Contextual object***: a physical object attached to the scene. The contextual object is usually static and whenever in motion, its motion can be foreseen using a priori information. For instance, it can be in motion such as a door, an elevator, a fountain, a tree or displaceable (by a human being) such as a chair, a luggage.

- ***Ground truth data***: data given by a human operator and which describe real world expected results (e.g. physical objects, events) at the output of a video understanding algorithm. These data are supposed to be unique and corresponding to end user requirements even if in many cases, this information may contain errors (annotation bias).

3.4 Application Domains

The proposed approach is mainly applied and evaluated on health care applications. For the sake of generality, the proposed approach is also tested on other public real-world datasets.

- **GERHOME**: The GERHOME[1] project consists in monitoring elderly observed in an experimental laboratory during 4 hours. The objective of GERHOME project is to develop, try out and certify technical solutions supporting the assistance services for enhancing independence of the elderly at home, by using intelligent technologies for house automation to ensure autonomy, comfort of life, security, monitoring and assistance to place of residence. The video dataset is available on
www-sop.inria.fr/members/Francois.Bremond/topicsText/gerhome Project.html.

- **SWEET-HOME**: SWEET-HOME[2] is a ANR TECSAN French project on long-term monitoring of elderly people at Hospital with Nice City Hospital, MICA Center in Hanoi, Vietnam, SMILE Lab at National Cheng Kung University, Taiwan and National Cheng Kung University Hospital. SWEET-HOME project aims at building an innovative framework for modelling activities of daily living (ADLs) at home. These activities can help assessing elderly disease (e.g. Alzheimer, depression, apathy) evolution or detecting precursors such as unbalanced walking, speed, walked distance, psychomotor slowness, frequent sighing and frowning, social withdrawal with a result of increasing indoor hours. The SWEET-HOME project focuses on two aspects related to Alzheimer disease: (1) to assess the initiative ability of patient and whether the patient is involved in goal directed behaviours (2) to assess walking disorders and potential risk of falls. In this focus, the goal is to collect and combine multi-sensor (audio-video) information to detect activities and assess behavioural trends to provide user services at different levels. In this project experimental rooms are used in Nice-Cimiez Hospital for monitoring Alzheimer patients.

[1] http ://gerhome.cstb.fr/en/home

[2] http://www-sop.inria.fr/pulsar/projects/Sweet-Home/SweetHomeProject.html

The proposed approach was also tested on other public real-world datasets:

- **ETISEO**: The ETISEO[3] project seeks to work out a new structure contributing to an increase in the evaluation of video scene understanding ; with the active participation of industrialists and many research laboratories, such as French, European and International partners. ETISEO project focuses on the treatment and interpretation of videos involving pedestrians and/or vehicles, indoors or outdoors, obtained from fixed cameras.

3.5 Thesis Hypotheses

This thesis assume the following hypotheses:

- **Available Detected and Tracked Objects**: In this work, we assume that tracked objects are available. We also assume that the tracking algorithm provides a measurement (i.e. trajectory probability) which is used during the event recognition process.

- **Fixed Video Camera**: In this work, we assume the use of fixed video cameras. The video cameras are fixed on a wall and without pan, tilt or zoom and without any restriction on the camera orientation. We assume availability of calibration to compute the transformation of the 2D image referential point to a 3D scene referential point.

- **Scene Context**: In this work, we assume the use of scene context. Scene context contains both geometric and semantic description of the specific zones, walls and equipments located in the observed scene.

- **Event models**: We assume the use of a set of event models needed to detect the activities of interest.

3.6 Architecture of the Proposed Activity Recognition Approach

Considerable researches have been devoted towards activity recognition [Fouhey et al., 2012], [Ryoo and Aggarwal, 2010], [Delaitre et al., 2012]. The design of generic and robust video understanding techniques is still an open problem. Providing robust information and recognizing complex activities from noisy videos can be a very complex problem, as several issues of different nature can complicate this task. For instance, these issues can be associated to the quality of the analysed video (e.g. bad contrast, illumination changes), the complexity of the scene (e.g.

[3]http ://www-sop.inria.fr/orion/ETISEO/.

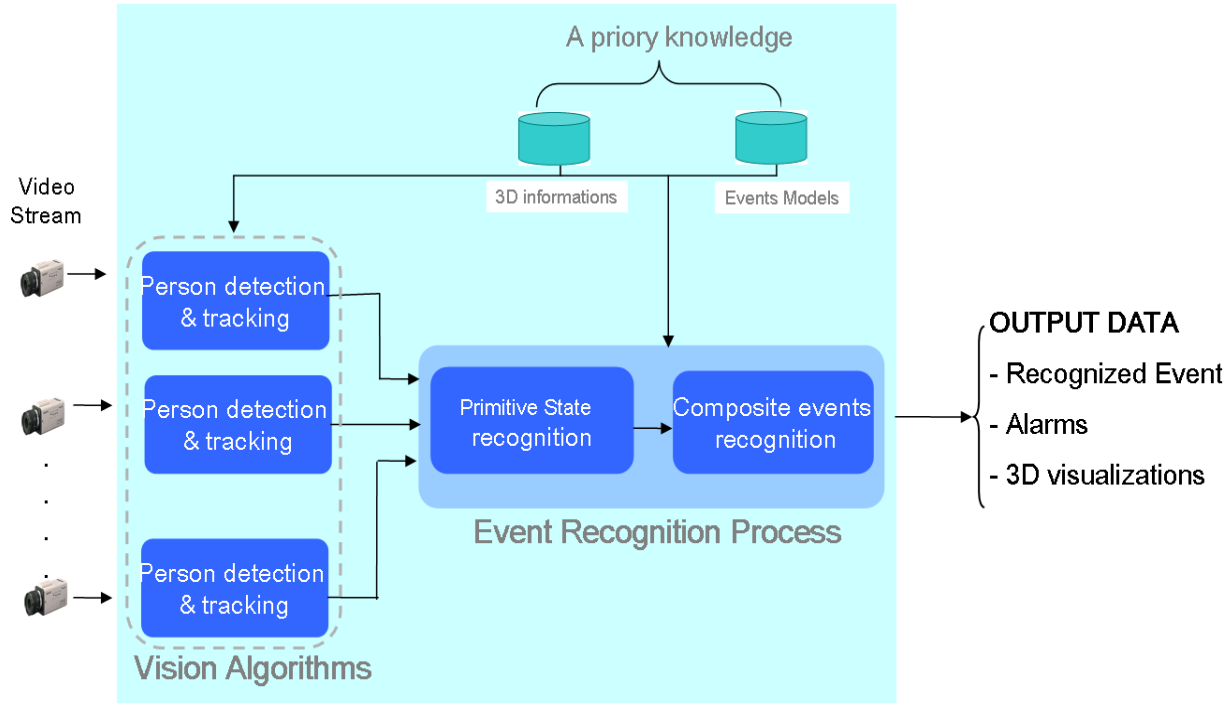


Figure 3.1: Overview of the Activity Recognition Framework.

cluttered scene, numerous mobile objects moving in the scene), or the interactions of the mobile objects with the scene and with other mobile objects (e.g. static and dynamic occlusion), among many other issues.

All these factors can induce to errors a video understanding approach due to the ambiguity of the visual evidence. Therefore, in order to achieve a robust video understanding process, it is necessary to handle the uncertainty and imprecision of low-level data and of the high semantic concepts. For coping with this problem, a new activity understanding framework for video event recognition is proposed. The proposed approach combine logic and probabilistic reasoning to handle the uncertainty of event recognition.

The proposed approach is a component of a video understanding framework (fig.3.1) which is the successor of the framework described in [Avanzi et al., 2005]. This framework is composed of:

- Video Analysis: detects, tracks mobile objects as they move in the scene (see section 3.6.1).

- Event Recognition: recognizes a set of predefined activities (see section 3.6.2).

3.6.1 Video Analysis

The activity recognition framework follows a bottom-up process to obtain high-level temporal information, starting from low-level image data. The first task of this framework is the segmentation task which is applied to each coming image to detect motion in the scene, obtaining a set of moving regions (called blobs) represented as the bounding boxes enclosing them. The algorithm segments moving pixels in the video into a binary image by subtracting the current image with the reference image. The reference image is updated along the time to take into account changes in the scene (e.g. light, object displacement, shadows). A Calibration task is used to compute the transformation of a 2D image referential point to a 3D scene referential point (fig.6.2). The calibration tool allows to obtain the calibration matrix which is used to compute the 3D position of mobile objects. The 3D position of the mobile object is computed from the detected blob and the calibration matrix associated with the video camera by considering that the bottom of the 3D mobile object is on floor level. When the blob representing the person is not completly visible (i.e. occluded by a specified contextual object), the person is supposed to be just behind the object. A blob merging task consists in assembling small 2D blobs to improve the classification task. The classification task uses the obtained 2D blobs, the calibration matrix of the camera and predefined 3D models of human to classify the mobile blobs. It adds a class label to each moving blob (e.g. person, vehicle). A set of 3D features such as 3D position, width and height are computed for each blob. A unique identifier is associated to each new classified blob by using the frame to frame tracker. A long term tracker allow to maintain this identifier throughout the whole video.

3.6.2 Activity Recognition Approach

In this thesis, the proposed event recognition approach is based on the constraint-based approach described in [Vu et al., 2003a]. This approach allows to detect in real-time which event is happening from the video stream of observed mobile objects (e.g. persons, vehicle) tracked by a vision module at each instant.

A major drawback of this algorithm is the deterministic aspect of event recognition which can lead to miss-detection of activities of interest for real-world videos. The event description formalism used in this approach has shown it success to describe high level activities but a major drawback is that this formalism does not handle the imperfect low-levels.

Handling the uncertainty of recognition is an important aspect of the recognition. Thus, in this thesis, we propose the management of uncertainty for video activity recognition. Figure 3.3 summarizes the main contributions proposed in this thesis for a reliable video activity recognition: *(i) At the level of event modeling*, we propose a set of event models for health care monitoring (see section 4.6, chapter 4) and we also propose an approach to address the

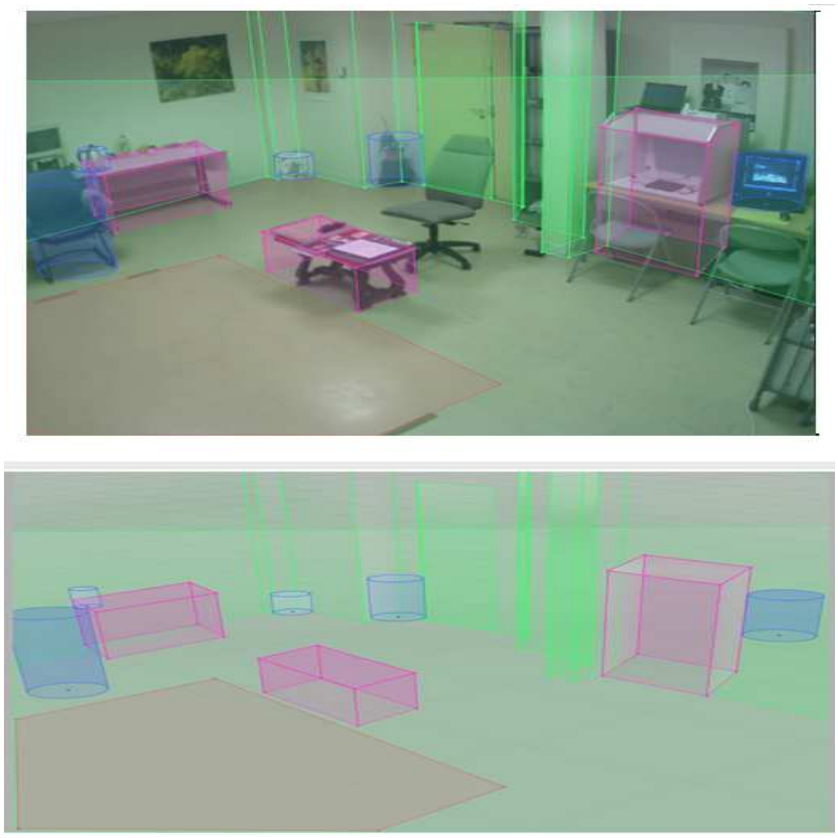


Figure 3.2: 3D geometrical information.

noise arising from low-level video analysis (see section 4.5, chapter 4). *(ii) At the level of 3D information modeling* , we have defined a new 3D model of the scene and 3D models of the mobile objects present in the observed scene, these 3D models are used during the event recognition process. *(iii) At the level of Event recognition process*, we first propose to formalize the probabilistic recognition of events by the mean of Bayesian probability theory, the Bayesian framework provides a strong tool to reason under uncertainty (see sections, 5.3 and 5.5). We second propose a probabilistic verification of the constraints (see section 5.6, chapter 5). Third, we present the proposed strategies to take into account the uncertainty inherent at the low-level observation.

3.6.2.1 Event Description Approach

We propose a generic event representation formalism that is capable to represent all types of events used for the automatic video recognition and that is able to manage the uncertainty of recognition at the event modeling level. This formalism contains the Event description Lan-

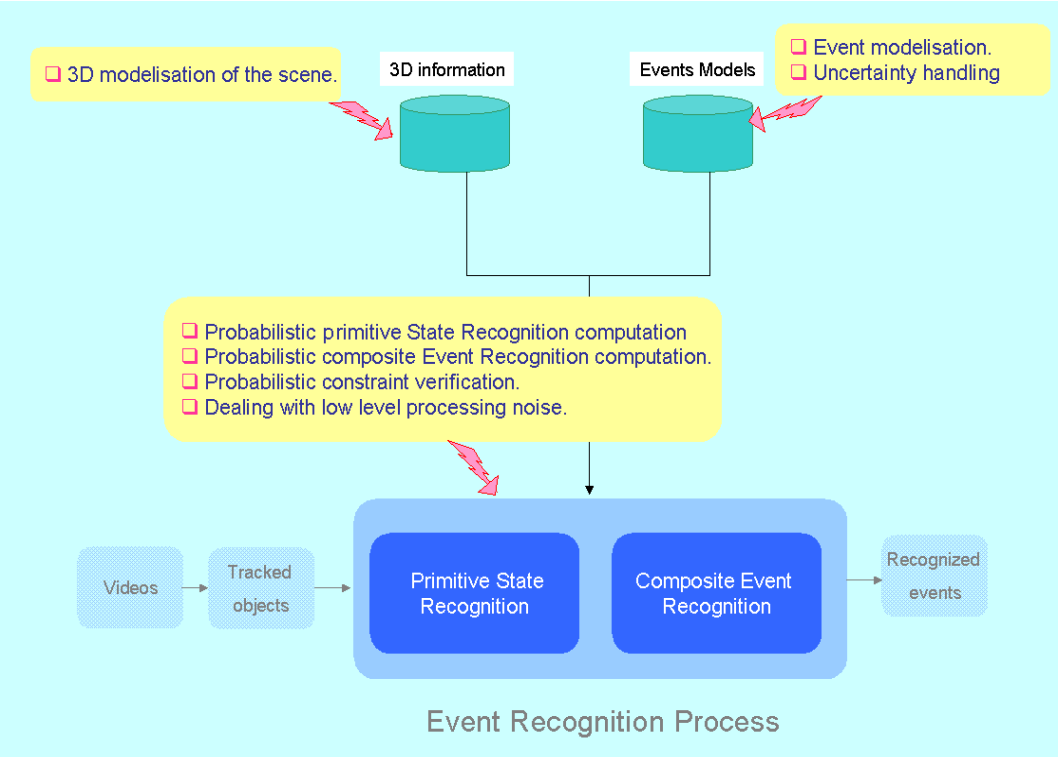


Figure 3.3: The proposed Contributions for Activity Recognition framework.

guage described in [Vu et al., 2003a] which is declarative and intuitive (in natural terms), so that the experts of the application domain can easily define and modify it. The main limitation of this language is the lack of mechanism to handle the uncertainty of recognition. For this, we proposed 2 extensions, mainly, we propose the notion of utility to deal with missed observations and we propose a specific relation in the representation of the event to manage the tracking identifier maintenance. More details are given in sections 4.5.1 and 4.5.2, chapter 4.

3.6.2.2 Event Description Language

The event description language uses a declarative representation of events. An ontology is defined [Vu et al., 2003a]. In this ontology, the physical objects (e.g. person, vehicle) are organized hierarchically (e.g. a car is defined as a sub-type of vehicle). We extend this ontology to be able to recognize all the event models for our activity monitoring (i.e. health care activity monitoring). Four types of event are defined [Vu et al., 2003a]: primitive state, composite state, primitive event and composite event (definitions and more details are available in chapter 4).

To model an event E, we need the set of physical objects (e.g. person, table) involved in the

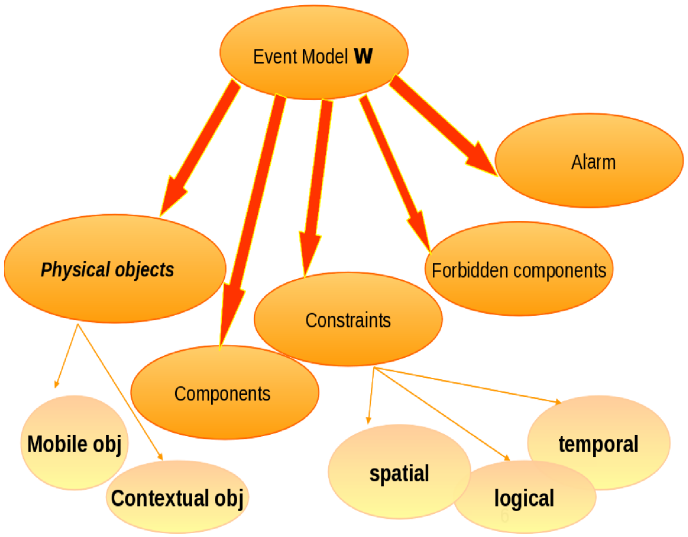


Figure 3.4: Event model structure.

event, the set of sub-events composing E and the set of the constraints on these physical objects and sub-events. A model of an event E is defined (figure. 3.4 and 3.5) based on :

PrimitiveState (Inside_zone,
PhysicalObjects ((p: Person), (z: Zone))
Constraints (p in z)
Alarm (Level: NOTURGENT))

Figure 3.5: Description of the primitive state model 'Inside-zone'.

- ***Physical objects***: is a non-empty set of non-temporal variables called physical object variables. The value of these variables correspond to real physical objects detected by the vision module. These objects are defined thanks to their attributes such as the type of object (i.e. person, table).

- ***Components***: is a set of temporal variables which value correspond to the sub-events composing the event E.

- ***Forbidden Components***: the set of variables corresponding to all event instances that are not allowed to be recognized during the recognition of the event E.

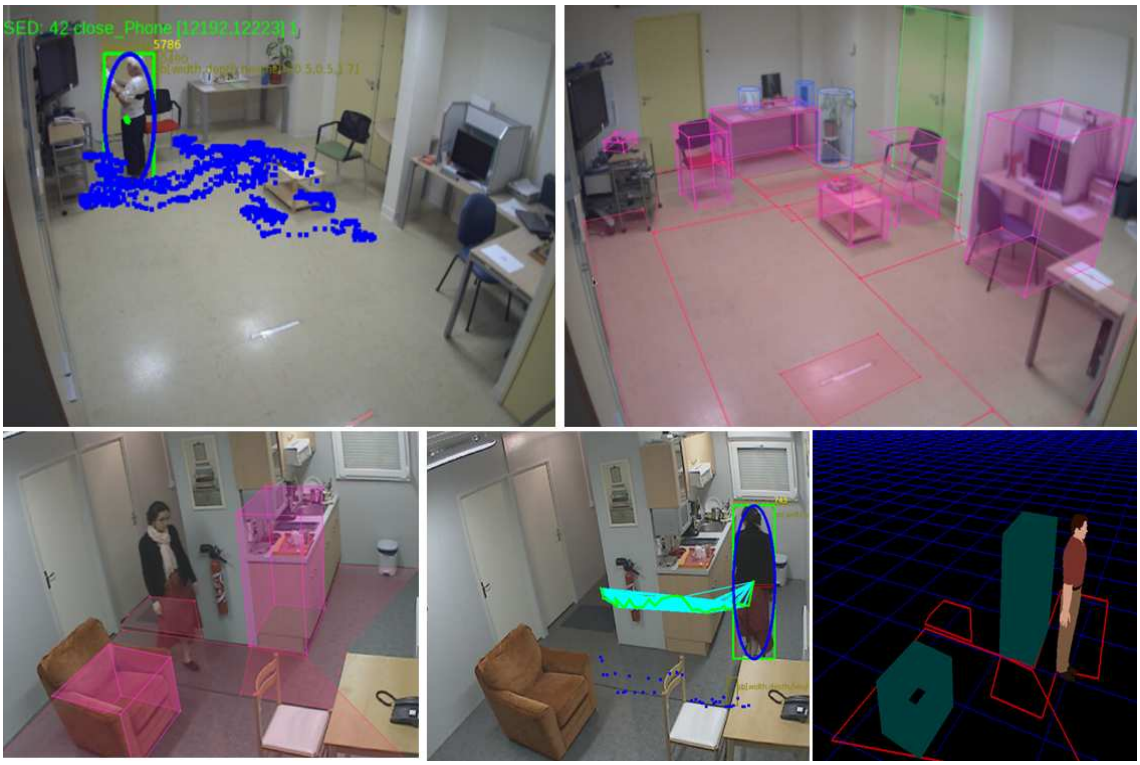


Figure 3.6: Illustration event recognition process: Scene context, alarms and 3D visualisation of the tracked persons and recognized events.

- ***Constraints***: set of constraints between the physical objects and/or the components including symbolic, logical, spatial and temporal constraints.
- ***Alarm***: The alarm information describes the importance of the scenario model in terms of emergency. Three values are possible, from less urgent to more urgent: NOTURGENT, URGENT, VERYURGENT. The alarm level can be used to filter the recognized events, for displaying only important events to the user.

3.6.2.3 Event Recognition Algorithm

The proposed video event recognition approach takes as input video streams and a priori knowledge. A priori knowledge consists of 3D geometrical information (i.e. camera calibration, empty scene) and pre-defined event models. The proposed approach gives as outputs, a set of XML files, alarms and also a 3D visualisation of the tracked persons and recognized events (figure. 3.6) .

For the recognition, the algorithm distinguishes two types of events: primitive state and composite event. The primitive states do not contain any sub-event (e.g. person close to a

machine). The composite events are composed of at least two sub-event.

The first step of the event recognition process is to recognize all the possible primitive states (most of these primitive states are based on vision primitives) by instantiating all the models with the detected objects (e.g. instantiating the event model Person-inside-Zone (taking as input one person and one zone) for all the detected mobile objects and for all the zones of the context). The second step consists in recognizing complex events according to the event model tree and the simple events previously recognized (figure 3.7). The final step checks whether the recognized event at time t has been already recognized previously to update the event end-time or create a new event instance.

In this thesis, we propose a formal probabilistic approach to reason under uncertainty for the recognition of primitive states and composite events. Probability is often considered as the best-known and most widely used formalism for quantitatively characterizing uncertainty. We propose to compute the conditional probability that an event B occur given the knowledge that an event A has already occurred. This probability is written $P(B|A)$. An important method for calculating conditional probabilities is given by Bayes's theorem. Bayesian probability theory is one of the major theoretical and practical framework for reasoning and decision making under uncertainty, using probability.

3.6.2.4 Probabilistic Primitive State Recognition

In this thesis, we propose to compute the conditional probability of the recognition of the event instance e belonging to an event model Ω given that the mobile physical objects in the model Ω have been observed and given that the constraints in the model Ω are satisfied by the observation O. More details are given in section 5.3.2, chapter 5.

3.6.2.5 Probabilistic Complex Event Recognition

The probabilistic recognition of complex event is defined as a hierarchical Bayesian inference. The objective is to recognize the complex event e given an observation O. What we want to calculate here is :
'The probability to recognize a complex event instance e belonging to an event model Ω given that the components (sub-events) defined through the model Ω are observed and the constraints in the model Ω are satisfied by the observation'. More details are given in section 5.5, chapter 5.

3.6.2.6 Probabilistic Constraint Verification

In this work, we consider the spatial constraints related to the mobile object speed, position and closeness to a given contextual objects (e.g. person near TV), the constraints related to

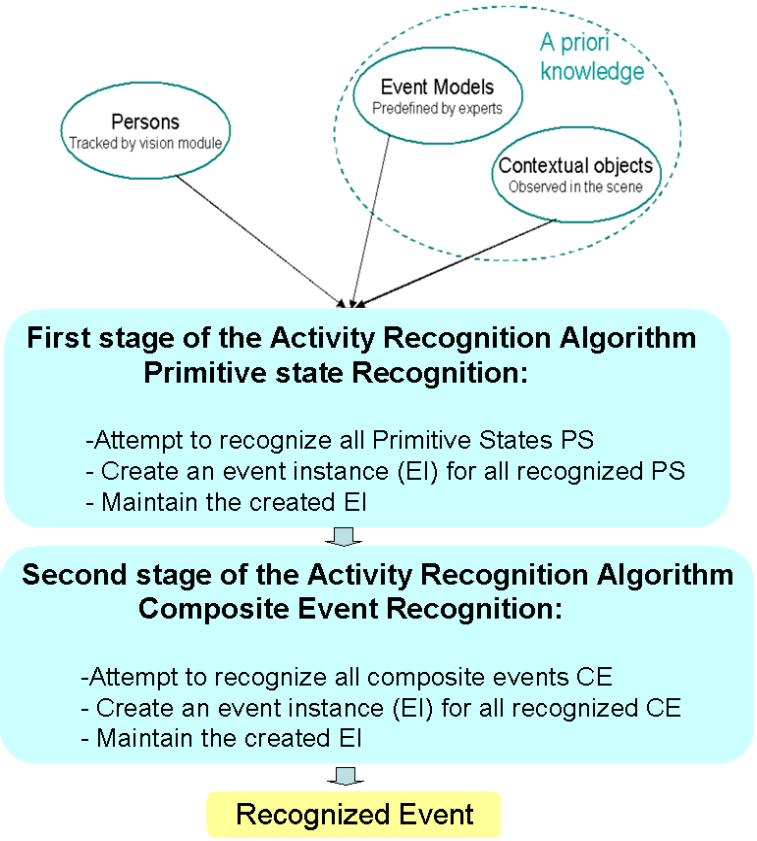


Figure 3.7: Event Recognition Process.

the posture (e.g. person is sitting) and we use the Allen[Allen, 1983] temporal constraints. We also use quantitative temporal relation (i.e. duration).

- **Spatial constraints**: A main problem for spatial constraints is the imprecision and uncertainty in the detection of the location of mobile objects due to low level detection errors (e.g. reflections, shadows or occlusions). Thus the verification of the constraint may fail. A solution to cope with this problem is to propose a probabilistic verification of the constraint. In the process of spatial constraint verification, we take into account:
  (i) the geometrical uncertainty which is related to the verification of the constraint (e.g. verifying the spatial constraint 'person-inside-zone' consists in the geometrical computation whether a point representing the person position is inside a polygon representing the zone). The first step consists in computing the distance of the person to the contextual objects (i.e. zone), the second step is to find a probability distribution function (PDF) that maximizes the value of probability when the person is inside the zone (fig.3.8). More details are given in section 5.6.1.1, chapter 5.

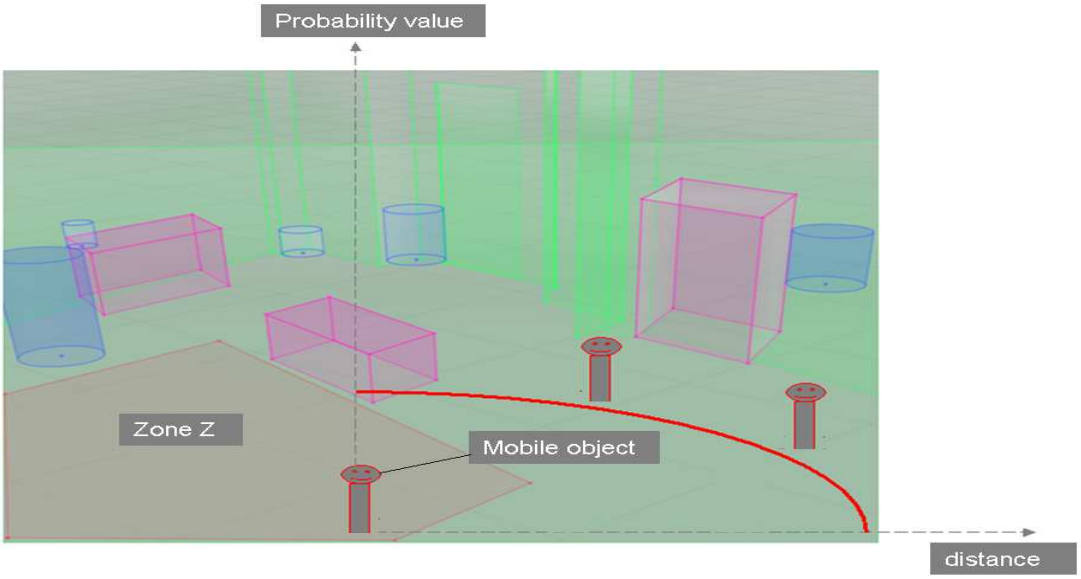


Figure 3.8: The probability computation of the spatial constraint 'inside zone' is distance-based. The probability decreases when the distance of the mobile object (e.g. person) to the zone z is big.

(ii) the uncertainty of the attribute values: due to noisy data and low-level algorithm errors. This type of uncertainty handling is described in chapter 5 by proposing a new dynamic model for the re-estimation of the attribute values to deal with noisy data..

- **Temporal constraints**: to improve the temporal constraint verification process, we add (i) the notion of *'tolerance'* to compare between the temporal intervals. We propose also (ii) to compute the probability that a constraint is verified. Table 3.1 illustrates six of the thirteen Allen temporal predicates. We take the temporal constraint *'A before B'* as

Predicat	Description
before(a, b)	$a_{end} < b_{start}$
meets(a, b)	$a_{end} = b_{start}$
overlaps(a, b)	$a_{start} < b_{start} < a_{end}$
starts(a, b)	$a_{start} = b_{start}$ and $a_{end} < b_{end}$
during(a, b)	$a_{start} > b_{start}$ and $a_{end} < b_{end}$
finishes(a, b)	$a_{start} > b_{start}$ and $a_{end} = b_{end}$

Table 3.1: Examples of Temporal Allen Predicates.

an example to illustrate how we propose to improve constraint verification. More details

about the other temporal constraints (e.g. meet, overlap) are given in chapter 5, section 5.6. For the verification of the temporal constraint *'A before B'* we need to find a time interval $[a_{start}, a_{end}]$ where the event A starts and ends and an event B ($[b_{start}, b_{end}]$) starts after A that verify: $a_{end} < b_{start}$.

To compute the probability of this temporal constraint, we have defined a function $\mathcal{F} : ([a_{start}, a_{end}] \times [b_{start}, b_{end}]) \rightarrow [0, 1]$ that maximizes the probability value when the difference Δ_t between the two time instants a_{end} and b_{start} is big. More details are given in section 5.6.3, chapter 5.

3.6.3 Dealing with low-level processing Noise

One of the major challenges to deal with, is the management of noise from low-level processing . This noise, is one of the first source of uncertainty in the event recognition process. In this section, we present the proposed strategies to take into account the uncertainty inherent in low level observations.

3.6.3.1 Visual Reliability

One of the first notion in the litterature to qualify uncertainty is the notion of reliability. In this thesis, we define the notion of 'visual reliability'. The visual reliability is intended to quantify how much the object can be seen from the camera point of view. The objective is to find a measure that gives a minimal value when the object is not visible, and a maximal value when the object is totally visible. We define the visual reliability for 2D and 3D attributes.

3.6.3.2 Dynamic Model for temporal attribute filtering

Observations in real world video sequences can be corrupted by noise, thus our goal is to estimate more accurately an attribute value given its observed value. We propose a dynamic linear model for reliability computing and updating the attributes value a and confidence based on a temporal history of the previous values (see chapter 5 for more details). We think that tacking into account an history of the previous values and not just the previous one help us to a better estimation. This process which follows the same strategy than a kalman filter works in two steps:
- *The first step (1)* consists in computing the expected value a_{exp} of an attribute a at the current instant t_c given the estimated value of a and its velocity at the previous time t_p.
- *The second step (2)* is to compute the estimated value a_{est} of the attribute based on the previous one.

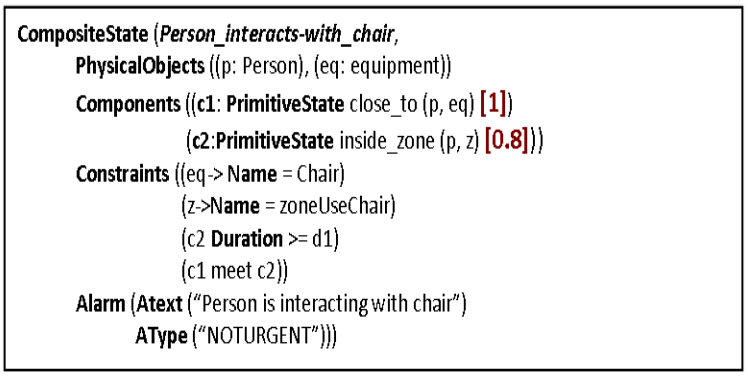


Figure 3.9: An utility coefficient is associated to each sub-event of the event model. The primitive state 'close-to' is associated with an utility equal to 1, that is mean that this primitive state is highly required to recognize the composite state 'Person-interacts-with-chair'. The primitive state 'inside-zone'is associated with an utility equal to 0.8.

- *The final value* $\bar{a}$ of the attribute is the mean between the expected and the estimated values of the attribute weighted by the expected and estimated reliability values

3.6.3.3 Missed Observation

Occlusion and poor imaging conditions (e.g. dark, shadowed areas of the scene) are common conditions that prevent us from observing the occurrence of some events. When we miss the recognition of one of the sub-events the whole event is missed. To prevent from this, we propose a notion of utility in the definition of the event model by associating a coefficient to each sub-event (fig. 3.9). Utility which is defined by a human expert expresses the importance or priority of sub-events for the recognition of the whole event. Its range is in the interval]0,1], higher is the utility value higher is the importance of the sub-event in the recognition of the whole event. The value 1 means that the sub-event is required for the recognition.

3.6.3.4 Dealing with the Tracking Identifier

One of the low-level errors which deeply affects the performance of the recognition is the changes of the tracking identifier. Identity maintenance is a primary source of uncertainty for activity recognition, it affects in particular the recognition of long-term events.

- **At the Event Modeling step**:

  Identity maintenance is necessary when there exist multiple identities that actually refer to the same mobile object. It is caused by lack of visual information (appearance, shape, etc.) to compute the correspondence between objects. Our approach to solve this issue at

```
PrimitiveEvent (PersonisStandingUp,
        PhysicalObjects ((p1: Person), (p2: Person))
        Components ((c1: PrimitiveState PersonSitting (p1)
                    (c2: PrimitiveState PersonStanding (p2))
        Constraints ((equal (p1, p2)),
                    (c1 before c2))
        Alarm (Atext ("Person is standing up")
              AType ("NOTURGENT")))
```

Figure 3.10: Illustration of the definition of the *'equal'* relation. The relation $equal(p1, p2)$ verify whether the identifiers of the two objects p1 and p2 refer to the same object.

the level of event modeling is to use specific relation in the representation of the event. More precisely, the identification whether two objects A and B refer to the same object is represented by the relation **equal(A, B)** (see section 4.5.2, chapter 4).

The evaluation of this relation is done using appearance matching (e.g. 3D height, 3D width, etc.). Different identifying contextual cues about identities can be discussed. These cues are based on the individual belongings, closed place activity, knowledge and appearance as pointed in [Tran and Davis, 2008].

- **At the Event Recognition step**

The recognition of an event over time needs maintaining the same identifier for each mobile object when recognizing its sub-events, otherwise it will be considered as a different object. To deal with this tracking error at the event detection level, we propose the use of the recognition history of an event e, $\{e^1, ..., e^{t-2}, e^{t-1}\}$: the recognized events over time are stored in a buffer and for each time t, and for each detected event e, we propose to look at the change of its physical object identifier. If the identifier of a physical object changes suddenly and/or for a short period of time, we do not consider the new identifier and we maintain the last identifier of the physical object.

3.7 Conclusion

We have presented in this chapter an overview of the proposed approach to recognize human activities. Our approach consists in combining logic and probabilistic approaches for event recognition. We have presented the different approaches proposed to manage the uncertainty of recognition during the event modeling step and the event detection step. We have also

presented the approaches proposed to manage the uncertainty of low-level data. In the next chapters, the proposed approach for activity recognition is described in details: chapter 4 describes the proposed activity modeling and chapter 5 describes the approach proposed for the probabilistic event recognition. Evaluation is detailed in chapter 6.

4

EVENT MODELING APPROACH

4.1 Introduction

Event modeling is an important step of event recognition process. The aim consists in modeling all the knowledge needed by the activity recognition algorithm to detect the activities of interest occurring in the scene. In this work, we propose to use and extend the event description language described in [Vu et al., 2003b] to be able to define all the event models of interest. We propose also to deal with imperfect low-levels during the event modeling.

In this chapter, we introduce first in sections 4.2 and 4.3 the video event ontology proposed in this thesis, and the hierarchical generic event model in section 4.4. We present the used 3D information in section 4.5. In section 4.6, we discuss the proposed approach for the verification of the constraints used in the event modeling. In section 4.7, we present the uncertainty representation in the event models and in section 4.8, we present the proposed event models elaborated for health care applications and finally we conclude the chapter in section 4.9.

4.2 Video Event Ontology

An ontology is the set of all the concepts and relations between the concepts shared by a community in a given domain. An ontology is useful for video interpretation. First, it enables the expert to clearly understand the term used for event modeling. Moreover, the ontology is useful for the system evaluation. It enables to understand exactly the type of events recognized by the system.

In this thesis we propose an ontology for health care monitoring based on the work proposed in [Vu et al., 2003a], [Zouba et al., 2009]. This ontology contains a set of physical objects (mobile objects and contextual objects) and a set of states and events which are interesting to recognize. We describe below the different concepts of this ontology. First, we present the general concepts as described in [Vu et al., 2003a], [Zouba et al., 2009] and then we detail the proposed extensions.

4.2.1 Concepts for Describing Physical objects

The physical object concept is a fundamental concept of the proposed ontology. Physical objects are all the real world objects observed by the camera sensors. The physical objects are characterized by their attributes which are essential for the recognition. We distinguish two types of physical objects corresponding to their nature and shape.

- **Mobile object** : is a physical object which is moving in the scene (e.g. person, group of person, animals, vehicle,...).
- **Contextual object** : is a physical object which is static or fixed in the scene (e.g. chair, table, tree,...).

4.2.2 Concepts for Describing Activities

The following concepts are defined in the event ontology described in [Vu et al., 2003a]. Four types of events have been designed. The first distinction lies on the temporal aspect of events : we distinguish states and events. A state is a spatio-temporal property characterizing one or several mobile objects at time t or a stable situation over a time interval. An event is one or several state transitions at two successive time points or in a time interval. The second distinction lies on the complexity aspect : a state/event can be primitive or composite.

- **A state** describes a stable situation in time characterizing one or several physical objects.
- **An event** is an activity containing at least a change of state value between two consecutive times (e.g. a person enters a zone of interest (kitchen): he/she is outside the zone and then inside).
- **A primitive state** (e.g. a person is located inside a zone) corresponds to a spatio/temporal property directly computed by the vision component.
- **A composite state** is a combination of primitive states.
- **A primitive event** corresponds to a change of primitive state value.
- **A composite event** is a combination of primitive/composite states/events.

4.2.3 Relation between Concepts

Many relations between all these concepts can be defined. The relations between the concepts for activities (i.e. states and events) and the physical objects express how the events are inferred from the physical objects and their attributes. The spatial relation includes the distance and geometrical relations. The temporal relations include Allen's interval algebra [Allen, 1983] and time duration. The logical relations includes 'and', 'or', conditional ('if ... then ...').

4.3 Ontology for Health Care Monitoring

Aging disorders represent a major challenge for health care systems. Many efforts are currently undertaken to investigate on psycho-behavioural disorders. In this thesis, we propose to investigate on the discorders of aging people with dementia especially with the Alzheimer's disease. To do that, we first need to define an ontology for health care monitoring. This ontology elaborated for health care monitoring contains a set of physical objects (mobile objects and contextual objects) and a set of states and events. The ontology is designed to allow its reutilization and to take into account future extensions without a need to revise existing concepts and definitions. We elaborate also in collaboration with clinician a number of criteria which will be computed by the proposed system that allow to detect early symptoms of Alzheimer's disease. In the next sections, we first define the vocabulary used for health care monitoring, then we present the different components of the ontology.

4.3.1 Health Care vocabulary

In this section, we present the definition of the vocabulary used for health care monitoring. These definitions are used during the thesis.

- **Activities of daily living (ADLs)** generally refer to basic activities of personal self-care. It consist of bathing, dressing, going to the toilet, transferring, continence, and feeding.
- **Instrumental ADLs (IADLs)** are associated with more complex tasks, they are the activities often performed by a person who is living independently in a community setting during the course of a normal day, such as using the telephone, shopping, food preparation, housekeeping, laundry, mode of transportation, responsibility for own medication, and ability to handle finances.
- **Alzheimer people (AD)**: are people with Alzheimer's disease which is a progressive disease that destroys memory and other important mental functions. It causes a group of brain disorders that results in the loss of intellectual and social skills. These changes are severe enough to interfere with day-to-day life.

- **Mild Cognitive Impairment (MCI)** is an intermediate stage between the expected cognitive decline of normal aging and the more serious decline of dementia. It can involve problems with memory, language, thinking and judgment that are greater than normal age-related changes. MCI is often a precursor to AD and other forms of dementia. Persons with MCI commonly have mild problems performing complex tasks. Nevertheless, they generally maintain their independence with minimal assistance.
- **Normal Control (NC)**: healthy person, aging normally which may forget things as well but they will typically remember them later.
- **Behavioral and psychological symptoms of dementia (BPSD)** are frequently associated with cognitive deficits during the progression of Alzheimer disease (AD) and other dementia. BPSD assessment is usually based on a structured interview, using subjective input from either the caregiver and/or the patient.
- **Mini Mental State Examination (MMSE) score** is the most commonly used test for complaints of memory problems. It is used by clinicians such as neuropsychologist to help diagnose dementia and to help assess its progression and severity. The MMSE is a series of questions and tests, each of which scores points if answered correctly. If every answer is correct, a maximum score of 30 points is possible. The MMSE tests a number of different mental abilities, including a person's memory, attention and language.
- **Neuropsychiatric Inventory (NPI)** a test of assessing neuropsychiatric symptoms and psychopathology of patients with Alzheimer's disease and other neurodegenerative disorders. The NPI originally examined 10 sub-domains of behavioral functioning: delusions, hallucinations, agitation/aggression, dysphoria, anxiety, euphoria, apathy, disinhibition, irritability/lability, and aberrant motor activity. A screening question is asked about each sub-domain. If the responses to these questions indicate that the patient has problems with a particular sub-domain of behavior, the caregiver is only then asked all the questions about that domain, rating the frequency of the symptoms on a 4-point scale, their severity on a 3-point scale, and the distress the symptom causes them on a 5-point scale [Cummings, 1997].
- **Behavioural disorders** include the following: delirious ideas, hallucinations, refusal to cooperate, agitation, aggression, abnormal motor behaviour, disinhibition, shouting, wake-sleep cycle disorders.
- **Apathy** is one of the most common psychological problems associated with dementia. Just over half of all people with Alzheimer's are emotionally blunted and lack motivation and initiative. As the disease worsens, apathy often grows worse. Apathy is usually assessed in clinical practice and research with the Neuropsychiatric Inventory (NPI) apathy domain.

4.3.2 Physical Objects for Health Care Monitoring

We have used a set of physical objects defined in the ontology for monitoring elderly activities at home proposed in [Zouba et al., 2009]. We have extended this ontology for Alzheimer people monitoring. A set of physical objects (fig. 4.1) and a set of state and event models (see section 4.6) have been proposed to monitor older people and in particular Alzheimer at hospital.

Figure 4.1 shows the physical object representation which is organized in a hierarchical way from more general to more specific. The general physical object can be re-used for any monitoring domain. More specific physical object can be re-used for health care monitoring in particular Alzheimer people monitoring.

