UNIVERSITÀ DEGLI STUDI
DI MILANO

SCUOLA DI DOTTORATO IN INFORMATICA

Tesi di Scuola di Dottorato

Semantic Analysis and Understanding of Human Behaviour in Video Streaming

Alberto Amato

Relatore: Prof. Vincenzo Piuri
Correlatore: Prof. Vincenzo Di Lecce

Coordinatore del Dottorato di Ricerca: Prof. Ernesto Damiani

Anno Accademico 2009/2010
UNIVERSITÀ DEGLI STUDI DI MILANO

Abstract

This thesis investigates the semantic analysis of the human behaviour captured by video streaming, both from the theoretical and technological points of view. The video analysis based on the semantic content is in fact still an open issue for the computer vision research community, especially when real-time analysis of complex scenes is concerned.

Automated video analysis can be described and performed at different abstraction levels, from the pixel analysis up to the human behaviour understanding. Similarly, the organisation of computer vision systems is often hierarchical with low-level image processing techniques feeding into tracking algorithms and, then, into higher level scene analysis and/or behaviour analysis modules. Each level of this hierarchy has its open issues, among which the main ones are:

- motion and object detection: dynamic background modelling, ghosts, suddenly changes in illumination conditions;
- object tracking: modelling and estimating the dynamics of moving objects, presence of occlusions;
- human behaviour identification: human behaviour patterns are characterized by ambiguity, inconsistency and time-variance.

Researchers proposed various approaches which partially address some aspects of the above issues from the perspective of the semantic analysis and understanding of the video streaming. Many progresses were achieved, but usually not in a comprehensive way and often without reference to the actual operating situations. A popular class of approaches has been devised to enhance the quality of the semantic analysis by exploiting some background knowledge about scene and/or the human behaviour, thus narrowing the huge variety of possible behavioural patterns by focusing on a specific narrow domain.

In general, the main drawback of the existing approaches to semantic analysis of the human behaviour, even in narrow domains, is inefficiency due to the high computational complexity related to the complex models representing the dynamics of the moving objects and the patterns of the human behaviours.

In this perspective this thesis explores an innovative, original approach to human behaviour analysis and understanding by using the syntactical symbolic analysis of images and video streaming described by means of strings of symbols. A symbol is associated to each area of the analysed scene. When a moving object enters an area, the corresponding symbol is appended to the string describing the motion. This approach allows for characterizing the motion of a moving object with a word composed by symbols. By studying and classifying these words we can categorize and understand the various behaviours. The main advantage of this approach consists in the simplicity of the scene and motion descriptions so that the behaviour analysis will have limited computational complexity due to the intrinsic nature both of the representations and the related operations used to manipulate them. Besides, the structure of the representations is well suited for possible parallel processing, thus allowing for speeding up the analysis when appropriate hardware architectures are used.

The theoretical background, the original theoretical results underlying this approach, the human behaviour analysis methodology, the possible implementations, and the related performance are presented and discussed in the thesis. To show the effectiveness of the proposed approach, a demonstrative system has been implemented and applied to a real indoor environment with valuable results. Furthermore, this thesis proposes an innovative method to improve the overall performance of the object tracking algorithm. This method is based on using two cameras to record the same scene from different point of view without introducing any constraint on cameras' position. The image fusion task is performed by solving the correspondence problem only for few relevant points. This approach reduces the problem of partial occlusions in crowded scenes. Since this method works at a level lower than that of semantic analysis, it can be applied also in other systems for human behaviour analysis and it can be seen as an optional method to improve the semantic analysis (because it reduces the problem of partial occlusions).

Dedication

To women of my life:
my wife Antonella for her love,
my daughter Giulia for her sweetness
my mother (it is not necessary to say because ;-)
and to my mother-in-law for her excellent cuisine.

Acknowledgements

I think that there are two types of world. A material world, that is our planet and a specific little world for each human being. The latter is composed of all the elements with which people interact during their life (people, places, and so on).

So, I would like to express my gratitude to my little world because it has made me what I am today.

Contents

Chapter 1

This thesis proposes a new methodology to automatically analyze human behaviour in narrow domains using video streaming. This chapter begins discussing the main technological innovations that are creating the premise to implement such kind of systems. Next, the principal application fields of these systems are briefly presented and then the main issues that the scientific community are facing to realize them are presented. Finally the proposed methodology is briefly presented and the remaining part of the thesis is outlined.

Introduction

Aim of this thesis is to investigate, both from a theoretical and a technological point of view, the problem of human behaviour analysis in video streaming. From a psychological point of view, the general concept of "behaviour" is the reaction of human beings to a set of external and internal impulses. These impulses may be caused by very different sources, and only the combination of them can represent an objective explanation of the observed behaviour.

In the latest years, information and communication technology (ICT) has had a strong improvement having significant influence on our every day life. The effects of the technological improvements in the fields of sensor manufacturing and communication networks are particularly relevant. For what concerns the latter, nowadays is the "Internet Age". Millions of computers are connected among them, sharing data and hosting services for a large number of users living everywhere in the world. Furthermore, the network is becoming pervasive. Indeed, thanks to the spread of wireless networks it is possible to be "connected" everywhere also by means of devices which are different from the traditional PC (notebook, netbook, smartphone, etc.).

For what concerns sensors, in these days there is a fast evolution toward smaller sensors and with increasing performance, e.g., in terms of precision, accuracy, and reliability. This has led to the development of a large number of applications. For example, nowadays the major part of videos and pictures are taken by using digital devices, while sensors and

actuators based on the Micro Electro-Mechanical Systems (MEMS) technology are the key elements of many applications for the modern smartphones.

These factors have also enabled and booted the development of sensor networks that are able to monitor wide areas where human beings perform their activities. Studying data recorded by these sensor networks we can develop technologies and systems for high semantic level analysis of the human behaviour [1]. The aim of these technologies consists of defining a description of the human behaviour that can be used in recognition tasks. The analysis from a psychological point of view of the reasons that have determined a given behaviour is beyond the scope of these technologies.

In the latest years, automatic human behaviour analysis has attracted the interest of the international scientific community. This strong interest is due to the many potential applications that such kind of systems can have. Essentially, these applications can be divided into three macro areas:

- *Surveillance*. This application area has a relevant interest also due to the facts characterizing the history of the last decade. In particular, the aim of the applications in this area is monitoring and understanding human behaviour in public and crowded areas (such as streets, bus and train stations, airports, shopping malls, sport arenas, and museums). These applications can be used both for management (for example: evaluation of crowding in the monitored areas, human flow analysis, and detection of the congestion points) and security (for example: human behaviour analysis, activity recognition of individuals and groups, and detection of suspicious persons) [2].
- *Control*. In this area the researchers try to develop systems that are able to evaluate some human motion parameters (i.e. speed, gait) and/or the poses (i.e. the mutual position of the harms, or the position of the head) to control actions and/or operations [3]. One of the most important fields of application for these systems is the human-machine interaction. On the market there are some interesting uses as input devices for videogames. These devices use both video and inertial sensors to evaluate the player's motion and understand his/her commands.
- *Analysis*. Applications of human motion analysis are used in various fields, including: sports (as an aid for evaluating the techniques and the performance of the

athletes), and medicine (as an aid in the diagnosis of problems in human postures and in the orthopaedic rehabilitation) [4].

The economic and social relevance of potential applications (especially the security, entertainment, and medical ones), the scientific complexity, the speed and price of current hardware intensified the effort within the scientific community towards automatic capture and analysis of human motion.

In the literature, various approaches have been proposed to capture the human behaviour. Most of them use video sensors since these technologies are not invasive and rather cheap, as well as they produce good-quality data suited for being processed by means of inference techniques mimicking the human ones. Some authors proposed alternative approaches by using other kind of sensors, such as inertial sensors, audio sensors, presence detectors; unfortunately, these sensors (often installed in devices as "wearable sensors") are characterised by some invasiveness and, therefore, can be used only in some specific control applications [5, 6]. Consequently, for these reasons, this thesis addresses the human behaviour analysis by using video streaming as the most appropriate technology to observe the human behaviour.

Automatic understanding of the human behaviour from video sequences is a very challenging problem since it implies understanding, identifying, and either mimicking the neuro-physiological and psychological processes, which are naturally performed in humans or creating similar outcomes by means of appropriate information and knowledge processing. In order to achieve this goal the problem has been split into two steps:

1. A compact representation of the real world is first defined by using the data sampled by cameras. This representation should be as close as possible to the reality, view invariant, and reliable for subsequent processing. A video streaming contains a large amount of data, but often they are redundant and/or useless for the human behaviour analysis (for example: the static data about the background scene is not helpful). Therefore, it is necessary to track the areas where a difference between two successive video frames has been detected in order to focus the attention on the areas in which there are moving objects (may be humans), while discarding the background with irrelevant information. Later these moving entities will be identified as human beings or unanimated objects (or animals). Moving entities will be traced through the various frames to characterize their movements.

2. By starting from the representation defined in the first step, the visual information will be interpreted for recognizing and learning the human behaviour. Various techniques have been proposed in the literature to define appropriate behaviour reference models. In these approaches, recognition will consist in comparing an unknown input to the models by using suitable distance functions. The nearest model to the input is considered as the class to which the observed behaviour belongs. The main limits of this approach are:
    a. there is a finite number of recognizable behaviours (as many as the number of the models defined before using the system);
    b. a large number of training sequences are usually required in order to define a model in a sufficiently accurate and recognizable way;
    c. it is not possible to automatically associate a meaning to an observed behaviour;
    d. it is not possible to generalize the learnt models and infer new behaviours from the learnt ones.

    Therefore, the efforts of the scientific community are focussed on studying systems in which the knowledge of the considered behaviour reference models is built incrementally, i.e., by starting from an empty set of models, new behaviour models are, first, identified by observing people's activities and considering the current behaviour knowledge and, then, added to the current behaviour knowledge. When models have been learnt, the system will be able to recognize them and similar behaviours in the future.