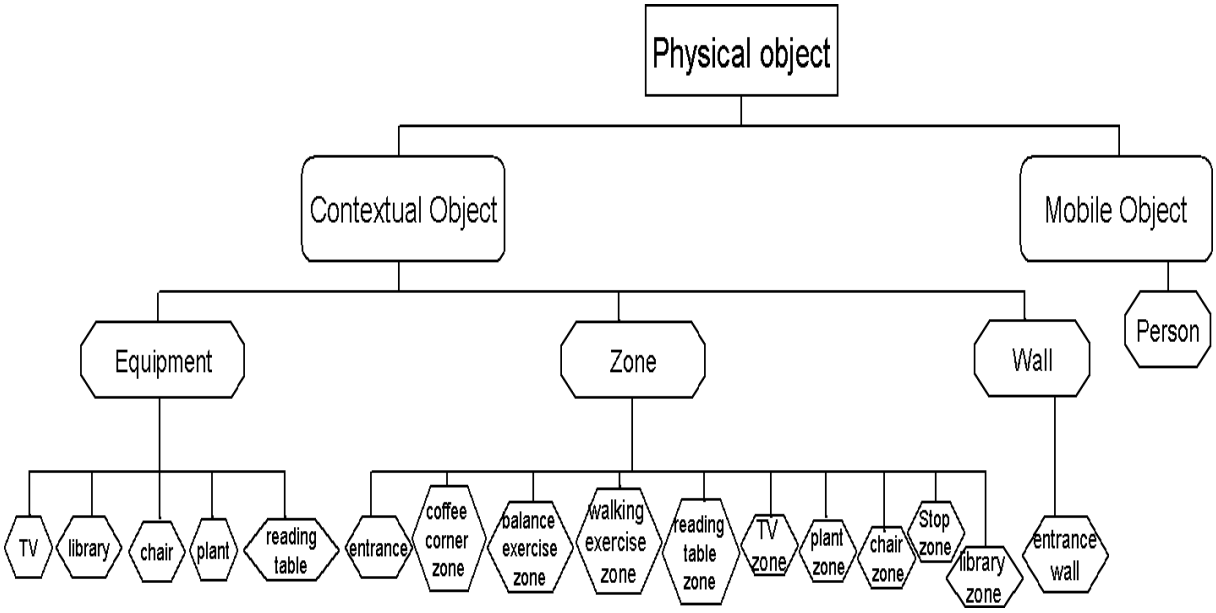


Figure 4.1: A graphical representation of physical objects organized in a hierarchical way. This hierarchy is built from more general to more specific: contextual object and mobile object are sub-type of physical object. In the same philosophy, equipment, zone and wall are sub-classes of contextual object.

4.3.3 Health Care Activities

The activity modeling has been done in close collaboration with clinicians. A strong effort have been done in this step of modeling the activities of interest for the purpose to meet the clinician requirements for patient monitoring. For medical needs, we have defined two clinical scenarios to be executed by the participants. The overall aim of the clinical scenarios was to enable the participants to undertake a set of daily tasks that could realistically be achieved in

the setting of an observation room and at the same time provide objective information about dementia symptoms. The clinical scenarios are intended to assess older people performance in IADLs (Instrumental Activities of Daily Living) and in gait analysis tests (e.g., performing a balance test). The first clinical scenario (S1) is elaborated to be executed by Alzheimer disease AD participants and to be compared with normal control participants (i.e. healthy older people). The second clinical scenario (S2) is elaborated to be executed by MCI (Mild Cognitive Impairment) participants and compared with normal control people. The activities in scenario (S1) are less complex than for scenario (S2) as the participants recruited for scenario (S1) are at an advanced stage of the illness and have more problems in memory and have more difficulties to perform goal-directed activities.

Both scenarios were divided in three parts covering basic to more complex activities: **(1)** Directed activities, **(2)** Semi-directed activities and **(3)** Undirected ('free') activities. Scenarios (S1) and (S2) have difference at the level of semi-directed activities and they keep the same content for directed activities and free activities.

- **Directed activities** (10 minutes duration): The aim of this part of the assessment is to identify characteristics of gait and walk parameters in activities with limited implication of cognitive capacities (table. 4.1). This part was based on short physical performance and required the examiner, who remained in the room, to verbally direct the participant to undertake various daily tasks. The examiner also scored the performance.

Balance testing The examiner asks the participant to perform several physical exercises: - Stand feet together side by side . - Semi tandem stand, stand with the side of the heel of one foot touching the big toe of the other foot - Tandem stand, with the heel of one foot in front of and touching the toes of the other foot **Up and Go Exercise** Patient is asked to Walk through the room, from the opposite side of the video camera for 4 meters and then go back to the starting point. **Transfer Exercise: Repeated chair stands testing** The examiner asks the participant to make a first chair stand, from sit to stand position without using his/ her arms. The examiner will then ask the participant to do the same action 5 times in a row.

Table 4.1: The clinical scenario: Description of directed activities.

- **Semi-directed activities** (20 minutes duration). The aim here is to determine the extent

to which the participant could undertake a list of daily activities in a given order, after having been given a set of instructions. Table 4.2 shows the semi-directed activities for the first clinical scenario (S1). For the clinical scenario (S2), participants were assessed in their ability to carry out a list of ten activities in a logical order respecting temporal execution constraints within a time frame of 15 minutes (see table. 4.3).

Prior to leaving the room, the examiner described each of the activities and the location and use of various objects needed to undertake the task. The examiner left the room only after it was clear that the participant understood the task. The participant was able to keep the instructions and refer to them at any point during the assessment. The participant was also told that the examiner would be available for questions on the other side of the door and that he/she could leave the room at any point should he/she choose to do so. During the clinical scenario, an examiner located outside of the room monitored the safety of the participants.

- Walk to the reading table and read something for 2 mn - Walk to the coffee corner where the kettle is and make warm some water. - Walk to the phone and compose this number: xxxxxx. - Take the watering can and water the plant. - Walk to the television and turn it on with the remote control. - Walk to the reading table, take the playing cards and classify them by color (reds with reds, blacks with blacks). - Take the green 'ABCD' folder on the desk with the A, B, C, D sheets in it. - Match the A, B, C, D sheets from the folder to one's dispersed all over the room; A with A, etc... - Put the 'ABCD' folder back on the desk. - Get out of the room.

Table 4.2: The clinical scenario: Description of semi-directed activities of the first scenario (S1).

- **Undirected ('free') activities** (30 minutes duration): The aim here was to assess how the participant spontaneously initiated activities and organized their time. Several items were at the participant's disposal, including magazines, newspapers, a book of photos, drinks (coffee, tea, fruit juice), plants, dominos, playing cards, TV and a telephone. During this period the participant was informed that the telephone might ring 30 minutes after the examiner had left the room and that the participant would be required to answer it. This was the only instruction given to the participant. They were otherwise free to do as they

- Read the newspaper
- Walk to the coffee corner where the kettle is and make warm some water.
- Water the plant.
- Answer the phone.
- Call the taxi.
- Prepare today's medication.
- Make the check for the Electricity Company.
- Leave the room when you have finished all activities
- Watch the TV.
- Prepare a hot tea.
- Write a shopping list for lunch.

1. watch the TV before the phone call.
2. water the plant just before leaving the room.
3. call a taxi, which will arrive in 10 minutes and ask the driver to drive you to the market.

Table 4.3: The clinical scenario: Description of semi-directed activities of the second scenario (S2).

pleased for the duration of the time. The participant was also told that the examiner would be available for questions on the other side of the door and that they could leave the room at any point should they choose to do so.

Tables (4.4, 4.5 and 4.6) summarize the set of daily living activities proposed for health care monitoring.

4.3.4 Visual Health Care Criterias

Patients with Alzheimer disease show cognitive decline commonly associated with psycho-behavioural disorders like depression, apathy and motor behaviour disturbances. One of the key clinical features of Alzheimer's disease (AD) is impairment in daily functioning. Patients with mild cognitive impairment (MCI) also commonly have mild problems performing complex tasks. However current evaluations of psycho-behavioural disorders are based on interviews and battery of neuropsychological tests with the presence of a clinician. These evaluations show limits of subjectivity (e.g., subjective interpretation of clinician at a date t).

The overall aim of this study is to demonstrate that it is possible using the proposed video monitoring system to improve the evaluations of psycho-behavioural disorders and obtain a quantifiable assessment of instrumental activities of daily living (IADLs) in Alzheimer people (AD) and in MCI (mild cognitive impairment).

Proposed Activities	Description
Entering Room	A person is entering the room.
Inside-zone	A person is inside a zone.
Change-zone	A person is changing his location from a place (i.e zone) to another.
Close-to	A person is close to a contextual object (e.g. person close TV).
Interact-with-Equipment	A person is close to an equipment, into the zone of use of this equipment.
Begin Balance Exercise	A person entering the room and moving to the zone use of Balance exercise with a standing posture.
Balance Exercise Step I	A person standing with feet stuck side by side.
Balance Exercise Step II	A person standing with staggered and stuck feet.
Balance Exercise Step III	A person standing with stuck feet, one foot before the other.
Up and go	A person walk through the room for 4 meters and then go back to the starting point.
Walking	A person is doing a displacement with a certain speed.
Reading	A person sitting close to reading table for a specific period of time.
Prepare collation	A person at coffee corner close to kettle preparing a collation.
Catching drugs in drug box	A person is catching prescription/ drugs in the drug box.
Selecting drugs	A person is selecting drugs.
Putting drugs into pillular	A person is putting drugs into pillular.
Writing shopping list	A person is writing a list for the shopping.
Paying Electricity bill	A person is paying electricity bill.
Using phone	A person taking the phone dialing number or answering a call.
Interacting with library	Standing close to the library and watching or reading books.
Playing card	Consists in assembling the black color card together and the red one together.
Matching Sheets	Matching sheets with letters to their corresponding sheets placed in specific places in the room.
Watering Plant	A person is close to a plant pot and watering the plant using a watering can.
Watching TV	A person standing or sitting at TV front view and looking to TV.

Table 4.4: List of activities of interest for Alzheimer people monitoring.

Proposed Activities	Description
Standing	A person with standing posture.
Standing with Arms Up	A person is standing with arms up for at least 2 seconds.
Standing Up from chair	A person is changing from sitting in a chair to posture standing.
Sitting	A person with sitting posture.
Sitting in a chair	A person is sitting in a chair.
Bending	A person with bending posture.
Slumping	A person is slumping on armchair for at least 3 seconds
Change from stand to sit	A person is changing posture from standing to sitting.
Change From sit to stand	A person is changing posture from sitting to standing.
Turning	A person is standing and turning at a different direction.
Repeated chair stands	A person makes a first chair stand, from sit to stand position and do the same action 5 times in a row.

Table 4.5: List of posture models activities of interest for Alzheimer people monitoring.

Proposed Activities	Description
Preparing a Meal	Staying in the kitchen and being close to equipment
Taking a Meal	Person already preparing a meal and going to eat in the living room and sitting at the chair for at least 10 minu
Watching TV	Person sitting or standing in the direct view and watching TV.
Using Phone	Taking phone and dialing numbers and then hanging up the phone.
Reading book/magazine	Using taking book/magazine and flipping throw the pages.
Relaxing	Staying at the armchair without engaging any other activities.
Falling down	Transition of postures from standing to bending to sitting and lying on the floor.
Entering house	Opening door from outside and entering the house.
Leaving the house	Opening door from inside and disappearing.

Table 4.6: List of activities of interest for monitoring elderly at home.

In this study, we work in close collaboration with clinician to propose a number of criteria which could be observed by camera sensors and be computed by the proposed system to allow detection of early symptoms of Alzheimer's disease in AD and MCI (mild cognitive impairment) compared with normal control group (NC).

- **Activity execution time** : It consists of the time that participant take to carry out goal-oriented activity. It corresponds to the time interval during which the activity is detected by our system.

- **Total execution time** : It corresponds to the percentage of time participant carried out goal-oriented behaviors. It was computed as follows:

$$RE_{ff} = \frac{total\ time\ spent\ by\ participant\ in\ performing\ listed\ activities}{total\ time\ spent\ in\ the\ room\ of\ experimentation} \quad (4.1)$$

  where, $RE_{ff} \in R_{[0,1]}$.

- **Up and Go criteria**: For Up and Go exercise, participants start from the sitting position. At the start signal given by the clinician, participants stand up, walk on 3 meters, make a U-turn in the center of the room, go-back on 3 meters, make a U-turn, and sit on the chair (fig.4.2). We compute 4 criteria which correspond to 4 execution durations:
  (1) D1 as the duration for walking on 3 meters from the standing position to the time when participant reach the zone where they undertake the U-turn. To compute this criteria by our system, we model the event model 'entering z2' and the event model 'entering z3'. The start time point of the duration D1 corresponds to the time point where the event model 'entering z2' is detected by the system and the end time point of D1 corresponds to the time point where the event model 'entering z3' is detected.

$$D1 = [enteringz2,\ \ enteringz3] \quad (4.2)$$

  (2) D2 as the duration for making the U-turn. To compute this criteria, we model the primitive state 'inside zone U-turn', the duration D2 corresponds to the time interval during which the primitive state is detected by the proposed video event recognition system.
  (3) D3 as the duration for going back on 3 meters from the zone U-turn to the zone where there is the chair. To compute this criteria, we model the event models 'leaving z3' and

'leaving z2', the start time point of the duration D3 corresponds to the time point where the event model 'leaving z3' is detected and the end time point of D3 corresponds to the time point where the event model 'leaving z2' is detected.
(4) D4 as the duration of the exercise (from the sitting position at the start point to the last position transfer at the end of exercise).

Figure 4.3, shows the zones of interest for this exercise and the criteria definitions.

Figure 4.2: Illustration of the Up and Go exercise.

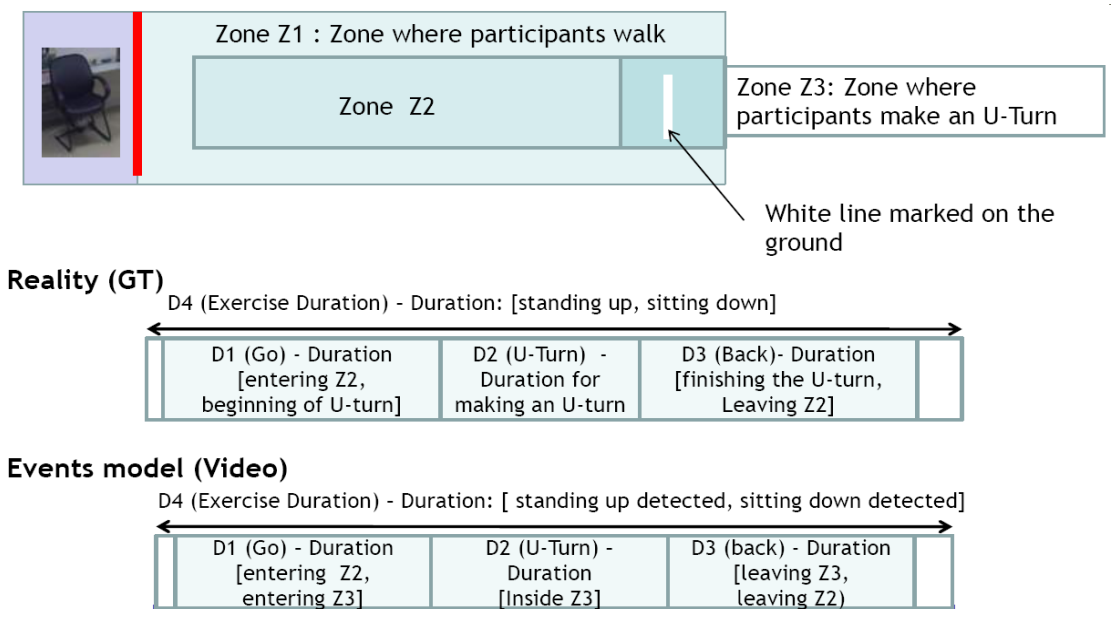


Figure 4.3: Up and Go criteria: Reality GT corresponds to the definition of criteria done in close collaboration with clinician. Event model (video) corresponds to the proposed method to compute them.

- **Walking Speed**: the speed of the walk of participant when doing the up and go exercise described in clinical scenario. It was computed in two ways:

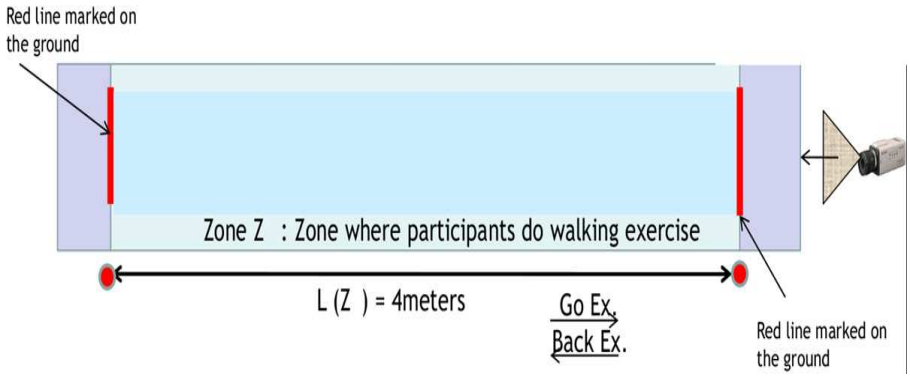


Figure 4.4: Definition of walking Exercise.

  - For the first method M_1: We define the composite event e_1 : 'is making displacement inside Exercise Walking zone'. $\Delta t_{e_1} = \max_k(\Delta t_{e_1(k)})$, k refers to an e_1 occurrence during go exercise (respectively go-back exercise). Then we define the distance d_{e_1} using the 3D coordinates of tracked person at the time points associated with the beginning and end of event $e_1(i)$ selected.

$$V_{M1,i} = \frac{d_{e_1}(i)}{\Delta t_{e_1(i)}} \tag{4.3}$$

    Where i refers to one walking exercise (i=1, 2, with 1 and 2 refer to the go and go-back exercise respectively).

  - For the second method M_2: we define a zone called 'walking exercise zone' which is bounded by two lines defining the start and the end points for the walking exercise (fig. 4.4). We model the primitive state e_2 : 'Inside Walking exercise zone'. The time interval Δt_{e_2} is the time when e_2 occurred (i.e. the person is tracked and recognized as inside this zone). The distance d_{e_2} is the constant length of 'Walking exercise zone' ($3m < d_{e_2} < 4m$, specific to each participant).

$$V_{M2,j} = \frac{d_{e_2}(j)}{\Delta t_{e_2}(j)} \tag{4.4}$$

    Where j refers to one walking exercise (j=1, 2, with 1 and 2 refer to the go and go-back exercise respectively).

To evaluate the results obtained by our system and compare them with the ground truth, we define a method to compute the ground truth walking speed V_{GT}.

– For the ground truth method M_{GT}: the starting time point for the time interval Δt_{GT} corresponds to the time point when participant cross the start line, and entering the zone of walking and the ending time point of Δt_{GT} corresponds to the time point when participant cross the end line marked on the ground (fig.4.4). The distance d_{GT} is the constant length of 'walking exercise zone' ($3m < d_{e_2} < 4m$, specific to each participant).

$$V_{GT,k} = \frac{d_{GT}(k)}{\Delta t_{GT}(k)} \tag{4.5}$$

Where k refers to one walking exercise (k=1, 2, with 1 and 2 refer to the go and go-back exercise respectively).

■ **Transfer position criteria**: participants have to execute consecutively several transfers of position from sitting to standing position (5 transfer positions, see fig. 4.5).

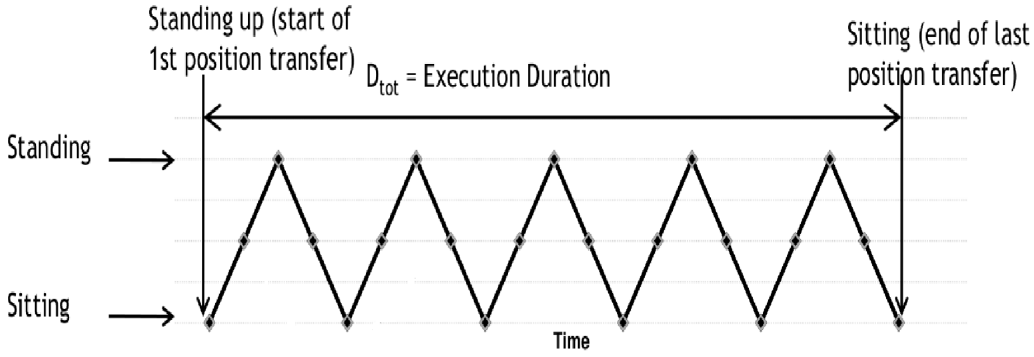


Figure 4.5: Transfer Exercise.

We calculate the duration of this exercise from the first position transfer to the last position transfer executed using two primitive states related to the posture, 'Person is standing' and 'Person is sitting'. We compute the number of transfer position $n_{transfer}$ (from sitting to standing position, and, from standing to sitting position) based on the historic of postures detected for the tracked participant. We model the event $e_{transfer}$: 'Transfer position exercise', which consists of a participant executing consecutively five transfers of position, we define the average transfer duration parameter defined as the ratio of the execution duration $\Delta t_{e_{transfer}}$ by the number of position transfer recognised by the system:

$$\mu = \frac{\Delta t_{e_{transfer}}}{n_{transfer}} \tag{4.6}$$

■ **Number of activities omitted**: The total number of activities that participants have forgotten or omitted to perform during the execution of the clinical scenario.

- **Total number of repetitions**: The total number of times where a participant repeat the same activity.
- **Order error**: when the participant make an error of the order of execution of the activities of the scenario.
- **Total number of attempts** before completing a given activity.
- **Stride length** it consists on calculating the length of stride of participant when doing the walking exercise. This criteria is defined in collaboration with clinician but not yet evaluated. Video camera sensor shows its limitation for the computation of this criteria. We think that this criteria could be accurately observed by type of sensors.

4.4 Hierarchical Generic Model of Event

The modeling of the activities has been done based on the event model language proposed in [Vu et al., 2003a]. Our objective is to have a generic event model that is capable to represent all types of events used for automatic video recognition. We have extended this event model with probabilistic notions to manage imperfect low-levels (see section 4.5). To model an event E, a set of physical objects (e.g. person, table) involved in the event E, a set of sub-events composing E and a set of the constraints on these physical objects and sub-events is needed. The model Ω of an event E (fig.4.6) is defined based on :

- ***Physical objects***: is a non-empty set of non-temporal variables called physical object variables. The values of these variables correspond to real physical objects detected by the vision module (i.e. person, table). These objects are defined thanks to their attributes such as the type of object.
- ***Components***: is a set of temporal variables which values correspond to the sub-events composing the event model.
- ***Forbidden Components***: the set of variables corresponding to all event instances that are not allowed to be recognized during the recognition of the event.
- ***Constraints***:set of constraints between the physical objects and/or the components including symbolic, logical, spatial and temporal constraints
- ***Alarm***: the alarm information describes the Name of the detected activity and its importance of urgency. For importance of urgency, three values are possible, from less urgent to more urgent: NOTURGENT, URGENT, VERYURGENT. The alarm level can be used to filter recognized events, for displaying only important events to the user.

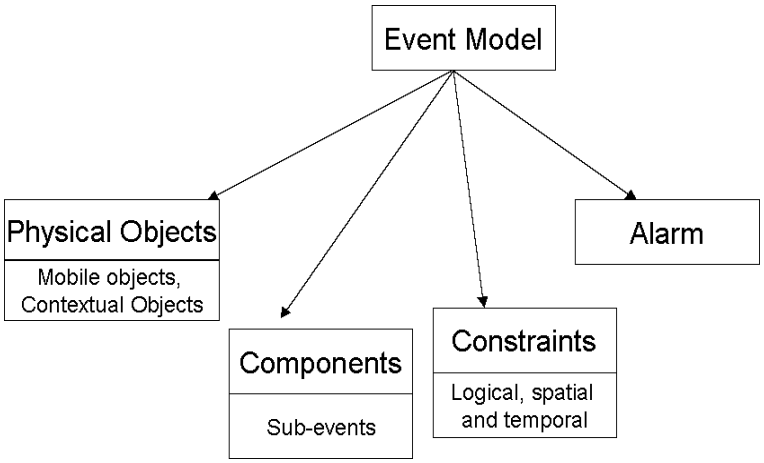


Figure 4.6: Description of the Event model.

Figure 4.7 represents the model of the primitive state 'inside-zone'. This event model is composed of 2 physical objects (i.e. person, zone) and a spatial constraint 'p in z' (i.e. person is inside a zone z).

PrimitiveState (Inside-zone,
PhysicalObjects ((p: Person), (z: Zone))
Constraints (p in z)
Alarm (Atext ('Person is inside zone')
AType ('NOTURGENT')))

Figure 4.7: Description of the primitive state model 'Inside-zone'.

4.5 Uncertainty Representation

In this section, we detail how we propose to handle the uncertainty of the knowledge at the level of event modeling for a more reliable event recognition.

4.5.1 Missed Observation

Occlusion and poor imaging conditions (e.g. dark, shadowed areas of the scene) are common conditions that prevent us from observing the occurrence of some events. When we miss the recognition of one of the sub-events the whole event is missed. To deal with this problem and be able to recognize an event even if one of its sub-events are not detected or detected

with a low probability, we propose a notion of ***'utility'*** in the definition of the event model by associating a coefficient to each sub-event.

Utility which is defined by a human expert expresses the importance or priority of sub-events for the recognition of the whole event. Its range is in the interval]0,1], higher is the utility value higher is the importance of the sub-event in the recognition of the whole event. The value 1 means that the sub-event is required for the recognition.

Figure 4.8 illustrates the use of the notion of utility in the event model ***'Person-interacts-with-chair'***. In this model, we choose an utility value equal to μ_{c1} for the primitive state 'close-to'. The primitive state 'inside-zone' is associated with an utility equal to μ_{c2}. For the detection of the primitive event ***'change-zone'*** (fig.4.9), the two sub-events are required with an utility equal respectively to μ_{c1} and μ_{c2} to recognize the event.

The utility values could be attributed either based on the domain expertise: knowing the recognition performance of the algorithm, we could attribute higher value of utility for the events which are highly recognised by the system and low utility value for the events which present a low recognition rate. The utility values could be also learned from a large and representative training data set. The training phase consists in evaluating the recognition of the activity on the training data when varying the utility values for each sub-events. The utility values which correspond to the highest rate of the recognition of the activity are then selected.

CompositeState (*Person-interacts-with-chair*,
PhysicalObjects ((p: Person), (eq: equipment), (z: zone))
Components ((c1: **PrimitiveState** close-to (p, eq) [μ_{c1}])
(c2:**PrimitiveState** inside-zone (p, z) [μ_{c2}]))
Constraints ((eq$\rightarrow$ **Name** = chair)
(z$\rightarrow$**Name** = chair zone)
(c2 **Duration** $\geq$ d1)
(c1 **meet** c2))
Alarm (Atext ('Person is interacting with chair')
AType ('NOTURGENT')))

Figure 4.8: A utility coefficient is associated to each sub-event of the event model. The primitive state 'close-to' is associated with an utility equal to μ_{c1}, it means that this primitive state is highly required to recognize the composite state 'Person-interacts-with-chair'. The primitive state 'inside-zone' is associated with an utility equal to μ_{c2}.

Figure 4.10 shows the composite event model ***'PersonStandingUp-FromChair'***, this video event involves two physical objects (a person p and an equipment, a chair), two temporal constraint, *Duration* and *after* and one logical constraint, the chair name. The utility coefficient

PrimitiveEvent(*Change-zone*,
PhysicalObjects ((p: Person), (z1: Zone), (z2: Zone))
Components ((c1: **PrimitiveState** Inside-zone (p, z1) [μ_{c1}])
(c2: **PrimitiveState** Inside-zone (p, z2)[μ_{c2}]))
Constraints (c1 **before** c2)
Alarm (Atext ('zone changing')
AType ('NOTURGENT')))

Figure 4.9: Utility coefficient is associated to each sub-event of the event model 'change-zone'.

associated to the primitive state *'PersonSlumping'* is choosen lower (i.e. 0.2) than the utility coefficient for the primitive state *'PersonStanding'* and the primitive state *'PersonSitting'* (i.e. 0.6). We make the choice to consider that the detection the primitive state *'PersonSlumping'* is not mandatory for the recognition of the event 'PersonStandingUp-FromChair'. This choice can be explained by the fact that the posture algorithm is more performant to detect the primitive state *'PersonStanding'* and *'PersonSitting'* than the primitive state *'PersonSlumping'*.

PrimitiveEvent (*PersonStandingUp-FromChair*,
PhysicalObjects ((p: Person), (eq: equipment))
Components ((c1: **PrimitiveState** stay-at (p, eq) **[0.6]**)
(c2: **PrimitiveState** PersonSlumping (p) **[0.2]**)
(c3: **PrimitiveState** PersonSitting (p) **[0.6]**)
(c4: **PrimitiveState** PersonStanding (p) **[0.6]**)
Constraints ((eq$\rightarrow$ **Name** = chair)
(c3 $\rightarrow$ **Duration** $\geq$ d1)
(c4 **after** c3)
Alarm (Atext ('Person is standing up from the chair')
AType ('NOTURGENT')))

Figure 4.10: Utility coefficient associated to each sub-event of the event model 'PersonStandingUp-FromChair'.

4.5.2 Identity Maintenance

Identity maintenance is necessary when there exists multiple identities that actually refer to the same mobile object. It is caused by lack of visual information (appearance, shape, etc.) to make unique identity connections across observation gaps. Identity maintenance is a primary

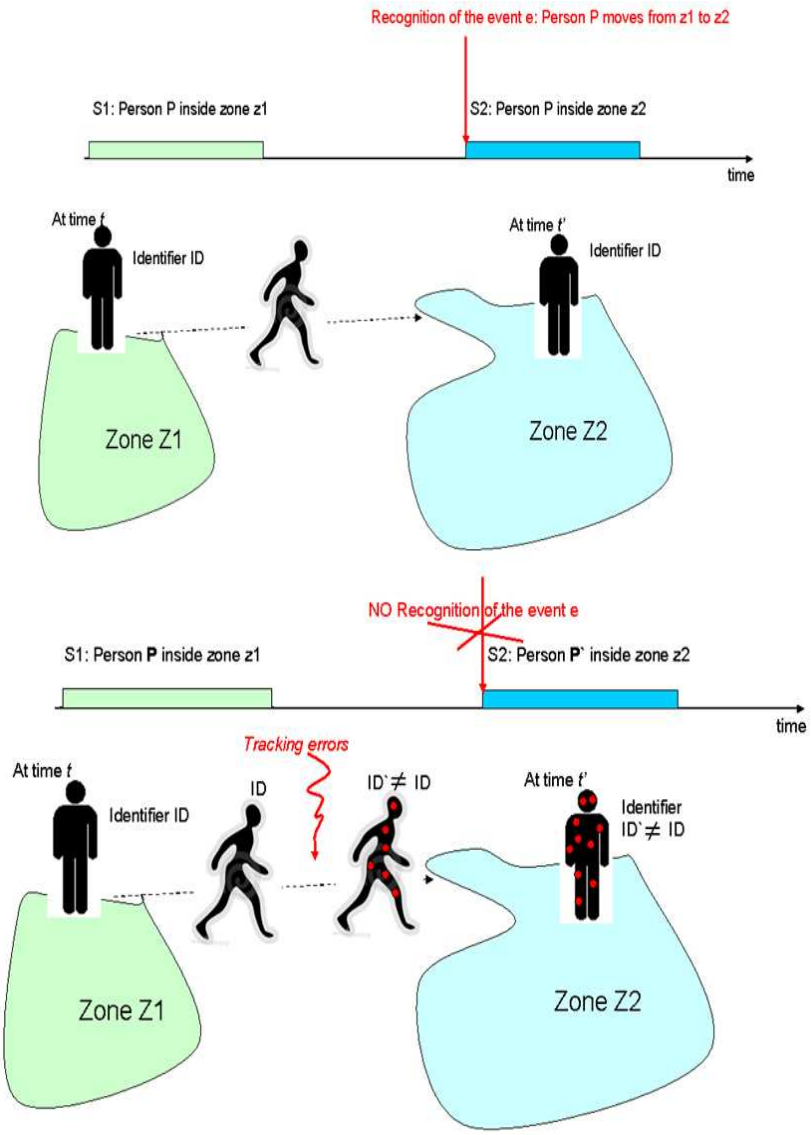


Figure 4.11: In the first figure, the event e is detected: Person moves from the zone z_1 to the zone z_2. In the second figure, there is no detection of the event e: the tracking identifier ID of the person has been changed from ID to ID$'$.

source of uncertainty for activity recognition. It affects more precisely the recognition of long-term events. The recognition of an event over time needs the maintaining of the same identifier for each mobile object when recognizing its sub-events. If the tracking identifier of a mobile object changes, it will not be considered as the same object (fig.4.11).

Our approach to solve this issue in the level of event modeling is to propose the use of specific relation ***'equal'*** in the representation of the event. More precisely, the identification

whether the identifier of two objects A and B refer to the same object is represented by the relation **A equal B**.

In this thesis, the evaluation of this relation is done using (i) appearance matching (e.g. 3D height, 3D width, 3D length, etc). We take also into account (ii) the closeness parameter, the object A is close to object B and are in the same place of activity or interacting with the same contextual equipment. This logic relation is very useful in the case where the vision algorithm (i.e. tracking algorithm) does not work well and does not consider the appearance attributes or fail to match and maintain the tracking identifier of the detected mobile object. Figure 4.12 shows an example of the use of the relation 'equal', other examples are available in section 4.6

PrimitiveEvent (*PersonisStandingUp*,
PhysicalObjects ((p1: Person), (p2: Person))
Components ((c1: **PrimitiveState** PersonSitting (p1)
(c2: **PrimitiveState** PersonStanding (p2))
Constraints((**p1 equal p2**),
(c1 **before** c2))
Alarm (Atext ('Person is standing up')
AType ('NOTURGENT')))

Figure 4.12: Illustration of the definition of the *'equal'* relation. The relation $equal(p1, p2)$ verify whether the identifiers of the two objects p1 and p2 refer to the same object.

4.6 Knowledge Base for Health Care Monitoring

In this thesis, we have modeled video event models for health care monitoring in particular Alzheimer monitoring at hospital. We have modeled 117 event models:

- 26 primitive states related to the location of the person in each zone (e.g. inside the exercise zone), location versus equipment in the observed scene (e.g. close to table, far from chair) and posture event models (e.g. person is sitting).

- 7 composite states related person interacting with contextual objects (i.e.equipment, zones, walls).

- 33 primitive events related to the moving from a place to another place (e.g. moving from entrance to coffee corner), the change of posture (e.g. change from posture sitting to standing).

- 51 composite events which combine primitive/composite states and primitive/composite events.

The proposed event models are organized from general to more specific or domain-dependent event models. Primitive states and primitive events are not domain-dependent and are considered as extendable basis. Composite event models in our knowledge base belongs to health care monitoring domain which is our domain of interest and can be shared by the specialists of the domain. The modeled activities are organized in an hierarchical way as shown in figure 4.13

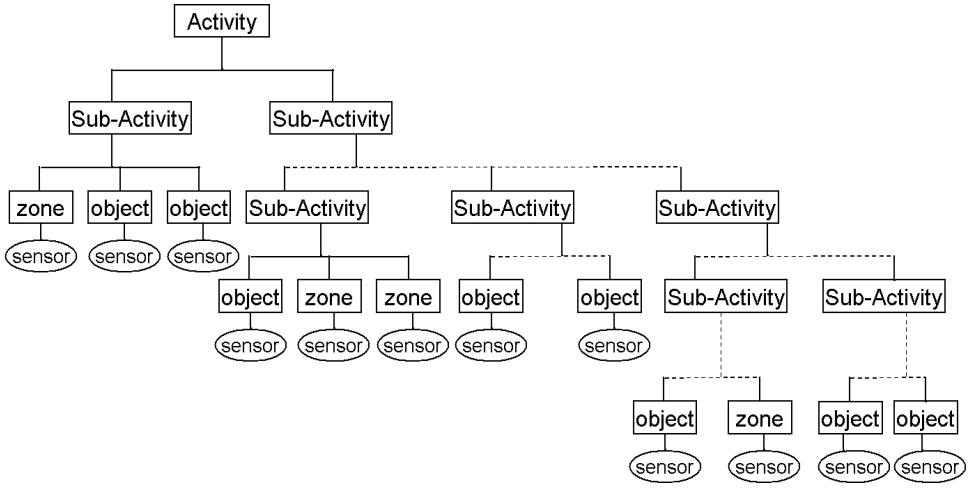


Figure 4.13: An hierarchical organization of activity model.

This section shows several examples of event models in our knowledge base.

- Figure 4.14 shows the model of a primitive state called ***'Inside-Coffee Corner'*** expressing the status of a person being inside a zone which name is 'Coffee Corner'. This video event involves two physical objects (a person p and a zone z), one spatial constraint and one logical constraint. The spatial constraint allows to verify whether the person p is geometrically inside the zone z and the logical constraint allows to verify the name of the zone z (i.e. zoneUseCoffeeCorner).

- Figure 4.15 shows the primitive event model ***'ChangePosture-StandingToBending'***, this video event involves one physical object (a person, p) and one temporal constraint. The utility coefficient associated to the primitive state *'PersonBending'* is choosen lower than the utility coefficient for the primitive state *'PersonStanding'*. That is explained by the fact

PrimitiveState (*Inside-CoffeeCorner*,
PhysicalObjects ((p: Person), (z: Zone))
Constraints ((p **in** z)
(z→ **Name** = CoffeeCorner zone))
Alarm (Atext ('Person is in the coffee corner')
AType ('NOTURGENT')))

Figure 4.14: Person inside Coffee Corner event model.

that the posture algorithm is more performant to detect the primitive state *'PersonStanding'* than *'PersonBending'*. In the case where the tracking algorithm looses the tracking identifier of the person p, we propose to use the ***equal*** relation to deal with this tracking error (fig.4.16).

PrimitiveEvent (*ChangePosture-StandingToBending*,
PhysicalObjects ((p: Person))
Components ((c1: **PrimitiveState** PersonStanding (p) [0.6])
(c2: PrimitiveState PersonBending (p) [0.4])
Constraints ((c1 **before** c2))
Alarm(Atext ('Person changes from posture standing to bending')
AType ('NOTURGENT')))

Figure 4.15: The event model based-posture *'ChangePosture-StandingToBending'*

.

PrimitiveEvent (*ChangePosture-StandingToBending-withEqualRelation*,
PhysicalObjects ((p1: Person),(p2: Person))
Components ((c1: **PrimitiveState** PersonStanding (p) [0.6])
(c2: PrimitiveState PersonBending (p) [0.4])
Constraints ((**p1 equal p2**)
(c1 **before** c2))
Alarm(Atext ('Person changes from posture standing to bending')
AType ('NOTURGENT')))

Figure 4.16: The event model based-posture ***'ChangePosture-StandingToBending'*** with the utility coefficient associated to the components(i.e. sub-events) and with the *equal* relation in the constraints.

- the event model ***'moves-away-from-person'*** describes the fact that a person moves away from another person. This event model illustrated in figure 4.17 is composed of 2 physical objects, 2 components (i.e. primitive states), and it includes 3 temporal constraints (i.e. *'Duration'*, *'before'* and *'meet'*).

PrimitiveEvent (*moves-away-from-person*,
PhysicalObjects ((p1: Person), (p2: Person))
Components ((c1: **PrimitiveState** close-to-person(p1, p2))
(c2: **PrimitiveState** far-from-person (p1, p2)))
Constraints ((c1 **meet** c2)
(c1 **Duration** $\geq$ d1)
(c1 **before** $\geq$ c2)
Alarm (Atext ('a person is moving away from another person.')
AType ('NOTURGENT')))

Figure 4.17: The event model 'moves-away-from-person'.

- Figure 4.18 shows the event model of the composite state ***'Person-intetacts-with-TV'***. The event model contains two logical constraints, the equipment name and the zone name and it contains two temporal constraints '*meet*' and '*Duration*'.

CompositeState (*Person-interacts-with-TV*,
PhysicalObjects ((p: Person), (eq: equipment))
Components ((c1: **PrimitiveState** close-to (p, eq))
(c2: **PrimitiveState** inside-zone (p, z)))
Constraints ((eq $\rightarrow$ **Name** = TV)
(z$\rightarrow$**Name** = TVzone)
(c2 **Duration** $\geq$ d1)
(c1 **meet** c2))
Alarm (Atext ('Person is interacting with TV')
AType ('NOTURGENT')))

Figure 4.18: Event model of the composite state 'Person-interacts-with-TV'.

- **'Start-WalkingTest'**: the event model (fig.4.19) of the composite event 'Start-WalkingTest' is composite of 3 physical objects, 2 components, 2 temporal constraints and 2 logical constraints. It describes that a patient is starting doing the walking activity defined in the clinical scenario.

CompositeEvent (*Start-WalkingTest*,
PhysicalObjects ((p: Person), (z1 : Zone), (z2: Zone))
Components ((c1: **PrimitiveState** Person-standing(p))
(c2: **PrimitiveEvent** change-zone (p, z1, z2)))
Constraints ((c1 **meet** c2)
(c1 **Duration** $\geq$ d1)
(z1$\rightarrow$ **Name** = Chair zone)
(z2$\rightarrow$**Name** = walking exercise zone))
Alarm (Atext ('Patient starts the walking test')
AType ('NOTURGENT')))

Figure 4.19: Description of the event model 'Start-WalkingTest'.

- **Begin Balance Exercise**: the event model *'Begin Balance Exercise'* is composed of five physical objects, four components and 5 constraints, 2 temporal constraints and 3 logical constraints and 1 alarm (fig. 4.20 and fig. 4.21). This event model describes the beginning of the physical exercise *'balance testing'* described in the clinical scenario: in this exercise, first, the nurse and the patient together enter the hospital's room dedicated for the medical experimentation, then the patient goes to a specific zone called the 'balance exercise zone' place marked by a red line and stop near a chair placed there. Then, the nurse walks to the end of the room at a zone named 'stop zone' and she/he stop there to begin to indicate to the patient the different physical activities to perform (figure 4.20).

- **Up and Go**: the event model 'Up-Go' corresponds to a medical exercise for testing the ability of the patient to perform physical activities. The model is composed of five steps: (1) the patient is standing at the chair of exercise for a predefined period of time,(2) he/she walks up to a stop zone marked by a red line, (3) goes back to the chair, (4) he/she sits at the chair and (5) gets up. This event model is composed of 4 physical objects, 6 components, 3 temporal constraints, 3 logical constraints and 1 alarm (fig. 4.22, and fig.4.23).

- **MatchingSheetsActivity**. This model (fig.4.24, 4.25 and 4.26) is one of the most complex event models in the proposed event models base. In this activity, the patient is asked to take a folder containing 4 sheets in which ones a alphabetical letter is written (i.e. A, B, C, D letters) and then the patient has to match each of these sheets to the ones which has the same letter written on it and which are distributed in different places in the room.

CompositeEvent (*BeginBalanceExercise*,
PhysicalObjects ((p1: Person), (p2: Person), (z1: zone) , (z2: zone) , (z3: zone))
Components ((c1: **PrimitiveState** at-Entrance (p1, z1))
(c2: **PrimitiveState** at-Entrance (p2, z1)
(c3: **PrimitiveEvent** change-zone (p1, z1, z2))
(c4: **PrimitiveEvent** change-zone (p2, z1, z3))
Constraints ((c1 **meet** c2)
(c3 **meet** c4)
(z1→**Name** = Entrance)
(z2→ **Name** = Balance exercise zone)
(z3→ **Name** = Stop zone))
Alarm (Atext ('Patient begin the Balance exercise ')
AType ('NOTURGENT')))

Figure 4.20: Example of the activity begin balance exercise model: the nurse and the patient entering together the room and then walk to different places.

Before the beginning of this exercise, the nurse has already indicated to the patient the different places where these sheets are placed in the room to facilitate to the patient the matching the corresponding letters. This event model is composed of 4 physical objects and 4 components and 2 temporal constraints.