Nowadays, both of the above steps are open research fields. At this time, there is not a comprehensive and universally valid method to obtain the representation of the real-world behaviours. Even the most suitable type of sensors to be used is still under discussion. This is basically due to the lack of a general solution to two important problems: the sensory gap and the semantic gap.

The sensory gap is the difference between the real world and its representation obtained by using data sampled by sensors [7]. For example, by using a camera, a bi-dimensional representation of the real world can be obtained, while our eyes give us a three-dimensional model. To address this problem researchers are proposing various data fusion

techniques based, for example, on two or more cameras, or on hybrid vision systems (cameras plus other kinds of sensor).

The semantic gap is the difference between the behaviour description used by human beings and the computational model used by the human behaviour analysis systems [8]. To solve this problem or at least to reduce the effects of this gap, researchers have been working on exploiting some knowledge about scene and/or the human behaviour, thus narrowing the huge variety of possible behavioural patterns by focusing on a specific narrow domain.

This thesis addresses therefore the analysis and understanding of the human behaviour by considering the various aspects, issues and implications depicted above, both from the theoretical perspective and from the point of view of technological supports. The aim is to provide solid foundations for implementing efficient and effective human behaviour analysis and understanding systems, suitable for a variety of applications encompassing, e.g., video surveillance, medical applications, human-machine interface, and entertainment.

In the above perspectives, this thesis gives two highly-significant, innovative, and original contributions to the human behaviour analysis and understanding in video streaming:

1. a new method to perform images fusion in multi-camera systems, by identifying the corresponding points in the various images without the assumption of epipolar geometry. This assumption is fundamental for many works on this topic but it imposes strong constraints on the camera positions (an in depth presentation of this problem is proposed in the chapter 4). The proposed method does not introduce this constraint. This fact makes this method applicable to many existing multi-camera acquisition systems. Using this method it is possible to mitigate some aspects of the sensory gap, like the target partial occlusions in crowding scenes. From this point of view, it can be seen as a method to improve the performance of the semantic analysis because it allows to improve the performance of the tracking algorithm (a step in the processing chain for human behaviour analysis). By the other hands, it should be stressed the fact that the kernel of the thesis is the semantic analysis of

the human behaviour in video streaming and that the proposed approach can work also without this module and so, also using a single camera system.

2. a new approach to analysis and understanding by using the syntactical symbolic analysis of images and video streaming described by means of strings of symbols. This allows for high simplicity in the scene and motion descriptions so that the behaviour analysis will have limited computational complexity, thanks to the intrinsic nature both of the representations and the related operations used to manipulate them. On the other hand, this approach has a great flexibility. Indeed, it allows for performing a hierarchical analysis of the recorded scenes defining a different grammar for each level. At the higher level, the behaviour of the human beings moving in the scene are analysed studying their motion parameters (the trajectory that they are following). In this way it is possible to have an analysis of scenario considering the moving objects and their interactions. At the lowest level, the system can produce a detailed analysis of the action taken by each single human being in the scene. According to the complexity of the task, between these two levels, it is possible to define as many intermediate levels as they are necessary. This hierarchical analysis can exploit the full potentiality of the modern video surveillance systems where there is a fixed camera of scenario and one or more moving cameras that can focus their attention on some areas of interest. In this application, the first level of the proposed hierarchy is applied to the camera of scenario and the second to the moving cameras.

To show the effectiveness of the proposed approach and technology, a demonstrative system has been implemented and applied to some real indoor environments with significant results.

The proposed approach is not tailored on a single specific operating environment, but has a high flexibility and is able to deal with a large number of applications, thus being of unique value as a new fundamental enabling technology. This has been possible since the proposed approach works at a high abstraction level on the semantic analysis of the human behaviours, also exploiting the advantages offered by the image fusion in multi-camera systems (which are becoming increasingly popular due to their decreasing costs). For example, in surveillance the proposed approach enables the creating of innovative systems which are able to evaluate the danger level of observed situations from a semantic point of view, thus significantly enhancing the correctness and completeness of alarms. In control

applications it supports the implementation of adaptive, advanced human-machine interfaces since the motion of human beings can be analysed in a fully automatic way. In analysis applications, the proposed approach can be used as an advanced enabling technology for implementing systems, e.g., for sport actions evaluation and automatic video indexing.

This thesis reports the accomplished research and, in particular, the theoretical foundations, the technology, the innovative approach, the experiments, and the achieved results. This thesis is structured as follows.

- Chapter 2 presents a brief overview of the sensors which can be used for human behaviour analysis.
- Chapter 3 reviews the current state of the art in sensor data processing, image fusion, and human behaviour analysis and understanding.
- Chapter 4 describes the proposed method for using multi-camera system in semantic analysis applications.
- Chapter 5 presents the method used for semantic analysis of human behaviour.
- The experimental evaluation and the obtained results are presented in Chapter 6.
- Conclusions and final remarks are reported in Chapter 7.