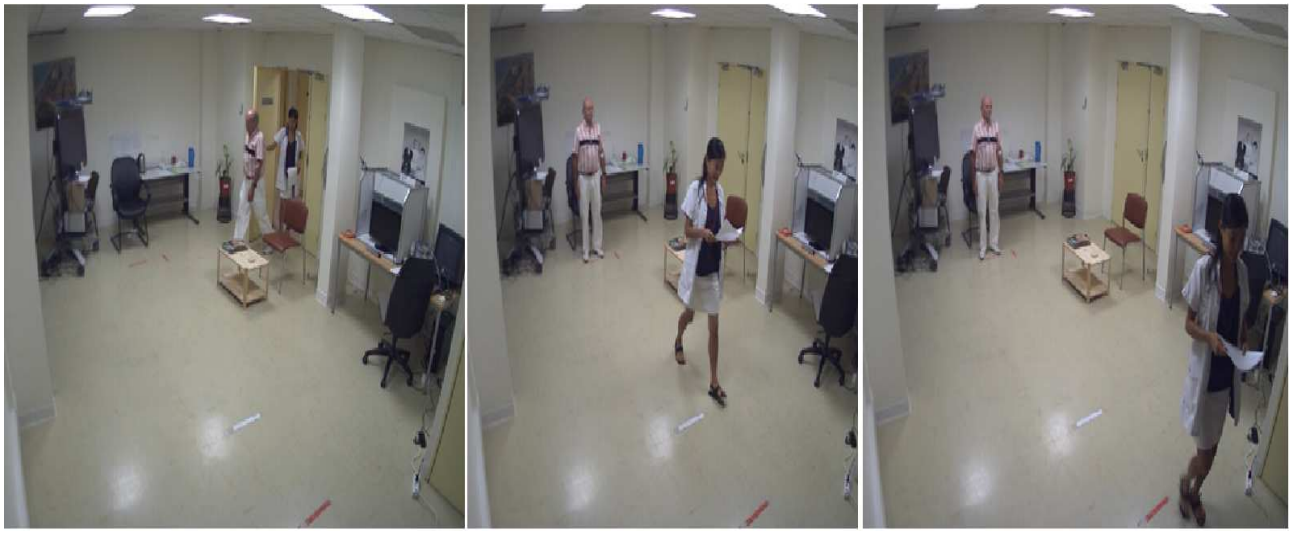


Figure 4.21: Illustration of the activity 'Begin Balance Exercise': the nurse and the patient together enter the hospital's room, then the patient goes to a specific zone called the 'balance exercise zone' place marked by a red line and stop near a chair placed there. Then, the nurse walks to the end of the room at a zone named 'stop zone'.

Figure 4.22: Illustration of the event Up-Go. (a) the patient is standing close the chair of exercise for a predefined period of time, (b) he/she walks up to a stop zone marked by a red line, (c, d) goes back to the chair, (e) he/she sits at the chair and (e) gets up.

CompositeEvent (*Up-Go*,
PhysicalObjects ((p1: Person), (eq: Equipment) (z1: Zone), (z2: Zone))
Components ((c1: (**CompositeState** interacts-with-chair (p1, eq, z1) **[1]**)
(c2: (**PrimitiveState** Person-walking (p1) **[0.2]**)
(c3: **PrimitiveState** inside-zone (p1, z2) **[1]**)
(c4: **PrimitiveState** inside-zone (p1, z1) **[1]**)
(c5: **PrimitiveEvent** change-posture-stand-to-sit (p1) **[0.5]**)
(c6: **PrimitiveEvent** change-posture-sit-to-stand (p1) **[0.5]**))
Constraints((c1 **before** c2; c3 **before** c4; c5 **before** c6)
(z1→**Name** = chair zone)
(z2→**Name** = Stop zone)
(eq→**Name** = chair))
Alarm (Atext ('Patient is doing Up-Go exercise')
AType ('NOTURGENT')))

Figure 4.23: The Event model: 'Up-Go' illustrates a medical exercise for testing the ability of the patient to perform several activities. The model is composed of five steps: (1) the patient is standing at the chair of exercise for a predefined period of time,(2) he/she walks up to a stop zone marked by a red line, (3) goes back to the chair, (4) he/she sits at the chair and (5) gets up.

CompositeEvent (*MatchingSheetsActivity*,
PhysicalObjects ((p: Person), (z1:Zone), (z2:Zone), (z3:Zone))
Components ((c1: **CompositeEvent** change-zones-coffee-Lib-TV (p, z1, z2, z3)
(c2: **PrimitiveEvent** moveFrom-TV-coffeeCorner (p, z3, z1)))
Constraints (c1 **before** c2)
Alarm (Level: NOTURGENT))

Figure 4.24: event model 'MatchingSheetsActivity'.

CompositeEvent (*change-zones-coffee-Lib-TV*,
PhysicalObjects ((p:Person), (z1:Zone), (z2:Zone), (z3:Zone))
Components ((c1: **PrimitiveEvent** moveFrom-coffeeCorner-Library (p, z1, z2))
(c2: **PrimitiveEvent** moveFrom-Library-TV (p, z2, z3)))
Constraints (c1 before c2)
Alarm (Level: NOTURGENT))

Figure 4.25: sub-event model of the event model 'MatchingSheetsActivity'.

PrimitiveEvent (moveFrom-coffeeCorner-Library,
PhysicalObjects((p: Person), (z1: Zone), (z2: Zone))
Components ((c1: **PrimitiveState** Inside-zone (p, z1))
(c2: **PrimitiveState** Inside-zone (p, z2)))
Constraints ((c1 **before** c2)
(z1→**Name** = coffee Corner zone)
(z2→**Name** = Library zone)))

Figure 4.26: sub-event model of the event model 'MatchingSheetsActivity'.

4.7 Conclusions

In this chapter, we have presented the proposed approach for event modeling based on the work of [Vu et al., 2003a],[Zouba et al., 2009]. Our objective is to have a generic event model that is capable to represent all types of activities used for automatic video event recognition. We have extended this event model with the notions of utility and identity maintenance to manage imperfect low-levels.

An ontology for health care monitoring was proposed, In this ontology, we presented several concepts useful for health care monitoring domain. Our goal is to improve the evaluations of behavioural disorders of Alzheimer people. Two clinical scenarios were elaborated. The overall aim of the clinical scenarios is to enable the participants to undertake a set of daily activities that could realistically be achieved in the setting of an observation room and at the same time provide objective information about dementia symptoms.

A knowledge base for health care monitoring was proposed, in this knowledge base, we define 117 activities of interest for Alzheimer people monitoring at hospital. Several criteria was elaborated in close collaboration with clinicians to obtain a quantifiable assessment of instrumental activities of daily living (IADLs) in Alzheimer people (AD) and in MCI (mild cognitive impairment) compared to healthy older people. We build this knowledge base in the ambition to be used as reference for other research in health care monitoring and in other countries.

In the next chapter, we present the proposed approach for automatic video activity recognition.

5

PROBABILISTIC ACTIVITY RECOGNITION

5.1 Introduction

Considerable researches have been devoted towards activity recognition [Vu et al., 2003a], [Rota and Thonnat, 2000], [Ryoo and Aggarwal, 2010], [Delaitre et al., 2012].

The automatic recognition of activities is a real challenge for cognitive vision research because it addresses the issue of handling the uncertainty of recognition. In real world scenes, there are several causes which affect the quality of the recognition of activities, among them we can cite:

- The change in brightness and shadows: change direction of the sun, clouds, adding lighting, reflections and flashes directly affect the appearance of objects and their shadows. These changes may be localized or may affect the whole observed scene.
- Poor lighting e.g. lighting in sub-ways is almost universally poor.
- Vibration of the camera: in surveillance applications, these vibrations can be generated by wind or by moving vehicles or other sources.
- Noise acquisition and sensor: like any electronic device, the sensor noise and system acquisition is a major issue. It can be very variable depending on conditions of acquisition or the type of sensor.
- The occlusion: a moving object may be fully or partially hidden by another object or static elements of the scene.

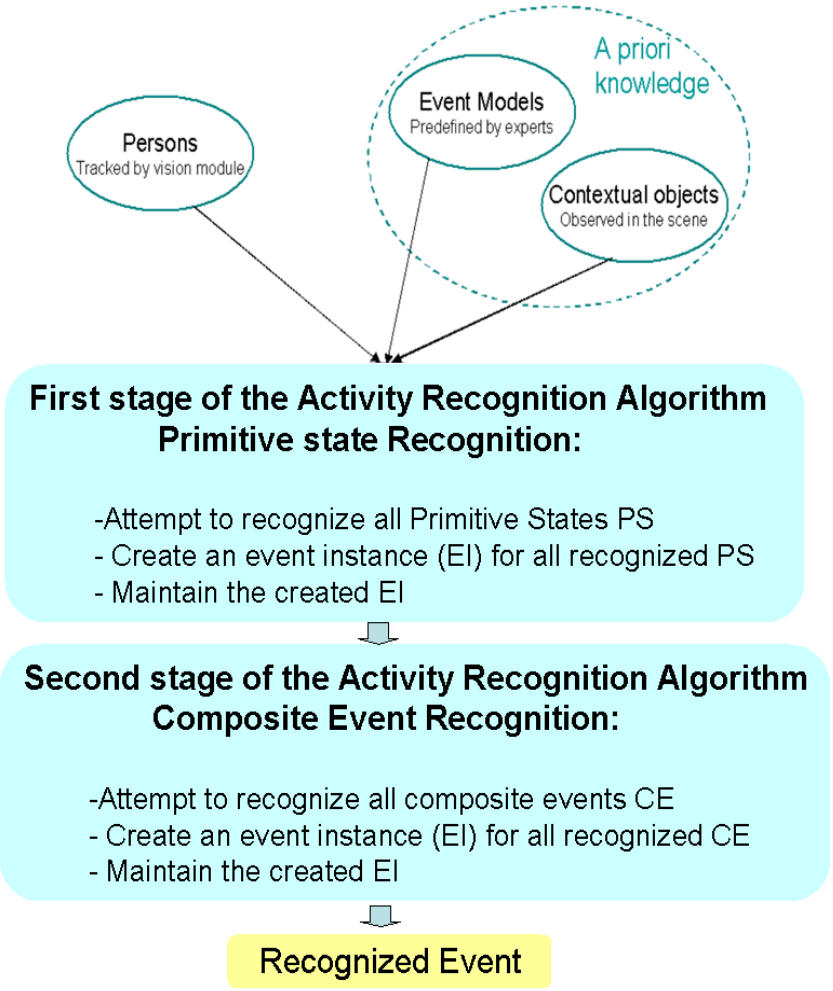


Figure 5.1: Event Recognition Process.

- The similarity: objects with similar appearance may cause confusion in the identities of objects.

Thus a mechanism to deal with noisy data and to handle the uncertainty of the recognition is more that needed. Designing a system that is able to overcome this issue is an interesting field of research in the recent years [Tran and Davis, 2008], [Ryoo and Aggarwal, 2009], [SanMiguel and Martinez, 2012].

In this chapter, we detail the proposed approach for activity recognition. The proposed activity recognition process uses as input the tracked mobile objects, a priori knowledge of the scene and predefined event models (fig. 5.1). The algorithm operates in 2 stages: (i) at each incoming frame, it computes all possible primitive states related to all mobile objects present in the scene, and (ii) it computes all possible events (i.e. primitive events, and then composite

states and events) that may end with the previously recognized primitive states.

We describe how we manage the uncertainty at the level of **(i)** the recognition of primitive states and at the level of **(ii)** the recognition of composed events (i.e primitive events, composite states, composite events).

5.2 First Stage of Activity Recognition

In this section, we describe the first stage in the algorithm of activity recognition, the recognition of primitive state. We also describe how we propose to handle the uncertainty of recognition. The primitive states do not contain any sub-event (e.g. person close to a machine). It can be calculated directly from geometrical/ physical attributes of physical objects (i.e. 3D position in the scene of the person, the speed, etc.). It can also be directly recognized by low-level algorithms (i.e. posture recognition algorithm). A time interval is associated with the recognized primitive state.

The algorithm (see algorithm 1) to recognize a primitive state e of event model Ω consists in a loop of two operations:
1. selection of a set of physical objects then,
2. verification of the corresponding constraints until all combinations of physical objects have been tested.

A ***solution*** for primitive state model Ω is a set of physical objects involved in the event and that are satisfying all the list of the constraints in the event model Ω.

In the proposed algorithm, the constraints are probabilistically verified based on Gaussian probability detailed in section 5.6. The Bayesian probability that the primitive state is recognized given that the mobile physical objects in the model Ω have been observed and given that the constraints in the model Ω are satisfied is then computed (see section 5.3 for more details).

Once a set of physical objects satisfies all constraints and the Bayesian probability is over a defined threshold, we consider that the event (i.e. primitive state) is recognized and we generate an event instance p of the event model with the recognition time interval T. The event instance is then stored in the list of recognized events.

```
begin
    Data:
    Ω : event model.
    List_Constraints : List of the event constraints.
    List_PO : List of physical objects in the event model Ω.
    Result: Find a solution for the event model Ω.
    if (Probabilistic Constraint Verification (List_PO , List_Constraints ) ) then
        if BayesianRecognition (List_PO , List_Constraints ) then
            Create a primitive state instance e.
            Store e to the list of recognized primitive states.
        end
    end
end
```

Algorithm 1: Probabilistic Solution for primitive state model. In the function *Probabilistic Constraint Verification*, the algorithm verifies probabilistically if the constraints in the event model Ω are satisfied. In function *BayesianRecognition*, the algorithm computes the Bayesian probability of the primitive state.

5.3 Probabilistic Primitive state Recognition

In this section, we present the probabilistic approach proposed to recognize primitive states. The event instance associated to the event model Ω is recognized from the observations O (e.g. video sequences). The observations are inherently uncertain, hence a formal probabilistic approach is needed to reason under uncertainty. In fact, probability is often considered as the best-known and most widely used formalism for quantitatively characterizing uncertainty. We propose to compute the conditional probability that an event B occur given the knowledge that an event A has already occurred. This probability is written $P(B|A)$. An important method for calculating conditional probabilities is given by Bayes's theorem. Bayesian probability theory is one of the major theoretical and practical framework for reasoning and decision making under uncertainty, using probability. Bayesian interpretation considers the probability of an event as the degree of belief that a person has that an event occurs, given the relevant information known to that person.

What we want to compute here is :
'The probability that an event instance e ***belongs to an event model*** Ω ***given that the mobile physical objects in the model*** Ω ***have been observed and given that the constraints in the model*** Ω ***are satisfied by the observation*** O***'.***

In mathematical words, the proposed way of calculating this is:

$$P(e \in \Omega|O) = P(e \in \Omega|\zeta(\Omega, O), V_\Omega = po_e^O) \tag{5.1}$$

- $e \in \Omega$, e is an instance of the event model Ω.
- $\zeta(\Omega, O)$, the constraints in event model Ω are satisfied by observation O.
- $V_\Omega = po_e^O$, the tracked physical objects in the observation O correspond to the physical object variables in the model Ω.
- Observation O is one observation of the event instance e.
- $P(e \in \Omega|\zeta(\Omega, O), V_\Omega = po_e^O)$ corresponds to the conditional probability.

5.3.1 Bayesian Probability Theory

Bayes rule involves the manipulation of conditional probabilities. Let's remember that the joint probability of two events, A and B, can be expressed as

$$\begin{aligned} P(A, B) &= P(A|B) \times P(B) \\ &= P(B|A) \times P(A) \end{aligned} \tag{5.2}$$

In Bayesian probability theory [Olshausen, 2004], one of these events is the hypothesis, H, and the other is data, D, and we wish to estimate the relative truth of the hypothesis given the data. According to Bayes rule, we do this via the relation:

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(D)} \tag{5.3}$$

Specifically Bayes's theorem states that the posterior probability of a hypothesis is equal to the product of **(a)** the prior probability of the hypothesis and **(b)** the conditional probability of the data (evidence) given the hypothesis, divided by **(c)** the probability of the data.

- The term $P(D|H)$ assesses the probability of the observed data arising from the hypothesis. Usually this is known by the experimenter, as it expresses one's knowledge of how one expects the data to look given that the hypothesis is true.
- The term $P(H)$ is called the prior, as it reflects one's prior knowledge before the data are considered. To evaluate the probability of a hypothesis, the Bayesian probabilist specifies some prior probability, which is then updated in the light of new, relevant data.

- The term $P(D)$ is obtained by integrating (or summing) $P(D|H).P(H)$ over all H, and usually plays the role of an ignorable normalizing constant.

- Finally, the term $P(H|D)$ is known as the posterior, and as its name suggests, reflects the probability of the hypothesis after consideration of the data.

Bayes theorem depends upon an ingenious turnabout[Paulos, 2011]: If you want to assess the strength of your hypothesis given the evidence, you must also assess the strength of the evidence given your hypothesis. In the face of uncertainty, a person with a Bayesian thinking asks three questions:
- How confident am I in the truth of my initial belief?
- On the assumption that my original belief is true, how confident am I that the new data (evidence) is accurate?
- And whether or not my original belief is true, how confident am I that the new data (evidence) is accurate?

Let's us now discuss the mathematical development of the probabilities of equation (5.1). It can be useful to bear in mind the more general equations for calculating the joint probability $P(A, B, C)$ and the conditional probability $P(A|B, C)$. Given that: $P(A, B) = P(A|B).P(B)$, we know from basic probability theory that we can factor the joint probability as a product of conditional probability :

$$P(A, B, C) = P(A|B, C) \times P(B, C) \tag{5.4}$$

$$P(A, B, C) = P(B|A, C) \times P(A, C) \tag{5.5}$$

From equations (5.4) and (5.5):

$$P(A|B, C) = \frac{P(B|A, C) \times P(A, C)}{P(B, C)} \tag{5.6}$$

since $P(A, C) = P(C|A) \times P(A)$ and the Bayesian denominator $P(B, C)$ is $P(B, C) = P(B, C|A)P(A) + P(B, C|not(A))P(not(A))$, these terms can be replaced in the previous equation as following (equation 5.7), $not(A)$ is denoted as $\neg a$:

$$P(A|B, C) = \frac{P(B|A, C) \times P(A, C)}{P(B, C)} = \frac{P(B|A, C) \times P(C|A) \times P(A)}{P(B, C|A)P(A) + P(B, C|\neg A)P(\neg A)} \tag{5.7}$$

To complete the mathematical development of $P(A|B, C)$, It can be noted that B, C are conditionally independent given A. Which means that:

$$P(B|A, C) = P(B|A) \tag{5.8}$$

In that case, we can simplify (5.7) using (5.8):

$$P(A|B,C) = \frac{P(B|A) \times P(C|A) \times P(A)}{P(B,C/A)P(A) + P(B,C/\rceil A)P(\rceil A)} \tag{5.9}$$

5.3.2 Probability of Recognizing an primitive state

Now that the mathematical development of $P(A|B,C)$ is complete, we can focus on equation (5.1) to compute the probability of the recognition of primitive states. A, B and C are the following:

- **A:** $e \in \Omega$, e is an instance of event model Ω.
- **B:** $\zeta(\Omega, O)$, the constraints of event model Ω are satisfied by observation O.
- **C:** $V_\Omega = po_e^O$, the tracked physical objects in the observation O match the physical object variables in the model Ω.

The equation (5.1) can be calculated as shown in (5.9):

$$P(e \in \Omega|\zeta(\Omega,O), V_\Omega = po_e^O) = \frac{P(\zeta(\Omega,O)|e \in \Omega) \times P(V_\Omega = po_e^O|e \in \Omega) \times P(e \in \Omega)}{P(\zeta(\Omega,O), V_\Omega = po_e^O|e)P(e) + P(\zeta(\Omega,O), V_\Omega = po_e^O|\rceil e)P(\rceil e)} \tag{5.10}$$

In the next section, we detail the computation of the probabilities for primitive states. To better explain the proposed equations, an example will be used.

5.3.2.1 Probability Computation

In this section, we specify how to compute each term of the equations (5.10). We attempt here to estimate the probability distributions from the available training data. The estimation of probability distributions, is a very large and complex domain. We describe here the most widely used method in the framework of Bayesian networks, the *maximum likelihood estimation* method. In the case where all variables are observed, the simplest method and the most widely used statistical estimation is to estimate the probability of an event by the frequency of occurrence of the event in the database. This approach, called maximum likelihood (ML) [Leray, 2006]:

$$\hat{P}(X_i = x_k|Y_j = y_e) = \hat{\theta}^{ML}_{i,j,k,e} = \frac{N_{i,j,K,e}\ \ (X_i = x_k, Y_j = y_e)}{\sum_k N_{i,j,K,e}} \tag{5.11}$$

Where $N_{i,j,K}$ is the number of events in the database for which the variable X_i is in state x_k and his parents Y_j are in the configuration y_e.

- The first probability to learn is $P(e \in \Omega)$:
  $P(e \in \Omega)$ is the prior probability that a certain scenario model Ω is detected. We can assume that all scenarios in a certain universe are equally probable, so as not to favor any scenario because it happens more often. By scenario models in the same universe, we mean a set of scenarios which are mutually exclusive (if any of them is happening the others can not occur) and include all possible situations so that in any observation one of them must be happening. For example, the universe of the scenario models that describe a person posture is: (*PersonStanding*,*PersonSitting* and *PersonBending*). The universe of scenario models that describe a person position in a certain area is: (*PersonInsideZoneTV*, *PersonInsideZoneEntrance*, *PersonInsideZoneUseReadingTable*, *PersonInsideZoneExerciseWalking*, *PersonInsideZoneUseCoffeeCorner*, *PersonInsideZoneUseChair*, *PersonInsideZoneUsePhone*, *PersonInsideZonePharmacy*, *PersonInsideZoneOfficeDesk*, *PersonInsideZoneWaterPlant*, *PersonInsideZoneEntrancePath*, *PersonInNoMarkedZone*).

$$P(e \in \Omega) = \frac{1}{Nbr.Scenario\Omega Universe} \tag{5.12}$$

  This prior probability has to be learned from a training set of video sequences which has to be large and representative (e.g. the same illumination conditions, etc.).

- The second probability to be computed is $P(\zeta(\Omega, O)|e \in \Omega)$:
  $P(\zeta(\Omega, O)|e \in \Omega)$ is the probability that the constraints of the event model are verified given that the event e is true (i.e. has been annotated as an instance of the event model Ω). This probability quantifies how likely it is that a constraint of event model Ω should be verified when an instance of Ω is taking place. In this section we explain how we compute this probability, we present the first formalization 5.13 and we detail based on other information, how we modify it to get the final equation 5.14. The first formulation of this probability is as following:

$$P(\zeta(\Omega, O)|e \in \Omega) = \prod_{i=1}^{n} \frac{\sharp(\zeta(\Omega, O)_i \wedge e \in \Omega)}{\sharp(e \in \Omega)} \tag{5.13}$$

  where n is the total number of constraints that are being considered. $\sharp(a)$ is the number of frames where a is verified. The term $\sharp(\zeta(\Omega, O)_i \wedge e \in \Omega)$ implies that only frames where event e has been identified (i.e. annotated) as an instance of Ω are considered, and for each constraint of event model Ω, the number of frames where it is satisfied are counted. $\sharp(e \in \Omega)$ is the total number of frames where the event e is annotated. Conditional independence among the constraints is assumed. Thus, the probability of all the constraints that are being considered is calculated as the multiplication of the probabilities of each of the constraints.

It is important that the frames where the scenario is said to be identified should be annotated (manually or automatically). Since we want to assess how much the verification of the constraint affects the event's detection, the cases where the event is present but the constraint is not verified must be identified. It is necessary to determine the event instances that are in the ground truth GT but are not detected, and assess what is causing this failure in detection.

The physical objects that intervene in the constraint should be added in the equation. Otherwise, a failure in detecting the physical objects that intervene in the constraint might result in low probabilities for that constraint, when the problem is in fact due to a failure of physical object detection. In other words, the equation as it is penalizes both following cases:
- The physical object is detected but the constraint is not verified,
- The physical object is not detected resulting that the constraint is not verified.

To take into account the intervention of the physical objects in the calculation of this probability, the final equation (5.14) to compute the probability $P(\zeta(\Omega, O)|e \in \Omega)$ becomes:

$$P(\zeta(\Omega, O)|e \in \Omega) = \prod_{i=1}^{n} \frac{\sharp(\zeta(\Omega, O)_i \wedge e \in \Omega)}{\sharp(e \in \Omega \wedge V_\Omega^{\zeta_i} \in po_e^O)} \tag{5.14}$$

Where V^{ζ_i} are the physical object variables that intervene in the constraint ζ_i.

The term $\sharp(e \in \Omega \wedge V_\Omega^{\zeta_i} \in po_e^O)$ indicates that we only consider the frames of the ground truth where the event e is annotated and the physical objects are correctly tracked. We do not take into account the frames of the ground truth where the physical object are not correctly tracked.

- The third probability to be computed is $P(V_\Omega = po_e^O|e \in \Omega)$:
  $P(V_\Omega = po_e^O|e \in \Omega)$ is the probability that the physical object variables in the event model Ω have been detected given that e is an event instance of the event model Ω. As in the previous factor, conditional independence among the physical objects is assumed and the probabilities obtained for each physical object are multiplied. This probability quantifies the likelihood of detecting a physical object when an event that involves it is actually happening. This probability is computed on-line and provided from the tracking algorithm as described in [Chau, 2012]. The algorithm in [Chau, 2012] computes the reliability of the trajectory of each mobile object po_t detected at instant t. For each object detected at instant t, the algorithm considers all its matched objects po_{t-n} (i.e. objects with temporarily established links) in previous frames that do not have yet official links

(i.e. trajectories) to any objects detected at the following frames. For such an object pair (po_t, po_{t-n}), the algorithm defines a global score $GS(po_t, po_{t-n})$ as follows:

$$GS(po_t, po_{t-n}) = \frac{\sum_k w_k GS_k(po_t, po_{t-n})}{\sum_k w_k} \tag{5.15}$$

where w_k is the weight of descriptor k, the descriptors are 2D, 3D positions, 2D shape ratio, 2D area, color histogram, histogram of oriented gradient (HOG), color covariance and dominant color. $GS_k(po_t, po_{t-n})$ is the global score for descriptor k between po_t and po_{t-n}, defined in function of link score and long-term score for descriptor k as follow :

$$GS_k(po_t, po_{t-n}) = (1-\beta).\ DS_k(po_t, po_{t-n}) + \beta.\ LT(po_t, \chi_{t-n}) \tag{5.16}$$

where β is the weight of long-term score, $DS_k(po_t, po_{t-n})$ is the similarity score for descriptor k between the two objects po_t and po_{t-n}; $LT(po_t, \chi_{t-n})$ is their long-term scores (for more detail see [Chau, 2012], sections 6.1.1.1, 6.1.2.2 and 6.1.2.3).

However in the case where the tracking algorithm does not provide this probability value, an alternative way to compute this probability is provided in the following equation:

$$P(V_\Omega = po_e^O | e \in \Omega) = \prod_{k=1}^{m} \frac{\sharp(V_\Omega^k \in po_e^O \wedge e \in \Omega)}{\sharp(e \in \Omega)} \tag{5.17}$$

Where m is the number of physical objects that intervene in constraint of Ω. $\sharp(V_\Omega^k \in po_e^O \wedge e \in \Omega)$ denotes the number of frames where the physical object is correctly tracked in the total frames where e is annotated. The training set is composed of the frames where the event e is annotated. $\sharp(e \in \Omega)$ is the total number of frames where the event e is annotated.

- $P(\zeta(\Omega, O), V_\Omega = po_e^O)$ is the Bayesian probability denominator. This probability is obtained by an integration (summation) over all the hypotheses. In this case, the first hypothesis is that a certain event instance e of event model Ω is detected, the other hypothesis is that an event instance $\rceil e$ of event model Ω' is detected. This probability is computed based on the following equation (5.18):

$$\begin{aligned} P(\zeta(\Omega, O), V_\Omega = po_e^O) &= P(\zeta(\Omega, O), V_\Omega = po_e^O | e \in \Omega) \times P(e \in \Omega) + \\ &\quad P(\zeta(\Omega, O), V_\Omega = po_e^O |\ e \in \rceil\Omega) \times P(e \in \rceil\Omega) \\ &= \frac{(\sharp(\zeta(\Omega, O), V_\Omega = po_e^O) \wedge e \in \Omega)}{\sharp(e \in \Omega)} \times P(e \in \Omega) + \\ &\quad \frac{(\sharp(\zeta(\Omega, O), V_\Omega = po_e^O) \wedge e \in \rceil\Omega)}{\sharp(e \in \rceil\Omega)} \times P(e \in \rceil\Omega) \end{aligned} \tag{5.18}$$

$\sharp(\zeta(\Omega, O), V_\Omega = po_e^O) \wedge e \in \Omega)$ corresponds to the number of frames where the constraint is verified and the physical object is detected, we consider the frames where the event e is annotated. $P(e \in \Omega)$ is the prior probability that the event e of event model Ω is detected. This probability is computed based on the equation 5.12.

$(\sharp(\zeta(\Omega, O), V_\Omega = po_e^O) \wedge e \in]\Omega)$ corresponds to the number of frames where the constraint is verified and the physical object is detected. We consider the frames where the other events of the same universe than e are annotated. The computation of this probability is detailed in the next section.

5.3.2.2 Probability Computation of an Example

We detail the probability computation of the primitive state 'PersonSitting' (fig.5.2) to better explain the presented equations above.

PrimitiveState (PersonSitting,
PhysicalObjects ((p: Person))
Constraints (p→ posture= Sit)
Alarm (Atext ('A person is sitting')
AType ('NOTURGENT')))

Figure 5.2: The event model of the primitive state 'a person is sitting'.

- $P(PersonSitting)$ is the prior probability that the event model $PersonSitting$ is detected. The universe of the scenario models that describes a person posture is: (*PersonStanding*,*PersonSitting* and *PersonBending*). Thus the prior probability $P(PersonSitting)$ is computed as follow:

$$P(PersonSitting) = \frac{1}{Nbr.ScenarioPersonSittingUniverse} = \frac{1}{3} \tag{5.19}$$

- $P(\text{'}Person \rightarrow Posture = Sit'|PersonSitting)$ is the probability that constraint 'Person → Posture = Sit' in the event model is verified given that the event $PersonSitting$ is true (has been annotated). It is computed based on the following equation:

$$P(\text{'}Person \rightarrow Posture = Sit'|PersonSitting) = \frac{\sharp((\text{'}Person \rightarrow Posture = Sit') \wedge PersonSitting)}{\sharp(PersonSitting \wedge po_e^O = Person)} = \frac{3242}{4228} \tag{5.20}$$

$\sharp(\text{`Person} \rightarrow \text{Posture} = \text{Sit}' \wedge \text{PersonSitting})$ is the number of frames where the constraint `Person $\rightarrow$ Posture = Sit' is verified. We consider only the frames where the primitive state PersonSitting is annotated. $\sharp(\text{PersonSitting} \wedge \text{po}_e^O = \text{Person})$ is the number of frames where the primitive state PersonSitting is annotated and the physical object $(\text{po}_e^O = \text{Person})$ is correctly tracked. In this probability learning step, 4 videos sequences are considered for the learning. A total of 4228 frames annotated with the primitive state PersonSitting are used. A total of 3242 frames correspond to situation where the constraint `Person $\rightarrow$ Posture = Sit' is verified. It gives us the probability of 0.76 to have the constraint `Person $\rightarrow$ Posture = Sit' verified given that the primitive state PersonSitting is detected.

- $P(\text{po}_e^O = \text{Person}|\text{PersonSitting})$ is the probability that the physical object 'Person' has been detected given an instance of the event model PersonSitting. It is computed based on the following equation:

$$P(\text{po}_e^O = \text{Person}|\text{PersonSitting}) = \frac{\sharp(\text{po}_e^O = \text{Person} \wedge \text{PersonSitting})}{\sharp(\text{PersonSitting})} = \frac{21268}{21534} \tag{5.21}$$

$\sharp(\text{po}_e^O = \text{Person} \wedge \text{PersonSitting})$ is the number of frames where the physical object 'Person' has been correctly tracked. We consider the frames where the primitive state PersonSitting is annotated. $\sharp(\text{PersonSitting})$ is the number of frames where the primitive state PersonSitting is annotated. We use 18 video sequences for this learning step, a total of 21534 frames were annotated with the primitive state PersonSitting. The physical object 'Person' has been correctly tracked for a total number of 21268 frames of the 21534 learning frames. It gives us the probability of 0.98 that the physical object 'Person' is correctly detected given that the event model PersonSitting is annotated.

- $P(\text{`Person} \rightarrow \text{Posture} = \text{Sit}', \text{po}_e^O = \text{Person})$ is the Bayesian probability denominator. It is computed based on the following equation (5.22):

$$\begin{aligned} P(\text{`Person} \rightarrow \text{Posture} = \text{Sit}', \text{po}_e^O = \text{Person}) = \\ P(\text{`Person} \rightarrow \text{Posture} = \text{Sit}', \text{po}_e^O = \text{Person}|\text{PersonSitting}) \times P(\text{PersonSitting}) + \\ P(\text{`Person} \rightarrow \text{Posture} = \text{Sit}', \text{po}_e^O = \text{Person}|\neg\text{PersonSitting}) \times P(\neg\text{PersonSitting}) \\ = \frac{3242}{4228} \times \frac{1}{3} + \frac{6175}{23217} \times \frac{2}{3} = 0.43 \end{aligned} \tag{5.22}$$

$P(\text{`Person} \rightarrow Posture = Sit', po_e^O = Person|PersonSitting)$ corresponds to the number of frames where the constraint is verified and the physical object is detected. We consider the frames where the primitive state $PersonSitting$ is annotated. $P(PersonSitting)$ is the prior probability of the primitive state $PersonSitting$ which is equal to $\frac{1}{3}$. $P(\text{`Person} \rightarrow Posture = Sit', po_e^O = Person|]PersonSitting)$ corresponds to the number of frames where the constraint is verified and the physical object is detected. We consider the frames in the learning set where the other primitive states of type posture (e.g. $PersonStanding$, $PersonBending$) than $PersonSitting$ are annotated. For the calculation of this probability, we use a learning set of 27445 frames: 4228 where the primitive state $PersonSitting$ is annotated and 23217 frames annotated with other primitive states. The constraint and the physical object are correctly detected during 3242 of the 4228 frames where the primitive state $PersonSitting$ is annotated.

Finally, the probability $P(PersonSitting|\text{`Person} \rightarrow Posture = Sit', po_e^O = Person)$ is computed as following:

$$P(PersonSitting|\text{`Person} \rightarrow Posture = Sit', po_e^O = Person) = \frac{0.76 \times 0.98 \times \frac{1}{3}}{0.43} = 0.58 \quad (5.23)$$

The probability that the primitive state $PersonSitting$ is detected given that its physical objects in the model $\Omega_{PersonSitting}$ have been observed and given that the constraints in the model $\Omega_{PersonSitting}$ are satisfied is equal to 0.58.

5.4 Second Stage of Activity Recognition

The second stage of the activity recognition algorithm is the recognition of composite events. The recognition of composite events requires two steps: first, the algorithm recognizes all the sub-events composing the composite event and second, the system has to check whether the time intervals of the detected sub-events satisfy the relationships in the event model as described in [Vu et al., 2004].

The recognition of a given event is triggered only if its last sub-event (called event terminaison) is recognized which avoids an exponential computation. Thus the algorithm runs in real time since only the events which their terminaison is recognized are processed. All the composite events are decomposed into states and events composed at the most of two components. Then the recognition of composite states and events is performed similarly to the recognition of primitive states. To recognize the predefined event models at each instant, we do first a selection of a set of event triggers that indicate which events can be recognized. These trig-

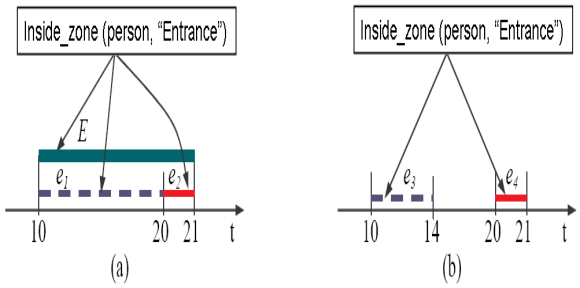


Figure 5.3: (a) an example of merging two event instances p_1 and p_2 of the same type into the same event instance. (b) two event instances e_3 and e_4 that can not be merged.

gers correspond to a primitive state or to a composite event that terminates with a component recognized at the previous or current instant.

For each of these event triggers, solutions are found by looking for component instances already recognized in the past to complete the recognition of event.

A *solution* of an event model Ω is a set of physical objects that are involved in the event and the list of corresponding component instances satisfying all the constraints of the event model Ω.

To prevent from event fragmentation, we consider that if at the previous instant, an event instance p' of the same type (same model, same physical objects) was recognized on a time interval $[t_0,t_1]$ with $|t_1 - t| < \delta$, the two event instances p_1 and p_2 are merged into an instance that is recognized on the time interval $|t_0 - t|$ (fig. 5.3).

5.4.1 Event Recognition Algorithm

We define a trigger as an event which can be potentially recognized. There are two kinds of triggers : the primitive state trigger (type 1), and the composite events trigger (type 2). A list LT of triggers is initialized with all triggers of type 1 (i.e. primitive states).

In the case of primitive state model Ω, a ***solution*** of Ω is a set of physical objects involved in the event and that are satisfying all the list of the constraints in the event model Ω. In the case of composite event model Ω', a ***solution*** of Ω is a set of physical objects that are involved in the event and the list of corresponding component instances satisfying all the constraints of the event model Ω'.

Once an event e is recognized, the algorithm try to extend it with a previously recognized event e'. If the event e is extensible, the algorithm adds then all the triggers of type 2 that terminate with e to the list of trigger and creates an event instance stored on the list of recognized events (see algorithm 2).

```
begin
    Initialization
    create a list of triggers LT including the triggers T_e for the primitive
    state models (type 1) and the triggers T_c for the composite event
    models (type 2)
    for each solution e of LT do
        if e is not extensible then
            Create an event instance e
            Store e to the list of recognized events
            add all triggers of type 2 that terminates with e to the list of
            trigger LT
        end
        if e is extensible with e' recognized at previous time t then
            Merge e with e'
            Create a composite event instance e_c
            Store e_c to the list of recognized events
            add all triggers of type 2 that terminates with e to the list of
            trigger LT
        end
    end
end
```

Algorithm 2: Event recognition algorithm

5.5 Probabilistic Composite Event Recognition

The probabilistic recognition of complex event estimation problem is defined as a hierarchical Bayesian inference. The objective is to recognize the complex event e given an observation O. That is, what we want to calculate here is :
'The probability that a complex event instance e ***belongs to an event model*** Ω ***given that the components (sub-events) in the model*** Ω ***are observed and the constraints in the model*** Ω ***are satisfied by the observation'***.
In mathematical words, the proposed way of calculating this is:

$$P(e \in \Omega, SE(\Omega, O), \zeta(\Omega, O)) = P(e \in \Omega | SE(\Omega, O), \zeta(\Omega, O)) \times P(SE(\Omega, O), \zeta(\Omega, O)) \quad (5.24)$$

- $e \in \Omega$, e is an instance of event model Ω.

- $\zeta(\Omega, O)$, the constraints in event model Ω are satisfied by observation O.
- $SE(\Omega, O)$, the components (sub-events) of the model Ω are observed.
- The first factor $P(e \in \Omega | SE(\Omega, O), \zeta(\Omega, O))$ corresponds to the Bayesian probability.

as demonstrated in section 5.3.2, the probability $P(e \in \Omega | SE(\Omega, O), \zeta(\Omega, O))$ is computed as follow:

$$P(e \in \Omega | SE(\Omega, O), \zeta(\Omega, O)) = \frac{P(SE(\Omega, O) | e \in \Omega) \times P(\zeta(\Omega, O) | e \in \Omega) \times P(e \in \Omega)}{P(SE(\Omega, O), \zeta(\Omega, O))} \quad (5.25)$$

5.5.1 Probabilistic Event Recognition Algorithm

To recognize an event model Ω, the algorithm tries to instantiate the components of Ω (sub-events) with event instances previously recognized. A combination of sub-events instances is a solution if it satisfies all the constraints of Ω.

In the proposed probabilistic event recognition algorithm (see algorithm 3 and algorithm 4), we propose (i) to probabilistically verify the constraint used in the event models and (ii) to compute the Bayesian probability of the event. The Bayesian probability computation is detailed in section 5.5.2. The probabilistic verification of constraints is detailed in section 5.6.

5.5.2 Probability Computation

In this section, it shall be specified how to compute each term of the equations (5.25)

- the first probability to be estimated is $P(e \in \Omega)$:
  $P(e \in \Omega)$ is the prior probability that a certain scenario model Ω is detected. We can assume that all scenarios in a certain universe are equally probable, so as not to favor any scenario just because it happens more often. By scenario models in the same universe, we mean a set of scenarios which are mutually exclusive (if any of them is happening the others can not happen) (5.26).

$$P(e \in \Omega) = \frac{1}{Nbr.Scenario\Omega Universe} \quad (5.26)$$

- The second probability to be estimated is $P(\zeta(\Omega, O) | e \in \Omega)$:
  $P(\zeta(\Omega, O) | e \in \Omega)$ is the probability that constraints of the event model are verified given that the event e is true (i.e. has been identified as an instance of the event model Ω). This probability quantifies how likely it is that a constraint of event model Ω should be verified when an instance of Ω is taking place. It is computed in the same way that in the case of primitive states:

```
begin
    Data:
    Ω : event model.
    List_Constraints : List of the event constraints.
    List_Components : List of components (i.e. sub-events) of the event model Ω with
          previously recognized instances.
    Result: Find a probabilistic solution for the event model Ω.
    if (Probabilistic Constraint Verification (List_Components , List_Constraints ) )
    then
        if BayesianRecognition (List_Components , List_Constraints ) then
            Create(e)     // a probabilistic solution for the event model Ω
        end
    end
end
```

Algorithm 3: Probabilistic Solution for composite event model. In the function *Probabilistic Constraint Verification*, the algorithm verifies probabilistically if the constraints in the event model Ω are satisfied. In function *BayesianRecognition*, the algorithm computes the Bayesian probability of the event.

$$P(\zeta(\Omega, O)|e \in \Omega) = \prod_{i=1}^{n} \frac{\sharp(\zeta(\Omega, O)_i \wedge e \in \Omega)}{\sharp(e \in \Omega \wedge V_{\Omega}^{\zeta_i} \in po_e^O)} \tag{5.27}$$

where $\sharp(a)$ is the number of frames where a is verified and i is the total number of constraints that are being considered. Only frames where event e have been identified as an instance of Ω are considered, and for each constraint in event model Ω, the number of frames where it is satisfied are counted. Conditional independence among the constraints is assumed (even if they are not actually independent, it would simplify the calculations). Thus, the probability of all the constraints that are being considered is calculated as the multiplication of the probabilities of each of the constraints.

It is important that the frames where the scenario is said to have been identified should be annotated (manually or automatically). Since we want to assess how much the verification of the constraint affects the event's detection, the cases where the event is present but the constraint is not verified must be identified. It is necessary to determine the event instances that are in the GT but are not detected, and assess what is causing this failure in detection.

The physical objects that intervene in the constraint are added in the equation. Otherwise,

```
begin
    Initialization
    create a list of triggers LT including the triggers T_e for the primitive
    state models (type 1) and the triggers T_c for the composite event
    models (type 2)
    for each Probabilistic solution e of LT do
        if e is not extensible then
            Create an event instance e
            Store e to the list of recognized events
            add all triggers of type 2 that terminates with e to the list of
            trigger LT
        end
        if e is extensible with e' recognized at previous time t then
            Merge e with e'
            Create a composite event instance e_c
            Store e_c to the list of recognized events
            add all triggers of type 2 that terminates with e to the list of
            trigger LT
        end
    end
end
```

Algorithm 4: Probabilistic Event recognition algorithm: once we compute the probabilistic Solution for composite event model detailed in algorithm 3, we try to recognize probabilistically the event.

a failure in detection of physical object that intervene in the constraint might result in low probabilities for that constraint, when the problem is in fact due to a failure of physical object detection.

- $P(SE(\Omega, O)|e \in \Omega)$ is the probability that the sub-event variables in the event model Ω have been detected given that e is an event instance of the event model Ω. As in the previous factor, independence among the sub-events is assumed and the probabilities obtained for each sub-event are multiplied. This probability quantifies the likelihood of detecting a sub-event when an event that involves it is actually happening. It is computed based on the following equation:

$$P(SE(\Omega, O)|e \in \Omega) = \prod_{k=1}^{m} \frac{\sharp(SE^k(\Omega, O) \wedge e \in \Omega)}{\sharp(e \in \Omega)} \tag{5.28}$$

Where m is the number of sub-events of the event e. The sub-events of an event e can be primitive states or composite events.

- $P(\zeta(\Omega, O), SE(\Omega, O))$ the Bayes denominator is computed based on the following equation in the same way that described in section 5.3.2.1:

$$\begin{aligned} P(\zeta(\Omega, O), SE(\Omega, O)) &= P(\zeta(\Omega, O), SE(\Omega, O)|e \in \Omega) \times P(e \in \Omega) + \\ &\quad P(\zeta(\Omega, O), SE(\Omega, O)|e \in \rceil\Omega) \times P(e \in \rceil\Omega) \\ &= \frac{(\sharp(\zeta(\Omega, O), SE(\Omega, O)) \wedge e \in \Omega)}{\sharp(e \in \Omega)} \times P(e \in \Omega) + \\ &\quad \frac{(\sharp(\zeta(\Omega, O), SE(\Omega, O)) \wedge e \in \rceil\Omega)}{\sharp(e \in \rceil\Omega)} \times P(e \in \rceil\Omega) \end{aligned} \tag{5.29}$$

5.5.3 Discussion

The proposed approach for probabilistic recognition of activities is based on Bayesian probability theory which provides a consistent framework for reasoning and decision making under uncertainty, using probability. The computed Bayesian probability corresponds to the degree of belief that an event occurs based on statistics about its constraints and its components. Based on the value of this probability, we make the decision that an event is recognized or not. The main problem is that based only on Bayesian probability computation, the algorithm will not take into account the uncertainty of the on-line sensors measurements. The sensors measurements can deeply affect the verification of the activity constraints (i.e. spatial and temporal constraints) which can lead to the failure of the detection of this activity. To deal with this problem and to enhance the recognition performance of the algorithm, we propose an approach for on-line

probabilistic constraints verification. In the following, we detail the proposed approach to study the uncertainty of the verification of the spatial and temporal constraints of activities.

5.6 Probabilistic Constraints Verification

Addressing the noise arising from low-level video analysis has inspired the proposed approach for probabilistic constraints verification. In this section we detail how we probabilistically verify the constraint used in the event models. We detail the verification of the spatial constraint, the posture constraint and the temporal constraints.

5.6.1 Probabilistic Spatial Constraint Verification

Typically in video interpretation, objects are identified and then tracked, such that a unique identifier can be associated with the different positions of the object at different times. However, the position/shape of the object may changes, either because it actually changes (e.g. a person moving, a person changing its posture), or because, in the image plane, an object appears to get larger as it moves closer to the camera, or because of the visual noise which results in the object detection/tracking processes assigning a different shape/position to the object in different frames.

It is this final problem which is of particular concern to us, since the changes are not 'real changes' but rather artefacts of vision algorithms. Often the size/position of the object changes rapidly from frame to frame. Such problem may cause undesirable changes of spatial relation and affects the activity recognition process.

In the process of spatial constraint detection, we take into account **(i)** the geometrical uncertainty which is related to the verification of the constraint (e.g. verifying the spatial constraint inside-zone consists in the geometrical computation if a point representing the person is inside a polygon representing the zone). This type of uncertainty handling depends on the type of the constraints as described in the sections 5.6.1.1 and 5.6.1.2. and **(ii)** the uncertainty of the attribute values due to noisy data and low-level algorithm errors. This type of uncertainty handling is described in the section 5.7.2 by proposing a new dynamic model for the re-estimation of the attribute value to deal with noisy data.

5.6.1.1 Inside-zone

In this section we study the relationship that can have a mobile object with the contextual object zone, the spatial relation *inside-zone* and we propose a probabilistic approach for the verification of this spatial relation. This approach can be generalized and extended for other spatial relationship between mobile objects and zones. The spatial relation *inside-zone* requires

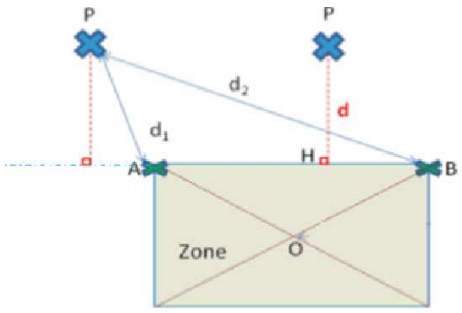


Figure 5.4: The computation of the distance of a person P to a zone.

two physical objects, a mobile object (e.g. person p) and a contextual object of type zone z. This relation ***(p in z)*** verifies wether the mobile object (e.g. person) is inside a zone z.

A main problem in the verification of this constraint is: **(1)** the imprecision and uncertainty in the detection of the relative location of a mobile object to the zone due to low level detection errors (e.g. reflections, shadows or occlusions). Thus, the verification of the constraint can fail. Another issue consists in **(2)** the value of position attribute itself (e.g. wrong value estimation, value change rapidly from frame to frame). A solution to cope with the first problem (1) is to propose a probabilistic distance-based analysis for the verification of the constraint. It is a two-step approach:

- The first step consists in computing the distance between the 3D position of the tracked mobile object (i.e. person) and the zone of interest. The person is represented by the position of its feet corresponding to the middle of the bounding box bottom segment. The zones of interest are represented as polygon. In the following we note (x_p, y_p, z_p) the 3D position of the mobile object and '$dist$' its distance to the zone. We have defined a method called border method, to compute this distance between the 3D position of the mobile objects with respect to the border of the zone of interest.

  The border method consists in computing the distance of the person P from each of the zone borders. As shown in figure 5.4, this step consists in calculating whether the orthogonal projection (noted H) of P on segment [AB] belongs to the same segment. If H $\in$ [AB] then the distance of the person to the zone is ||PH||, it is the minimal one. Otherwise, we calculate the distance of P to the two extremities of the segment [AB] and the minimal distance of P to [AB] is the Min (d1,d2); with d1= ||AP|| and d2 = ||PB||.

- For the second step, we have to choose the exact shape of the distribution, which depends on the exact situation we want to model. The goal is to find a function that describes the likelihood that a person is inside a zone given the distance of the person to this zone. In

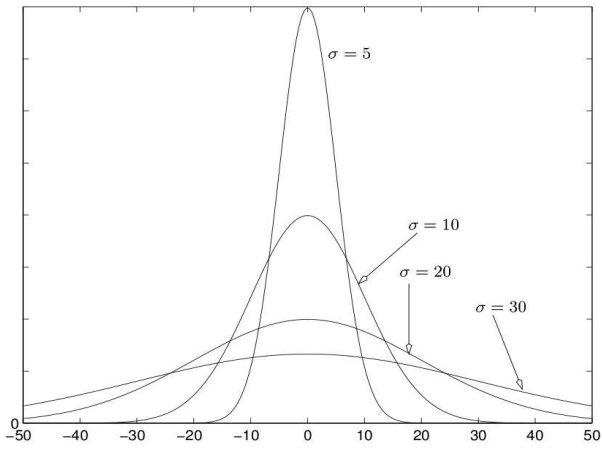


Figure 5.5: Gaussian probability distributions, with mean 0 and different standard deviations σ. The parameter μ determines the location of the distribution while σ determines the width of the bell curve. The standard deviation σ shows how much variation or "dispersion" exists from the average. A low standard deviation indicates that the data points tend to be very close to the mean; high standard deviation indicates that the data points are spread out over a large range of values.

probability theory, it is to find a probability distribution function (PDF) that maximizes the value of probability when the distance '$\mathtt{dist}'$ is small (i.e. person inside the zone/ person near the zone) and a minimum value when the distance '$\mathtt{dist}'$ is big (i.e. person far from the zone).

There are different probability models proposed in the literature, and the appropriate model depends on the distribution of the outcome of interest. A very common distribution that is often used in probability theory is the Gaussian distribution (fig.5.5), it needs only two parameters to represent the normal distribution denoted by $\mathcal{N}(\mu, \sigma)$: the mean μ of the distribution and the standard deviation σ (Eq. 5.30).

The normal (Gaussian) probability model applies when the distribution of the outcome conforms reasonably well to a normal or Gaussian distribution, which resembles a bell shaped curve. Note, normal probability model can be used even if the distribution of the continuous outcome is not perfectly symmetrical; it just has to be reasonably close to a normal or Gaussian distribution [Sullivan and LaMorte, 2013].

$$\mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(\mathtt{dist} - \mu)^2}{2\sigma^2}) \tag{5.30}$$

In this context, the outcomes represent the value of the distance of the person to the zone of interest. Figures 5.6, 5.7 and 5.8 are graphical representations showing a visual impression of

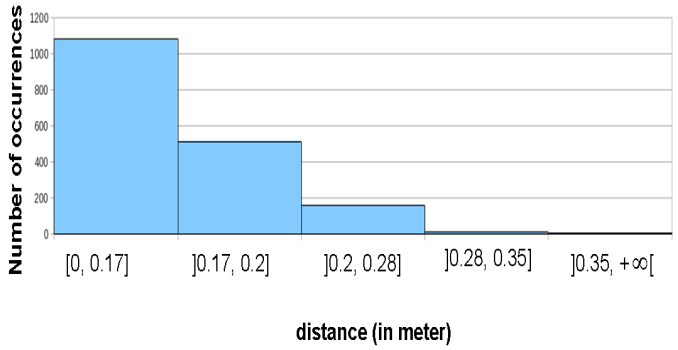


Figure 5.6: Histogramme of the distance distribution; distance(person, reading zone).

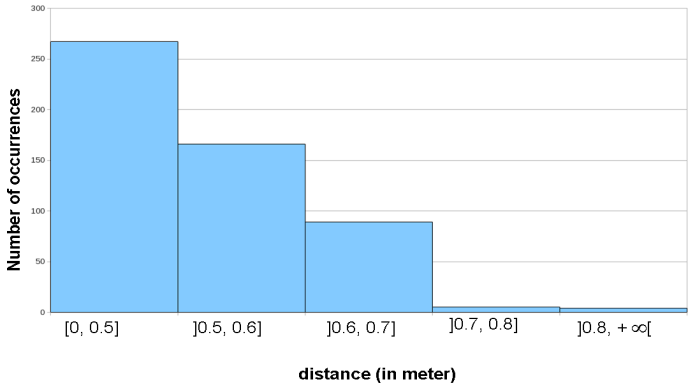


Figure 5.7: Histogramme of the distance distribution; distance(person, coffee corner zone).

the distribution of data. These histograms show the underlying frequency distribution (shape) of the distance values shown as rectangles, erected over discrete intervals, with an area equal to the frequency of the observations in the interval. The height of a rectangle is equal to the frequency density of the interval. These histograms show the proportion of cases that fall into each of several categories. The categories are usually specified as consecutive, non-overlapping intervals of a variable in our context the distance variable. These figures show and let us discover that the distribution of these distances fit into a normal distribution.

In figure 5.6, we take a video sequence where the ground truth indicates that a person is inside a zone called 'reading zone', we compute and store at each frame the distance of the person to this zone. We plot the histogram that is used for estimating the probability density function of the variable distance. In the same manner we plot the histogram of the distance of a person to the zone called 'coffee corner zone' and the zone called 'TV zone'.

Once, we decide for the density distribution function, we have to learn the parameters of

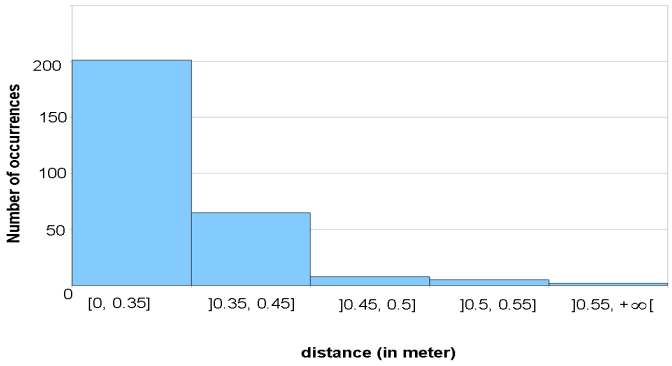


Figure 5.8: Histogramme of the distance distribution; distance(person, TV zone).

this function, i.e. the Gaussian parameters (μ, σ). That is, having a sample $(x_1, ..., x_n)$ from a normal $\mathcal{N}(\mu, \sigma^2)$ population we would like to learn the approximate values of parameters. This calculation is based on equations:

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{5.31}$$

n is the sample size

In our context, to learn the Gaussian parameters, we divide the video database: two-third for learning database and one third for testing. For the learning dataset which is selected to be representative to the whole video database, we calculate and store the distances $dist_i$ of a tracked person to each contextual zone $(dist_1, ..., dist_n)_{z_1}, ..., (dist_1, ..., dist_k)_{z_i}$, etc. Having these samples, we calculate the Gaussian parameters based on the following equations:

$$\hat{\mu} = \bar{dist} = \frac{1}{n}\sum_{i=1}^{n} dist_i, \qquad \hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(dist_i - \bar{dist})^2 \tag{5.32}$$

n is the sample size. More details of the step of Gaussian parameters learning are presented in section 6.7.1.2, chapter 6.

5.6.1.2 Close to

In this section, we study the relationship that can have a mobile object (i.e. person) with the contextual object equipment (e.g. chair, TV), the spatial relation*close-to*. The probabilistic verification of the constraint *Close to* is defined in the same way than the spatial relation*inside-zone*. The proposed approach can be generalized and extended for other spatial relationship between mobile objects and equipment. As described above, to verify the spatial constraint, we compute the distance dist of the person to the objects (i.e. person, equipment), we choose the

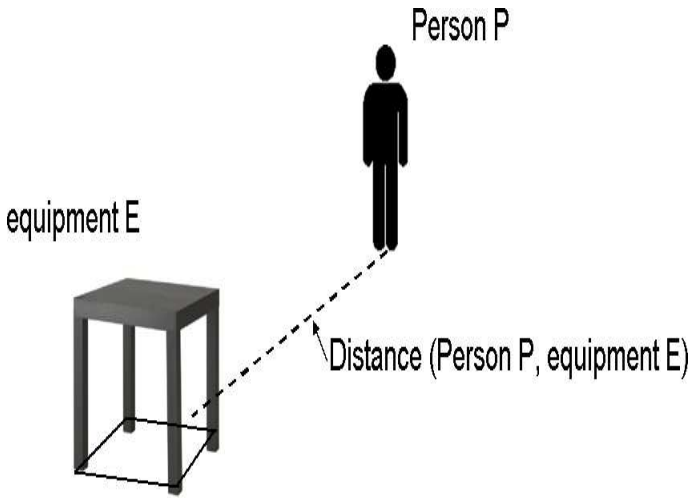


Figure 5.9: Computation of the distance od a person to the equipement.

Gaussian function $\mathcal{N}(\mu, \sigma)$ that gives a maximal value of probability (i.e. 1) when the distance is small and a minimun value (i.e. 0) when the distance is big.

- The first step consists in computing the euclidean distance of the person to the contextual object (i.e. equipment) (fig. 5.9). The tracked person is represented by the position of its feet corresponding to the middle of the bounding box bottom segment. The equipment of interest is represented as 2D polygon. The distance of the person to the equipement is calculated in the same manner than for the spatial relation *inside-zone* (see section 5.6.1.1).

- For the second step, we have to choose the distribution function that describes the likelihood that a person close to the equipment of interest given the distance of the person to this equipment. In probability theory, it is to find a probability distribution function (PDF) that maximizes the value of probability when the distance '$\mathtt{dist}'$ is small (i.e. person near the equipment) and a minimum value when the distance '$\mathtt{dist}'$ is big (i.e. person far from the equipment). As described in section 5.6.1.1, we use the Gaussian distribution function. For each equipment, we calculate and store the distances $\mathtt{dist}_i$ of a tracked person to each contextual equipment $(\mathtt{dist}_1, ..., \mathtt{dist}_n)_{equipment_1}$, ..., $(\mathtt{dist}_1, ..., \mathtt{dist}_k)_{equipment_i}$ etc. For the learning step, we divide the video database: two-third for learning and one third for testing. The learning of the Gaussian parameters (μ, σ) is based on equation 5.31.

5.6.1.3 Discussion

We discussed the approach that we propose for a probabilistic verification of spatial constraints based on the Gaussian probability. We study the relationship that can have a mobile object with the contextual object zone, the spatial relation *inside-zone* and compute the Gaussian probability of this constraints which allows us to decide if this constraint is satisfied or not. This approach can be generalized and extended for other spatial relationship between mobile objects and zones.

We also study the relationship that can have a mobile object (i.e. person) with the contextual object equipment (e.g. chair, TV), the spatial relation *close-to*. The probabilistic verification of the constraint *Close to* is defined in the same way than the spatial relation *inside-zone*. The proposed approach can be generalized and extended for other spatial relationship between mobile objects and equipment. In the following, we discuss the posture and temporal relations and we detail the way that we propose to deal with uncertainty.

5.6.2 Posture

We have selected a set of specific postures which are representative of typical applications in video understanding. These postures are classified in a hierarchical way. We have four general posture categories and eight detailed posture sub-categories:
- Standing postures: standing with one arm up, standing with arms along the body and T-shape posture,
- Sitting postures: sitting on a chair and sitting on the floor,
- Bending posture,
- Lying postures: lying with spread legs and lying with curled-up legs.

The human posture recognition algorithm [Boulay et al., 2007] determines the posture of the detected person using the detected silhouette and its 3D position. The 3D human model silhouettes are obtained by projecting the corresponding 3D human model on the image plane using the 3D position of the person and a virtual camera which has the same characteristics (position, orientation and field of view) as the real camera. A dedicated 3D engine (virtual 3D scene generator) can animate and display a 3D human model. It also extracts the generated model silhouette. Finally, 3D human model silhouettes are compared with the detected silhouette to recognise the posture of the detected person.

The management of uncertainty of this constraint is given from the posture recognition algorithm based on a computed score of the most likely postures (see algorithm 5). The temporal coherence of posture is exploited to compute the posture recognition score: the identifier of the recognized person is used to retrieve the previous detected postures in a temporal window.

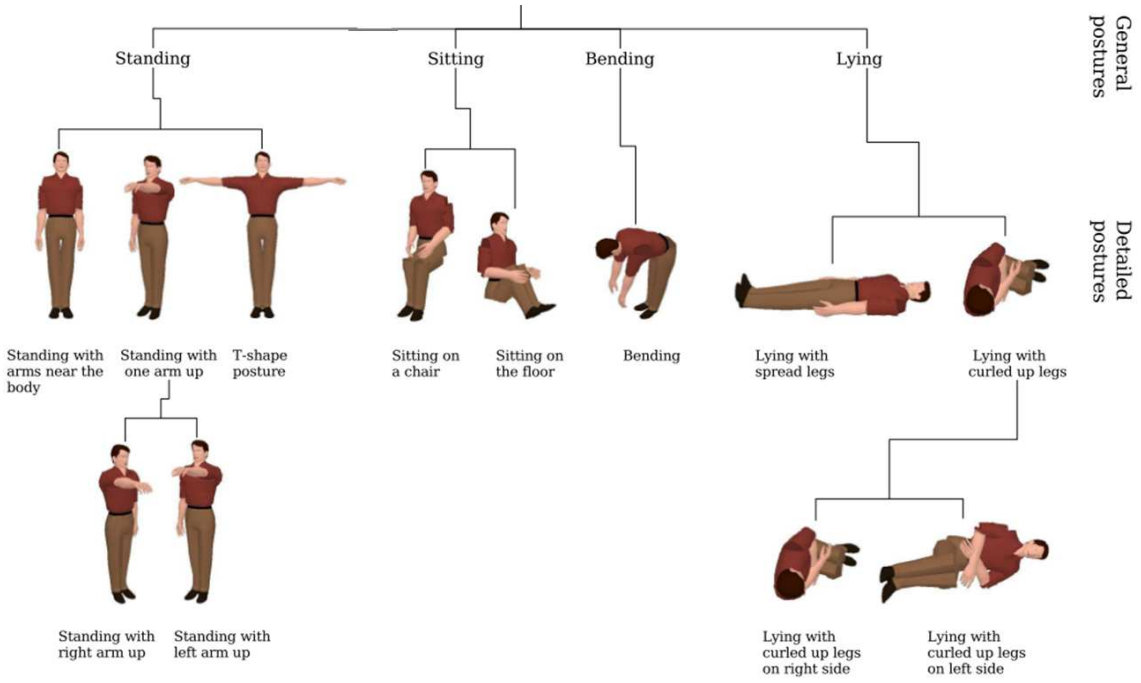


Figure 5.10: Hierarchical representation of the postures of interest [Boulay et al., 2007].

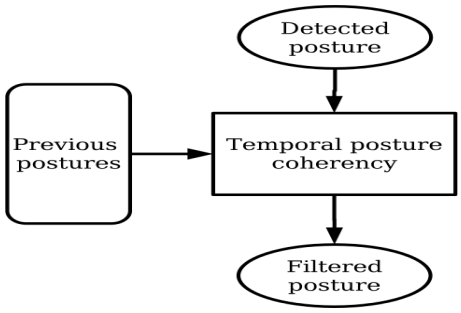


Figure 5.11: Detected posture is compared with the previous detected postures to verify the temporal coherency.

These postures are then used to compute the filtered posture (i.e. the main posture) by searching the most probable posture which corresponds to the most frequent posture for a certain period of time (fig. 5.11).

This smoothing algorithm reduces the misdetection and allows activity recognition algorithm to benefit from reliable filtered postures, more details are available in [Boulay et al., 2007], section 5.4.

```
begin
    Data:
    postureList ← NULL the list which contains the quantity of occurrence of the postures
    for i = -windowSize to windowSize do
        postureList[detectedPOsture[t + i]] += weightList[i]
    end
    return indexOfTheMaximum (postureList) return the posture which occurs the most
    frequently as the filtered posture at time t
end
```

Algorithm 5: Determination of the most probable posture in a temporal window of $2 * windowSize + 1$. The weight list $weightList$ determines how probable the i^{th} posture occurs in the window of $2 * windowSize + 1$.

5.6.3 Probabilistic Temporal constraint verification

We adopt the Allen's temporal predicates [Allen, 1983] which have been widely adopted for the description-based approaches to specify relationships between time intervals. The proposed approach can be generalized for all Allen temporal relations. In this section, we will discuss the temporal relation that have been implemented.

Allen Relation. Allen introduced a calculus for representing the temporal relations of events delimited by time intervals. In particular, he defined 13 qualitative relations between intervals summarized in figure 5.12. For example, if X is a time interval with boundaries $X = [10sec, 12sec]$ and Y is a time interval with boundaries $Y = [2sec, 40sec]$, the only relation hold is: 'X *during* Y'.

However, using temporal relations under real world applications, we often need to deal with uncertain information. The qualitative nature of Allen's relations may not fit very well to temporal segments that result from automated analysis, it may not be robust enough to small temporal perturbations. This can lead to failure of temporal constraint verification.

Lack of Robustness. Automated analysis results are inherently uncertain due to the limited accuracy of the vision algorithms. One source of uncertainty is the inaccuracy of time segment boundaries. It is important that small differences in segmenting interval boundaries should not results in significant difference in their inferred relations. This can happen with Allen's relations, for two reasons. First, several relations (e.g. meet) requires equality of time points, something which can only hold approximately in real-world time intervals. Second, negligible changes to one of the end-points, with respect to the size of the time interval may result in a different qualitative relation as pointed in [Mouhoub and Liu, 2008], [Sergios et al., 2010].

To illustrate more clearly that, let us examine the case of two intervals of approximately the

name	definition	example
Before (b)	$X_b < X_e < Y_b < Y_e$	
Meets (m)	$X_b < X_e = Y_b < Y_e$	
Overlaps (o)	$X_b < Y_b < X_e < Y_e$	
Starts (s)	$Y_b = X_b < X_e < Y_e$	
During (d)	$Y_b < X_b < X_e < Y_e$	
Finishes (f)	$Y_b < X_b < Y_e = X_e$	
Equals (=)	$X_b = Y_b < Y_e = X_e$	
Finishes-i (fi)	$X_b < Y_b < Y_e = X_e$	
During-i (di)	$X_b < Y_b < Y_e < X_e$	
Starts-i (fi)	$X_b = Y_b < Y_e < X_e$	
Overlaps-i (oi)	$Y_b < X_b < Y_e < X_e$	
Meets-i (mi)	$Y_b < Y_e = X_b < X_e$	
After (a)	$Y_b < Y_e < X_b < X_e$	

Figure 5.12: Allen's qualitative relations between two time intervals $X = [X_b, X_e]$ and $Y = [Y_b, Y_e]$.

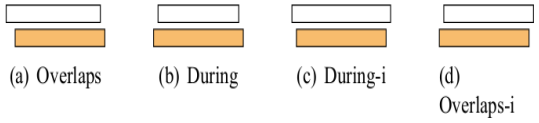


Figure 5.13: Allen's reations lack of robustness.

same size. Let a referred time interval $X = [X_b, X_e]$ and let $Y = [X_b + \epsilon_1, X_e + \epsilon_2]$, ϵ_1 and ϵ_2 are time durations that can be chosen to be arbitrary small. By definition, interval Y is almost equal to X but not necessarily exactly equal. In fact, as depicted in figure 5.13, several of Allen relations may hold between X and Y, depending on ϵ_1 and ϵ_2. Consequently the information that a particular relation holds between two intervals looses its significance, since most other could characterize approximately the same situation.

Inadequacy. Another issue with Allen's relations is their inadequacy in discriminating between very different situations occurring between intervals. In particular, there are cases where a different relation would result from the constraint verification due to the fact that end-point of an interval is slightly perturbed. This failure can be explained by the lack of quantitative information taken into account when verifying the constraint as pointed by [Petridis and Psomas, 2012]. Figure 5.14 depicts this failure in the case of three quite different situations, all being qualified as Overlaps.

Addressing the uncertainty issue, we propose a probabilistic extension of Allen algebra by

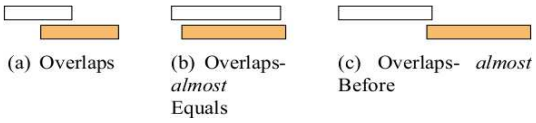


Figure 5.14: Example of Allen relation inadequacy.

providing a formal way for assigning uncertainty value to Allen's relations. This approach is motivated mostly by the need to impose soft temporal constraints to compensate for the inherent uncertainty from vision algorithms and improve the robustness and applicability of Allen's relations.

5.6.3.1 The relations During and During-i

In this section, we study the uncertainty of the temporal relations *'During'* and *'During-i'* and we detail how we propose to compute the probability of these relations. Let's see how the *'During'* relation is defined and see how we extend it and attach a probability measure. *'X During Y'* holds if X is entirely into Y. As shown in figure 5.12, even if a small part of X is not into Y, then this relation will not hold.

In the case the relation holds, we may say that the probability that 'X is during Y' equals 1. However, we may also relax the definition so that when a part of X not in Y is very small, then the relation almost holds.

In this work, The computation of the probability of the temporal relations *'during'* and *'during-i'* (see equations 5.33 and 5.34) depends of **(i)** the size of the two temporal intervals X and Y: i.e. (1) the first interval is smaller than the second, $P(X < Y)$, (2) the same size for the two intervals, $P(X \sim Y)$, (3) the first interval is larger than the second $P(X > Y)$. The probability computation depends also of **(ii)** the the position of an interval with respect to the other interval in the temporal axis, $P(X \asymp Y)$. We note that the probabilities relative to the size of intervals are independent of the probability relative to the intervals position.

$$P(X \ \mathrm{during} \ Y) = P(X < Y).P(X \asymp Y) \tag{5.33}$$

and

$$P(X \ \mathrm{during-i} \ Y) = P(X > Y).P(X \asymp Y) \tag{5.34}$$

In the following, we detail the computation of these probabilities (i.e. $P(X < Y)$, $P(X \sim Y)$, $P(X > Y)$ and $P(X \asymp Y)$). In the purpose to compute each of these probabilities, we propose to define two interval relations.

- **Relation for relative position: the intersection relation**. The intersection relation denoted by $\asymp$ define the intersection interval that can have two temporal intervals. In case

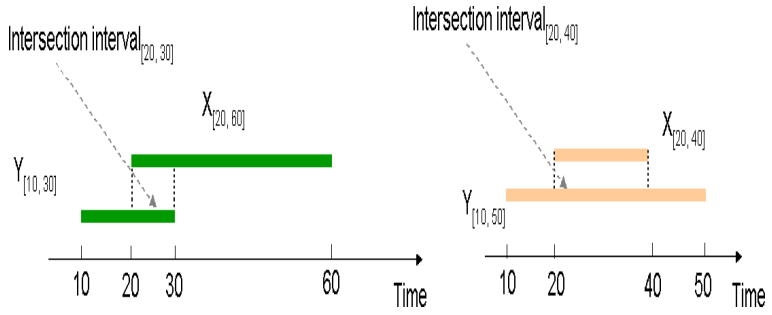


Figure 5.15: Illustration of the intersection relation between two temporal intervals.

this relation holds, we say that the probability of X intersects with Y is equal to 1 (fig. 5.15). Otherwise, we relax the definition of this relation:

Let a point t chosen randomly in X, *what is the probability that* t *is also in* Y? In other words, we measure the probability $P(X \quad intersects \ with \ Y)$ by considering $P(t \in Y | t \in X)$. In order to define a symmetric relation, instead of chosen t from X in all cases, we will be choosing t randomly from the smallest interval, and thus define $P(X \quad \asymp \quad Y)$ as:

$$P(X \quad \asymp \quad Y) \stackrel{def}{=} \begin{cases} P(t \in Y | t \in X) & |X| \leq |Y| \\ P(t \in X | t \in Y) & |X| > |Y| \end{cases} \tag{5.35}$$

Where $|.|$ denote the size of the interval, $|X| = X_e - X_b$. Based on the assumption that there is no reason to choose one time point over another, equation 5.35 becomes:

$$P(X \quad \asymp \quad Y) \stackrel{def}{=} \begin{cases} \frac{|X \cap Y|}{|X|} & |X| \leq |Y| \\ \frac{|X \cap Y|}{|Y|} & |X| > |Y| \end{cases} \tag{5.36}$$

When the ratio value equals 1, it means that the smallest interval is entirely into the larger interval. Small probability value indicates that some part of the smallest interval is not into the larger one. The value 0 means that no part of the interval is within the other.

- **Relation for relative size**. We discuss now the second relation for interval, namely the relation between the size of intervals. Three cases can be enumerated:

  - both intervals have the same size,

  - the first is smaller than the second,

  - the first is larger that the second.

  If we assume that the boundaries of intervals are uncertain, the estimation of the size of intervals is done with some uncertainty.

Considering uncertainty, we may compute the probability that **(1)** an interval X have the same size than Y, $P(X \sim Y)$; **(2)** X is smaller than Y, $P(X < Y)$ and **(3)** X is greater than Y, $P(X > Y)$.

In this work, we propose the computation of these probabilities based on the cumulative probability. In the following we define some useful basis for cumulative probability and then we define the way that we propose for the computation of the temporal relation probability.

Cumulative distribution function

In probability theory and statistics, the cumulative distribution function, usually denoted by Φ, describes the probability that a real-valued random variable $\mathcal{V}$ with a given probability distribution f will be found at value less or equal to v, in case of a continuous distributions, it gives the area under the probability distribution function [Zhang, 2008] (fig.5.16).

$$\Phi(v) = P(\mathcal{V} \leq v) = \int_{\infty}^{v} f(t)dt \tag{5.37}$$

We consider $\hat{v}$ the mean value of a random variable $\mathcal{V}$, following normal distribution.

$$\mathcal{V} \sim \mathcal{N}(\hat{v}, \sigma) \tag{5.38}$$

The probability that $\mathcal{V}$ is smaller or larger a value v is defined using the cumulative probability of $\mathcal{V}$:

$$P(\mathcal{V} < v) = \Phi(v|\hat{v}, \sigma) \tag{5.39}$$

$$P(\mathcal{V} > v) = 1 - \Phi(v|\hat{v}, \sigma) \tag{5.40}$$

Where $\hat{v}$ and σ are the parameters of the normal distribution.

To compute $P(X \sim Y)$, let n be a parameter, $n \in \Re^{+}$, we define the probability of being equal to some value v as:

$$P(\mathcal{V} \sim v) = P(V < v + n) - P(V < v - n) \tag{5.41}$$

The values $v + n$ and $v - n$ may be seen as defining a neighborhood around v.

Now, once we have defined the cumulative probability, we describe in the following the relation between the size of intervals. Let us consider the ratio,

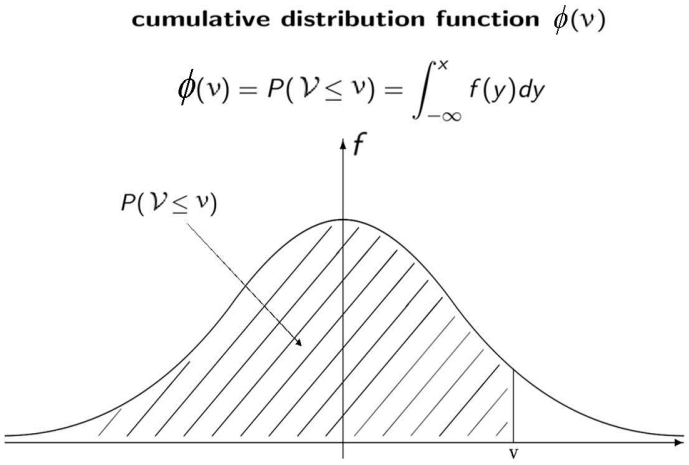


Figure 5.16: The cumulative distribution function Φ describes the probability that a real-value random variable $\mathcal{V}$ with a given probability distribution f will be found at value less or equal to v. it gives the area under the probability distribution function f.

$$\hat{v} = \frac{|X| - |Y|}{|X| + |Y|} \tag{5.42}$$

This ratio takes value from $[-1, +1]$, if the value of this ratio is positive, it means that the interval X has greater size than the interval Y. If the value of this ratio is negative, it means that the interval X has smaller size than the interval Y and if the value of the ratio is null, the interval X has the same size than the interval Y. The upper value limit corresponds to the case where the interval X has infinitely greater size value than Y, the lower value limit corresponds to the case where the interval X has infinitely smaller size value than Y. Thus when evaluating the value of this ratio, we can know the relation between the size of the two intervals X and Y.

Considering uncertainty in the boundaries of intervals, the value of this ratio $\hat{v}$ is uncertain and corresponds more to a region than a single point.

To evaluate the uncertainty of measurement of this ratio, we model this quantity (i.e. ratio value) by a normal distribution ($\mathcal{V} \sim \mathcal{N}(\hat{v}, \sigma)$), the best estimated value of the ratio is the center of two limits v^- and v^+ as described in figure 5.17 [Taylor and Kuyatt, 1994], [Nist, 2000].

Based on all the above definitions and equations 5.39, 5.40 and 5.41, we compute the probability $P(X < Y)$ as the cumulative probability of $\mathcal{V}$ taking negative values. $P(X > Y)$ is the cumulative probability of $\mathcal{V}$ taking positive value and $P(X \sim Y)$ is the cumulative probability of $\mathcal{V}$ taking close to zero values. Figure 5.18 illustrates all these definitions.

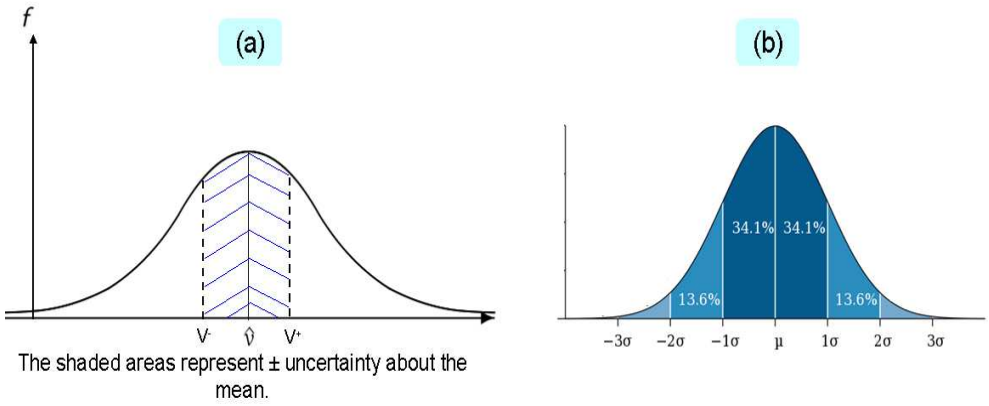


Figure 5.17: In figure **(a)**: To evaluate the uncertainty of measurement of the ratio $\hat{v}$, we model this ratio by a normal distribution $\mathcal{V} \sim \mathcal{N}(\hat{v}, \sigma)$ and estimate lower and upper limits v^- and v^+ such that the best estimated value of this ratio is $(v^+ + v^-)/2$ (i.e. the center of the limits). Figure **(b)** illustrates the repartition of values in a normal distribution: about 68% of values drawn from a normal distribution are within one standard deviation σ away from the mean; about 95% of the values lie within two standard deviations; and about 99.7% are within three standard deviations.

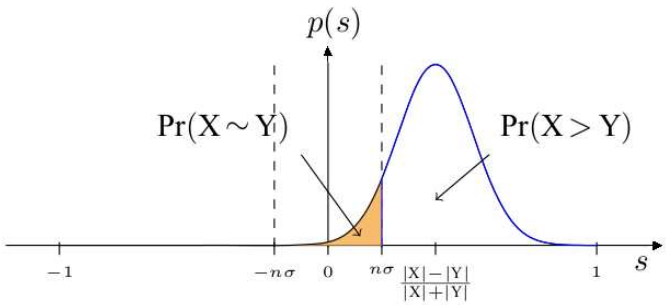


Figure 5.18: Probability definition as areas of $\mathcal{N}(\hat{s}, \sigma)$, delimited by $-n.\sigma$ and $n.\sigma$.

More formally, the above probabilities are expressed as follow:

$$P(X < Y) = \Phi(-n.\sigma|\hat{s}, \sigma) \tag{5.43}$$

$$P(X \sim Y) = \Phi(n.\sigma|\hat{s}, \sigma) - \Phi(-n.\sigma|\hat{s}, \sigma) \tag{5.44}$$

$$P(X > Y) = 1 - \Phi(n.\sigma|\hat{s}, \sigma) \tag{5.45}$$

To compute the cumulative distribution function Φ, we use usually the complemantary error function $erfc$ (eq. 5.46)

$$erfc(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} e^{-t^2}\, dt \tag{5.46}$$

based on the equation 5.46, the probability equations are reduced to:

$$P(X < Y) = \frac{1}{2}\ erfc\ \left(n + \frac{\hat{s}}{\sigma}\right) \tag{5.47}$$

$$P(X \sim Y) = 1 - \frac{1}{2}\left(\ erfc\ \left(n - \frac{\hat{s}}{\sigma}\right) +\ \ erfc\ \left(n + \frac{\hat{s}}{\sigma}\right)\right) \tag{5.48}$$

$$P(X > Y) = \frac{1}{2}\ erfc\ \left(n - \frac{\hat{s}}{\sigma}\right) \tag{5.49}$$

One can easily verify that $P(X < Y) + P(X \sim Y) + P(X > Y) = 1$, i.e. they form a complete probability base.

Dependence on the parameters σ and n: These parameters allow us to model our uncertainty towards the measurement and our tolerance for considering that two intervals are equisized. Namely, σ quantifies the uncertainty regarding the relative size measurement: small value of σ, (e.g.$\sigma \ll 0.1$), should be used when there is a high confidence regarding the measurement, whereas, large value of σ should be used when there is a low confidence regarding the measurement. The parameter n regulates our tolerance in favor of the equisized relation. In this work, the calculation of these values were done experimentally: in our experimentation, we vary the value of σ and n and compare the results that we obtain (in term of the recognition rate of our algorithm) and finally we opted for σ and n that present the highest recognition rate. Figure 5.19 illustrates how the probability distribution of $P(X < Y)$, $P(X \sim Y)$ and $P(X > Y)$ are affected when varying the parameters σ and n.

5.6.3.2 The Before and After relations

We discuss here how we extend and attach a probability measure to the relation *'before'* and *'after'*. Let $Y = [Y_b, Y_e]$, the left complement of the interval Y is denoted by Y^- i.e. the intervals that covers the entire time axis until Y_b:

$$Y^- = (-\infty, Y_b)$$

Similarily, the right complement of the interval Y is denoted by Y^+ i.e. the intervals that covers the entire time starting from Y_e:

$$Y^+ = (Y_e, +\infty)$$

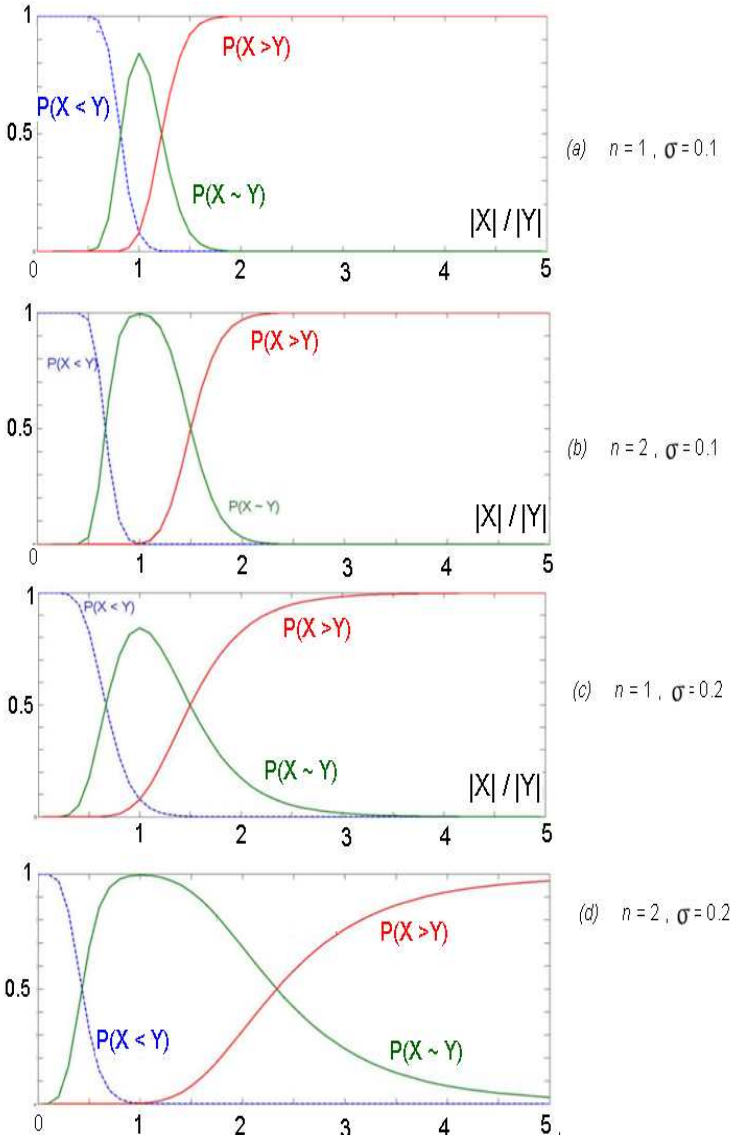


Figure 5.19: Graph of the probability distribution of $P(X < Y)$ (in blue), $P(X \sim Y)$ (in green) and $P(X > Y)$ (in red) in function of the σ and n parameters: The first figure (a), illustrates the variation of the probability distribution with parameters $\sigma = 0.1$, $n = 1$. The second figure (b), the variation of the probability distribution with parameters $\sigma = 0.1$, $n = 2$. The third figure (c), the variation of the probability distribution with parameters $\sigma = 0.2$, $n = 1$ and the last figure (d), the variation of the probability distribution with parameters $\sigma = 0.2$, $n = 2$.