Chapter 2
Sensors for Human Behaviour Analysis

This chapter proposes a brief description of the most common sensors used in literature to perform the automatic analysis of the human behaviour. For each kind of sensor, at least a scientific work using it for human behaviour analysis is presented. This chapter helps the readers to understand because the author has focused his attention on the analysis of human behaviour using video streaming.

2.1 Motivation

Automatic human behaviour analysis and recognition are complex tasks that have attracted a lot of researchers in the latest years. One of the primary tasks to perform implementing such analysis is to define a representation of the real world using the data sampled by some sensors.

Nowadays the most frequently used sensors are camera devices, but in literature there are also approaches based on the employment of other kinds of sensor. These approaches achieve good results in some specific domains but often they are not generalizable for other contexts.

In this chapter a brief overview about the following sensors and of the works using them is presented: Radio Frequencies Identifier (RFID), pressure sensors, Micro-Electro-Mechanical Systems (MEMS) and image sensors.

The aim of this brief overview is to show the most important used sensors in human behaviour analysis, their applications and their limits. This discussion should help the readers to understand because the author has focused his attention on the analysis of human behaviour using video streaming. The main goal of the thesis is the semantic analysis of video streaming. From this point of view, since the sensors play a secondary role in this dissertation, a critical overview of the literature about sensors is beyond the scope of this thesis.

2.2 Radio Frequencies Identifier Technology

RFID is the acronym of Radio Frequencies Identifier. This technology is based on four key elements: the RFID tags themselves, the RFID readers, the antennas and choice of radio characteristics, and the computer network (if any) that is used to connect the readers.

RFID tags are devices composed of an antenna and a small silicon chip containing a radio receiver, a radio modulator for sending a response back to the reader, control logic, some amount of memory, and a power system. A RFID tag transmits the data stored inside its memory module when it is exposed to radio waves of the correct frequency sent by the reader.

According to the used power system there are two kinds of tags: passive tags where the power system can be completely powered by the incoming RF signal and active tags where the tag's power system has a battery. Passive tags are cheaper than active tags but they have a shorter range of action (the reader must be positioned very close to the tag).

Figure 1 shows a diagram of the power system for a passive inductively coupled transponder-RFID tag. An inductively coupled transponder comprises an electronic data-carrying device, usually a single microchip, and a large area coil that functions as an antenna. The reader's antenna coil generates a strong, high frequency electromagnetic field, which penetrates the cross-section of the coil area and the area around the coil. The antenna coil of the transponder and the capacitor C_1 form a resonant circuit tuned to the transmission frequency of the reader. The voltage U at the transponder coil reaches a maximum due to resonance step-up in the parallel resonant circuit. The layout of the two coils can also be interpreted as a transformer (transformer coupling), in which case there is only a very weak coupling between the two windings.

The simplest RFID chips contain only a serial number. This serial number is written into the chip by the manufacturer but there are also tags where this code can be written by the end user. An example of this code is the EPC (Electronic Product Code) that is a number composed of 96 bits. This number is the kernel of an international standard making RFID technology a pillar element of the international logistic chain. This standard is under the oversight of EPCglobal Inc™ a not-for-profit joint venture between GS1 (formerly EAN International) and GS1 US (formerly the Uniform Code Council).

More sophisticated RFID chips can contain read-write memory that can be programmed by a reader.