What we want to calculate here is ***the probability that*** 'X ***is before*** Y', this probability is defined as the probability of a chosen random point t in X that lies before Y, i.e. in Y^{-}. More formally:

$$\begin{aligned} \forall |X| \leq |Y| \quad P(\text{`X} \quad \text{before } Y') &\stackrel{\text{def}}{=} P(t \in Y^{-} | t \in X) \\ &= \frac{|X \cap Y^{-}|}{|X|} \end{aligned} \tag{5.50}$$

In the same manner, ***the probability that 'X is after Y'*** is defined as:

$$\begin{aligned} \forall |X| \leq |Y| \quad P(\text{`X} \quad \text{after } Y') &\stackrel{\text{def}}{=} P(t \in Y^{+} | t \in X) \\ &= \frac{|X \cap Y^{+}|}{|X|} \end{aligned} \tag{5.51}$$

Note that when we have $|X| > |Y|$, the same definitions hold but with the argument inverted. Namely:

$$\begin{aligned} \forall |X| > |Y| \quad P(\text{`X} \quad \text{before } Y') &\stackrel{\text{def}}{=} P(t \in X^{+} | t \in Y) \\ &= \frac{|Y \cap X^{+}|}{|Y|} \end{aligned} \tag{5.52}$$

$$\begin{aligned} \forall |X| > |Y| \quad P(\text{`X} \quad \text{after } Y') &\stackrel{\text{def}}{=} P(t \in X^{-} | t \in Y) \\ &= \frac{|Y \cap X^{-}|}{|Y|} \end{aligned} \tag{5.53}$$

5.6.3.3 The relation Overlaps and Overlaps-i

Let's now discuss about the relations *'Overlaps'* and *'Overlaps-i'*. The relations *'Overlaps'* holds if $X_b < Y_b < X_e < Y_e$ as shown in figure 5.12. Similarly to the probability computation of the *'during'* and *'durig-i'*, we compute the probability of *'Overlaps'* and *'Overlaps-i'* as follow:

$$P(\text{`X} \quad \text{overlaps } Y') = \begin{cases} P(X \sim Y) P(X \asymp Y) & P(X \text{ after } Y) = 0 \\ 0 & P(X \text{ after } Y) > 0 \end{cases} \tag{5.54}$$

$$P(\text{`X} \quad \text{overlaps} - \text{i } Y') = \begin{cases} P(X \sim Y) P(X \asymp Y) & P(X \text{after} Y) > 0 \\ 0 & P(X \text{ after } Y) = 0 \end{cases} \tag{5.55}$$

We notice that, (i) the probability of overlaps (respectively overlaps-i) are correlated to the probability of before (respectively after) and (ii) the overlaps and overlaps-i can not be simultaneously non-zero. Thus we conclude the following relations:

$$\begin{aligned} &P(X \text{ after } Y) = 0 \\ &\lim_{(X \sim Y) \to 1} P(X \text{ overlaps } Y) = P(X \asymp Y) \\ &\lim_{(X \sim Y) \to 1} P(X \text{ overlaps} - \text{i } Y) = 0 \end{aligned} \tag{5.56}$$

and

$$\begin{aligned} &P(X \ after \ Y) > 0 \\ &\lim_{(X \sim Y) \to 1} P(X \ overlaps \ Y) = 0 \\ &\lim_{(X \sim Y) \to 1} P(X \ overlaps - i \ Y) = P(X \asymp Y) \end{aligned} \tag{5.57}$$

5.6.3.4 The relation Meet and Meet-i

In this section, we define how we compute the probability of the *'meet'* and *'meet-i'*. The computation of the probability of the temporal relations *'meet'* and *'meet-i'* (see equations 5.70 and 5.71) depends of **(i)** the size of the two temporal intervals X and Y and **(ii)** the the position of an interval with respect to the other interval in the temporal axis. We note that the probabilities relative to the size of intervals are independent of the probability relative to the intervals position. We propose first the following notions and definitions:

Definition 1. (Relative position). The relative position of intervals X and Y is defined as the function $p : R^4 \to [-1, 1]$,

$$p(X, Y) = \begin{cases} \frac{|X_c - Y_c|}{l(X,Y)} & \text{if } \ l(X, Y) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.58}$$

Where,

X_c is the center of interval X, defined as following:

$$X_c = \frac{X_b + X_e}{2} \tag{5.59}$$

Y_c is the center of interval Y, defined as following:

$$Y_c = \frac{Y_b + Y_e}{2} \tag{5.60}$$

$l(X, Y)$ is the total span of intervals defined as :

$$l(X, Y) = max\{X_e, Y_e\} - min\{X_b, Y_b\} \tag{5.61}$$

The nominator of the relative position function is a signed index of how far from each other the centers of the intervals are. On the other hand, the denominator acts as a normaliser so that the relative position is always within $[-1, 1]$.

Definition 2. (Relative size). The relative size of intervals X and Y is defined as the function $s : R^4 \to [-1, 1]$,

$$s(X, Y) = \begin{cases} \frac{|X_m - Y_m|}{d(X,Y)} & \text{if } \ d(X, Y) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.62}$$

where

the interval length is defined as following:

$$X_m = X_b - X_e \tag{5.63}$$

$d(X, Y)$ is the total size of intervals, defined as:

$$d(X, Y) = X_m + Y_m \tag{5.64}$$

The nominator of the relative size function is a signed index of how the size of intervals differ from each other. On the other hand, the denominator acts as a normaliser so that the relative size is always within [-1, 1].

Let's now examine the case of *'meet'* and *'meet-i'*. For *'meet'* case, we know that the interval Y begins exactly the moment that interval X ends, which implies that: $X_e = Y_b$ and $Y_e > X_b$. This means that:

$$p = \frac{\frac{X_b + X_e}{2} - \frac{Y_b + Y_e}{2}}{l(X, Y)} = \frac{X_b - Y_e}{2.l(X, Y)} = -\frac{Y_e - X_b}{2.l(X, Y)} = -\frac{1}{2} \tag{5.65}$$

Similarly, *'meet-i'* corresponds to $p = 1/2$.

As described in section 5.6.3.1, we define the probability of being smaller of larger to some value v using the cumulative probability of a random variable $\mathcal{V}$ as:

$$P(\mathcal{V}) < v) \stackrel{def}{=} \Phi(v|\hat{v}, \sigma) \tag{5.66}$$

$$P(\mathcal{V}) > v) \stackrel{def}{=} 1 - \Phi(v|\hat{v}, \sigma) \tag{5.67}$$

we let n be a parameter, $n \in \Re^+$, and define the probability of being equal to some value v, as:

$$P(\mathcal{V}) \sim v) \stackrel{def}{=} P(V < v + n) - P(V < v - n) \tag{5.68}$$

The values $v + n$ and $v - n$ may be seen as defining a neighbourhood around v . Moreover for two value v_1 and v_2, the probability of being between these two values is defined as:

$$P(v_1 < V < v_2) \stackrel{def}{=} P(V < v_2) - P(V < v_1) \tag{5.69}$$

Based on the above definitions, and given that for the *'meet'* relation, $p = -1/2$ and for *'meet-i'*, $p = 1/2$, the following definitions are straightforward:

$$P(X \ \ meet \ \ Y) \stackrel{def}{=} P(p \sim -\frac{1}{2}).P(s \in (-1, 1)) \tag{5.70}$$

$$P(X \ \ meet - i \ \ Y) \stackrel{def}{=} P(p \sim \frac{1}{2}).P(s \in (-1, 1)) \tag{5.71}$$

5.6.3.5 Examples

To illustrate how the proposed approach improves the robustness and expressiveness of the temporal relations between intervals, we present the application of these interval relations in some examples. Figure 5.20 illustrates cases of different temporal relations (e.g. overlaps, during). Probabilities are calculated using the proposed approach and obtained for several relation in each case. The most probable relation is highlighting in each case. In all these cases, the most probable relation coincides with the true one (i.e. coincides with the ground truth). However, these probabilistic extension provides a more expressiveness to the temporal relation, for example, in case (2), the 'overlap' relation has a greater value than the 'overlap-i' in case (5). This is because in case (2) the intervals overlap more than for the case (5). Furthermore, by applying the probabilistic extension, the relation becomes more robust with respect to the boundaries uncertainties. For instance, in case (1), when we apply Allen definitions, the relation between X and Y would be overlaps, however, this hide the fact that X is almost before Y and the fact that they overlaps is may be due to noisy boundaries measurement. By applying the Allen extension, the 'before' relation is more probable than 'overlaps'.

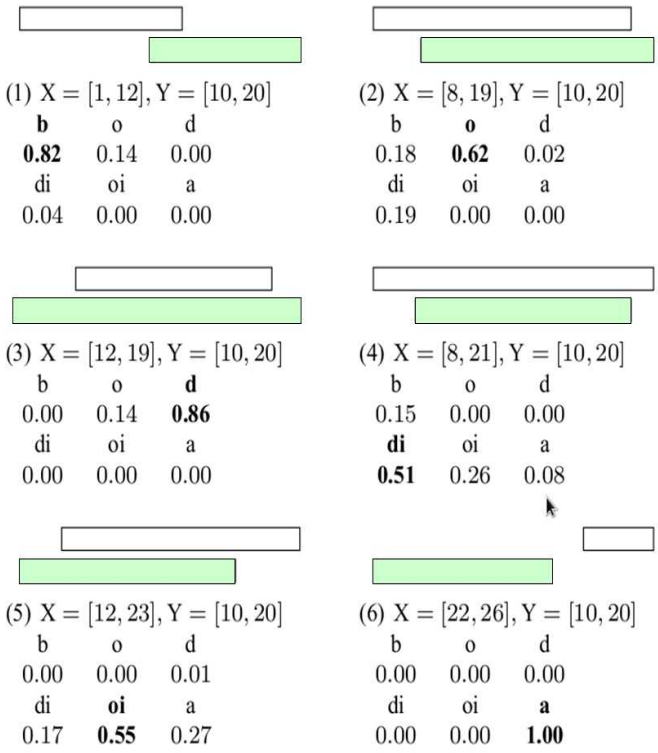


Figure 5.20: Expressiveness and robustness of probabilistic Allen temporal relations. The letter **b** is for the relation 'before', **a**, for 'after', **o**, for 'overlaps', **oi** for 'overlaps-i', **d**, for 'during' and **di**, for 'during-i', for $\sigma = 0.1$, and $n = 1.0$.

5.7 Low-level attribute Noise Processing

In video interpretation objects are tracked and associated with different positions at different times. However, the position of the object may change rapidly from frame to frame, because of visual noise which results in the object tracking processes assigning a different position to the object in different frames. These changes are not 'real changes' but rather artefacts of vision algorithms. Such problem affects considerably the activity recognition process.

In this section, we detail the proposed methods to take into account this kind of uncertainty inherent in low level observations by proposing a dynamic model for temporal attributes filtering.

5.7.1 Visual reliability Estimation of Attributes

The visual reliability for a dimension $d \in \{width(w), length(l), height(h)\}$ is intended to quantify the visual evidence for the dimension by analysing how much the object can be seen

from the camera point of view. The objective is to find a measure that gives a minimal value (e.g. 0) when the object is not visible, and a maximal value (e.g. 1) when the object is totally visible.

The visual reliability RD_a estimation of the object attribute a changes according to its type. The attributes considered belong to the set $A = \{X, Y, W, H, x_p, y_p, z_p, w, l, h, \beta\}$.
- (X, Y) is the 2D position of the object,
- (W, H) are the 2D blob width and height in image plane coordinates,
- (x_p, y_p, z_p) is the 3D position of the object,
- (w, l, h) correspond to the 3D width, lenght and height.

For 2D attributes W, H, X and Y a reliability measure inversily proportional to the distance to the camera d_{cam} is calculated, accounting to the fact that the segmentation errors increase when the objects are farther the camera (5.72).

$$RD_a = \frac{1}{d_{cam}} \tag{5.72}$$

The reliability measures of the 3D attributes w, l and h are obtained with the equation (5.73). It represents the maximal magnitude of projection of a 3D attribute onto the image plane in proportion with the magnitude of each 2D blob limiting segment.

The concept of visibility is not necessary for describing the reliability of the 3D position (x_p, y_p, z_p) and orientation β, because these attributes depend on w and l. The proposed reliability is calculated as the mean between the reliability of w and l.

$$RD_a = min(\frac{dY_a}{H} + \frac{dX_a}{W}, 1) \tag{5.73}$$

dX_a and dY_a represent the lenght in pixels of the projection of the dimension a on the X and Y reference axes of the image plane, respectively.

5.7.2 The Proposed Dynamic Model for temporal attributes filtering

As we work with real video sequences, observations can be corrupted by noise. This noise which results in the object detection/tracking softwares assigning a different attribute values (e.g. position, shape) to the object in different frames. These attribute values changes are not 'real changes' but rather artefacts of the software system. Often the size/position of the object will change rapidly from frame to frame. Thus our job is to make the best estimation of an attribute value given its previous observed value. The aim is to use a smothing technique to smooth and reduce this rapidly changes of attribute value.

We propose a dynamic linear model for reliabily computing and updating the attributes value on a temporal history of the previous values stored in a short-term history buffer. We

think that tacking into account an history of the previous values and not just the previous one help us to a better estimation. The proposed process of the temporal filtering of attributes works in two steps:
- **First step (1)** consists in computing the expected value a_{exp} of an attribute a at the current instant t_c given the estimated value of a and its velocity at the previous time t_p.
- **Second step (2)** is to compute the estimated value a_{est} of the attribute at the current instant t_c based on the previous value.
- **The final value** $\tilde{a}$ of the attribute is the mean between the expected and the estimated values of the attribute weighted by the expected and estimated reliability values (R_{aexp}, R_{aest}) (5.74).

$$\tilde{a}(t_c) = \frac{a_{exp}(t_c).R_{aexp}(t_c) + a_{est}(t_c).R_{aest}(t_c)}{R_{aexp}(t_c) + R_{aest}(t_c)} \tag{5.74}$$

- $a_{exp}(t_c)$ is the expected value of an attribute a at the current instant t_c. The expected value a_{exp} of an attribute a at the current instant t_c is computed based on the value of a at previous time t_p and its velocity at previous time t_p(Eq.5.75).

$$a_{exp}(t_c) = \tilde{a}(t_p) + V_a(t_c)(t_c - t_p); \tag{5.75}$$

V_a corresponds to the estimated velocity value of the attribute a described in the equation (5.76)

$$V_a(t_c) = \frac{V_{a_c}.R_v + e^{-\lambda(t_c - t_p)}.V_a(t_p)S_{V_a}(t_p)}{S_{V_a}(t_c)}; \tag{5.76}$$

$$S_{V_a}(t_c) = R_v + e^{-\lambda(t_c - t_p)}.S_{V_a}(t_p) \tag{5.77}$$

- V_{a_c} corresponds to the instantaneous velocity of the attribute a at time instants t_{c-1} and t_c,
- R_v is the instantaneous reliability of the velocity computed as the mean between the visual reliability of a at time instants t_{c-1} and t_c.
- $V_a(t_p)$ is the estimated velocity at the previous time t_p.
- S_{V_a} is the temporal reliability of velocity.
- The value $e^{-\lambda(t_c - t_p)}$ corresponds to the cooling function of the previously observed attribute values. It can be interpreted as a forgetting factor for reinforcing the newer information.

- R_{aexp} and R_{aest} are respectively the reliability measure of the expected and estimated value of the attribute.

- $R_{aexp}(t_c)$ is the expected reliability of an attribute a at the current instant t_c. It is determined as the mean of R_v and the global reliability of the attribute R_a at the previous time t_p (5.78). The global reliability R_a is the mean between the expected and estimated reliability.

$$R_{aexp}(t_c) = Mean(R_v, R_a)_{t_p} = \frac{R_v(t_p) + R_a(t_p)}{2} \tag{5.78}$$

- $a_{est}(t_c)$ is the estimated value of an attribute a at the current instant t_c. It is computed based on the equation (5.79):

$$a_{est}(t_c) = \frac{a_c.RD_{a_c} + e^{-\lambda(t_c - t_p)}.a_{est}(t_p).S_a(t_p)}{S_a(t_c)} \tag{5.79}$$

$$S_a(t_c) = RD_{a_c} + e^{-\lambda(t_c - t_p)}.S_a(t_p) \tag{5.80}$$

Where a_c is the value of the attribute given by vision algorithm and RD_{a_c} is the visual reliability of this attribute a extracted from visual evidence at time t_c.

- $R_{aest}(t_c)$ is the estimated reliability of an attribute a at the current instant t_c. R_{aest} is calculated as the mean between the coherency reliability RC_a(5.84) and the visual reliability RD_a (5.83).

$$R_{aest}(t_c) = Mean(RC_a, RD_a)_{t_c} = \frac{RC_a(t_c) + RD_a(t_c)}{2} \tag{5.81}$$

$$RC_a(t_c) = 1.0 - min(1.0, \frac{\sigma_a(t_c)}{a_{max} - a_{min}}) \tag{5.82}$$

σ_a corresponds to the standard deviation of the attribute a. The value a_{max} and a_{min} correspond to a pre-defined minimal and maximal value for a.

$$RD_a(t_c) = \frac{S_a(t_c)}{sumCooling(t_c)} \tag{5.83}$$

With

$$sumCooling(t_c) = sumCooling(t_p) + e^{-\lambda(t_c - t_p)} \tag{5.84}$$

The visual reliability RD_a represents the mean between the reliability measures of RD_{a_i} $i \in History_a$.

5.8 Conclusion

In this chapter, we have shown how we propose to handle the recognition uncertainty in the high level of event recognition process: first, in the step of primitive states detection and second in the step of composite event recognition. we show how the Bayesian probability allows us to reason under uncertainty and to compute the probability that an event occurs given the observation.

Addressing the noise arising from low-level video analysis, we propose a probabilistic spatial constraint verification approach based on Gaussian probability model applied on distance distributions. We propose also temporal constraint verification approach applied on Allen temporal constraints.

A dynamic linear model for attributes values estimation is proposed to deal with the noise which result from the object detection/tracking algorithms assigning wrong attribute values to the object in different frames. The goal is to make the best estimation of an attribute value given its previous observed value using smoothing technique.

In the next chapter, we evaluate the proposed approach using a set of real world videos from health care applications.

6

EVALUATION AND RESULTS OF THE PROPOSED APPROACH

6.1 Introduction

In order to evaluate the proposed activity recognition approach, several experiments have been performed. The main objectives of these experiments are to validate the different phases of the activity recognition approach, to highlight interesting characteristics, and to evaluate the potential of the framework for real world applications.

In this chapter, we first introduce in sections 6.2 the automatic video monitoring goals, and in section 6.3, the details concerning the implementation. We describe the evaluation process in section 6.4. In section 6.5, we present the experimental video datasets used for evaluation. In section 6.6, we detail the step of activity annotation. In section 6.7, we present the evaluation results for video activity recognition and the medical results for health care monitoring in section 6.8, and finally we conclude the chapter in section 6.9.

6.2 Automatic Video Monitoring Goals

In this work, we are interested of an accurate recognition of both simple and complex video activities. For simple activities, we interested to:

- Determine whether one or several individuals are present in the environment.
- Determine the location of each person (e.g. person close TV).

- Recognize body postures such as standing, bending, sitting.

For complex activities, we interested to:

- Recognize how a person interacts with the environment (e.g. use kettle, sit on a chair).
- Recognize activities of daily living (ADLs) such as eating, dressing, manage medication for monitoring people with dementia and in particular Alzheimer people.

6.3 Implementation

The proposed algorithms in this thesis are implemented in SUP[1] (i.e. Scene Understanding Platform) developed by STARS team. SUP platform is written in C++ language. It provides algorithms for video interpretation systems (e.g. image acquisition, image segmentation, object tracking). All experiments presented in this chapter have been performed in a machine of Intel(R) Xeon(R) CPU E5430 @ 2.66GHz (4 cores) and of 4GB RAM. In the following, we detail the video analysis algorithms used during our experimentation.

6.3.1 Video Analysis Algorithms

Video analysis aims at segmenting, detecting and tracking people moving in the scene. To achieve this task, we have used a set of vision algorithms coming from the video interpretation platform SUP (Figure. 6.1).

The first task is the segmentation task which is applied to each coming image to detect motion in the scene, obtaining a set of moving regions represented as the bounding boxes enclosing them (called blobs). The algorithm segments moving pixels in the video into a binary image by subtracting the current image with the reference image. The reference image is updated along the time to take into account changes in the scene (e.g. light, object displacement, shadows).

A Calibration task is used to compute the transformation of a 2D image referential point to a 3D scene referential point (fig.6.2). The keeneo[2] calibration tool allows to obtain the calibration matrix which is used to compute the 3D position of mobile objects.

The 3D position of the mobile object is computed from the detected blob and the calibration matrix associated with the video camera by considering that the bottom of the 3D mobile object is on floor level. When the blob representing the person is not completly visible (i.e. occluded by a specified contextual object), the person is supposed to be just behind the object.

[1] https://team.inria.fr/stars/software/ platform

[2] http://www.genetec.com/Publications/technicalnotes/Pages/keeneo-analytics-en.aspx

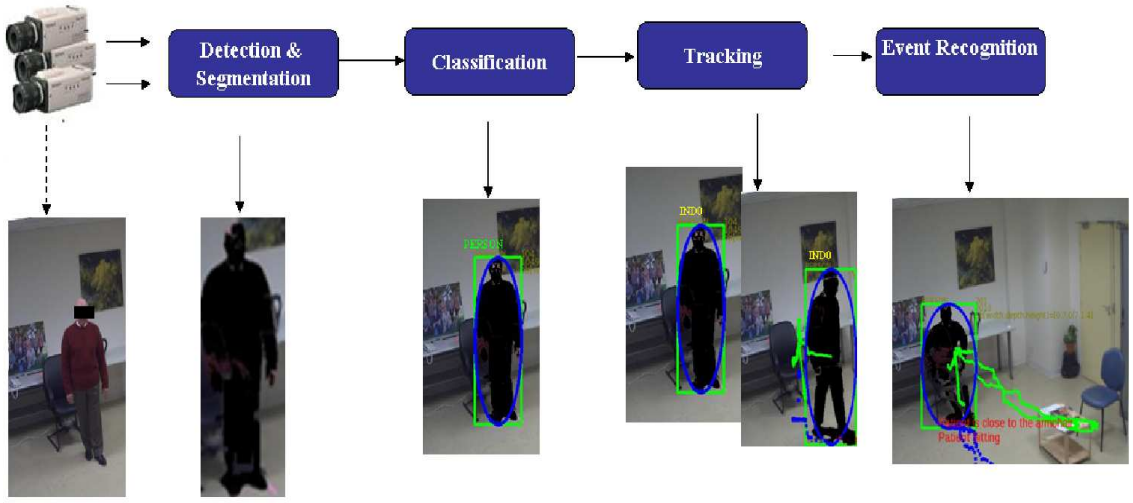


Figure 6.1: Visions Algorithms of SUP platform.

A blob merging task consists in assembling small 2D blobs to improve the classification task. The classification task uses the obtained 2D blobs, the calibration matrix of the camera and predefined 3D models to classify the mobile blobs. It adds a class label to each moving blob (e.g. person, vehicle). A set of 3D features such as 3D position, width and height are computed for each blob. A unique identifier is associated to each new classified blob by using the frame to frame tracker. A long term tracker allow to maintain this identifier throughout the whole video.

6.3.1.1 Background Substraction Algorithm

The algorithm first constructs a background representation for each pixel as described in [NGHIEM, 2010]. Then to detect foreground pixels in the current frame, the algorithm compares the current frame with the background representation of the algorithm. The pixels in the current frame which are different from the background representation are classified as foreground pixels. The background representation is updated regularly after each incoming frame. When the scene conditions change, the background subtraction algorithm has to adapt itself to the new conditions. However, working only at the pixel level, it is difficult for the background subtraction algorithm to fulfill this work. In this algorithm, a controller for the background subtraction algorithm is used to help the background subtraction algorithm to adapt to the current scene conditions. To do this, the controller uses the feedback from the classification task and the information about the background subtraction algorithm and the scene. With the controller for the background subtraction algorithm, the object detection framework works as follows:
- The framework takes as input a video sequence from a single and fixed camera. From this

Figure 6.2: 3D geometrical information (empty scene, calibration matrix) computed from the Keeneo calibration tool.

video sequence, for each frame, the background subtraction algorithm produces a list of potential foreground pixels.

- The algorithm to remove shadow and highlight receives this list and it removes from the list the pixels corresponding to shadow or highlight. The results are sent to the blob construction task and the controller.

- The blob construction task constructs the blobs from the foreground pixels, then sends the blobs to the blob classification task.

- The blob classification task classifies these blobs and sends the list of blobs together with their types to the higher tasks and to the controller.

6.3.1.2 Tracking Algorithm

The tracking algorithm tracks mobile objects based on their trajectory properties as described in [Chau, 2012]. The used tracker includes two stages : tracking and global tracking. The tracking stage follows the steps of a Kalman filter including estimation, measurement and correction. First for each tracked object, its state including position and 2D bounding box is estimated by a Kalman filter. Second, in the measurement step, this tracked object searches for the best matching object based on four descriptors : 2D position, 2D area, 2D shape ratio and color histogram. Third, the best matching object and its estimated state are combined to update the position and 2D bounding box sizes of the tracked object. However, the mobile object trajectories are usually fragmented because of occlusions and misdetections. Therefore, the global tracking stage aims at fusing the fragmented trajectories belonging to the same mobile object and removing the noisy trajectories. The advantages of this algorithm [Chau, 2012] over the existing state of the art ones are : (1) no prior knowledge information is required (e.g. no calibration and no contextual models are needed), (2) the tracker can be effective in different scene conditions : single/several mobile objects, weak/strong illumination, indoor/outdoor scenes, (3) a global tracking stage is defined to improve the object tracking performance.

6.3.1.3 Posture Recognition

We have used the posture recognition algorithm proposed in [Boulay et al., 2006] in order to recognize in real-time a set of human postures from any position of the camera and for any orientation of the person.

The goal of the human posture recognition approach is to provide accurate information about the people observed in the scene. Human posture recognition is a difficult and challenging problem due to the huge quantity of possible cases. The number of postures depends on the degree of freedom of the human body. Moreover the morphology of the person influences the appearance of the same posture. Finally, clothes can give different types of appearance for the same posture. The posture recognition task combines the 2D techniques and 3D posture models to generate silhouettes. The generated silhouettes are then compared with the silhouettes of the detected objects moving in the scene based on 2D techniques to determine the posture. The algorithm takes as input the contextual base to generate a virtual camera (using camera calibration information) and uses the information given by the detection task to generate silhouettes (Figure. 6.3).

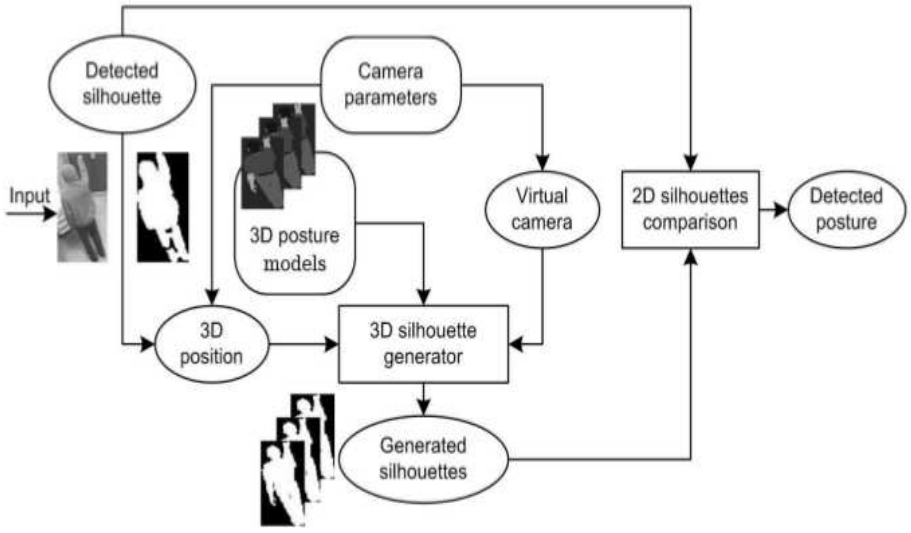


Figure 6.3: Posture Recognition Approach [Boulay et al., 2006].

6.4 Evaluation Process

In this section we describe the different evaluations of the proposed activity recognition approach. For the algorithm evaluation, we use the evaluation tool **ViSEvAl**[3] described in section 6.4.2. Figure 6.4 illustrates the evaluation process. The proposed activity recognition algorithm takes as input (1) the event model, the (2) 3D geometric information and (3) the tracked objects. The output is the set of recognized events. To evaluate the recognition performance of the proposed algorithm, we compare the recognized events with the events annotated by human expert (i.e. ground truth). The metrics used for the evaluation are described in section 6.4.1.

6.4.1 Evaluation Metrics

To evaluate the performance of the proposed approach, we have used different metrics according to the nature of the experiment. For the recognition of events, the adopted metrics are:

- True Positive (TP): An event Ei is correctly detected according to the ground truth.
- False Positive (FP): An event Ei is wrongly detected according to the ground truth.
- False Negative (FN): An event Ei present in the ground truth is not detected.

[3]https://team.inria.fr/stars/2012/02/02/viseval-software/

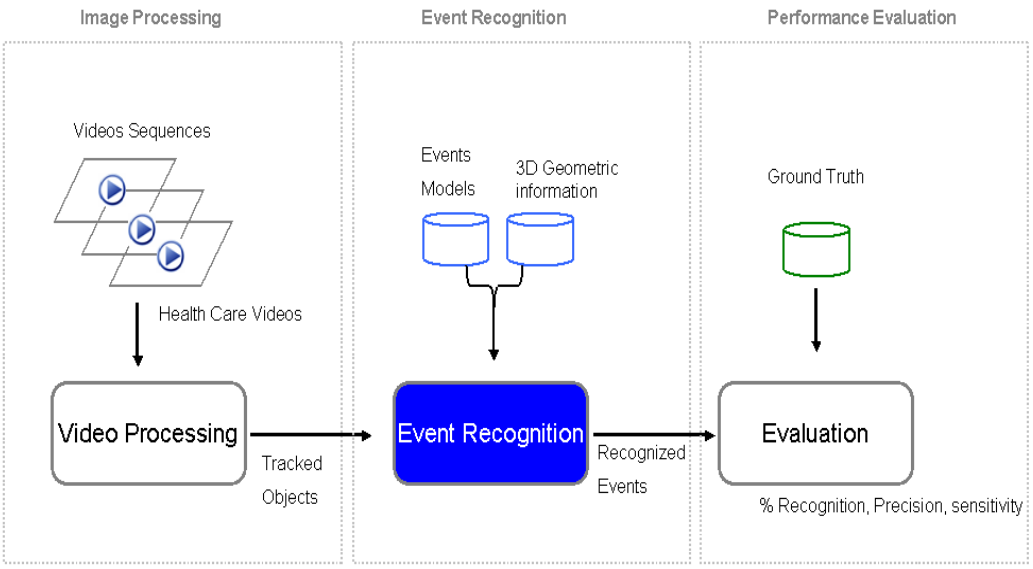


Figure 6.4: The Evaluation Process: The proposed activity recognition algorithm takes as input (1) the event model, the (2) 3D geometric information and (3) the tracked objects from video processing algorithms. The output is the set of recognized events. To evaluate the recognition performance of the proposed algorithm, we compare the recognized events with the events annotated by human expert (i.e. ground truth). The comparison is done using evaluation metrics described in section 6.4.1.

- Precision (P): The precision metric can be seen as a measure of exactness or fidelity. The precision corresponds to the number of events correctly detected divided by the total number of detected events. This metric is formally defined as:

$$P = \frac{TP}{TP + FP} \tag{6.1}$$

- Sensitivity (S): A sensitivity corresponds to the number of events correctly detected divided by the total number of occurred events. A sensitivity of 100% means the recognition of the all occurring events. This metric is formally defined as:

$$S = \frac{TP}{TP + FN} \tag{6.2}$$

- Recall (R): is the same metric than the sensitivity. The recall metric corresponds to the number of events correctly detected divided by the total number of occurred events. This metric is formally defined as:

$$R = \frac{TP}{TP + FN} \tag{6.3}$$

6.4.2 Evaluation platform: VisEval

The evaluation of video processing algorithm results is an important step in video analysis research. For the evaluation of the proposed approach, we use the evaluation tool ViSEvAl. **ViSEvAl**[4] is a software developped in STARS team and dedicated to the evaluation and visualization of video processing algorithm outputs. It respects three important properties:(i) visualize the algorithm results, (ii) visualize the metrics and evaluation results and (iii) possibility to add new evaluation metrics. The GUI is composed of several parts (fig. 6.5). This tool has 5 windows, the first window depicted by (3), displays the current image and information about the detected and ground-truth objects/events (e.g. bounding-boxes, identifier, type,...), the second window depicted by (5), displays a 3D view of the scene (3D avatars for the detected and ground-truth objects, events, context, ...), the third window depicted by (7), displays the temporal information about the detected and ground truth objects, and about the recognized and ground-truth events. The fourth window depicted by (8), the description part gives detailed information about the objects and the events and the last window, depicted by (9), displays the evaluation results of the frame metrics.

[4]https://team.inria.fr/stars/2012/02/02/viseval-software/

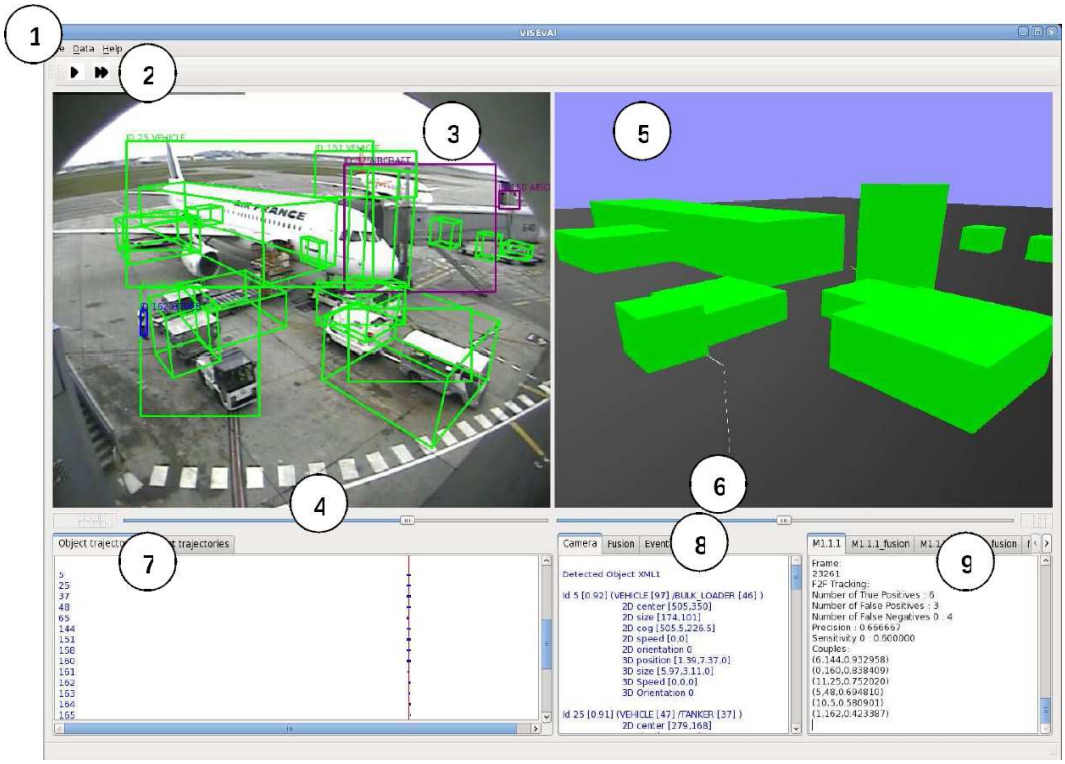


Figure 6.5: ViSEvAl Evaluation Tool: (1) the menu, (2) tool bar, (3) real video, (4) real video slider, (5) 3D visualization, (6) scale slider, (7)display of the detected objects and events, (8) object and event description and (9) evaluation metrics.

6.5 Experimental Dataset Presentation

The proposed approach is mainly applied and evaluated on health care applications. For the sake of generality, the proposed approach is also tested on other public real-world datasets.

6.5.1 GERHOME Dataset

The GERHOME[5] project consists in monitoring older people observed in an experimental laboratory (fig.6.6). Fourteen volunteers (i.e. 6 women and 8 men aged from 60 years to 85 years) were recruited for the experiments. The volunteers have been observed, each one during 4 hours. Atotal of 56 video sequences have been acquired by 4 video cameras (about ten frames per second)(fig.6.7), each video sequence contains about 144 000 frames. The collected data includes the 56 video streams, and also sensors data provided by the 24 environmental sensors

[5]http ://gerhome.cstb.fr/en/home

(a) The kitchen (b) The livingroom (c) The bathroom

Figure 6.6: Internal views of the Gerhome laboratory.

(a) Video camera 1 in the kitchen (b) Video camera 2 in the living-room

(c) Video camera 3 in the living-room (d) Video camera 4 in the bedroom

Figure 6.7: Views from the installed video cameras in the Gerhome laboratory.

(fig.6.8) . The access of the video dataset is public and available on a web site[6].

[6]www-sop.inria.fr/members/Francois.Bremond/topicsText/gerhome Project.html

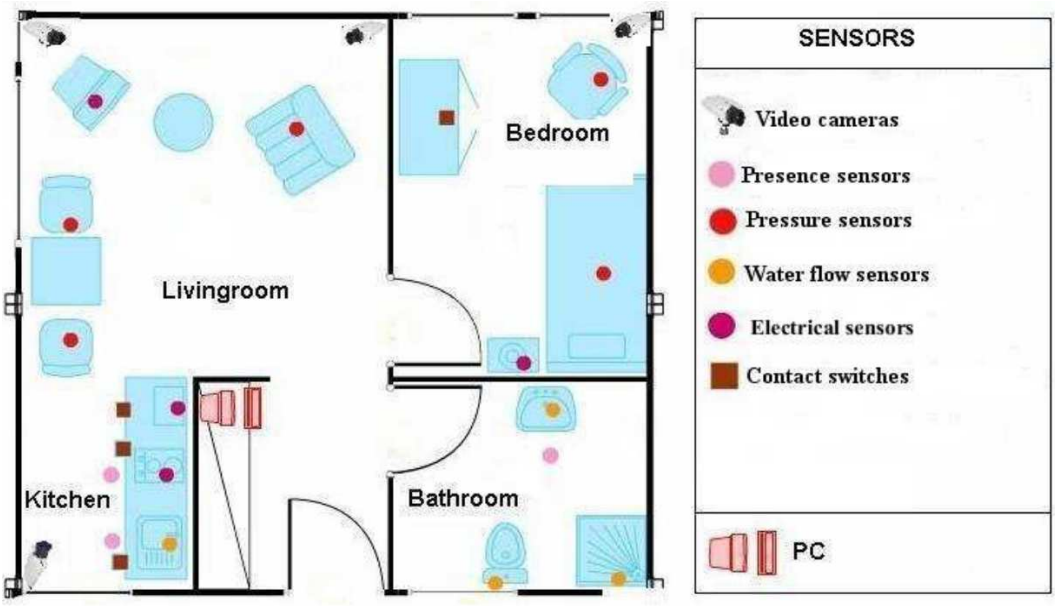


Figure 6.8: Architecture of Gerhome laboratory, sensors positions.

6.5.2 SWEETHOME Dataset

SWEET-HOME[7] is a project on long-term monitoring of elderly people at Hospital with Nice Hospital research memory center. The experimentation was performed in a room of Nice Research Memory Center (fig. 6.9). This room was equipped with everyday household appliances for use in activities of daily living ADLs (fig. 6.10), e.g. an armchair, a table, a coffee corner, a TV. Experimental data was recorded using a 2D video camera (AXIS, Model P1346, 8 fps (frames per second), and an ambient audio microphone (Tonsion, Model TM6, Software Audacity, WAV file format, 16bit PCM/16kHz). A motion sensor (i.e., MotionPod) was fixed on the chest of the participant to quantify their movements. Two protocols of experimentation were conducted and for each protocol, a clinical activity scenario was elaborated. A total of 64 participants (22 NC (i.e. normal control), 30 MCI (i.e. mild cognitive impairment), 12 AD (i.e. Alzheimer)) aged more than 65 years were recruited by Nice Research Memory Center.

6.5.3 ETISEO Dataset

ETISEO[8] project focuses on the treatment and interpretation of videos involving pedestrians and/or vehicles, indoors or outdoors, obtained from fixed cameras. This project seeks to work out a new structure contributing to an increase in the evaluation of video scene understanding

[7]http://www-sop.inria.fr/pulsar/projects/Sweet-Home/SweetHomeProject.html
[8]http ://www-sop.inria.fr/orion/ETISEO/.

Figure 6.9: Internal views of the SWEETHOME experimentation room.

with the active participation of industrialists and many research laboratories, such as French, European and International partners. The ETISEO videos are provided by the ETISEO project. We use the Etiseo building entrance dataset which is a publicly available database of real world videos (fig. 6.11).

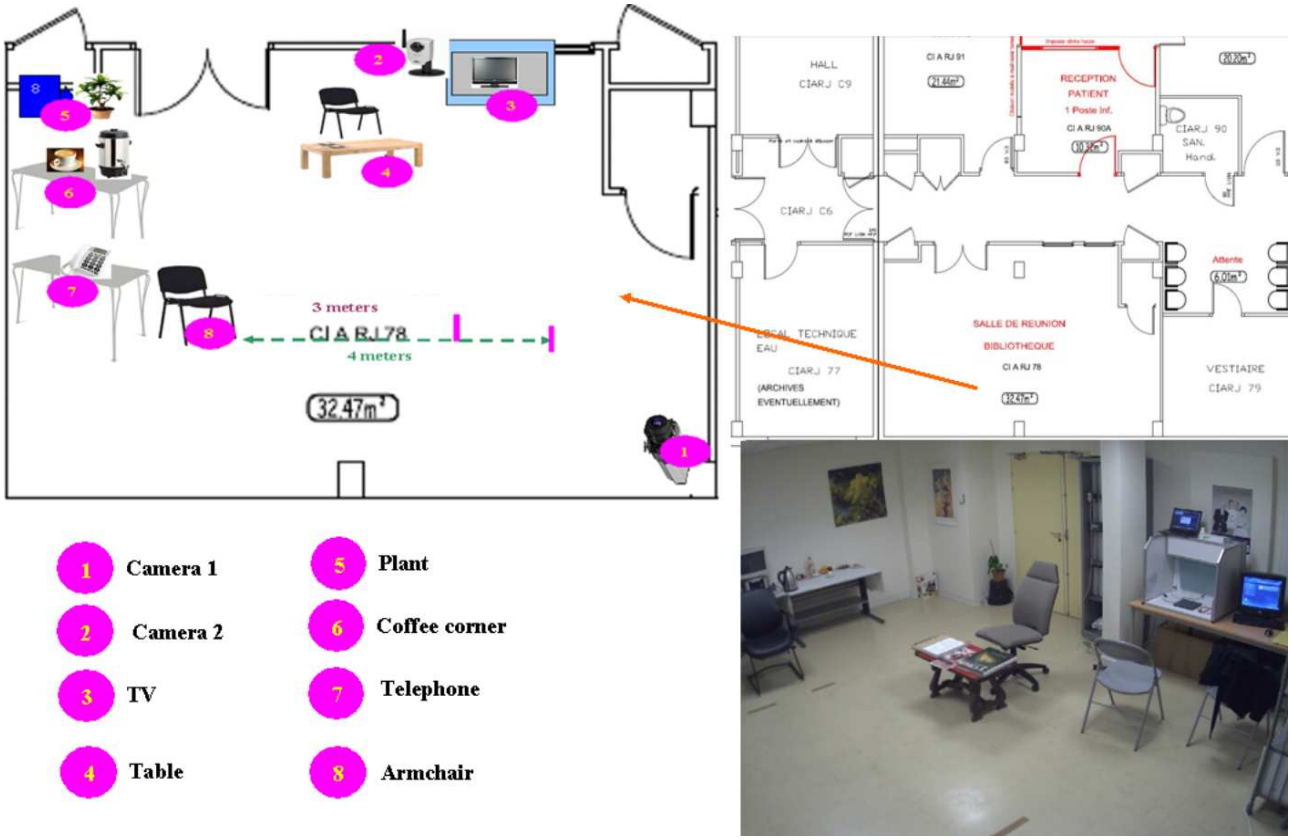


Figure 6.10: Architecture of SWEETHOME experimentation room, video sensors positions.

Figure 6.11: Examples of Etiseo dataset.

6.6 Annotation

For SWEETHOME videos, we work in close collaboration with clinicians to annotate the videos. The annotation has been done using VIPER-GT tool[9] which is developped by Maryland university and is a standard for video surveillance community. It enables the visualization and the markup of visual data ground truth for evaluating how closely sets of result data approximate that truth. To avoid inter- and intra-rater variability due to subjective interpretation of performed activities, we define with the help of clinicians the beginning and the end of each activity using specific criteria. Figure 6.12 illustrates an example of the observable criteria used to annotate video sequences. For example, to annotate the activity 'read the newspaper', we define exactly the beginning of this activity (i.e. taking for the first time the newspaper) and we define the end of this activity (i.e. putting on the table the last newspaper taken). This strict definition of the beginning and ending of activity allows us to limit variability of annotation between the human expert who annotate. We annotate the video sequences of 48 patients, a total of 144 ground truth files were generated. Figure 6.13 shows an example of a ground truth file. For Gerhome videos, we have used the ground truth of 5 observed older persons among the experimental data and the ground truth of 20 video sequences of one human actor [Zouba, 2010], the total duration is 20 hours. For Etiseo videos, annotation file for each video sequence is provided by the project.