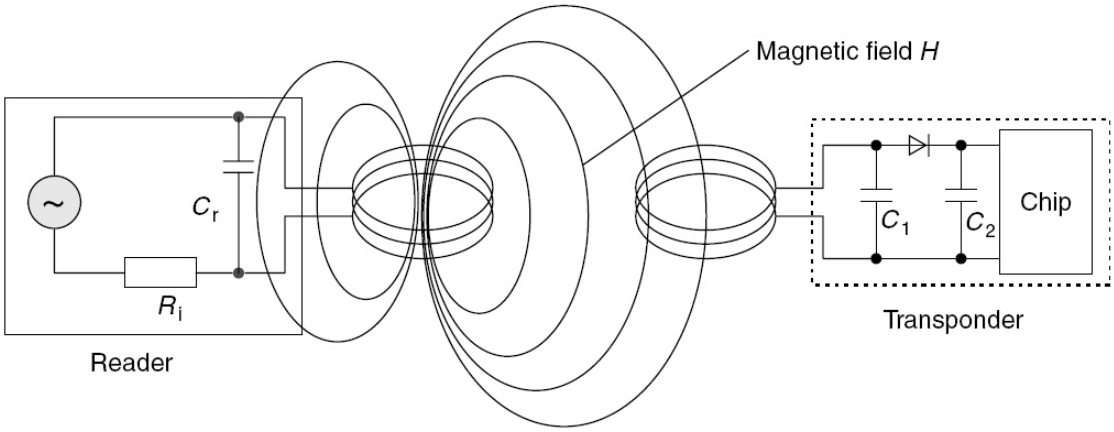


Figure 1 - A block diagram of a RFID tag

Among the various applications where RFID technology is used, there is also the human activity detection and monitoring, an example of such applications is [9]. In this work the authors propose a system to detect the activity of daily living performed at home. This is a challenging task for various reasons such as: each one can perform the same action in different ways, there are many possible activities that a system should model with minimum human effort, etc. The key observation done by the authors is that the sequence of objects that a person uses performing a given activity is a good marker both of the action type and quality. Starting from this observation they propose a system composed of three modules: specialized sensors to detect object interactions, a probabilistic engine that infers activities given observations from sensors, and a model creator to create probabilistic models of activities. In this work the used sensors are RFID tags attached on each object of interest. The RFID reader is built inside a glove that the user should wear while performing his activities of daily living. The system was tested on 14 predefined activities and the obtained results showed good performance both in terms of precision (88%) and recall (73%).

Despite these good results the system presents various limits. From a technological point of view, water and metal absorb the radio waves that most RFID tags use; metal can also short-circuit the tag antenna. This fact limits the number of correctly observable

actions. But the most important limitation of this approach is that it is too invasive. Indeed it requires that the user must wear a glove. This can be a serious problem due to the fact that many people are not too attracted by using gloves while performing activities of daily living.

2.3 Pressure sensors

Pressure transducers are very common and cheap. They are used in various applications and they work using various principles (variation of capacity, variation of resistance, piezoelectric, etc.).

In [10] the authors propose a human behaviour recognition system using a set of pressure sensors based on the variation of resistance. For this kind of sensors the conversion of pressure into an electrical signal is achieved by the physical deformation of strain gages which are bonded into the diaphragm of the pressure transducer and wired into a Wheatstone bridge configuration. Pressure applied to the pressure transducer produces a deflection of the diaphragm which introduces strain to the gages. The strain will produce an electrical resistance change proportional to the pressure.

Figure 2 shows a block diagram of the whole pressure measurement system proposed in [10]. The transducer is connected to a microcontroller that is able to communicate with a standard PC using a Bluetooth link. A serious constrain for this system is its power supply module that is composed of a 9V alkaline battery. This fact introduces the well known limitations due to the lifecycle of the battery, maintenance, etc.

The main idea at the base of this work is that measuring the plantar pressure distribution it is possible to infer information about the actions performed by the user. Four sensors were installed in each shoe. According to the authors, it is possible to classify fifteen different behaviours: walking (slow, normal, fast), running (slow, normal, fast) standing (leaning forward, load on tiptoe, upright), leaning standing to one foot (leaning forward, normal), sitting (bending forward, normal), floating, and no wearing. The parameters used to classify the various actions are pressure values and length of time where a given pressure is measured. These parameters are compared to a fixed set of thresholds to identify the various actions.

According to the authors, the system achieves good classification rate (about 90% of successfully classifications) but the experiment settings are not well described. The main limits of this approach are the necessity of a calibration stage for each person to define the various thresholds and its invasivity (indeed the measuring system is installed in the shoes and it is quite visible). These considerations make the system not suitable for applications in every day life situations.

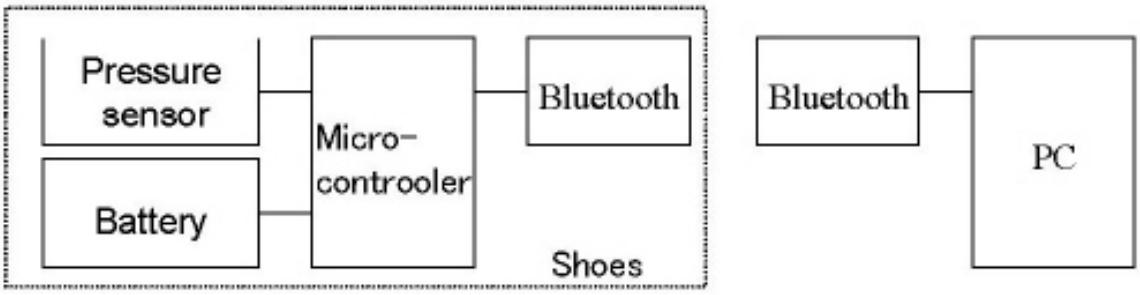


Figure 2 - A block diagram of the pressure sensor proposed in [10]

2.4 (Micro Electro-Mechanical Systems sensors

Micro-Electro-Mechanical Systems (MEMS) are devices built by means of the integration of mechanical elements, sensors, actuators, and electronics on a common silicon substrate through micro-fabrication technology. The current leaders in commercially successful MEMS technology are accelerometers. These devices are used in a large number of applications such as: automotive industry, inertial navigation systems, cellular phones, etc.

The physical mechanisms underlying MEMS accelerometers include capacitive, electromagnetic, piezoelectric, ferroelectric, optical, etc. The most successful types are based on capacitive transduction; the reasons are the simplicity of the sensor element itself, no requirement for exotic materials, low power consumption, and good stability over temperature.

These sensors are used in many applications involving the so-called “wearable-sensors”. For example, in [11] the authors propose a system to classify the human pose in “sitting”, “standing” and “walking” using a bi-axial accelerometer attached to the user’s thigh. The sensor is positioned in order to measure the gravity acceleration using the Y-axis when the

user is in “standing” position and the X-axis when she/he is in “sitting” position. The “walking” action is detected when the variance of acceleration is greater than a predefined threshold.

[12] presents a method to detect physical activities from data acquired using five small biaxial accelerometers worn simultaneously on different parts of the body. In particular the sensors are placed on each subject’s right hip, dominant wrist, non-dominant upper arm, dominant ankle, and non-dominant thigh to recognize ambulation, posture, and other everyday activities.

The experiments carried-out aim at identifying twenty different actions of every day life (see [12] for further details). The system was tested on twenty subjects from the academic community volunteered. Data was collected from 13 males and 7 females. Data were classified using various methods but decision tree classifiers showed the best performance recognizing everyday activities with an overall accuracy rate of 84%. Interestingly, the obtained results show that some activities are recognized well with subject-independent training data while others appear to require subject-specific training data.

2.5 Image sensors

This kind of sensors are used in all the cameras and camcorders used to create still images and video streaming. The kernel of this kind of sensors consists of an array of tiny pixels (PIcture ELements). Sensor pixels are composed of photodiodes.

During the imaging process, the light starts to fall on photodiodes, and they convert photons into electric charge. Photodiodes are not sensible to colour, so digital cameras use different colour filters to transmit light through. The most common ones are filters for three basic colours: red, green and blue. So, the camera is able to calculate the number of photons of three basic colours that fell on each photodiode while the camera shutter was open. To calculate all colour components around each photodiode, red, green and blue filters should be situated adjacently. This makes it possible to convert raw image data into a full-colour image in RGB (Red Green Blue) space.

The most common type of colour filter array is called a "Bayer array". This filter gives priority to the green mimicking the behaviour of human eyes. A schematic overview of the imaging process is shown in Figure 3. The light is filtered by an infra red filter (Figure

3.a). The filtered light goes through a Bayer array filter and finally hits the pixels of the sensor.

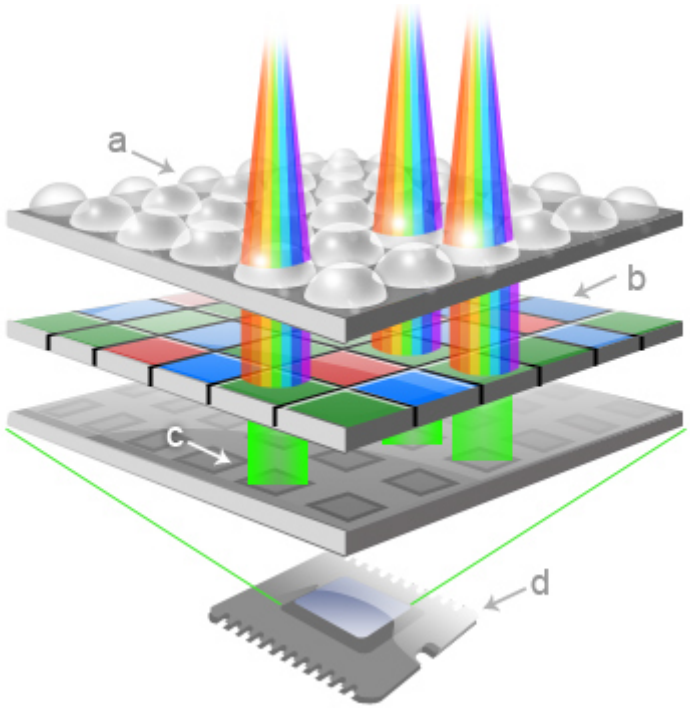


Figure 3 – A schematic view of an image sensor. a – IR-Blocking Filter, b – Colour Filters, c – Colour blind sensors, d – The image sensor composed of millions of light sensors [13]

Nowadays, the image sensors are built using two technologies: CCD (Charge-Coupled Device) and CMOS (Complementary Metal–Oxide–Semiconductor). Both the technologies use photodiodes to convert light in electrons, but they use two different methods to read these values from each pixel of the sensor. In a CCD device, the charge is actually transported across the chip and read at one corner of the array. A special manufacturing process is used to create the ability of transporting charges across the chip without distortion. This process leads to very high-quality sensors in terms of fidelity and light sensitivity. In most CMOS devices, there are several transistors at each pixel that amplify and move the charge using more traditional wires. The CMOS approach is more flexible because each pixel can be read individually. CCD sensors give better images than CMOS sensors but the latter are cheaper then the former. Furthermore, in the last years the quality difference between the images sampled by these family of sensors are becoming smaller indeed, ever more often CMOS sensors are used in good quality cameras.

The major part of the works present in literature on human behaviour analyse video streaming sampled using these image sensors. The next chapter proposes a review of such works.