[9] http://viper-toolkit.sourceforge.net/products/gt/

Activity	Activity period		Achievement
	Start	End	
Read the newspaper	Taking for the first time the newspaper	Putting on the table the last newspaper taken	Opening a newspaper taken to read the content
Water the plant	Taking the watering can	Putting the watering can on the table	Making the gesture of watering the plant
Answer the phone	Taking the handset in the hand	Putting the handset on the phone base	Speaking on the phone
Call the taxi	Taking the handset in the hand	Putting the handset on the phone base	Dialing the correct phone number and speaking on the phone
Prepare the medication for today	Taking the pillbox located inside a basket with the medication prescription inside	Putting the pillbox inside the basket	Correct use of medication prescription (both dose and timetable)
Make the check for the Electricity Company	Taking the pen to write on the check	Putting the pen on the desk	Correct amount, date, signature and recipient name on the check
Leave the room when you are finished with all activities	Taking the handle of the exit door	Closing the door	Closing the door as the last activity undertaken in the scenario
Watch the TV	Taking the remote call	Returning for the last time the remote control	TV must have been switched off
Prepare a hot tea	Turning on the tea kettle	Returning the tea kettle after having poured the water in the cup of tea	Brewed tea
Write the shopping list for the lunch	Taking the pen to write on the shopping list	Replacing the pen	Write at least one item to eat or drink

Figure 6.12: Example of the observable criterias used to annotate SWEETHOME video sequences. For example, to annotate the activity 'read the newspaper', we define exactly the beginning of this activity (i.e. taking for the first time the newspaper) and we define the end of this activity (i.e. putting on the table the last newspaper taken). This strict definition of the beginning and ending of activity allows us to limit variability of annotation between the human expert who annotate.

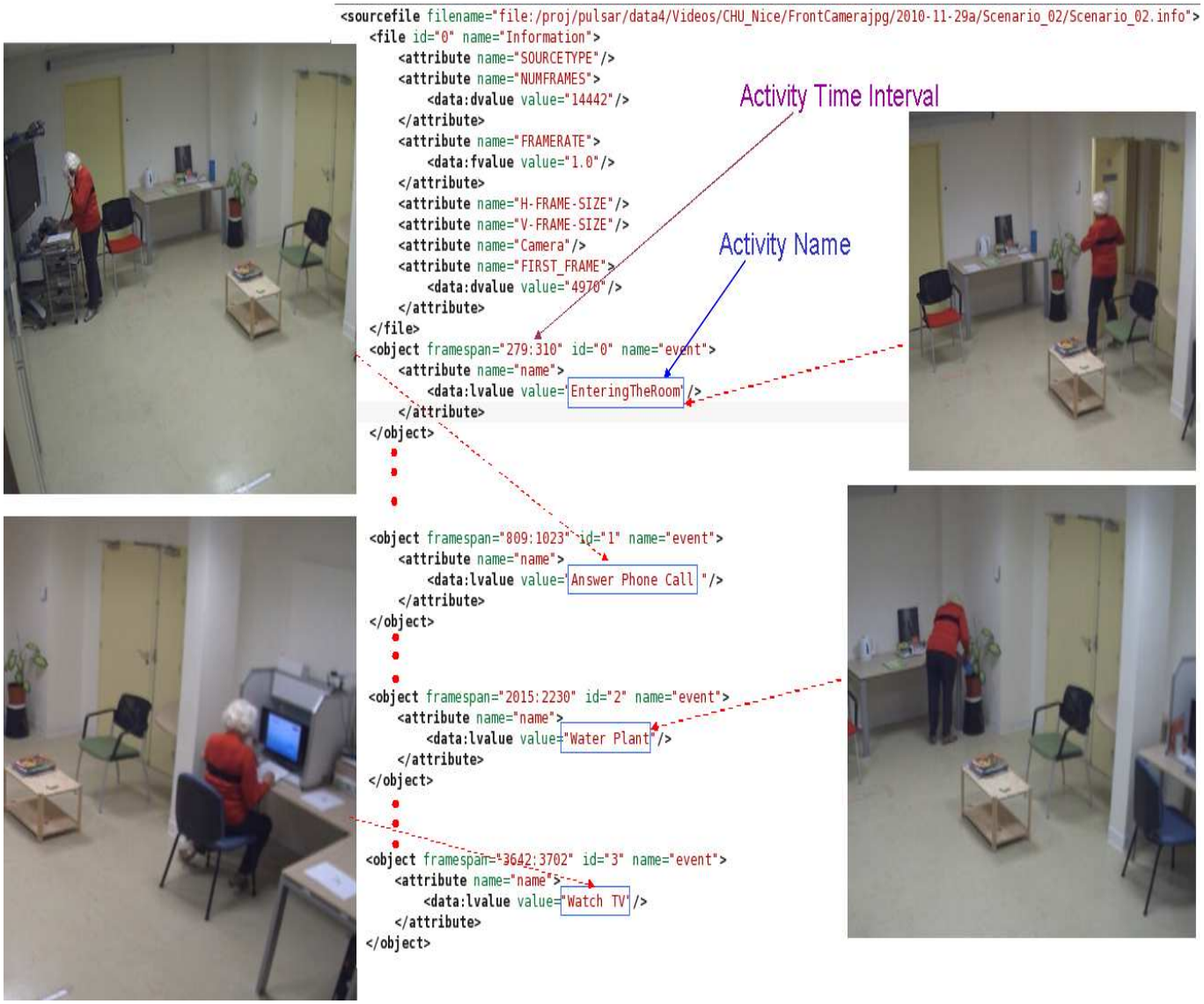


Figure 6.13: Example of a ground truth file: the annotated activities with the associated key frames.

6.7 Performed experiments

To evaluate the proposed activity monitoring framework we have tested a set of human activities on the SweetHome, Gerhome and Etiseo datasets.

6.7.1 Learning Step

In an off-line learning step, we have used 24 video sequences representative to the whole SweetHome video dataset. The length of each video sequence is about 12 ± 5 min. The training dataset is composed of a total of 305 min (146400 frames).

6.7.1.1 Bayesian Probabilities Learning

We estimate from training data the Bayesian probabilities (see tab.6.1) using the maximum likelihood estimation method based on the formulas detailed in chapter 5 ((5.12), (5.13), (5.17) and (5.18), chapter 5).

Probability	**Semantic**
$P(e \in \Omega)$	- Prior probability that a certain scenario model Ω is detected.
$P(\zeta(\Omega, O)\|e \in \Omega)$	- Probability that the constraints of the event model are verified given that the event e is true.
$P(V_\Omega = po_e^O\|e \in \Omega)$	- Probability that the physical objects variables in the event model Ω have been detected given that e is an event instance of the event model Ω.
$P(SE(\Omega, O)\|e \in \Omega)$	- Probability that the sub-event variables in the event model Ω have been detected given that e is an event instance of the event model Ω.
$P(\zeta(\Omega, O), V_\Omega = po_e^O)$	- Bayesian probability denominator for elementary events.
$P(\zeta(\Omega, O), SE(\Omega, O))$	Bayesian probability denominator for composite events.

Table 6.1: Bayesian probabilities.

Table 6.2 illustrates some learned Bayesian probabilities. For instance, to compute the conditional probability of the primitive state 'person inside coffee corner' we have to learn the Bayesian probabilities:

- The prior probability P(person inside coffee corner), is the prior probability that the event model 'person inside coffee corner' could be detected. The universe of scenario models that describe a person position in a certain zone is: (*PersonInsideZoneTV*, *personInsideCoffeeCorner*, *PersonInsideZoneReading*, *PersonInsideEntrance*, *PersonInsideZoneLibrary*.

$$P(\text{person inside coffee corner}') = \frac{1}{\text{Nbr.Scenario}\Omega\text{Universe}} = \frac{1}{5} \tag{6.4}$$

Annotated Events	$P(e \in \Omega)$	$P(\zeta(\Omega, O)\|e \in \Omega)$	$P(V_\Omega = po_e^O\|e \in \Omega)$	$P(\zeta(\Omega, O), V_\Omega = po_e^O)$
Person inside coffee corner	0.2	0.858	0.997	0.493
Person inside reading zones	0.2	0.92	0.96	0.342
Person close TV	0.25	0.702	1.00	0.403
Person standing	0.25	0.76	0.75	0.5
Annotated Events	$P(e \in \Omega)$	$P(\zeta(\Omega, O)\|e \in \Omega)$	$P(SE(\Omega, O)\|e \in \Omega)$	$P(\zeta(\Omega, O), SE(\Omega, O))$
change zone	0.2	0.81	0.75	0.5
interacts with TV	0.33	0.78	0.27	0.5

Table 6.2: Examples of Learned Bayesian probabilities.

- P(person **in** coffee corner zone | person inside coffee corner) is the probability that the spatial constraints (person **in** coffee corner zone) in the event model is verified given that the event 'person inside coffee corner' is true (has been annotated). It is computed based on the following equation:

$$P(person\ in\ coffee\ corner\ zone|person\ inside\ coffee\ corner) = \frac{\sharp((person\ in\ coffee\ corner\ zone) \wedge person\ inside\ coffee\ corner)}{\sharp(person\ inside\ coffee\ corner \wedge po = Person)} = \frac{4837}{5633} = 0.858 \tag{6.5}$$

  $\sharp((person\ in\ coffee\ corner\ zone) \wedge person\ inside\ coffee\ corner)$ is the number of frames where the constraint (person **in** coffee corner zone) is verified. We consider only the frames where the primitive state 'person inside coffee corner' is annotated. $\sharp(person\ inside\ coffee\ corner \wedge po = Person)$ is the number of frames where the primitive state is annotated and the physical object $(po = Person)$ is correctly tracked. We note that to complete the computation of the probability of this constraint, we need also the on-line probability of the constraint which is calculate as described in section 5.6.1.1, chapter 5.

- The probability of the physical object P(po = Person | person inside coffee corner) is the probability that the physical object 'Person' have been detected given an instance of the event model 'person inside coffee corner'. It is computed based on the following equation:

$$P(po = Person|\ person\ inside\ coffee\ corner) = \frac{\sharp(po = Person \wedge person\ inside\ coffee\ corner)}{\sharp(person\ inside\ coffee\ corner)} = \frac{4876}{4888} = 0.997 \tag{6.6}$$

$\sharp(po = Person \wedge person\ inside\ coffee\ corner)$ is the number of frames where the physical object 'Person' have been correctly tracked. We consider the frames where the primitive state 'person inside coffee corner' is annotated. $\sharp(person\ inside\ coffee\ corner)$ is the number of frames where the primitive state 'person inside coffee corner' is annotated.

- The Bayesian denominator P(person in coffee corner zone, po = Person) is computed based on the following equation (6.7):

$$\begin{aligned}
&P(person\ in\ coffee\ corner\ zone\ , po = Person) = \\
&P(person\ in\ coffee\ corner\ zone\ , po = Person | person\ inside\ coffee\ corner) \times \\
&P(person\ inside\ coffee\ corner) + \\
&P(person\ in\ coffee\ corner\ zone\ , po = Person | \neg person\ inside\ coffee\ corner) \times \\
&P(\neg person\ inside\ coffee\ corner) \\
&= \frac{4837}{9477} \times \frac{1}{5} + \frac{4640}{9477} \times \frac{4}{5} = 0.493
\end{aligned} \tag{6.7}$$

$P(person\ in\ coffee\ corner\ zone\ , po = Person | person\ inside\ coffee\ corner)$ corresponds to the number of frames where the constraint is verified and the physical object is detected, we consider the frames where the primitive state $person\ inside\ coffee\ corner$ is annotated. $P(person\ inside\ coffee\ corner)$ is the prior probability of the primitive state $person\ inside\ coffee\ corner$ which is equal to $\frac{1}{5}$.

$P(person\ in\ coffee\ corner\ zone\ , po = Person | \neg person\ inside\ coffee\ corner)$ corresponds to the number of frames where the constraint is verified and the physical object is detected, we consider the frames in the learning set where other primitive states of type location in certain zone than $person\ inside\ coffee\ corner$ are annotated.

6.7.1.2 Gaussian parameter Learning

We learn also the Gaussian parameters (μ, σ) used in the process of calculating the spatial constraint probability. To do that, we calculate and store the distances d_i of a tracked person to each contextual zone $(d_1, ..., d_n)_{z_1}, (d_1, ..., d_m)_{z_2}, ..., (d_1, ..., d_k)_{z_i}$, $z_1, ...z_i$ are the zones defined in the scene context. (see section 5.6.1.1, chapter 5). Having these samples, we calculate the Gaussian parameters based on the formula (5.31, chapter 5). The finale parameter value is the mean of these computed one. In our experiment, we use the values ($\mu = 0$ and $\sigma = 0.2$) because they allow us in one hand, to have a high percentage of true positive detection and in the other hand to have low false positive detections. Figure 6.14 shows the percentage of true

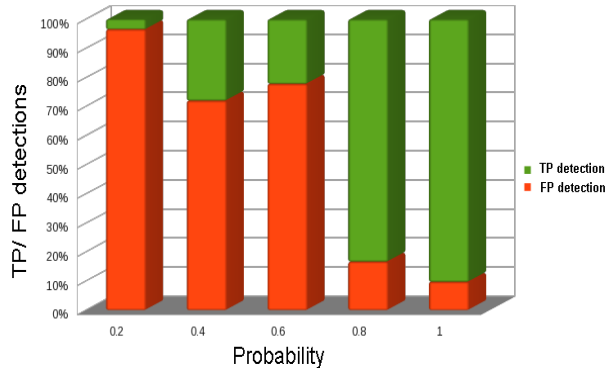


Figure 6.14: Percentage of true positive TP (green) and false positive FP (red) detections with respect to the probability values when using the Gaussian parameters ($\mu = 0$ and $\sigma = 0.2$).

positive and false positive detections of the spatial constraint 'inside-zone' of the reading zone with respect to the probability values when using the Gaussian values $\mu = 0$ and $\sigma = 0.2$. The percentage of TP detections has progressively increased as the value of probability becomes higher.

Figure 6.15 shows the value of probability at each frame of the spatial constraint 'inside-zone' compared with the ground truth (GT). This graphic shows that when the probability of GT is equal to zero (i.e. the spatial constraint is annotated as not detected), the value of the probability computed by the proposed system is either zero or a lower value of 0.4.

6.7.1.3 Detection Threshold Learning

We learn the threshold value for event detection. For that, We have tested the recognition performance of the proposed system by varying the decision threshold value.

Figyre 6.16 shows the recognition performance (precision and recall) of the primitive states when varying the decision threshold value. The primitive states are sometimes wrongly recognized due to video noise and vision errors. However, by fixing for all experiments the threshold of detection of primitive states to 0.75 we manage to successfully decrease the false detection of primitive states by avoiding miss detections of primitive states.

6.7.2 Experimental results

We have evaluated the event recognition accuracy of our algorithm on SweetHome real world health care application (fig.6.17) and have compared our results with the approaches proposed in [Vu et al., 2003a] and [Zouba, 2010].

Video recordings of *37* patients are used to assess the proposed framework performance.

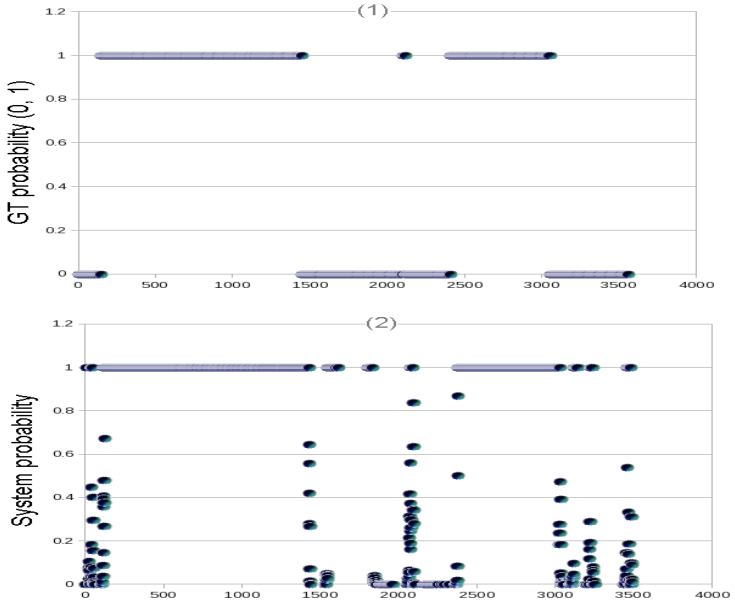


Figure 6.15: Probability value variation at each frame of the spatial constraint inside zone reading (figure(2)), compared with the ground truth GT (figure(1)) ($\mu = 0$ and $\sigma = 0.2$).

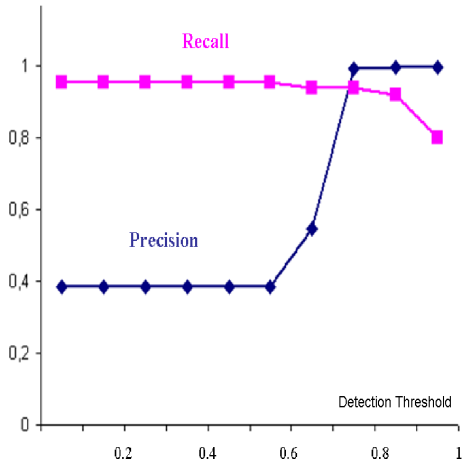


Figure 6.16: The performance of primitive states detection is measured depending on the threshold defining the level of likelihood to decide that an event is recognized. With the threshold equal to 0.75, the performance of our system is 0.96 for precision and 0.93 for recall.

These patients are part of a clinical trial for Alzheimer study and they are asked to perform a set of activities. The length of each video sequence is about 12($\pm$5) min, 8 fps.

Table. 6.3 shows that we manage to successfully recognize primitive state (e.g. 'Close phone': 85.7%, 'Person inside coffee corner': 98%) with a low false detection rate. By avoiding

Events	GT	% R	FP	FN
Person sitting	21	76.2	6	5
Close phone	14	85.7	0	2
Person inside coffee corner	51	98	3	1
Person inside reading zone	28	100	5	0
Person inside zone library	19	95	2	1
Person inside zone TV	14	100	2	0
change-zone from coffeeCorner to reading zone	17	100	0	0
change-zone from reading zone to coffeeCorner	15	100	2	0
change-zone from coffeeCorner to library zone	11	90	2	1
change-zone from library zone to coffeeCorner	10	100	0	0
Person reading	14	92	2	1
MatchingSheetsActivity	31	58	1	13

Table 6.3: Recognition Results of the proposed algorithm using Bayesian probability with temporal filtering: the recognition rate (% R), the false positive (FP) and the false negative (FN). *37* patients, 12($\pm$5) min, 8 fps.

miss detections of primitive states, the proposed system recognizes the complex events with a recognition rate about 58% for the MatchingSheetsActivity' event and 100% for the event 'change-zone reading zone to coffeeCorner'.

Figure 6.18 shows the value of probability of the composite state 'Person interacts with reading table' compared to the ground truth. This graph shows that the probability computed by the proposed algorithm is concordant with the GT probability: when the composite state is annotated (i.e. the ground truth probability is equal to 1), the probability computed by the system has a high value and when the ground truth probability is equal to 0, the probability computed by the system has a low value.

The comparison of the performance of the proposed algorithm with/without probabilistic constraint verification algorithm (table 6.4) shows that the complex event 'Up-Go' in the case of the algorithm with probabilistic constraints (92.59 %) is higher than the recognition rate of the deterministic algorithm (59.25%). This can be explained by the fact that the deterministic algorithm fails to recognize the primitive state 'Person-inside-Stop-zone' because the person was not correctly detected, the bounding box representing the person is not well detected (see fig.6.19). This error causes a wrong person coordinates and affects deeply the process of the spatial constraint verification based on people and zone coordinates. However, the probabilistic algorithm manages to recognize this primitive state and as a consequence the complex event.

The comparison with [Vu et al., 2003a](table 6.5 and table6.6) shows that the proposed

Figure 6.17: Activity detection evaluated on Health care videos. The top left image illustrates the recognition of the primitive state 'close-Phone', the right bottom image illustrates the recognition of the primitive state 'inside-zone-coffeeCorner' and the primitive event 'change-zone-reading-to-coffeeCorner'.

algorithm recognizes activities with a higher rate of precision and sensitivity than the algorithm [Vu et al., 2003a]. Table 6.6) shows that the recognition rate of the complex event 'Matching-SheetsActivity' in the case of the proposed algorithm (58%) is higher than the deterministic algorithm [Vu et al., 2003a] (38%). This can be explained by the fact that the deterministic algorithm fails to recognize some primitive states mainly due to multiple occlusions and blob segmentation errors. However, the proposed algorithm manages to recognize the primitive state and as a consequence the complex event. Table 6.5 shows that the proposed algorithm recognizes the primitive event 'Using phone' with a precision of 100% and a sensitivity of 93.3% whereas the algorithm [Vu et al., 2003a], recognizes this event with a precision of 85.50% and a sensitivity of 72.83%.

Events	♯videos	♯actor	% R	FP	FN
Deterministic constraints algo					
Up-Go	27	1	59.25	3	11
Begin-Guided-test	9	2	88.9	1	1
Interacts-with-chair	10	1	100	0	0
Probabilistic constraints algo					
Up-Go	27	1	92.59	5	2
Begin-Guided-test	9	2	100	1	0
Interacts-with-chair	10	1	100	0	0

Table 6.4: Comparison of algorithm performance with/without probabilistic reasoning: recognition rate (% R), the false positive (FP) and the false negative (FN) of our algorithm with probabilistic reasoning (probabilistic) and without probabilistic reasoning (deterministic).

Events	Precision (P)	Sensitivity (S)
[Vu et al.**, 2003**a**]**		
Using Phone	85.50%	72.83%
watching TV	71.42%	80.0%
Using reading table	58.62%	92.72%
In zone coffee Corner	69.44%	90.36%
Proposed Bayesian Approach with temporal filtering		
Using Phone	100%	92.3%
watching TV	87.5%	100%
Using reading table	84.8%	100%
In zone coffee Corner	94.3%	98%

Table 6.5: Comparison of recognition results (sensitivity (S) and precision (P)) of the proposed algorithm with the state of the art algorithm[Vu et al., 2003a].

Algorithm	GT	% R	FP	FN
[Vu et al.**, 2003**a**]**				
MatchingSheetsActivity	31	38	0	19
Proposed algorithm				
MatchingSheetsActivity	31	58	1	13

Table 6.6: Comparison of recognition rate (% R), the false positive (FP) and the false negative (FN) of the proposed algorithm with the state of the art algorithm [Vu et al., 2003a]. The ground truth GT correspond to 31 videos sequences (12($\pm$5) min, 8 fps)

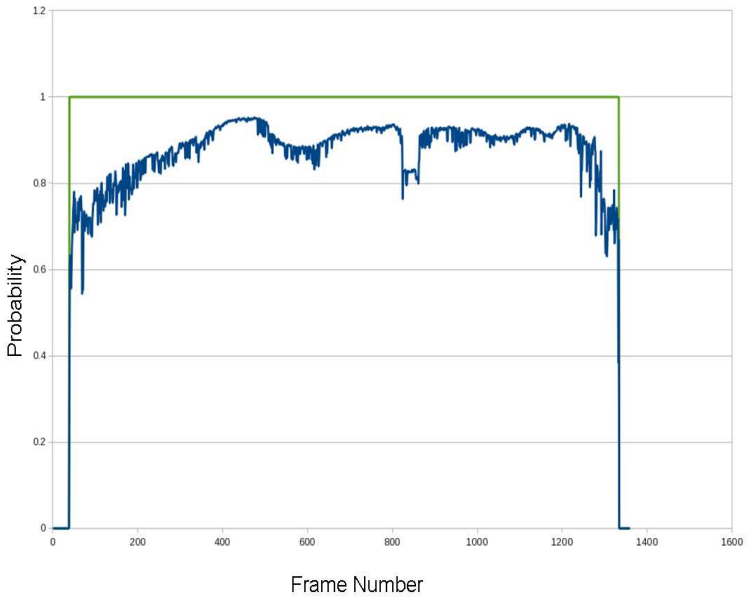


Figure 6.18: Figure shows probability values of the composite state 'Person interacts with reading table' computed by the system (blue graph) and the ground truth probabilities(green graph) (i.e. 1 when the primitive state is detected (annotated) and 0 when it is not annotated as recognized).This graph shows that the probability computed by the proposed algorithm is concordant with the GT probability: when the composite state is annotated (i.e. the ground truth probability is equal to 1), the probability computed by the system has a high value and when the ground truth probability is equal to 0, the probability computed by the system has a low value.

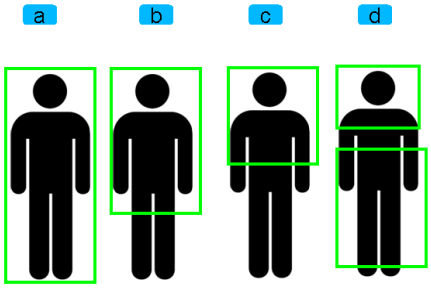


Figure 6.19: Figure (a) shows for a person an example of a well detected bounding box, in figures (b), (c) and (d), the person is not well detected.

6.7.2.1 GERHOME experiments

Table 6.7 shows the results of the proposed algorithm for Gerhome database. This table summarizes the ground truth (GT), the recognition rate (%R), the false positive (FP) and the

Events	GT	% R	FP	FN
Close Armchair	9	88	1	1
Sitting at armchair	4	100	3	0
Move from kitchen to livingRoom	4	100	3	0
Move from livingRoom to kitchen	6	100	2	0

Table 6.7: Recognition Results of the proposed algorithm on Gerhome dataset: the recognition rate (% R), the false positive (FP) and the false negative (FN). The ground truth GT corresponds to 4 videos sequences, with a total of 9452 frames, 8 fps.

Events	♯videos	TP	FP	FN
Deterministic algo				
Stay-at-kitchen	15	13	1	2
prepare-meal	8	6	1	2
Probabilistic algo				
Stay-at-kitchen	15	14	1	1
prepare-meal	8	7	3	1

Table 6.8: Comparison of algorithm performance with/without probabilistic reasoning: recognition rate (% R), the false positive (FP) and the false negative (FN) of our algorithm with probabilistic reasoning (probabilistic) and without probabilistic reasoning (deterministic).

false negative (FN) of 4 videos sequences, with a total of 9452 frames, 8 fps.

Table 6.8 shows the comparison of our algorithm performance with/without probabilistic reasoning. The comparison shows that our approach with probabilistic mechanism shows better performance than without probabilistic reasoning.

Table 6.9 shows a comparison of the performance of the proposed algorithm with the state of the art algorithm[Zouba, 2010], the comparison shows that we manage to outperform the results of the state of the art algorithm. For the primitive state 'In livingRoom' we have almost the same precision rate (93.3% vs. 93%) but our algorithm has better results in term of sensitivity rate (93% vs. 88%). False negative detection were clearly reduced by the use of the proposed probabilistic spatial constraint verification algorithm. In fact, the spatial constraint of this primitive state

For the primitive state 'In Kitchen', we have better results for both precision (87.2% vs. 86%) and sensitivity (97.1% vs. 91%) for camera video sensor. The authors [Zouba, 2010] obtain better results for precision (i.e. 89%) when using environmental sensor to detect the primitive state 'In Kitchen' but to the detriment of sensitivity (i.e. 55%).

Events	Precision (P)	Sensitivity (S)
[Zouba, 2010] using camera video		
In livingRoom	93%	88%
In Kitchen	86%	91%
[Zouba, 2010] using environmental sensor		
In Kitchen	89%	55%
Proposed algorithm using camera video		
In livingRoom	93.3%	93%
In Kitchen	87.2%	97.1%

Table 6.9: Comparison of recognition results (sensitivity (S) and precision (P)) of the proposed algorithm with the state of the art algorithm[Zouba, 2010].

Events	Accuracy (Ac)	Recall (Re)
inside zone corridor	75%	100%
close stairs	67%	67%
enter zone	85%	100%
change zone	100%	100%
overall	81.75%	91.75%

Table 6.10: Recognition Results of the proposed algorithm on Etiseo dataset: the accuracy, Ac = TP/(TP+FP+FN) and the recall, Re = TP/(TP+FP).

6.7.2.2 Etiseo experiments

We carry experiment on Etiseo dataset which is a publicly available video database composed of real world videos which include multiple camera views (figure 6.20). We have selected this database because it was used previously in several work [Lavee et al., 2010a], [Albanese et al., 2008], [Vu et al., 2003a], which allows us to compare our work with state of the art methods. We use the Building Entrance videos composed of 4 video sequences of about 924-1027 frames (30-41 seconds in length). We define 5 activities of interest and each video sequence contains one or more of these activities. Table 6.10 describes the results obtained for Etiseo videos. The algorithm recognizes the primitive state 'inside zone corridor' with an accuracy of 75% and a recall of 100% for the event enters zone, the accuracy is 85% and the recall 85%.

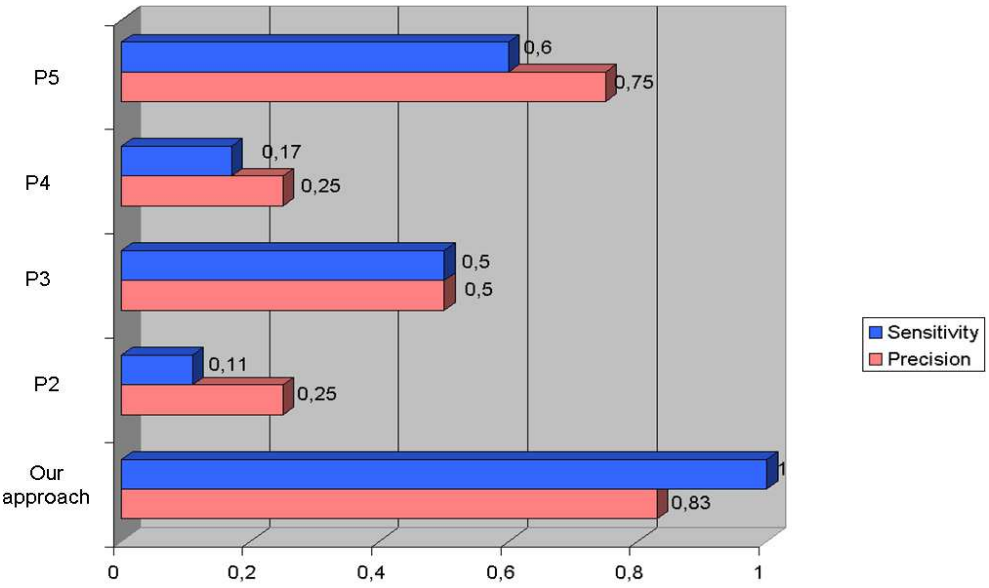
We compare the recognition performance of our approach with the results of the partners who participate in the Etiseo project. We note that the partner results are kept anonymous as the goal of the project was to provide a platform where every one can evaluate and com-

Figure 6.20: Etiseo dataset.

Approach	Accuracy (Ac)	Recall (Re)
Our approach	81.75%	91.75%
Probabilistic Petri Net [Lavee et al., 2010a]	89%	87%

Table 6.11: Comparison of recognition results (the accuracy, Ac = TP/(TP+FP+FN) and the recall, Re = TP/(TP+FP)) of the proposed algorithm with the state of the art algorithm[Lavee et al., 2010a].

pare anonymously his work on video understanding. We also note that not all partners have given results for each video understanding task (detection, tracking, activity recognition). We compare also our performance with the probabilistic approach based on probabilistic Petri Net [Lavee et al., 2010a]. Table 6.11 shows the comparison of the overall results with the probabilistic Petri Net [Lavee et al., 2010a]. We have better results for recall (91.75% vs. 87%), for accuracy, the results are (81.75% vs. 89%).



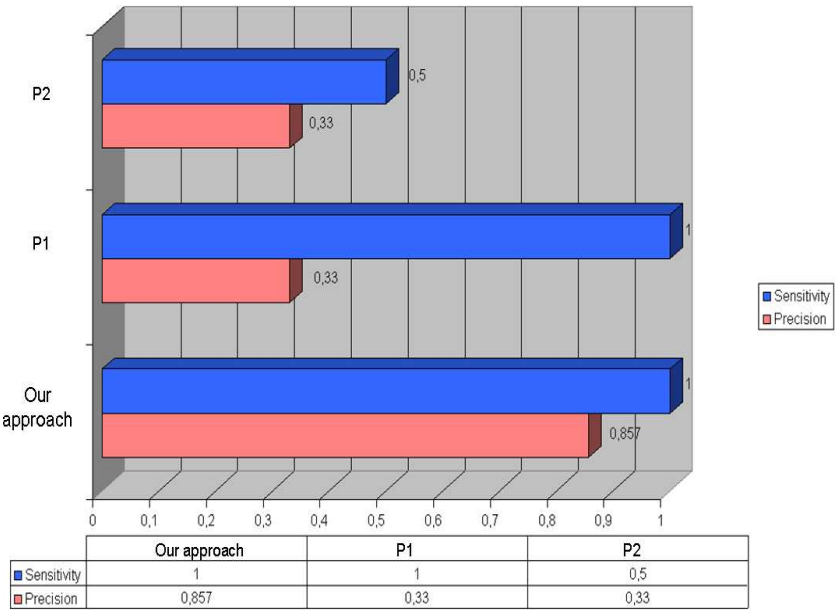
	Our approach	P2	P3	P4	P5
Sensitivity	1	0,11	0,5	0,17	0,6
Precision	0,83	0,25	0,5	0,25	0,75

Figure 6.21: The performance of the detection of the primitive state 'inside-zone' evaluated on Etiseo dataset, compared with 4 partners (2006).

Figures 6.21 and 6.22 show the comparison of the detection performance of the primitive state 'inside-zone' and the primitive event 'enters-zone' evaluated on Etiseo public database. We use the Building Entrance video composed of 924 frames and about 30 seconds in length. We compare our results with results of 4 partners of Etiseo project which are anonymous. The comparison show that we have better results in term of precision and sensitivity.

6.7.2.3 Discussion

We have evaluated the proposed activity recognition approach on SweetHome, Gerhome and Etiseo real world videos. The results show that the proposed algorithm manages to recognize primitive states and complex events with a high recognition rate. First, the probabilistic constraint verification algorithm allows us to avoid miss-detections (i.e. false negatives) of primitive states which allows us to improve in one hand the recognition rate of primitive state and in other hand improve the recognition rate of composite events. Second, Bayesian probability allows us know how probable is the detection of each event, event which are uncertain are detected with low probability and event with a high certainty are detected with a a high probability value. The re-estimation of low level attributes based on temporal filtering allows us to have more robust values and thus improve the detection of events. The comparison of the



	Our approach	P1	P2
Sensitivity	1	1	0,5
Precision	0,857	0,33	0,33

Figure 6.22: The performance of the detection of the primitive event 'enter zone' evaluated on Etiseo dataset, compared with 4 partners (2006).

performance of the proposed algorithm with/without probabilistic constraint verification algorithm shows that the recognition rate of complex events for the algorithm with probabilistic constraint verification is higher that the recognition rate of the algorithm without probabilistic constraint verification algorithm.

We have compared our results with state of the art approaches [Zouba, 2010], [Vu et al., 2003a], [Lavee et al., 2010a]. The comparison shows that the proposed algorithm outperform in term of precision and sensitivity the algorithms described in [Zouba, 2010], [Vu et al., 2003a] for Gerhome and SweetHome Videos. We compare our results on Etiseo videos with the algorithm described in [Lavee et al., 2010a]. The comparison shows that the two algorithms have very close performance which is a good performance on Etiseo database: we have better results for recall (91.75% vs. 87%), for accuracy, the results are (81.75% vs. 89%). But the main limitation of the work described in [Lavee et al., 2010a] lies in the way that the authors propose to integrate the probability in the deterministic Petri Net: the authors have not described how they have calculated the different parameters of the algorithm. For example, to calculate the transition probability, the authors use this equation: $\hat{P}(x_t = b|x_{t-1} = a) = (\alpha)^{\tau}$. The authors have not specified how the minimum distance τ is calculated and they have not detailed their choice of α, which value can deeply change the final value of the probability (for more details, see section 2.1.4, chapter 2). Another limitation which is more specific to the formalism of Petri Net, is the non-modular aspect of the Petri Network, which makes difficult any modification of

the structure of network. In the following, we detail the several results obtained for health care monitoring and in particular Alzheimer monitoring at hospital.

6.8 Health Care Monitoring

Alzheimer Disease (AD) and related disorders represent a major challenge for health care systems with aging populations. Alzheimer Disease is associated with neurodegenerative changes that compromise cognitive and functional abilities and may result in behavioral and NeuroPsychiatric Symptoms (NPS). Many efforts are currently undertaken to investigate Alzheimer disease pathology and develop appropriate treatment strategies. These strategies focus on preserving cognitive and functional abilities along with reducing NPS and maintaining quality of life in the Alzheimer disease sufferer.

Rating scales are essential tools for the diagnosis of Alzheimer disease, the assessment and careful monitoring of symptoms, as well as the evaluation of treatment effects.

However these standard rating scales do not fully capture the complexity of the disease. In fact, Alzheimer disease includes deterioration in cognitive, behavioral and functional domains that do not always progress in parallel and may change idiosyncratically according to the individuality of a given patient. The current evaluations of psycho-behavioural disorders are based on interviews and battery of neuropsychological tests with the presence of a clinician. These evaluations show limits of subjectivity (e.g., subjective interpretation of clinician at a date t).

For these reasons, we propose to overcome these limitations by capturing more fully the functional, as well as behavioral and cognitive disturbances associated with Alzheimer disease. The overall aim of this study is to demonstrate that it is possible using the proposed video monitoring system to improve the evaluations of psycho-behavioural disorders and obtain a quantifiable assessment of instrumental activities of daily living in Alzheimer and in mild cognitive impairment (MCI).

6.8.1 Health Care Monitoring Goals

In this section we detail the main goals for health care monitoring.

- Assess the initiative ability of the patient and whether the patient is involved in goal directed behaviors
- Assess walking disorders and potential risk of falls.
- Differentiate of early stage Alzheimer with patients with mild to moderate stage of the disease.

- Assess the impact of behavioral disturbances and apathy in particular on the completion of the proposed activities of daily living.

- Assess the participant acceptability to introduce a follow-up video monitoring system within their own house

6.8.2 Assessment Feasibility Study: 3 clinical cases

The first aim of this study was to demonstrate the feasibility of the proposed new assessment method from both the patient and the technical points of view. We demonstrate here using 3 clinical cases the feasibility results of the proposed automatic video monitoring system aiming to assess subjects involved in a clinical scenario.

6.8.2.1 Experimental Protocol

This part is based on the (1) short physical performance (i.e. directed activities, see section 4.3.3, chapter 4) and requires the examiner, who remains in the room, to verbally direct the participant to undertake various daily tasks. (2) Semi-directed activities (see section 4.3.3, chapter 4). The aim here is to determine the extent to which the participant could undertake a list of daily activities in a given order, after having been given a set of instructions and (3) Undirected ('free activities) (see section 4.3.3, chapter 4). Prior to leaving the room, the examiner describes each of the activities and the location and the use of various objects needed to undertake the task.

6.8.2.2 Participants

Two AD patients and one normal control participant (NC) are included in the study. Quantitative clinical characteristics of the participants are shown in figure 6.23. Brief case histories are outlined here below:

- **Participant 1**: Mr M. was a 77-year-old retired engineer. He was diagnosed with AD seven years ago. He has a major depressive disorder associated with AD. At the time of the assessment, he had some depressive symptoms but was not apathetic. He was no longer independent and his wife assisted him with most ADLs. He demonstrated difficulties in memory and executive function especially in attention, inhibitory control, planning and mental flexibility. MMSE score was 21. As already described in section 4.3.1, chapter 4, the MMSE is a series of questions, each of which scores points if answered correctly. If every answer is correct, a maximum score is 30 points.

	Mr M Participant 1 - AD	Mrs B Participant 2 - AD	Mr D Participant 3 - Control
Sex	M	F	M
Age	77	76	77
Level of Education	High	High	High
MMSE score	21	20	29
NPI total score	17	9	0
NPI domain score	17	9	0
NPI *Delusion*	0	0	0
Hallucination	0	0	0
Agitation	0	0	0
Depression	6	6	0
Anxiety	0	0	0
Euphoria	0	0	0
Apathy	2	4	0
Desinhibition	3	1	0
Irritability	3	0	0
A M B	0	0	0
Sleep disorder	3	0	0
Appetite	0	0	0
MADRS	32	7	0
GDS	14	4	0
AI caregiver	2	6	0
AI patient	0	6	0
AI clinician	2	6	0

Figure 6.23: Clinical characteristics of the 3 subjects participating in the experiment. MADRS: Montgomery Asberg depression Scale; GDS: Geriatric Depression Scale; AI: Apathy Inventory; NPI: Neuropsychiatric Inventory.

- **Participant 2**: Mrs B was a 76-year-old woman. She was diagnosed with AD 7 years ago. On neuropsychological testing, her profile was similar to Mr M., but with additional deficits in episodic memory and executive function. MMSE score was 20. She was generally independent with ADLs, but required some assistance for more complex tasks. She had clinically significant apathy.

- **Participant 3**: Mr D was a, 77-year-old man, with no significant medical history and not on any medication. He was cognitively intact on testing and fully independent in ADLs.

6.8.2.3 Judgment Criteria

To assess patients, we compute several criteria. For the directed activities part, these criteria are: (1) the speed of execution of each activity (seconds), which corresponds to the time spent undertaking each activity; (2) the speed of displacement of the participant (cm/seconds) computed on the basis of the 3D coordinates of the person over time, and (3) the step stride length (centimetres).

For the semi-directed activities part, these criteria are: (1) speed of execution(seconds), (2) number of tasks done in the correct order, which corresponds to the correct order of recognized activities in the video monitoring system, (3) number of errors in the order of the tasks undertaken by the participant , and (4) task omissions, referring to any task that the participant was unable or forgot to undertake.

For the free activities part these indicators are: (1) Number of occurrences of an activity, (2) time spent in each activity chosen by the participant, and (3) proportion of free activity time not recognized by the video monitoring system.

6.8.2.4 Results

The first result of the study is to demonstrate the feasibility of this new assessment method from both the patient and the technical points of view. During the first step, the control participant performed all these activities faster than the two AD participants. During the second step of the clinical scenario (i.e. Semi-directed activities), the two AD participants were not able to follow the correct order of the tasks and even omitted some of them. Finally during the last step of the scenario devoted to free activities the control participant chose one of the proposed activities (reading) and undertook this activity for almost the entire duration. In contrast, the two AD participants had more difficulties choosing one of the suggested activities and were not able to undertake any one activity in a sustained manner.

- Automatic video monitoring results

The results show that we could differentiate between the three participants using the proposed video monitoring system. Table 6.12 shows that, for the first step of the clinical scenario (i.e. directed activities), Participant 1 (AD) took the longest time (134 seconds) to perform the balance and sitting-to-standing (39 seconds) tasks. Participant 1 walked also slower (18.18$\mathrm{cm/s}$) compared to Participant 2 (54.5$\mathrm{cm/s}$) and Participant 3 (100$\mathrm{cm/s}$). There were also important differences among the three participants in the gait parameter of step length.

Total duration of the semi-directed activities also differs significantly among the three participants: 4:21 minutes for the control participant, 15:54 minutes for Participant 1 and 12:44 for Participant 2. The speed of execution and data regarding the tasks undertaken is outlined in table 6.13. For each activity defined in the semi-guided part of the clinical scenario, the table describes the time spent by each patient to execute the activity. The table describes also the omission, error of order done by each patient when executing the activities. Briefly, this revealed that the control participant was able to undertake the semi-directed activities in the correct order, in contrast to the two AD participants who made several errors and had omissions.

Finally, the activities detected for each participant during the undirected or 'free' part are presented in table 6.14. The total duration of this part was 30 minutes. However, whereas

Criteria computed by system	**P1- AD**	**P2- AD**	**P3- NC**
Time execution of balance exercise (s)	134	52	62
Time execution of walking exercise (s)	33	34	26
Time execution of sit-to-stand exercise (s)	39	11	8
Speed of walk (cm/s)	18.18	54.5	100
Criteria computed by GT			
Step-length	40.8	46.1	88.9

Table 6.12: directed activities step video monitoring results: the speed execution (seconds) of each part of the directed activities (balance, walk, sit-to-stand), the speed of displacement (cm/s) and the step length (cm) of the participants P1, P2 and P3.

the video monitoring system captured 25 minutes of the control participant in only one activity (reading), only 11 minutes and 13 minutes of activity were captured for Participants 1 and 2 respectively doing goal directed activities.

■ Discussion

Alzheimer disease combines symptoms belonging to different domains, including cognition, behaviour and daily functioning. These domains are usually assessed separately with specific scales or interviews involving the clinician, the patient and the caregiver. In general, these evaluations provide an overview of the patient's ability and, ideally, offer a correct understanding of the level of severity of the disease. In contrast, the objective of the current study was to assess patients in situations closer to real life by using a new assessment approach. In other words, patients are invited to spend some time and to undertake daily activities in a room equipped with the proposed automatic video activity monitoring system.