Summary

This chapter has presented a brief and not exhaustive overview of the sensors used in some applications of human behaviour analysis. Some sensors have been omitted because they are often used in conjunction with video analysis by means of data fusion techniques. Two relevant examples of such sensors are: audio and multi-spectral sensors.

Audio sensors are becoming ever more interesting due to progress in speech recognition. In literature it is possible to find their applications for the recognition of specific human behaviours. For example, in [14] a system to detect aggressions in trains is proposed while in [15] a method for action detection in action movies is described. Furthermore, audio-vision data are used to detect human emotions as shown in [16] where the system is able to detect 4 cognitive states (interest, boredom, frustration and puzzlement) and 7 prototypical emotions (neural, happiness, sadness, anger, disgust, fear and surprise).

Multi-spectral sensors are used in many remote sensing applications. In human behaviour analysis there is a strong interest to infrared sensors because they can operate in total darkness allowing for the person detection during the night. In [17] a data fusion approach working on infrared images and classical CCD camera images is presented.

The sensors presented in this chapter are used in systems called wearable sensors. These kind of systems are intrusive (because the user are required to wear the sensors) and so they can be applied only in some restricted applications (such as human-machine interface applications).

Multi-spectral sensors can give excellent results but they are too expensive to be used in real world applications.

Audio sensors are cheap but they are not able to give good results in terms of human activity detection without using some data fusion techniques with video streaming. But these systems have a high computational cost due to the complexity of the used algorithm.

For these reasons, in this thesis, only approaches based on video streaming analysis will be considered.

Chapter 3

Related Works

This chapter proposes an overview of the literature about the automatic human behaviour analysis using streaming video. The discussion starts presenting the processing chain used to implement such kind of systems and then the attention is focused only on the works aiming at semantic analysis. The works presented in this chapter are classified in: scene interpretation, human recognition and action primitives and grammars. For each class a brief introductive description is provided and some relevant works are analyzed to give an idea of the proposed approaches and of the difficulties that they face. Finally, the chapter ends with a discussion about the state of the art in this field and with a brief overview of the specific challenges that this thesis is addressing.

3.1 Introduction

Even though in literature there are many works on human behaviour analysis and recognition, this is still an open research field. This is due to the inherent complexity of such task. Indeed, human behaviour recognition can be seen as the vertex of a computational pyramid as shown in Figure 4. Each level of this pyramid takes in input the output of the lower one and gives an output that can be used as input for the upper level or as a stand alone application. The lowest level takes in input the raw video streams and gives in output a map of the image region where a moving object is detected. Climbing up this pyramid, the semantic level of performed tasks grows up. The processes at the lowest level work with **moving region** in a single frame while those at the second level work identifying **objects** in the same frame. At the third level a new parameter plays an important role: the **time**. Indeed, the processes at this level work associating the detected moving objects in the current frame with those in the previous one, providing temporal **trajectories** through the state space. The output of this level is sent to the human behaviour analysis module.

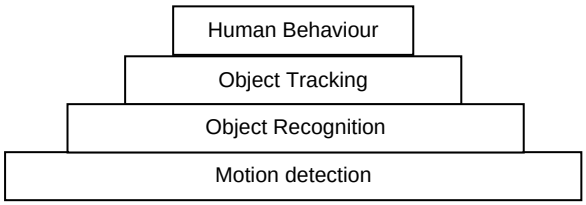


Figure 4 - A hierarchical overview of the computational chain for human behaviour recognition

Each level of this processing pyramid has its own characteristics and difficulties often due the partial completeness of the used data (for example: the tentative to extract 3D data about moving objects working on 2D images):

- **Motion detection**: These algorithms find the moving areas computing the difference at pixel level between the current frame and a background model. This model can be a fixed frame (useful for indoor applications) or a complex model where each pixel is defined by a Gaussian probability distribution [18]. A good background model should be enough robust to handle rapid illumination changes in the scene and at the same time should be able to recognize permanent changes (for example a chair that was moved from its previous position) as quick as possible.
- **Object detection**: These algorithms identify moving objects analysing the moving areas detected by the previous module. Typically, they classify the pixels in the moving areas in: shadows, moving object, cast shadow from moving object, ghost object (false positive), and ghost shadow. Some examples of criterions used to classify the pixels are colour and gradients [19] and motion flow analysis [20].
- **Object tracking**: Difficulties in tracking objects can arise due to abrupt object motion, changing appearance patterns of the object and/or the scene, object-to-object and object-to-scene occlusions, etc. An interesting survey of the approaches proposed in literature to perform this task is [21].
- **Human behaviour analysis and recognition**: This is one of the most challenging tasks for the international researchers' community. Here the problem of the semantic gap must be faced to obtain satisfying results. Furthermore, the errors done at the earlier stages have strong effects here (for example, the object tracking

module can lose a person due to an occlusion while this person is performing an interesting act for the behaviour analysis application).