The first result of the study is to demonstrate the feasibility of this new assessment method from both the patient and the technical points of view. It is interesting to note that following a full explanation of the automatic video monitoring system, none of the AD patients nor caregivers refused to participate. The assessment is also interesting from the caregiver's point of view in that they are able to observe the AD participant's behaviour by means of a monitor located outside the study room. This is meaningful from a qualitative perspective in understanding the AD patients behavioural disturbances.

The scenario is composed of three distinct steps. The first step is devoted to directed motor activities such as standing and sitting. The control participant has performed all these activities faster than the two AD participants. Results also indicated that the motor parameters obtained allow the two AD participants to be differentiated. Participant 2 was faster than Participant 1 during the walking task and for the balance task. These parameters are interesting to observe

Participant 1 (AD)	**Right Order**	**Error of Order**	**Omission**	**Speed**
1. read something for 2mn	(×)			2:20
2. make warm some water	(×)			0:28
3. compose phone number			(×)	-
4. water plant			(×)	-
5. turn TV on		(×)		0:31
6. classify playing cards by color		(×)		1:06
7. take 'ABCD' folder			(×)	-
8. Match the A, B, C, D sheets			(×)	-
9. Put the 'ABCD' folder back			(×)	-
Participant 2 (AD)	**Right Order**	**Error of Order**	**Omission**	**Speed**
1. read something for 2mn	(×)			2:18
2. make warm some water	(×)			1:16
3. compose phone number		(×)		1:04
4. water plant			(×)	-
5. turn TV on			(×)	-
6. classify playing cards by color		(×)		0:57
7. take 'ABCD' folder		(×)		
8. Match the A, B, C, D sheets		(×)		3:36
9. Put the 'ABCD' folder back		(×)		-
Participant 3 (NC)	**Right Order**	**Error of Order**	**Omission**	**Speed**
1. read something for 2mn	(×)			1:46
2. make warm some water	(×)			0:15
3. compose phone number			(×)	-
4. water plant	(×)			0:05
5. turn TV on	(×)			0:25
6. classify playing cards by color	(×)			0:55
7. take 'ABCD' folder	(×)			0.02
8. MAtch the A, B, C, D sheets	(×)			0.19
9. Put the 'ABCD' folder back			(×)	-

Table 6.13: semi directed activities video monitoring results. Activities done by each participant in the correct order, the activities done but not in the right order (error of order), the activities that participants have not done or forget to do (omission) and the speed of execution of the activity (minutes: seconds).

Activity	Participant 1- AD	Participant 2- AD	Participant 3- NC
reading	-	1mn(1)	24mn(2)
making tea/coffee	-	12mn(3)	-
watching TV	10mn(4)	-	-
looking at library	1mn(2)	-	1(2)

Table 6.14: Activities recognized by the video monitoring system during the free activities part. For each activity: time spent for the activity (in minutes) and the number of different occurrences of this activity.

considering that recent studies show that ambulatory velocity decreases with the severity of the disease [Nordin et al., 2006], [Ries, 2009].

The second step of the scenario iss composed of semi-directed cognitive activities. Here the speed of processing also differed among participants. Furthermore, AD participants were not able to follow the correct order of the tasks and even omitted some of them. These types of disturbances have previously been demonstrated using multitasking [Esposito et al., 2010] in AD. Indeed, multitasking is characteristic of complex situations with few external constraints, and requires a number of executive competencies, such as selecting, organizing, and executing various tasks and impact on daily functioning.

The last step of the scenario is devoted to free activities and the data collected are particularly interesting from a qualitative point of view. The control participant chooses one of the proposed activities (reading) and undertakes this activity for almost the entire duration. In contrast, the two AD participants had more difficulties choosing one of the suggested activities and were not able to undertake any one activity in a sustained manner for the entire duration (30 minutes) of the step.

Hence, these results demonstrate that assessments using video monitoring system are not only feasible and well-tolerated, but can also provide useful information concerning motor, cognitive and behavioural dimensions of a disease such as AD.

There are, however, a main limitation to this first study. This was a pilot study and it is therefore impossible to clearly demonstrate that the technology can differentiate significantly between AD participants and controls on the basis of only three clinical cases. In next section, we conduct experimentation with higher number of patients.

6.8.3 Assessing Motor Behavioral disorders in Alzheimer Disease: 28 clinical cases

In this work, we study the ability of the proposed automatic video activity monitoring system to detect activity changes between older people subjects with and without dementia during a

clinical experimentation. A total of 28 volunteers participate to the experimentation, composed of 11 healthy elderly subjects (healthy control group, G1) and 17 AD patients (AD group, G2).

Table 6.15: Demographic and characteristics of volunteers.

Characteristics	**Control Group (G1)**	**AD patients (G2)**
Sample size (N)	11	17
Age, years (Mean SD)	74.8 ($\pm$6.67)	77 ($\pm$7.43)
Sex Ratio (F/M)	6/5	(12/5)

6.8.3.1 Judgment Criteria

For this work, we focus on the automatic detection of motor disturbances. For that, we analyse different sub-activities composing three main guided physical activities (see section 4.3.3, chapter 4). For walking exercise, we compute the walking speed, the transfer position criteria and the up and go criteria. These criteria are defined in chapter 4, section 4.3.4.

6.8.3.2 Performance system evaluation

For each video record, events modeled were annotated by the same expert in blind test. Then execution duration of activities and gait parameters associated with event of interest were performed. To compute the walking speed by GT method V_{GT} , we use as start time the instant when participant cross the start line, and, as end time the instant when they cross the end line or make a halt in front of the end line. Reference value used as walked distance is the same for all participants, such as $d_{referencevalueGT} = 4m$. Mean of speed parameters is used to compare performance between estimations provided by the proposed automatic video system and the ones provided by GT (i.e., position and event annotations).

To compare the ability of the proposed system to detect differences between patients profiles compared to GT, statistics analysis are conducted with SPSS[10] release 19.0 software using the non parametric Mann-Withney test and their associated p-value determined from Monte-Carlo simulation in order to have more meaningfull conclusion given the small sample size of the two groups.

6.8.3.3 Results

The proposed study [Joumier et al., 2011] shows that we could differentiate the two profiles of participants based on motor activity parameters, computed from the proposed automatic video activity recognition system. Differences in motor disturbances between the two profiles

[10]http://www-01.ibm.com/support/docview.wss?uid=swg2402867

of participants are concordant between the one identified by the ground truth GT and the one identified by the proposed automatic video system. We have similar results in terms of statistics differences detected from automatic video system and ground truth results performed: walking speed, the duration to execute one position transfer during the transfer position exercise (μ), and D3 (i.e., parameter related to the up and go exercise). Tables 6.16 shows the parameters results for the Control Group (G1) and the AD patients group (G2). Differences in the estimation of walking speed are (i) higher with M2 method than with M1 method (see chapter 4, section 4.3.4). These differences can be explained as follows: *(i)* in most cases, the participant l walks a distance inferior to the landmark distance when the event 'Inside Walking exercise zone' is recognized. Thus, the walking speed computed with M2 method is overestimated compared with the one computed by M1 method, *(ii)* some AD patients stop before the end line and are detected inside the zone Exercise Walking (compared to other participants who cross the end line), that's why the walking speed of these participants is underestimated (see [Joumier et al., 2011]).

Mean, ±sd	G1	G2	p-value, [CI(95%)] (&)
1. Walking Exercise			
Walking Speed(cm/s)			
1st Exercise (go)			
- video, $V_{M1,1}$	95.8, (21.7)	74.9, (22.3)	0.014, [0.012, 0.016](*)
- video, $V_{M2,1}$	100.1, (23.7)	75.7, (24.5)	0.013, [0.011, 0.015](*)
- GT, $V_{GT,1}$	85.5, (22.3)	66.3, 21.4[]	0.034, [0.030, 0.037](*)
2nd Exercise (go)			
- video, $V_{M1,2}$	109.9, (30.7)	88.6, (23.7)	0.089, [0.083, 0.094]
- video, $V_{M2,2}$	113.9, (36.4)	85.8 (25.9)	0.021, [0.018, 0.024](*)
- GT, $V_{GT,2}$	94.1 (23.0)	74.4, (19.2)	0.012, [0.010, 0.014](*)
Mean on the exercise			
- video, mean($V_{M1,1}$,$V_{M2,1}$)	102.9 (24.7)	81.7, (20.5)	0.013, [0.011, 0.015](*)
- video, mean($V_{M1,2}$,$V_{M2,2}$)	107.0, (28.5)	80.7, (21.5)	0.003, [0.002, 0.004](*)
- GT, mean($V_{GT,1}$, $V_{GT,2}$)	89.8, (21.2)	70.4, (18.1)	0.005, [0.004, 0.006](*)
2. Transfer Position Exercise			
Execution Duration Δ t			
- video	13.60, (5.70)	17.43, (6.85)	0.170, [0.162, 0.177]
- GT	15.72, (6.20)	19.96, (7.31)	0.141, [0.135, 0.148]
Transfer parameter μ			
- video	1.26, (0.57)	1.71, (0.61)	0.019, [0.016, 0.021](*)
- GT	1.46, (0.61)	1.95, (0.66)	0.012, [0.010, 0.014](*)
3. Up and Go Exercise			
Execution Duration(s)			
(D1) For walking 4m (go)			
- video	2.31, (0.96)	2.81, (1.22)	0.058, [0.054, 0.063]
- GT	2.67, (1.16)	3.37, (1.43)	0.051, [0.046, 0.055]
(D2) For doing U-turn			
- video	2.31, (1.13)	3.77, (2.65)	0.099, [0.093, 0.104]
- GT	1.68, (0.65)	2.56, (1.79)	0.053, [0.049, 0.057]
(D3) For walking 4m (back)			
- video	2.38, (0.81)	3.01, (1.27)	0.020, [0.017, 0.023](*)
- GT	2.95, (0.84)	4.19, (1.90)	0.002, [0.001, 0.003](*)
(D4) between posture changes			
- video	10.13, (3.95)	12.52, (5.83)	0.13, [0.123, 0.137]
- GT	10.19, (3.85)	12.64, (5.97)	0.071, [0.066, 0.076]

Table 6.16: Parameter estimation for healthy older participants and AD patients. (&) Non parametric Mann-Withney test is used to compare the results between both groups G1 vs G2. Bilateral p-value associated with the Mann-Whitney test and its 95% confidence interval [CI (95%) are estimated using Monte-Carlo simulation based on a sample size of $10,000$. (*) Inter group comparisons: differences between healthy older participants and AD patients, using a significance level of $.05(p - value < .05)$.

Figure 6.24 shows the comparison of the speed performance of Alzheimer people (AD) and healthy older people compared to the ground truth GT. Figure 6.25 shows the accuracy of our

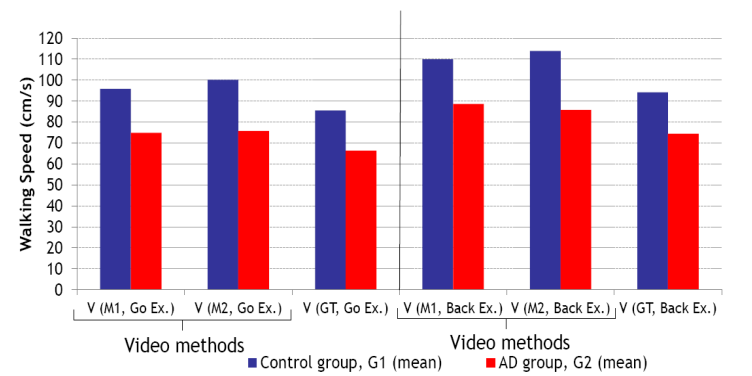


Figure 6.24: Comparison with healthy older people (group, G1) and AD people (group, G2) for walking exercise: the walking speed computed with methods M1 during the walking exercise (the Go exercise $V(\mathrm{M1}, \mathrm{GoEx})$ and the Go back exercise $V(\mathrm{M1}, \mathrm{BackEx})$), and M2. The walking speed is computed with methods M2 during the walking exercise (the Go exercise $V(\mathrm{M2}, \mathrm{GoEx})$ and the Go back exercise $V(\mathrm{M2}, \mathrm{BackEx})$). The two methods are compared with the ground truth GT method.

approaches for walking speed computation. Figure 6.26 shows the comparison between healthy

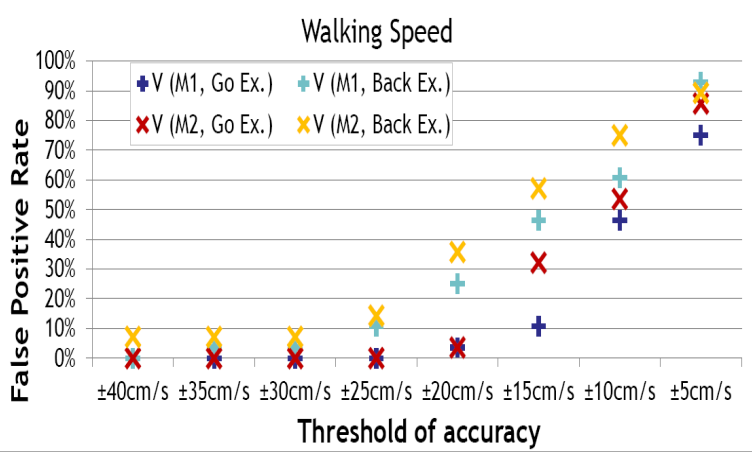


Figure 6.25: Performance evaluation of walking speed.

older people (group, G1) and AD people (group, G2) performance for the transfer exercise and figure 6.27 shows the accuracy of our approach for calculating the duration of the transfer exercise.

Figure 6.28 shows the comparison between healthy older people (group, G1) and AD people

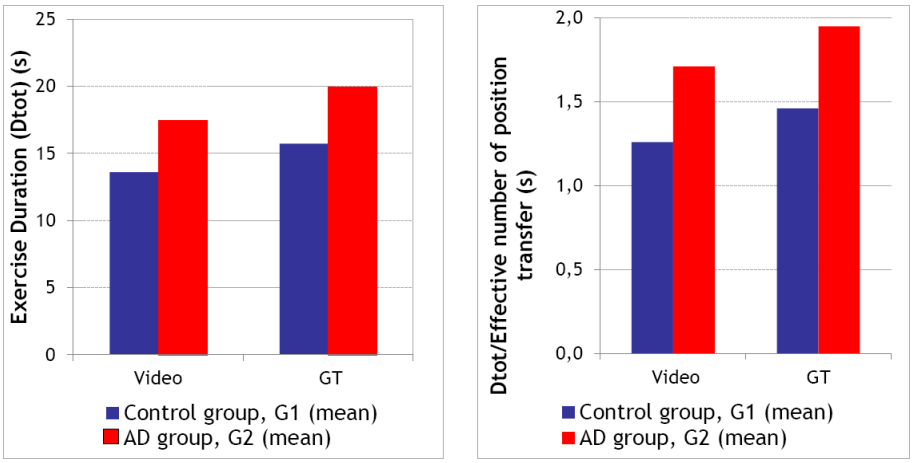


Figure 6.26: Comparison between healthy older people (group, G1) and AD people (group, G2) performance for transfer exercise: the transfer parameters (i.e the duration that takes each group to perform the exercise, and the Comparative measurement = $\mathrm{Dtot/numberofposition}$) are compared with the ground truth GT parameters. the results shows that we could differentiate between the two groups, the group G1 is faster than G2 for transfer exercise.

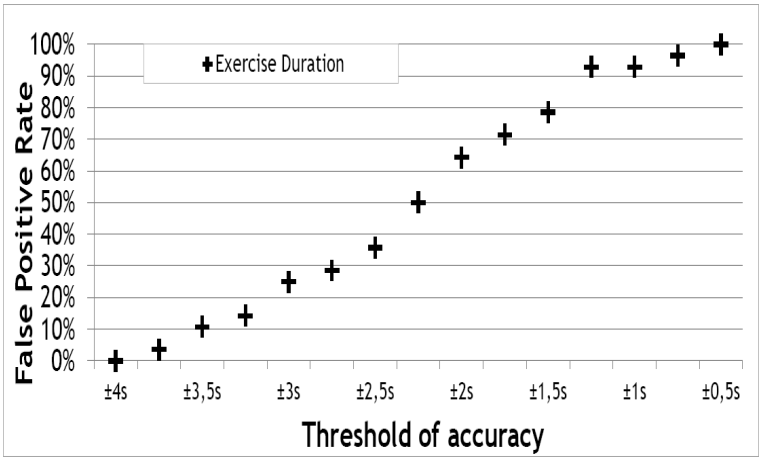


Figure 6.27: Performance evaluation of transfer exercise duration.

(group, G2) performance for up and go exercise. Figure 6.29 shows the performance of the evaluation of the duration parameters of up and go exercise.

6.8.4 Alzheimer disease patient activity assessment: 44 clinical cases

In this work, we assess older people performance on Instrumental Activities of Daily Living (IADL) and motor tests in a clinical protocol developed and executed in the Memory Center

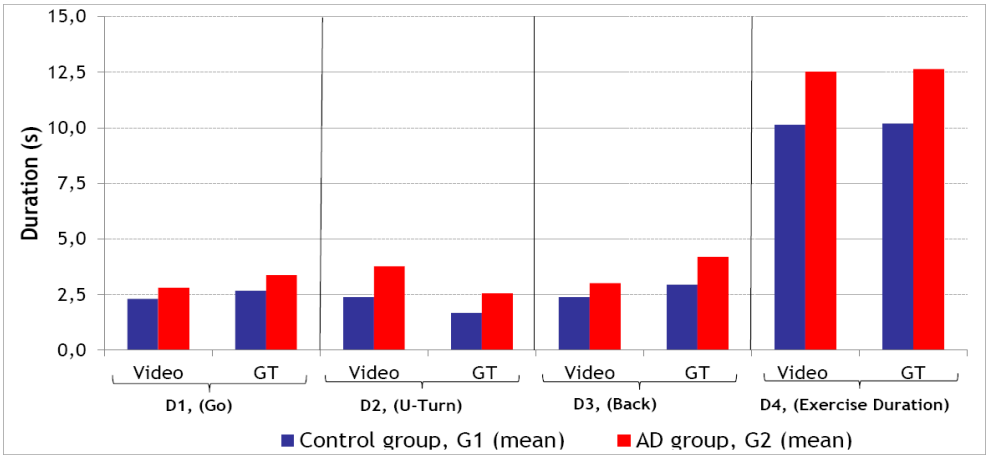


Figure 6.28: Comparison between healthy older people (group, G1) and AD people (group, G2) performance for up and go exercise: G1 is faster than G2 to do successively different actions during this exercise (transfering, walking, making U-turn).

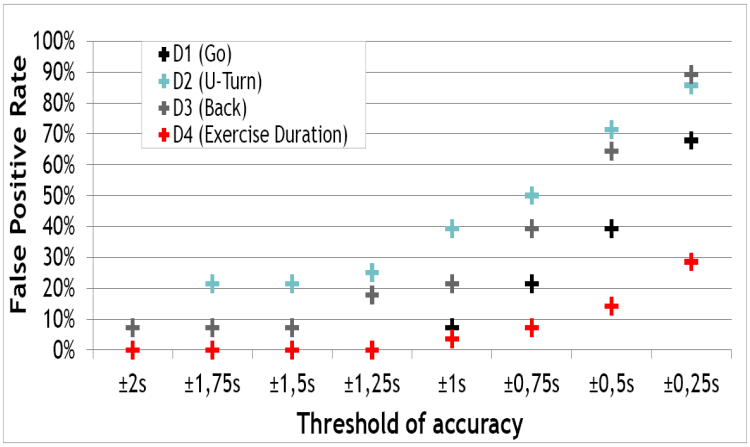


Figure 6.29: Performance evaluation of up and go exercise duration parameters.

of the Nice Hospital. In this clinical protocol, a total of 44 volunteers participate, composed of 21 healthy elderly subjects and 23 AD patients (see tables 6.17 and 6.18). The proposed monitoring system detected the full set of activities of the first part of the clinical protocol (e.g., balance test, repeated sequence of sitting-standing positions) with a detection rate of 96.9% to 100% (true positive rate).

Table 6.17: characteristics of participants on part 1 of the clinical scenario.

Characteristics	Healthy Control	AD patients
Sample size (N)	21	23
Age, years (Mean SD)	73.4 ($\pm$6.09)	76.7 ($\pm$7.84)
Female, N(%)	12 (50%)	16(68.75%)
MMSE(mean $\pm$ SD)	28.1$\pm$ 0.98	21.35$\pm$3.97

Table 6.18: characteristics of participants on part 2 of the clinical scenario.

Characteristics	Healthy Control	AD patients
Sample size (N)	10	16
Age, years (Mean SD)	73.9 ($\pm$6.24)	76.7 ($\pm$7.56)
Female, N(%)	5 (50%)	11(68.8%)
MMSE(mean $\pm$ SD)	28.1$\pm$ 1.85	20.7$\pm$3.70s

6.8.4.1 Statistical Analysis

Results are presented in terms of mean ($\pm$ standard deviation) for continuous variables. For categorical variables, results are presented in terms of frequency for each modality. Intergroup comparisons for continuous variables used the parametric Student's t-test or the non-parametric Mann-Whitney test ($p < 0.05$) if one of the assumptions of the Student's t-test is not satisfied. Intergroup comparisons for categorical variables (2 modalities) used the Fisher's exact test ($p < 0.05$). All calculations were performed using SPSS software (version 19.0).

6.8.4.2 Results for part 1 of the clinical scenario

In this section, we present the results obtained from the first part (i.e. directed activities) of the clinical scenario. These results are compared with the ground truth GT results obtained by annotation.

Table 6.19 shows the results of the participants performing the activities of the first part (i.e. directed activities) of the clinical scenario. The parameters were calculated using the human expert annotation about the activities in the video sequences. The results shows that AD participants needed more time to complete the different physical activities due to a lower displacement speed.

Table 6.19: Performance of participants on part 1 of the clinical scenario.

Parameter	NC	AD	p-value
Walking Exercise			
walking speed (Go) (m/s)	0.88±0.25	0.69±0.20	0.009 (**)
walking speed (Go back) (m/s)	1.02 (±0.21)	0.77 (±0.18)	<0.001 (**)
Transfer position Exercise			
duration (s)	14.90± 5.75	19.7±6.79	0.012 (*)
duration/number of transfers	1.50± 0.58	1.9±0.63	0.006 (**)
Up and Go Exercise			
duration (s)	10.30± 4.12	14.6±6.16	0.002 (**)

Table 6.20 shows the same activities parameters than table 6.19, but with activity parameters calculated using the activities detected by the video activity monitoring system (36/44 video sequences, where HC=16 and AD=20). Although the absolute values of the parameters calculated using our system results are different from the parameters values obtained from human annotations, the statistically significant differences between healthy participants and AD patients group were preserved.

Table 6.20: Performance of participants on part 1 of the clinical scenario.

Parameter	NC	AD	p-value
Walking Exercise			
walking speed (Go) (m/s)	1.06±0.23	0.79±0.23	0.001 (**)
walking speed (Go back) (m/s)	1.20 (± 0.31)	0.89 (± 0.23)	<0.002 (**)
Transfer position Exercise			
duration (s)	12.8± 5.40	17.7±6.31	0.006 (**)
duration/number of transfers	1.3± 0.53	1.7±0.56	0.002 (**)
Up and Go Exercise			
duration (s)	8.8± 3.80	12.1±5.64	0.007 (*)

6.8.4.3 Results for part 2 of the clinical scenario

Tables 6.21 and 6.22 present participants performance in part 2 (i.e. semi-directed activities) of the clinical scenario. Table 6.21 shows global results according to activity parameters: duration (seconds) spent inside the room to perform the part 2 of the clinical scenario and organizational errors in activity ordering. Activity ordering errors are presented as the number of participants who at least once omitted, repeated, or changed the expected temporal order of activities.

Activity	NC	AD	p-value
Total time spent in the room (s)	454±160.4	715±352	0.060
Number of participnats presenting errors about:			
activity omission (n, %)	1(3.2%)	2(12.5%)	0.508
activity repetition (n, %)	0(0%)	6(37.5%)	0.053
activity order (n, %)	0(0%)	4(25%)	0.106 (**)
at least one error at activities organization (n, %)	0(0%)	8(50%)	0.008 (**)

Table 6.21: Global performance for semi-directed activities. P-values for continuous variables were computed using Wilcoxon test; p-values for categorical variables (2 modalities) were computed using Fisher's exact test; (*) Statistical significance at $p < 0.05$; (**) Statistical significance at $p < 0.01$.

Table 6.22 shows the participants' performance for each activity in terms of speed (seconds), omission, and repetition parameters. The speed term was used instead of activity time duration to imply that lower values of this attribute highlight the ability of a participant at performing the activity faster. AD participants spent more time performing activities that involve sorting or classifying objects (classify card, up an go), and they had difficulty to manage the time of reading activity compared to NC participants.

Tables 6.23 and 6.24 present examples of an AD patient and a normal control (NC) participant performance in semi-directed activities. In this example the AD participant forgot to perform 3 activities, and performed 2 activities in the wrong order. Comparatively, NC participant performed the activities in the correct order, only omitting one.

Activity	Right Order	Order Error	Omission	Duration
reading for 2mn	OK			2:20
warming water	OK			0:28
making call			X	
watering plant			X	
watching TV		X		00:31
classifying card by color		X		01:06
Up and Go	OK		X	

Table 6.23: AD Participant's performance on semi-directed activities of the clinical scenario.

Activity	Right Order	Order Error	Omission	Duration
reading for 2mn	OK			1:45
warming water	OK			0:16
making call			X	
watering plant	OK			00:05
watching TV	OK			00:25
classifying card by color	OK			00:55
Up and Go	OK			00:19

Table 6.24: Nornal control (NC) Participant's performance on semi-directed activities of the clinical scenario.

6.8.4.4 Results Comparison

In this section, we compared the results obtained by the proposed video activity monitoring framework with the results obtained by a Taiwanese team who collaborate with our team in SWEETHOME project. The Taiwanese experiments took place in indoor and outdoor environments. For the indoor experiments a room equipped with household appliances was used and experimental data was recorded using eight ambient 2D video cameras (AXIS, Model 215PTZ,30 fps). For outdoor experiments a tri-axial accelerometer mounted on the shoes of the participants was used to analyze their gait parameters.

A stride detection algorithm was developed by the Taiwan team for the automatically acquisition of gait information using a triaxial accelerometer embedded in a wearable device. It acquires data about the participant locomotion (e.g., walking time, stride length, stride frequency). It was tested with 33 participants (healthy=17, AD=16), on a 40-meter walking test.

Taiwanese participants aged more than 50 years were recruited by the Department of Neurology at National Cheng Kung University Hospital. The Inclusion criterion of the AD group was a MMSE score value above 16. (see tab. 6.25)

Table 6.25: characteristics of Taiwanese participants on.

Characteristics	Healthy Control	AD patients
Sample size (N)	45	36
Age, years (Mean SD)	64.51($\pm$8.33)	70.25($\pm$9.25)
Female, N(%)	24 (53.3%)	21 (58.3%)
MMSE(mean $\pm$ SD)	27.60$\pm$ 2.04	23.44$\pm$3.32

Table 6.26 shows the results of Taiwanese participants performing part 1. Activity param-

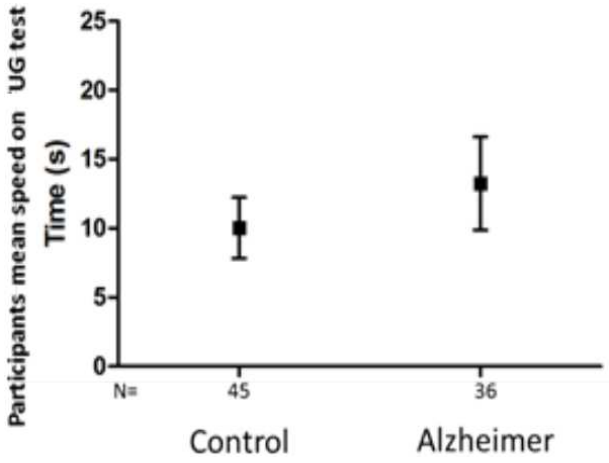


Figure 6.30: Mean duration of Taiwanese participants in the up and Go test [Crispim-Junior et al., 2012].

eters were calculated based on the annotations of Human expert. Activity parameters of Taiwanese participants agree with our experimental results in the sense that AD participants took more time to perform the selected activities, probably due to their lower speed of displacement when compared to NC participants.

Table 6.26: Taiwanese participants' performance for part 1 of the clinical scenario [Crispim-Junior et al., 2012].

Parameter	**NC**	**AD**	**p-value**
Walking Exercise			
walking speed (Go) (m/s)	0.38± 0.08	0.32 ±0.74	0.001 (**)
walking speed (Go back) (m/s)	0.41 (± 0.07)	0.34(± 0.08)	<0.001 (**)
Transfer position Exercise			
duration (s)	14.5 ± 3.33	20.0 ±9.41	0.001 (**)
duration/number of transfers	1.4 ± 0.33	2.00 ±0.94	0.001 (**)
Up and Go Exercise			
duration (s)	10.2 ± 2.42	13.4 ±3.34	0.001 (*)

Figures 6.30 and 6.31 show the mean speed of participants in the up and Go test for AD and healthy participants at the Taiwanese and our experimental sites, respectively. In both sites, AD patients presented a significantly lower speed compared with healthy controls in the up and Go test $(p < 0.01)$.

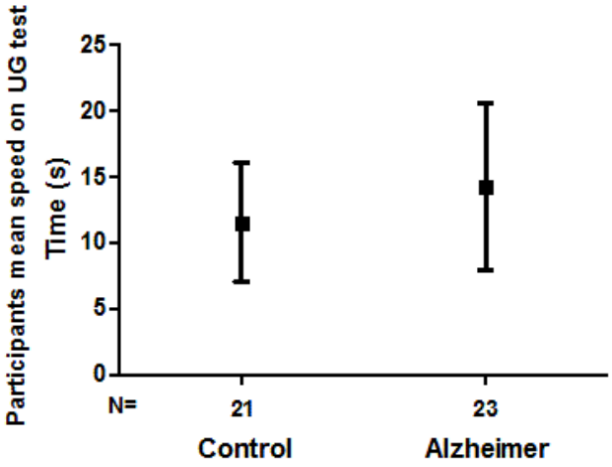


Figure 6.31: Mean duration of participants in Nice hospital France in the up and Go test.

Table 6.27: Taiwanese Results: Mean and standard deviation of patients speed in activities of part 2 [Crispim-Junior et al., 2012].

Activity	**reading for 2mn**	**warming water**	**making call**
AD	129.8±57.8	41.4±21.0	39.7±22.5
NC	121.1± 49.2	23.8 ±12.1	21.7±6.9
Activity	**watering plant**	**watching TV**	**classifying card**
AD	52.3±28.7	36.4±24.4	98.8 ±35.5
NC	25.3± 10.4	23.0 ±11.7	69.5±21.0

Tables 6.27 and 6.28 show the mean and standard deviation of the speed of participants' activities and their number of errors at performing the activities proposed in the given order (number of activities skipped or performed in an temporal order different from the expected), respectively. Statistical differences were found in the comparison between AD and NC groups ($p < 0.01$, **;$p < 0.05$, *). For instance water the plant activity show a significant differences between AD and NC for both activity parameters analyzed (speed and frequency).

6.8.4.5 Results Discussion

Certain similarities are seen between Taiwanese and our results, although a direct comparison of the results of both sites is not possible due to differences in participant population inclusion criteria. Alzheimer's patients can be characterized by several criteria. AD participants presented a lower balance and a shorter gait length frequency. Similar findings were also found by Gillain et al. [Gillain et al., 2009] who pointed out lower gait speed and lower stride length

of AD patients when compared to normal controls, in single and in dual tasks (gait speed, stride length, stride cycle frequency, and stride regularity).

Alzheimer's patients have also omitted activities and changed their temporal order indicating a decline in IADLs performance (semi-directed, part 02 of clinical scenario). Statistically significant differences among AD and NC participants in activity like 'watering the plant' could be an indicator of AD participants difficulty at performing unusual activities.

The proposed automatic monitoring system provided activity values similar to the ones calculated from events annotated by a human assessor. Although their absolute values differ, they follow the same tendency, and the statistical differences found among AD and NC groups are preserved. These findings highlight the use of the proposed approach as a support platform for clinicians to objectively measure AD patients performance in IADLs and gait analysis. Among the advantages of the proposed system are the stability of its results over time (as it does not suffer from emotional state conditions or biases like stress and fatigue), and its quantitative measurement of patient performance.

Activity	NC	AD	p-value
reading for 2mn			
Speed (s)	138±79.6	88±215.2	0.001(**)
omitted, n(%)	0(0%)	0(0%)	X
repeated, n(%)	0(0%)	2(12.5%)	0.508
warming water			
Speed (s)	4±3.8	8±13.4	0.660
omitted, n(%)	0(0%)	0(0%)	X
repeated, n(%)	0(0%)	2(12.5%)	0.508
making call			
Speed (s)	25±12.3	28±20.3	0.979
omitted, n(%)	0(0%)	0(0%)	X
repeated, n(%)	0(0%)	2(12.5%)	0.508
watering plant			
Speed (s)	9±4.4	11±7.2	0.856
omitted, n(%)	0(0%)	0(0%)	X
repeated, n(%)	0(0%)	3(18.75%)	0.262
watching TV			
Speed (s)	32±24.3	57±57.9	0.165
omitted, n(%)	0(0%)	0(0%)	X
repeated, n(%)	0(0%)	2(12.5%)	0.508
classifying card by color			
Speed (s)	78±30.9	143±176	0.216
omitted, n(%)	0(0%)	1(6.25%)	1.00
repeated, n(%)	0(0%)	3(18.75%)	0.262
Up and Go			
Speed (s)	49±21.3	80±48.5	0.129
omitted, n(%)	0(0%)	1(6.25%)	1.00
repeated, n(%)	0(0%)	0(0%)	X

Table 6.22: Participants performance for each activity of the semi-directed part of the clinical scenario. P-values for continuous variables were computed using Wilcoxon test; p-values for categorical variables (2 modalities) were computed using Fisher's exact test; (*) Statistical significance at $p < 0.05$; (**) Statistical significance at $p < 0.01$.

Table 6.28: Taiwanese Results: Mean and standard deviation of participants number of errors in the order of activities for part 02 [Crispim-Junior et al., 2012].

Activity	**reading for 2mn**	**warming water**	**making call**
AD	0.20±0.41	0.08±0.28	0.08±0.28
NC	0.26± 0.44	0.03 ±0.18	0±0
Activity	**watering plant**	**watching TV**	**classifying card**
AD	0.23±0.43	0.26±0.45	0.19 ±0.40
NC	0.03± 0.18	0.03 ±0.18	0.03±0.18

6.9 Conclusion

In this chapter, we have presented the several evaluations done in computer vision domain, in term of activity recognition and uncertainty handling and we present the evaluation in health care monitoring domain. We demonstrate that the proposed approach allows to recognize reliably with a low false alarm rate a set of activities of interest. In health care domain, the obtained results highlight the advantages of the use of the proposed approach as a support platform for clinicians to objectively measure AD patients performance in IADLs and gait analysis.

In the next chapter, we conclude our work and we propose and discuss future work to improve the proposed activity recognition framework.

7

CONCLUSIONS AND FUTURE WORK

In this thesis we have proposed a new approach for monitoring human activities. This approach includes an algorithm for real-time recognition of primitive and composite activities that have occurred in the scene observed by video cameras. The proposed approach is based on combining constraint-based approach and a probabilistic approach to recognize human activities. The proposed approach takes as input the data provided by video sensors and uses three major sources of knowledge: the 3D model of the scene, the 3D model of mobile objects (e.g. person), and the models of activities. An overview of the contributions of this work is described in section 7.1. Then a discussion is made to show the limitations of the proposed approach in section 7.2. Finally, future works are proposed in section 7.3 to improve the proposed approach.

7.1 Overview of the contributions

The main contributions of this work are the following:

- **A new video activity recognition approach combining semantic event modeling and probabilistic event inference to cope with uncertainty** (chapters 4 and 5). The proposed algorithm performs probabilistic event detection based on conditional probabilities and relies on probabilistic constraint verification. Considering uncertainty is crucial in order to maintain a robust recognition performance for real-world videos.

- **Knowledge base for health care monitoring** (chapter 4). We have worked in close collaboration with clinicians to define an ontology and a knowledge base composed of a total of 117 event models for Alzheimer monitoring at hospital. The defined ontology contains several concepts useful for health care. We have also defined a number of criteria which

could be observed by camera sensors to allow detection of early symptoms of Alzheimer's disease.

- **Assessing Behavioral disorders in Alzheimer Disease**: in chapter 6, we have studied the ability of the proposed automatic video activity monitoring system to detect activity changes between older people subjects with and without dementia. This study shows that we could differentiate the two profiles of participants (Alzheimer and normal control patients) based on daily living activities and gait parameters. These findings highlight the use of the proposed approach as a support platform for clinicians to objectively measure Alzheimer patients performance in daily living activities and gait analysis. Among the advantages of the proposed system are the stability of its results over time and its quantitative measurement of patient performance.

7.2 Limitations

The proposed activity recognition approach shows the ability to help experts to represent easily interesting events and the capacity of reliably recognize daily living activities.

However, the approach has some limitations and can be extended in a number of new studies and new research directions. These limitations are mainly divided on three main parts:

At the level of event modeling, the proposed event description method is formal. Events are described in a natural language which is intuitive. Nevertheless, event modeling is time consuming and an error prone process. Automatic event learning can improve the recognition performance (see sections 7.3.1.4 and 7.3.1.5).

At the level of event recognition, The proposed event recognition algorithm shows its success to recognize complex activities. However, many improvement can be explored for the management of uncertainty (see sections and 7.3.2.1 and 7.3.2.2). Detection of more complex situations could be improved (see sections 7.3.1.1, 7.3.1.2, 7.3.2.4 and 7.3.2.5). Actually, there is no cooperation and feed back between event recognition process and vision algorithms. Feed backs could improve the performance of the recognition (see section 7.3.1.6).

At the level of health care, the proposed approach shows its ability to help clinicians to objectively measure Alzheimer patients performance. However, the approach shows its limitation in term of finer activities detection. For instance, the detection of the change of the walk style of one patient over time could be of interest for clinicians to detect the change of patient's behaviour. Investigation on automatic detection of finer activities change is a new research direction (see sections 7.3.2.4, 7.3.2.5).

In the following, we describe propositions for future work.

7.3 Future Work

The purpose of this section is to analyze the future work, as extensions to the proposed approach and as possible solutions to its limitations. In this section we present firstly the proposed short-term perspectives, after that we present the proposed long-term perspectives.

7.3.1 Short-Term Perspectives

In short term, the activity monitoring approach can be extended in several ways:

7.3.1.1 Testing with other sensors

The proposed approach was tested mainly on 2D videos. One short term work consists on testing the performance of the algorithm using RGB-D approach for instance by using Kinect[1].

7.3.1.2 Multi-sensors Event Detection

The proposed approach for uncertainty handling was mainly applied on camera sensors. Future work consists in handling uncertainty of other sensors and combining them for a more rich multi-sensor event detection.

7.3.1.3 Group Activity Recognition

The proposed approach was evaluated mainly with one person or two at most (i.e. patient and nurse). We propose the extension of our work for group detection like the work described in [Zaidenberg et al., 2012]. Future work consists in evaluating the integration of uncertainty management in the performance of group activity detection.

7.3.1.4 Learning utility values

In this work, the utility values during the event modeling process were defined according to the user interest. Further analysis on different utility values can be performed in order to establish the best values for each event model. A learning phase of these values could be of interest for a more reliable event detection.

7.3.1.5 Learning Temporal Event Models and Uncertainty handling

The modeling of behaviors can be expressed by experts by the mean of the presented event description language and a dedicated ontology. Nevertheless, event modeling is time consum-

[1] http://www.microsoft.com/en-us/kinectforwindows/

ing and an error prone process. Thus, it will be interesting to learn automatically normal behaviors of every day data, because normal behaviors are frequent and can be extracted from everyday activities. Of course, in the process of event learning we could handle uncertainty. A potential future direction of our work might be to investigate learning techniques to deal with the uncertainty of recognition at the level of event modeling.

7.3.1.6 Cooperation and feed back between event recognition process and vision algorithms

In the current video recognition framework, there is no dynamic cooperation between the algorithms, the event detection algorithm takes simply the output of vision algorithms without sending feed backs. It will be interesting for event recognition algorithm based on the probability of the detected activities to send feedback to vision algorithms to perform certain vision tasks for instance to track again people in a time interval. This could improve the performance of the recognition. A future research work consists in studying the criteria to establish cooperation and feed backs between vision modules and event recognition.

7.3.2 Long-Term Perspectives

In long term, the activity monitoring approach can be extended in several ways:

7.3.2.1 Low level Uncertainty Management

The calculation of probability measures to assess the reliability of the low level algorithms can be an interesting extension of the approach. These measures could be associated to the detected moving regions in order to account for the quality of low level and its robustness faced to low level noise (e.g. illumination changes, contrast between the moving objects and the background of the scene, shadows, among other noise aspects).

7.3.2.2 Uncertainty Management Techniques Comparison

In this work, we have presented a number of techniques for the management of uncertainty during the different levels of the event recognition process, among many other state of the art techniques. Further analysis and comparison with other state of the art techniques is of interest. Evaluating the most appropriate techniques for the context of health care monitoring could enrich the proposed Alzheimer monitoring platform.

7.3.2.3 Context Learning

In this work, the scene context was manually annotated. Learning the scene context could be of interest to gain in flexibility in context definition like the work described in [Pusiol, 2012]. Future work can focus on exploring (i) different context learning techniques like detecting the most frequented areas based on person presence and (ii) the impact of different learned contexts of the recognition performance.

7.3.2.4 Study interaction

Currently the interactions between physical objects are studied based on the position/closeness of physical objects to each other. This definition is coarse and a study of finer interactions could enrich our event recognition framework and could be of interest for health care monitoring. Further analysis on the results from the low level algorithm which gives actually a coarse description of the object is needed. In order to evolve in the interpretation of more complex interactions, more detailed and class-specific object models could be utilised. Future work can point to the utilisation of more specific object representations as articulated models, object contour, or appearance models, among others.

7.3.2.5 Finer activities detection

The detection of the change of the walk style of a patient over time or the detection of way how patient perform balance exercise could be of interest for clinicians to detect the change of patient's behaviour. Detecting the way a person is moving his/her hand, legs and each part of his body could help us to detect for instance walk style. Learning techniques could be also used to learn some behaviors specific to Alzheimer disease patients.

7.3.2.6 Activity Monitoring in other Environment

In the current work, the proposed activity recognition approach was evaluated mainly on health care applications. The next step requires to test this approach on other environment for instance metro surveillance, parking lot, etc.

Publications of the Author

■ **International Journal:**

1. Automatic Video Monitoring system for assessment of Alzheimer's Disease symptoms. R Romdhane, E Mulin, A Derreumeaux, N Zouba, J Piano, L Lee, I Leroi, P Mallea, R David, M Thonnat, F Bremond and P H Robert. The journal of nutrition, health and aging 2012, (JNHA' 2012), 16(3): pp 213-219.
2. Functional dementia assessment using a video monitoring system: Proof of concept. E Mulin, V Joumier, I Leroi, J H Lee, J Piano, N Bordone, A Derreumeaux, P Mallea, P Brocker, A Dechamps, R Romdhane, M Thonnat, F Bremond, R David and P H Robert. Gerontechnology, International journal on the fundamental aspects of technology to serve the aging society, 2012; 10(4): pp 244-247 doi: http://dx.doi.org/10.4017/gt.2012.10.4.005.00

■ **International Conferences** (Review committee and proceedings)

1. Handling Uncertainty for Video Event Recognition. R Romdhane , F Bremond and M Thonnat. The 3rd International Conference on Imaging for Crime Detection and Prevention, London, United Kingdom, 3 December 2009 (ICDP'2009), pp 1-6.
2. A Framework dealing with Uncertainty for Complex Event Recognition. R Romdhane, F Bremond, and M Thonnat. The 7th IEEE International Conference on Advanced Video and Signal-Based Surveillance, Aug. 29 2010-Sept. 1, 2010, Boston, USA, (AVSS'2010), pp 392-399. ISBN: 978-0-7695-4264-5.
3. Probabilistic Recognition of Complex Events. R Romdhane, B Boulay, F Bremond and M Thonnat. 8th International Conference on Computer Vision Systems, Sept 20-22, 2011 (ICVS'2011), Volume 6962, 2011, pp 122-131. ISBN:978-3-642-23967-0.
4. Measurement instrument for assessing functional abilities of elderly people with and without dementia using a video monitoring system. V Joumier, E Mulin, J H Lee, J Piano, A Derreumaux, R David, P Mallea, A Dechamps, P H Robert, R Romdhane, M Thonnat and F Bremond. 6th International Conference on Bioinformatics and Biomedical Engineering (ICBBE). mai 2012, pp 1-4 Shanghai (China).
5. Activity Recognition and Uncertain Knowledge in Video Scenes. R Romdhane, C-J Crispim, F Bremond and M Thonnat. 10th International Conference on Advanced Video and Signal-Based Surveillance, Aug 27-30 2013 (AVSS'2013).
6. Evaluation of a Monitoring System for Event Recognition of Older People. C-J Crispim, V Bathrinayanan, B Fosty, R Romdhane, A Konig, M Thonnat and F Bremond. 10th International Conference on Advanced Video and Signal-Based Surveillance, Aug 27-30 2013 (AVSS'2013).

- **International congress:** (Abstract with review committee)

    1. Clinical assessment using information and communication technologies in Alzheimer disease: interest for clinical trials. P Robert, R Romdhane, F Bremond and E Mulin. The third Conference Clinical Trials on Alzheimer's Disease, November 3-5, 2010. Toulouse France (CTAD'2010), published in a special supplement of the the Journal of Nutrition Health and Aging (JNHA' 2010), Vol 14, suppl 2, 2010.
    2. Interet des gerontechnologies dans la mesure de l'apathie chez les patients atteints de maladie d'Alzheimer. E Mulin, R Romdhane, R David, J Lee, N Zouba, J Piano, A Derremaux, F Bremond, P H Robert. 9eme Congres International Francophone de Geriatrie et Gerontologie (CIFGG 2010), France (Nice) October 2010.

- **International Workshop:**

    1. Video Activity Recognition Framework for assessing motor behavioural disorders in Alzheimer Disease Patients. V Joumier, R Romdhane, F Bremond, M Thonnat, E Mulin, PH Robert, A Derreumeaux, J Piano and L Lee. International Workshop on Behaviour Analysis and Video Understanding, joint to the 8th International Conference on Computer Vision Systems, pp 1-9, Sept 20-22, 2011.

    1. Combining Multiple Sensors for Event Recognition of Older People. Ca F CRISPIM-JUNIOR, B Fosty, Q Ma, R Romdhane, F Bremond and M Thonnat. 1st ACM MM Workshop on Multimedia Indexing and information Retrieval for Healthcare. October 22, 2013, Spain.

A

PRÉSENTATION DES TRAVAUX DE THÈSE EN FRANCAIS

La recherche sur la reconnaissance de l'activité reçoit une attention croissante de la communauté scientifique aujourd'hui. Il est l'un des problèmes les plus difficiles en vision par ordinateur et de la recherche en intelligence artificielle. L'objectif principal de la recherche actuelle de reconnaissance de l'activité consiste à reconnaitre des activités complexes temporelles de court terme et de long terme. Les progrès dans le traitement des données vidéos bas niveau telles que la détection de mouvements, classification et suivi d'objets ont permis de mettre l'accent sur l'analyse de la reconnaissance de l'activité de niveau supérieur [Vu et al., 2003a], [Ryoo and Aggarwal, 2009], [Brendel et al., 2011], [Kwak et al., 2011]. Les travaux actuels s'intérèssent à la gestion de l'incertitude pour faire face à la demande de robustesse requise par les applications récentes [Tran and Davis, 2008], [Ryoo and Aggarwal, 2010], [Lavee et al., 2010a]. Il ya plusieurs domaines d'application intéressants pour les systèmes de reconnaissance d'activités, principalement la vidéo-surveillance et le domaine de la santé. Les systèmes de surveillance automatiques dans les lieux publics comme les aéroports et stations de métro nècessitent la détection des activités anormales et suspectes. La surveillance dans le domaine de santé consiste à surveiller les activités d'une personne ou des personnes âgées grâce à des caméras et de détecter automatiquement les premiers symptômes de certaines maladies. Il est bien connu

que même les changements subtils dans le comportement de la personne âgée peut donner des signes importants de la progression de certaines maladies. Les perturbations du sommeil qui pourraient être causés, par exemple, par une insuffisance cardiaque et les maladies chroniques. Les changements dans la démarche, d'autre part, peut être associée à des signes précoces de troubles neurologiques liés á plusieurs types de démences. Ces exemples mettent en évidence l'importance de l'observation en continu des changements de comportement chez les personnes âgées afin de détecter la détérioration de la santé avant qu'elle ne devienne critique. Ainsi, un système permettant d'analyser le changement des comportements des personnes âgées dans leurs activités est plus que nécessaire. Bien sûr, pour qu'un tel système soit efficace, la tâche de reconnaissance de l'activité doit fournir des résultats très précis. En effet, si les activités sont mal reconnus, le système de surveillance peut tirer des conclusions erronées sur l'adhésion réelle du patient pour les prescriptions du praticiens, ainsi que fournir des statistiques sujettes à l'erreur sur l'état de santé du patient.
Les sections suivantes décrivent les motivations, les objectifs de cette thèse, le contexte de l'étude, mes hypothèses, mes contributions et la description des chapitres de la thèse.

A.1 Motivations

Ce travail a été grandement motivé par les études réalisées dans la compréhension de l'activité humaine. Au cours des dernières années, beaucoup d'efforts ont été mis dans le développement de systèmes pour la reconnaissance de l'activité et les employant dans une variété de domaines de surveillance.

Une question centrale est la question de la robustesse. La robustesse est définie comme la mesure dans laquelle un système ou un composant peut fonctionner correctement en présence d'entrées invalides ou des conditions d'environnement stressantes. La plupart des systèmes de reconnaîssance d'activités qui ont été construits ont été limitées dans la variété des activités qu'ils reconnaissent et / ou dans la gestion de l'incertitude de la reconnaissance.

Dans ce travail, nous proposons une approche pour la reconnaissance d'activité vidéo qui

aborde ces questions en combinant la modélisation sémantique avec un raisonnement probabiliste pour faire face (i) aux erreurs des détecteurs de bas niveau (au niveau d'événements primitif) et (ii) de gérer l'incertitude de la reconnaissance des événements haut niveau pour soutenir les demandes du domaine de santé. Notre approche vise à fournir plusieurs services pour les patients atteints de la maladie d'Alzheimer afin de les aider à conserver leur indépendance et de vivre en toute sécurité.

A.2 Objectifs

Le principal objectif de cette thèse est de proposer une nouvelle méthode de reconnaissance d'activité en mesure de gérer les erreurs de détecteurs de bas niveau et l'incertitude de reconnaissance d'événements de haut niveau pour détecter des activités intéressantes dans le domaine de la surveillance de la santé. Nous nous concentrons sur la représentation sémantique des événements vidéos et sur la reconnaissance des activités de la vie quotidienne pour construire un système automatique de surveillance vidéo capable d'aider les cliniciens à détecter les premiers symptômes de la maladie d'Alzheimer.

A.3 Contexte de l'étude

La recherche dans le domaine de la surveillance des personnes âgées atteintes de démence est un axe de recherche très populaire de nos jours. La maladie d'Alzheimer (AD) et les troubles connexes représente un défi majeur pour les systèmes de soins de santé avec le vieillissement des populations. La maladie d'Alzheimer est associé à des changements neuro-dégénératifs qui compromettent les capacités cognitives et fonctionnelles et peut entrainer des symptômes comportementaux et neuro-psychiatriques. Beaucoup d'efforts sont actuellement déployés pour analyser la pathologie d'Alzheimer et développer des stratégies de traitement appropriées. Ces stratégies mettent l'accent sur la préservation des capacités cognitives et fonctionnelles et le maintien de la qualité de vie chez le malade. Les échelles mèdicales d'évaluation sont des outils essentiels pour la diagnostic de la maladie d'Alzheimer, l'évaluation et le suivi at-

tentif des symptômes, ainsi que l'évaluation des effets du traitement. Toutefois, ces échelles d'évaluation standard ne rendent pas pleinement compte de la complexité d'une maladie. En effet, la maladie d'Alzheimer comprend les détériorations cognitive, comportementale et fonctionnelle qui ne progrèssent pas en parallèle et peut changer selon l'individualité d'un patient donné [Romdhane et al., 2011]. Pour cette raison, le suivi automatique des activités de la vie quotidienne comme préparer à manger, utilier le téléphone et la gestion des médicaments a été un centre d'intêret en gérontechnologie.

La détection de ces activités permettrait aux systèmes de suivi et de reconnaître les changements dans les habitudes de comportementd qui pourraient être indicateurs de développer des problèmes de santé physique ou mentale. De même, il pourrait aider à déterminer le niveau d'indépendance des personnes âgées. S'il est possible de développer des systèmes qui reconnaissent ces activités automatiquement, les experts médicaux peuvent être en mesure de détecter les changements dans les habitudes de comportement des personnes qui indiquent un déclin de la santé. C'est un domaine difficile pour de multiples raisons. Tout d'abord, pour les cliniciens, il est important d'établir le type exact d'indicateurs qui sont cliniquement pertinente et peut fournir des informations qui peuvent être utilisés dans la pratique quotidienne. Deuxièmement, dans le domaine de la vision par ordinateur, le défi est d'adapter les contraintes techniques et les besoins du clinicien. Pour les patients et les soignants, participer d'une manière active et en donnant un avis sur la faisabilité et la tolérance de l'étude est important.

Le principal moteur de cette étude provient de l'objectif sociétal d'assister et de maintenir les personnes âgées dans leur milieu familial, ou pour leur permettre de 'vieillir sur place'. Plus précisément, l'objectif global est de développer une approche de reconnaissance automatique des activités pour une évaluation comportementale et pour adapter les soins préventifs au début ou à un stade modérée de la maladie d'Alzheimer.

Ce travail de thèse a été menée dans l'équipe STARS à l'INRIA Sophia Antipolis, en France. Stars est une équipe pluri-disciplinaire à la frontière de la vision par ordinateur, l'intelligence artificielle et le génie logiciel. Les travaux de STARS se concentrent sur deux domaines d'application principaux: la sécurité / sûreté et la surveillance dans le domaine de la santé. Ce travail se situe

dans ce contexte et vise à reconnaître les activités humaines pour les applications de la santé. Dans cette étude, nous collaborons avec des cliniciens de l'hôpital de Nice pour déterminer les scénarios d'activités à reconnaître qui sont les plus importants à surveiller pour la maladie d'Alzheimer.

A.4 Contributions

Les principales contributions de ce travail sont les suivants:

- **La première contribution principale consiste en une nouvelle approche de reconnaissance des activités combinant représentation sémantique de l'événement et inférence probabiliste pour gérer l'incertitude.** (chapitre 5) Nous avons utilisé un cadre générique pour représenter la sémantique liées à des événements (par exemple, des modèles d'événements, des informations contextuelles). Ensuite, nous avons combiné une méthode sémantique pour modéliser des activités avec un raisonnement probabiliste basé sur deux couches (une première couche des événements primitifs et une couche des événements composites) pour effectuer une reconnaissance probabiliste compte tenu de l'incertitude des analyses de bas niveau. L'incertitude est gérée en combinant la sortie probabiliste des deux couches. Bien que notre approche est démontrée dans le domaine de la surveillance vidéo de la santé, elle n'est pas limité à un domaine ou une application spécifique.

- **Base de connaissances pour la surveillance des personnes âgées**. (chapitre 4) Nous avons travaillé en étroite collaboration avec les cliniciens pour définir une ontologie et une base de connaissances composée d'un total de 117 modèles d'événements pour la surveillance des patients atteints de la maladie d'Alzheimer à l'hôpital. L'ontologie définie contient plusieurs concepts utiles pour le domaine de la surveillance médicale. Nous avons également défini un certain nombre de critères visuels qui pourraient être observées par les capteurs de caméra pour permettre la détection des premiers symptômes de la maladie d'Alzheimer.

- **Evaluation des troubles du comportement dans la maladie d'Alzheimer**: dans le chapitre 6, nous avons étudié la capacité du système de surveillance automatique des activités vidéos proposé afin de détecter des changements d'activité entre les personnes âgées avec et sans démence. Cette étude montre que nous puissions différencier les deux profils des participants (Alzheimer et les patients témoins normaux) basé sur les activités de la vie quotidienne et les paramétres de la marche. Ces résultats mettent en évidence l'utilisation de l'approche proposée comme une plate-forme de soutien pour les cliniciens permettant de mesurer objectivement les performances des patients Alzheimer dans les activités de la vie quotidienne et l'analyse de la démarche. Parmi les avantages du systéme proposé sont la stabilité de ses résultats au fil du temps et sa mesure quantitative de la performance des patients.

- **Une Nouvelle base de données vidéos.** Cette nouvelle base contenant des séquences vidéos qui ont été enregistrés à l'hôpital de Nice. Ces enregistrements consistent en des personnes âgées saines et des personnes âgées atteintes de la maladie d'Alzheimer en train d'effectuer des activités de la vie quotidienne (manger, lire, regarder la télé). La base de données contient plus de 375 séquences vidéos de 125 participants et un total de plus de 125 heures d'enregistrements.

A.5 Plan de travail

Cette thèse est organisé comme suit:

1. **chapitre1** , nous avons introduit la motivation, l'objectif et le contexte de l'étude.

2. **chapitre 2** , nous examinons d'abord les travaux connexes dans le domaine de la reconnaissance d'activité et puis nous décrivons les différentes technologies pour surveiller les activités humaines dans le domaine de santé.

3. **chapitre 3** , avons présenté un aperçu de la démarche de reconnaissance d'activité proposée. Nous donnons une architecture générale de l'approche proposée. Nous décrivons

les entrées, les sorties et les principales sources de connaissance de notre approche. Nous définissons le problème de la reconnaissance d'activité comme un élément clé de la surveillance automatique dans le domaine de la santé.

4. **chapitre 4** , nous avons présenté l'ontologie et la base de connaissances proposées qui contient un total de 117 modèles d'activités. Nous avons travaillé en étroite collaboration avec les cliniciens pour définir une ontologie et une base de connaissances pour la surveillance à l'hôpital de patients atteints de la maladie d'Alzheimer. La définition de cette ontologie nous permet de représenter les connaissances sur les activités d'intérêt. Nous avons également défini un certain nombre de critères médicaux qui peuvent être observés par les caméras pour permettre la détection des premiers symptômes de la maladie d'Alzheimer.

5. **chapitre 5** , nous avons présenté l'approche proposée pour la reconnaissance des événements primitifs et des événements temporels composites. Il s'agit d'une nouvelle approche combinant la représentation sémantique et le raisonnement probabiliste pour la détection d'événements. Nous avons adopté la théorie des probabilités de Bayes pour calculer la probabilité conditionelle de la reconnaissance des activités. Nous avons adopté égalemet un raisonnement probabiliste pour résoudre et vérifier les constraintes spatiales et temporelles des modèles d'évènements. Nous avons également abordé la gestion de bruits des détecteurs bas niveau.

6. **chapitre 6** , nous évaluons l'approche proposée et nous avons testé les modèles d'activités proposées dans les scénarios médicales avec des vidéos réalisés dans un site expérimental situé à l'hôpital de Nice. Nous présentons les résultats obtenus de la performance de notre algorithme ainsi que les résultats médicaux.

7. Enfin, dans **chapitre 7**, nous concluons ce travail, en résumant les contributions de cette thèse, et en présentant des perspectives à court terme et à long terme.

BIBLIOGRAPHY

[Achard et al., 2008] Achard, C., Qu, X., Mokhber, A. and Milgram, M. (2008). A novel approach for recognition of human actions with semi-global features. In Machine Vision and Applications vol. 19, pp. 27–34,. 18, 28

[Albanese et al., 2008] Albanese, M., Chellappa, R., Moscato, V., Picariello, A., Subrahmanian, V. S., Turaga, P. and Udrea, O. (2008). A constrained probabilistic petri net framework for human activity detection in video. In IEEE Trans. on Multimedia vol. 10, pp. 1429–1443,. 32, 36, 166

[Allen, 1983] Allen, J. F. (1983). Maintaining knowledge about temporal intervals. In Communications of the ACM vol. 11, pp. 832–843,. xii, 20, 25, 55, 63, 120

[Allen and Ferguson, 1994] Allen, J. F. and Ferguson, G. (1994). Actions and events in interval temporal logic. In Journal of Logic and Computation vol. 4, pp. 531–579,. 20

[Alwan et al., 2006] Alwan, M., J.Rajendran, P., Kell, S., Mack, D., Dalal, S., Wolfe, M. and Felder, R. (2006). A smart and passive floor-vibration based fall detector for elderly. In In Proc. Information and Communication Technologies (ICTTA) vol. 1, pp. 1003–1007,. xii, 38, 39

[Ascough et al., 2008] Ascough, J., HR, M., JK, R. and MW, S. (2008). Future research challenges for incorporation of uncertainty in environmental and ecological decision-making. In ecological modelling pp. 383–399,. 44

[Avanzi et al., 2005] Avanzi, A., Bremond, F., Tornieri, C. and Thonnat, M. (2005). Design and assessement of an intelligent activity monitoring platform. In EURASIP Journal on Applied Signal Processing, Special Issue on Advances in Intelligent Vision Systems: Methods and Applications. 48

[Biswas et al., 2010a] Biswas, J., M.Mokhtari, Dong, J. S. and Yap, P. (2010a). Mild dementia care at home - integrating activity monitoring, user interface plasticity and scenario verification. In ICOST'10 Proceedings of the Aging friendly technology for health and independence, and 8th international conference on Smart homes and health telematics pp. 160–170,. 38

[Biswas et al., 2010b] Biswas, J., Sim, K., Huang, W., Tolstikov, A., Aung, A., Jayachandran, M., Foo, V. and Alexandra, P. Y. (2010b). Sensor based micro context for mild dementia assistance. In Proceedings of the 3rd International Conference on PErvasive Technologies Related to Assistive Environments, (PETRA'10). 37

[Biswasa et al., 2011] Biswasa, J., Waia, A. A. P., Tolstikova, A., Kennetha, L. J. H., Maniyeria, J., Victora, F. S. F., Leea, A., Phuaa, C., Jiaqia, Z., Hoaa, H. T., Tiberghienb, T., Alouloub, H. and Mokhtaria., M. (2011). From Context to Micro-context â Issues and Challenges in Sensorizing Smart Spaces for Assistive Living. In ANT2011, the 2nd International Conference on Ambient Systems, Networks and Technologies pp. 288–295,. 38

[Blunsom, 2004] Blunsom, P. (2004). Hidden Markov Models. 16

[Boulay et al., 2006] Boulay, B., Bremond, F. and Thonnat, M. (2006). Applying 3d human model in a posture in a posture recognition system. In Pattern Recognition Letter vol. 27, pp. 1785–1796,. xvi, 143, 144

[Boulay et al., 2007] Boulay, B., Bremond, F. and Thonnat, M. (2007). Human Posture Recognition for Behaviour Understanding. In PhD, ORION, INRIA Sophia Antipolis. xv, 118, 119

[Brendel et al., 2011] Brendel, W., Fern, A. and Todorovic, S. (2011). Probabilistic Event Logic for Interval-Based Event Recognition. In CVPR pp. 3329–3336,. 3, 33, 36, 201

[Buxton and Gong, 1995] Buxton, H. and Gong, S. (1995). Advanced visual surveillance using Bayesian Networks. In In International Conference on Computer Vision. 14

[Cardinaux et al., 2011] Cardinaux, F., Bhowmik, D., Abhayaratne, C. and Hawley, M. (2011). Video Based Technology for Ambient Assisted Living: A review of the literature. In School of Health and Related Research, University of Sheffield, UK. 37

[Chau, 2012] Chau, D. P. (2012). Dynamic and robust object tracking for activity recognition. In PhD thesis. 101, 102, 143

[Chen et al., 2007] Chen, D., Bharucha, A. J. and Wactlar, H. D. (2007). Intelligent video monitoring to improve safety of older persons. In In Proc. IEEE International Conference on Engineering in Medicine and Biology Society (EMBS) pp. 3814–3817,. 40

[Chen and Zhang, 2006] Chen, X. and Zhang, C. (2006). An Interactive Semantic Video Mining and Retrieval Platform Application in Transportation Surveillance Video for Incident Detection. In Sixth International Conference on Data Mining (ICDM 06) pp. 129–138,. xi, 10, 11, 12

[Chen and Zhang, 2007] Chen, X. and Zhang, C. (2007). Interactive mining and semantic retrieval of video. In 8th international workshop on Multimedia data mining (MDM 07) pp. 1–9,. 10, 11, 28, 43

[Chiang et al., 2007] Chiang, C. L., Lien, C. C. and Lee, C. H. (2007). Scene-based event detection for baseball videos. In Journal of Visual Communication and Image Representation pp. 1–14,. 18

[Chleq and Thonnat, 1996] Chleq, N. and Thonnat, M. (1996). Realtime image sequence interpretation for video-surveilance applications. In International conference on Image Processing (ICIP'96) vol. 2, pp. 801–804,. 23

[Clarkson et al., 1998] Clarkson, B., Sawhney, N. and Pentland, A. (1998). Auditory context awareness via wearable computing. In Proceedings of the Perceptual User Interfaces Workshop (PUI). 37

[Crispim-Junior et al., 2012] Crispim-Junior, C. F., Joumier, V., Hsu, Y.-L., Pai, M.-C., Chung, P.-C., Dechamps, A., Robert, P. and Bremond, F. (2012). Alzheimer's patient activity assessment using different sensors. In Gerontechnology. xviii, 1, 187, 188, 191

[Cummings, 1997] Cummings, J. L. (1997). The Neuropsychiatric Inventory: Assessing psychopathology in dementia patients. In Neurology. 64

[Cuntoor et al., 2005] Cuntoor, N., Yegnanarayana, B. and Chellappa, V. (2005). Interpretation of state sequences in hmm for activity representation. In Proc. of IEEE Int. Conf. Acoustics, Speech and Signal Processing vol. 2, pp. 709–712,. 18, 28

[Darwiche, 2000] Darwiche, A. (2000). Constant-Space Reasoning in Dynamic Bayesian Network. In University of California, Computer Science Departement, Los Angelos. 16

[Davis and Shet, 2005] Davis, L. and Shet, V. D. (2005). Vidmap: Video monitoring of activity with prolog. In In Proceedings of Advanced Video and Signal-Based Surveillance (AVSS). 20

[Dechter et al., 1991] Dechter, R., Meiri, I. and Pearl, J. (1991). Temporal constraint networks. In Artificial Intelligence - Special issue on knowledge representation archive vol. 49, pp. 61–95,. xi, 21, 23

[Delaitre et al., 2012] Delaitre, V., Fouhey, D., Laptev, I., Sivic, J., Gupta, A. and Efros, A. A. (2012). Scene semantics from long-term observation of people. In ECCV. 47, 93

[Demiris et al., 2001] Demiris, G., Speedie, S. and Finkelstein, S. (2001). Change of patients perception of telehomecare. In Telemedicine Journal and e-Health vol. 7, pp. 241–248,. 41

[Dousson and Ghallab, 1993] Dousson, P. G. and Ghallab, M. (1993). Situation recognition: Representation and algorithms. In 13th International Joint Conference on Artificial Intelligence (IJCAI). 20

[Duong et al., 2005] Duong, T., Bui, H., Phung, D. and Venkatesh, S. (2005). Activity recognition and abnormality detection with the switching hidden semi-markov model. In CVPR. xi, 18, 19

[Esposito et al., 2010] Esposito, F., L, R., der Linden AC, V., F, L., A, Q., A, C. and der Linden M., V. (2010). Apathy and executive dysfunction in Alzheimer disease. In Alzheimer Dis Assoc Disord pp. 131–137,. 176

[Foroughi et al., 2008] Foroughi, H., Naseri, A., Saberi, A. and Yazdi, H. S. (2008). An eigenspace-based approach for human fall detection using integrated time motion image and neural network. In International Conference on Signal Processing (ICSP) pp. 1499–1503,. 38

[Fouhey et al., 2012] Fouhey, D. F., Delaitre, V., Gupta, A., Efros, A., Laptev, I. and Sivic., J. (2012). People Watching: Human Actions as a Cue for Single-View Geometry. In ECCV. 47

[Funtowicz and Ravetz, 1990] Funtowicz, S. and Ravetz, J. (1990). Uncertainty and Quality in Science for Policy. In kluwer Academic Publishers, Dordrecht. 44

[Ghallab, 1996] Ghallab, M. (1996). Representation, on-line recognition and learning. In 5th International Conference on Principales of knowledge Representation and Resonning, (KR'96). 20, 28

[Ghanem et al., 2004] Ghanem, N., DeMenthon, D., Doermann, D. and Davis, L. (2004). Representation and recognition of events in surveillance video using Petri nets. In In IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW). 26, 28, 43

[Gillain et al., 2009] Gillain, S., Warzee, E., Lekeu, F., Wojtasik, V., Maquet, D., Croisier, J.-L., Salmon, E. and Petermans, J. (2009). The value of instrumental gait analysis in elderly healthy, MCI or Alzheimer's disease subjects and a comparison with other clinical tests used in single and dual-task conditions. In Annals of Physical and Rehabilitation Medicine vol. 52, pp. 453–474,. 188

[Haas, 2002] Haas, P. (2002). Stochastic Petri NetsâModelling, Stability, Simulation. In in: Springer Series in Operations Research and Financial Engineering. xii, 26

[Hongeng and Nevatia, 2001] Hongeng, S. and Nevatia, R. (2001). Multi-agent event recognition. In International Conference on Computer Vision (ICCV'01). 14

[Howell and Buxton, 2001] Howell, A. J. and Buxton, H. (2001). Time delay RBF networks for attentional frames in Visually Mediated Interaction. In Neural Processing Letters. 10, 11, 28

[Ivanov et al., 2005] Ivanov, Y., Bobick, A. and Mihailidis (2005). Recognition of visual activities interactions by stochastic parsing. In EEE Trans. Patt. Anal. Mach. Intel vol. 1, pp. 838–845,. 20

[Joumier et al., 2011] Joumier, V., R., R., F., B., M., T., E., M., PH., R., A., D., J., P. and J., L. (2011). Video Activity Recognition Framework for assessing motor behavioural disorders in Alzheimer Disease Patients. In ICVS workshop. 177, 178

[Kim et al., 2010] Kim, S., Schap, T., Bosch, M., Maciejewski, R., Delp, E., Ebert, D. and Boushey, C. (2010). Development of a mobile user interface for image-based dietary assessment. In 9th International Conference on Mobile and Ubiquitous Multimedia. xii, 37, 38

[Kumar et al., 2005] Kumar, P., Ranganath, S., Weimin, H. and Sengupta, K. (2005). Framework for real-time behavior interpretation from traffic video. In IEEE Trans. on Intelligent Transportation Systems vol. 6, pp. 43–53,. 14

[Kwak et al., 2011] Kwak, S., Han, B. and Han, J. H. (2011). Scenario-Based Video Event Recognition by Constraint Flow. In CVPR pp. 3345–3352,. 3, 34, 36, 201

[Lavee et al., 2007] Lavee, G., Borzin, A., Rivlin, E. and Rudzsky, M. (2007). Building Petri Nets from Video Event Ontologies. In ISVC07 pp. 442–451,. 26, 28

[Lavee et al., 2010a] Lavee, G., Michael, R. and Ehud, R. (2010a). Propagating Uncertainty in Petri Nets for activity Recognition. In ISVC. xii, xx, 3, 32, 33, 36, 166, 167, 169, 201

[Lavee et al., 2010b] Lavee, G., Rudzsky, M., Rivlin, E. and Borzin, A. (2010b). Video event modeling and recognition in generalized stochastic petri nets. In IEEE Trans. on Circuits and Systems for Video Technology pp. 102–118,. 32

[Laxton et al., 2007] Laxton, B., Lim, J. and Kriegman, D. (2007). Leveraging temporal,contextual and ordering constraints for recognizing complex activities in video. In CVPR pp. 1–8,. 15, 16

[Leray, 2006] Leray, P. (2006). RÃ©seaux bayÃ©siens : apprentissage et modÃ©lisation de systÃ¨mes complexes. In Habilitatin a diriger les travaux de reherche, UniversitÃ© de Rouen. 99

[Londei et al., 2009] Londei, S. T., Rousseau, J., Ducharme, F., St-Arnaud, A., Meunier, J., Saint-Arnaud, J. and Giroux, F. (2009). An intelligent videomonitoring system for fall detection at home: perceptions of elderly people. In Journal of Telemed Telecare vol. 15, pp. 383–390,. 41

[Lv et al., 2006] Lv, F., Song, X., Wu, V., Kumar, B. and Nevatia, R. (2006). Left luggage detection using bayesian inference. In Proc. of IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance pp. 83–90,. 15

[Marquis-Faulkes et al., 2005] Marquis-Faulkes, F., McKenna, S. J., Newell, A. F. and Gregor, P. (2005). Gathering the requirements for a fall monitor using drama and video with older people. In Journal Technology and Disability vol. 17, pp. 227–236,. 41

[Minnen et al., 2003] Minnen, D., Essa, I. and Starner, T. (2003). Expectation grammars: Leveraging high-level expectations for activity recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 626–632,. 28

[Mouhoub and Liu, 2008] Mouhoub, M. and Liu, J. (2008). Managing Uncertain Temporal Relations using a Probabilistic Interval Algebra. In Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on pp. 3399–3404,. 120

[Nait-Charif and McKenna, 2004] Nait-Charif, H. and McKenna, S. J. (2004). Activity summarisation and fall detection in a supportive home environment. In IEEE International Conference on Pattern Recognition (ICPR) vol. 4, pp. 323–326,. 37

[Nevatia et al., 2004] Nevatia, R., Hongeng, S. and Bremond, F. (2004). Video-based event recognition : activity representation and probabilistic recognition methods. In Computer Vision and Image Understanding vol. 2, pp. 129–162,. xi, 14, 15, 28, 43

[Nevatia et al., 2003] Nevatia, R., Zhao, T. and Hongeng, S. (2003). Hierarchical language-based representation of events in video streams. In In IEEE Workshop on Event Mining. 20, 27

[NGHIEM, 2010] NGHIEM, A.-T. (2010). Adaptive algorithms for background estimation to detect moving objects in videos. In PhD thesis. 141

[Nicholson and Korb, 2006] Nicholson, A. E. and Korb, K. B. (2006). Bayesian AI Tutorial. In Faculty of Information Technology Monash University Clayton AUSTRALIA. xi, 15, 16

[Nist, 2000] Nist (2000). Uncertainty of Measurement Results. In http://physics.nist.gov/cuu/Uncertainty/index.html. 125

[Nordin et al., 2006] Nordin, E., Rosendahl, E. and Lundin-Olsson, L. (2006). Timed ”Up and Go” test: reliability in older people dependent in activities of daily living-focus on cognitive state. In PhysTher pp. 646–655,. 176

[Oliver et al., 2002] Oliver, N., Horvitz, E. and Garg (2002). Layered representations for human activity recognition. In IEEE International Conference on Multimodal Interfaces (ICMI) pp. 3–8,. 18, 27, 43

[Olshausen, 2004] Olshausen, B. A. (2004). Bayesian probability theory. 97

[Park and Aggarwal, 2004] Park, S. and Aggarwal, J. (2004). A Hierarchical Bayesian Network for Event Recognition of Human Actions and Interactions. In pecial issue on Video Surveillance vol. 2, pp. 164–179,. 14

[Paulos, 2011] Paulos, J. A. (2011). The Mathematics of Changing Your Mind. In The New york times. 98

[Perse et al., 2010] Perse, M., Kristan, M., Pers, J., Music, G., Vuckovic, G. and Kovacic, S. (2010). Analysis of multi-agent activity using petri nets. In Pattern Recognition pp. 1491–1501,. 26

[Petridis and Psomas, 2012] Petridis, S. and Psomas, C. (2012). Allen’s Hourglass: Probabilistic Treatment of Interval Relations. In 24th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 12 pp. 87–94,. 121

[Pinhanez and Bobick, 1998] Pinhanez, C. S. and Bobick, A. F. (1998). Human action detection using PNF propagation of temporal constraints. In IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 20

[Pusiol, 2012] Pusiol, G. (2012). Discovery of human activities in video. In PhD manuscrit. 197

[Reddy et al., 2003] Reddy, S., Gal, Y. and Shieber, S. (2003). Recognition of group activities using dynamic probabilistic networks. In The 9th International Conference on Computer Vision. 15, 27

[Reddy et al., 2009] Reddy, S., Gal, Y. and Shieber, S. (2009). Recognition of Usersâ Activities Using Constraint Satisfaction. In Springer Berlin / Heidelberg vol. 5535, pp. 838–845,. 20

[Ries, 2009] Ries, J. (2009). Test-retest reliability and minimal detectable change scores for the timed ”up and go” test, the six-minute walk test, and gait speed in people with Alzheimer disease. In Phys Ther pp. 569–579,. 176

[Romdhane et al., 2011] Romdhane, R., Mulin, E., Derreumeaux, A., Zouba, N., J.Piano2, L.Lee2, I.Leroi, MallÃ©a, P., David, R., M.Thonnat, F.Bremond and Robert., P. H. (2011). Automatic video monitoring system for assessment of Alzheimer's disease symptoms. In Journal of Nutrition, Health and ging, (JNHA2011). 5, 204

[Rosario et al., 2000] Rosario, B., Oliver, N. M. and Pentland, A. (2000). A Bayesian computer vision system for modelling human interactions. In IEEE transactions on Pattern Analysis and Machine Intelligence (PAMI) pp. 831–843,. 18

[Rota and Thonnat, 2000] Rota, N. and Thonnat, M. (2000). Activity recognition from video sequences using declarative models. In 14th European Conference on Artificial Intelligence (ECAI). xi, 20, 21, 22, 28, 93

[Ryoo and Aggarwal, 2006] Ryoo, M. S. and Aggarwal, J. K. (2006). Recognition of composite human activities through context-free grammar based representation. In CVPR pp. 1709–1718,. 27, 28

[Ryoo and Aggarwal, 2007] Ryoo, M. S. and Aggarwal, J. K. (2007). Hierarchical Recognition of Human Activities Interacting with Objects. In CVPR. 28

[Ryoo and Aggarwal, 2009] Ryoo, M. S. and Aggarwal, J. K. (2009). Semantic representation and recognition of continued and recursive human activities. In International Journal of Computer Vision (IJCV) pp. 1–24,. xii, 3, 28, 29, 30, 36, 94, 201

[Ryoo and Aggarwal, 2010] Ryoo, M. S. and Aggarwal, J. K. (2010). Stochastic Representation and Recognition of High-Level Group Activities. In International journal of computer Vision. 3, 28, 35, 43, 47, 93, 201

[SanMiguel and Martinez, 2012] SanMiguel, J. C. and Martinez, J. M. (2012). A semantic-based probabilistic approach for real-time video event recognition. In CVIU. 28, 94

[Sergios et al., 2010] Sergios, P., Paliouras, G. and Perantonis, S. J. (2010). Allen's Hourglass: Probabilistic Treatment of Interval Relations. In Temporal Representation and Reasoning (TIME), 2010 17th International Symposium on. 120

[Shi et al., 2006] Shi, Y., Bobick, A. and Essa, I. (2006). Learning temporal sequence model from partially labeled data. In CVPR. 16

[Shi et al., 2004] Shi, Y., Huang, Y., Minnen, D., Bobick, A. and Essa, I. (2004). Propagation networks for recognition of partially ordered sequential action. In CVPR pp. 1631–1638,. 16

[Sidenbladh and Black, 2001] Sidenbladh, H. and Black, M. (2001). Learning image statistics for bayesian tracking. In IEEE International Conference on Computer Vision (ICCV). 37

[Sigel et al., 2010] Sigel, K., Bernd, K. and Claudia, P.-W. (2010). Conceptualising uncertainty in environmental decision-making: The example of the EU water framework directive. In ecological Economics pp. 502–510,. 44

[Siskind, 2001] Siskind, J. M. (2001). Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. In Journal of Artificial Intelligence Research (JAIR) vol. 15, pp. 31–90,. 20

[Squire, 2004] Squire, D. M. (2004). Tutorial: The Naive Bayes Classifier. In Faculty of Information Technology, Monash university. 14

[Sullivan and LaMorte, 2013] Sullivan, L. and LaMorte, W. W. (2013). The Role of Probability. In Boston University School of Public Health. 114

[Taylor and Kuyatt, 1994] Taylor, B. N. and Kuyatt, C. E. (1994). Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results. In United States Department of Commerce Technology Administration National Institute of Standards and Technology. 125

[Thonnat and Rota, 1999] Thonnat, M. and Rota, N. (1999). Image Understanding for Visual Surveillance Applications. In Third International Workshop on Cooperative Distributed Vision, Invited paper pp. 51–82,. 20, 25

[Tolstikov et al., 2008] Tolstikov, A., Biswas, J., Tham, C. and Yap, P. (2008). Eating Activity Primitives Detection - a Step Toward ADL Recognition. In HealthCom. 39

[Town, 2006] Town, L. (2006). Ontological inference for image and video analysis. In Machine Vision and Applications pp. 94–115,. 15

[Tran and Davis, 2008] Tran, S. and Davis, L. S. (2008). Event modeling and recognition using Markov logic networks. In European Conference on Computer Vision, ECCV 08. xii, 3, 30, 31, 35, 36, 59, 94, 201

[Vincent et al., 2002] Vincent, C., Drouin, G. and Routhier, F. (2002). Examination of new environmental control applications. In Assistive Technology Journal vol. 14, pp. 98–111,. 41

[Vu et al., 2003a] Vu, T., Bremond, F. and Thonnat, M. (2003a). A Novel Algorithm for Temporal Scenario Recognition. In The Eighteenth International Joint Conference on Artificial Intelligence. xix, 3, 6, 23, 24, 28, 43, 49, 51, 62, 76, 91, 93, 159, 161, 162, 163, 166, 169, 201

[Vu et al., 2003b] Vu, T., Bremond, F. and Thonnat, M. (2003b). Automatic Video Interpretation: A Recognition Algorithm for Temporal Scenarios Based on Pre-compiled Scenario Models. In International Conference on Computer Vision Systems pp. 523–533,. 24, 61

[Vu et al., 2004] Vu, T., Bremond, F. and Thonnat, M. (2004). Temporal Scenario for Automatic Video Interpretation. In Phd. xii, 25, 105

[Vu et al., 2002] Vu, V., Bremond, F. and Thonnat, M. (2002). Temporal Constraints for Video Interpretation. In 15th European Conference on Artificial Intelligence ECAI. xii, 23, 24

[Walker et al., 2003] Walker, W., P, H., J, R., der Sluijs J, V., M, V. A., P, J. and von Krauss M, K. (2003). Defining Uncertainty A Conceptual Basis for Uncertainty Management in Model-Based Decision Support. In Integrated Assessment vol. 4, pp. 5–17,. 44

[Wang et al., 2007] Wang, S., Pentney, W., Popescu, A., Choudhury, T. and Philipose, M. (2007). Common sense joint training of human activity recognizers. In IJCAI. 37

[Warmink et al., 2010] Warmink, J., JAEB, J., MJ, B. and MS, K. (2010). Identification and classification of uncertainties in the application of environmental models. In Environmental Modelling and Software vol. 25, pp. 1518–1527,. 44, 45

[Wells, 2001] Wells, T. (2001). Neural Networks Tutorial:The Single Layer Perceptron and Multilayer Perceptron. xi, 10, 11

[Zaidenberg et al., 2012] Zaidenberg, S., Boulay, B. and Bremond., F. (2012). A generic framework for video understanding applied to group behavior recognition. In 9th IEEE International Conference On Advanced Video and Signal Based Surveillance (AVSS 12). 195

[Zaidi, 1999] Zaidi, A. K. (1999). On temporal logic programming using Petri nets. In IEEE Transactions on Systems, Man and Cybernetics pp. 245–254,. 26

[Zhang, 2008] Zhang, L. (2008). Cumulative Distribution Functions. In Applied Statistics I. 124

[Zouba, 2010] Zouba, N. (2010). Multisensor Fusion or Monitoring Elderly Activities at Home. In PhD manuscrit. xx, 153, 159, 165, 166, 169

[Zouba et al., 2009] Zouba, N., Bremond, F., Thonnat, M., Anfonso, A., Pascual, E., Mallea, P., Mailland, V. and Guerin, O. (2009). A Computer system to monitor older adults at home: preliminary results. In the international journal Gerontechnology vol. 8,. 37, 62, 65, 91

RESUMÉ

Cette thèse aborde le problème de la reconnaissance d' activités. Elle est fortement motivée par la recherche dans le domaine de la reconnaissance des activités vidéo appliquée au domaine de la surveillance de personnes agées. Dans ce travail, nous proposons deux contributions principales. La première contribution consiste en une approche pour la reconnaissance d'activitResumé vidéo avec gestion de l'incertitude pour une détection précise d'événements. La deuxième contribution consiste à définir une ontologie et une base de connaissances pour la surveillance dans le domaine de la santé et en particulier la surveillance à l'hôpital de patients atteints de la maladie d'Alzheimer. L'approche de reconnaissance d'activité proposée combine une modélisation sémantique avec un raisonnement probabiliste pour faire face aux erreurs des détecteurs de bas niveau et pour gérer l'incertitude de la reconnaissance d'activité. La reconnaissance probabiliste des activités est basée sur la théorie des probabilités bayésienne qui fournit un cadre cohérent pour traiter les connaissances incertaines. L'approche proposée pour la vérification probabiliste des contraintes spatiale et temporelle des activités est basée sur le modèle de probabilité gaussienne. Nous avons travaillé en étroite collaboration avec les cliniciens pour définir une ontologie et une base de connaissances pour la surveillance à l'hôpital de patients atteints de la maladie d'Alzheimer. L'ontologie définie contient plusieurs concepts utiles dans le domaine de la santé. Nous avons également défini un certain nombre de critères qui peuvent être observés par les caméras pour permettre la détection des premiers symptômes de la maladie d'Alzheimer. Nous avons validé l'algorithme proposé sur des vidéos réelles. Les résultats expérimentaux montrent que l'algorithme de reconnaissance d'activité proposé a réussi à reconnaître les activités avec un taux élevé de reconnaissance. Les résultats obtenus pour la surveillance de patients atteints de la maladie d'Alzheimer mettent en évidence les avantages de l'utilisation de l'approche proposée comme une plate-forme de soutien pour les cliniciens pour mesurer objectivement les performances des patients et obtenir une évaluation quantifiable des activités de la vie quotidienne.

Mots-clés: reconnaissance d'événement vidéo, modélisation d'événements, incertitude, probabilité bayésienne, densité de probabilité gaussienne, probabilité cumulative, surveillance des personnes âgées, activités de la vie quotidienne, maladie d'Alzheimer.

ABSTRACT

This work deals with the problem of human activity recognition. It is greatly motivated by research in video activity understanding applied to the domain of health care monitoring. Research on activity recognition is receiving an increasing attention from the scientific community today. It is one of the most challenging problem in computer vision and artificial intelligence domains. The main goal of the current activity recognition research consists in recognizing and understanding short-term action and long-term complex activities. In this work, we propose two main contributions. The first contribution consists of an approach for video activity recognition that addresses the uncertainty management issues for accurate event detection. The second contribution consists in defining an ontology and a knowledge base for health care monitoring and in particular Alzheimer monitoring at hospital. The proposed activity recognition approach combines semantic modelling together with a probabilistic reasoning to cope with the errors of low-level detectors and to handle activity recognition uncertainty. The probabilistic recognition of activities is based on Bayesian probability theory which provides a consistent framework for dealing with uncertain knowledge. The proposed probabilistic constraint verification approach based on Gaussian probability model enforces the accuracy of the activity recognition algorithm. We work in close collaboration with clinicians to define an ontology and a knowledge base for Alzheimer monitoring at hospital. The defined ontology contains several concepts useful for health care. We also define a number of criteria which could be observed by camera sensors to allow detection of early symptoms of Alzheimer's disease. We validate the proposed algorithm on real-world videos. The experimental results show that the proposed activity recognition algorithm can successfully recognize activities with a high recognition rate. The obtained results for health care monitoring highlight the advantages of the use of

the proposed approach as a support platform for clinicians to objectively measure patient performance and obtain a quantifiable assessment of instrumental activities of daily living and gait analysis.

Keywords: video event recognition, event modeling, uncertainty, Bayesian probability, Gaussian probability density, cumulative probability, health care monitoring, activities of daily living, Alzheimer disease.