Since this thesis focuses on human behaviour analysis and recognition, this chapter overviews the approaches about this problem proposed in literature. In various works, terms like actions, activities, complex or simple actions, and behaviours are often used in an undifferentiated way. This fact makes their comparison and/or classification difficult. In this thesis the following action hierarchy is used: action primitives, actions, and activities. At the highest level of the hierarchy there are the activities (for example dancing), each activity is composed of a set of actions (for example: pirouette, etc.) and each action is in turn divided into action primitives (the set of sequential elementary steps to perform to do a pirouette). The works presented in this chapter are classified according to the following visual abstraction hierarchy:

- **scene interpretation**: these works try to interpret the whole image without identifying particular objects or humans;
- **human recognition**: where either the entire human body or individual body parts are identified and used for the recognition task
- **action primitives and grammars**: these works try to obtain a semantic description of the scene starting from the identification of an action hierarchy.

3.2 Scene interpretation

The works falling in this framework try to interpret the whole scene. The moving objects are detected and tracked without considering their identity (they can be humans or other moving objects). Typically, these systems have two working stages:

- **Learning stage**: where the observed trajectories of the moving objects are classified using various methods to build the system knowledge base. This stage is not trivial because human motion is highly non-linear, a-priori unknown, and it is always subject to sudden and unforeseeable changes of orientation and speed;
- **Operating stage**: where the knowledge base is used to classify the runtime observed object trajectories.

These systems are used in frameworks where a well defined set of situations are allowed while others are forbidden or in applications where the goal is to classify the scenes/behaviours in usual or unusual. At this level of visual abstraction, the moving

objects are considered in their completeness. This allows for building systems that analyse at a high level the objects trajectories and their interactions.

The main drawbacks of the works falling in this class are:

- They produce a poor semantic level analysis of the scene: indeed, they are only able to state if a given behaviour is usual or unusual or in other words if a scene is known or unknown

- The number of recognizable actions is limited to that of the actions learnt in the learning stage. Furthermore, this stage is not trivial due to issues cited above.

In literature there are many works using this approach.

In [22] an approach is proposed to classify scenes in usual and unusual ones by starting from the analysis of a single frame or object rather than from a sequence of frames. This method is tested in various environments both indoor and outdoor. It is designed to work on huge datasets, indeed, the results presented in the paper are obtained on a database composed of 4 years of continuous video acquisition (that give millions of records). Despite the fact that this work is quite old, it still remains a key work because it represents clearly the potentiality of the methods using scene interpretation.

They use an advanced motion detection method modelling the background as a Gaussian mixture [18]. The detected moving objects are described by means of a feature vector composed of: position, dimension, speed, moving direction and a binary representation of the object silhouette. The feature space is quantized with an online vector quantization method to obtain a codebook (namely a set of prototype representations which approximate the density of the input representations). This method starts selecting K random prototypes among the existing data. Then, each point in the dataset is associated to the nearest prototype that is adapted toward the data point using a learning factor. Some constraints from [23] are applied to avoid the problems that occur when an outlier is selected as initial prototype. Once that the codebook has been generated, it is used in all the successive classification processes instead of the whole original dataset obtaining a strong improvement in the classification performance in terms of processing time.

Since this system works by classifying a single object and or a single frame rather than recognizing sequences, it considers a sequence as a multi-set of symbols. A multi-set is a

set that can contain multiple instances of the same element. Each pair of objects in a sequence belonging to the same multi-set is evidence that those two prototypes' appearances resulted from the same underlying class. The multi-sets of prototypes are used to estimate the co-occurrence statistic over the codebook. All the prototypes and the co-occurrence matrix defined on them are used to define a binary tree structure by recursively defining two probability mass functions across the prototypes of the code book that best explain the co-occurrence matrix. The leaf nodes of the binary tree are probability distributions of co-occurrences across the prototypes and at a higher tree depth define simple scene activities like pedestrian and car movement. These can then be used for scene interpretation. The proposed results are obtained testing the system on a streaming video recording a scene that consists of a road with adjacent parking spots and a path through the grass near the loading bay of a building. The camera is placed far from the scene so that the moving objects appear to be quite small.

The strong points of this method are:

- the **tracking method** is based on a Gaussian mixture model of the background that has been proven to be quite reliable also in outdoor scenes because it is able to model the slight changes in luminosity. The same is not valid for sudden changes due, for example, to the car's headlight in the night. Nevertheless, the system can work also at night because the scene is taken from far and a car with headlight turned-on occupies a small percentage of the frame.

- The **ability to handle huge datasets**. The used concept of "codebook" can be ideally considered as a "summary" of the all recorded scenes. This approach is often applied also in other field of applications such as image database indexing [24] to speed up the search process in complex database. Using this technique is useful also for the successive step of co-occurrence statistic definition and it introduce a strong improvement in the system performance. The key element of this approach is the size of the codebook. A codebook with many elements can be used to classify more actions but the needed data to build the co-occurrence statistic accumulation on it grows-up as the square of its dimension. This point can become a limit to the applicability of this system in frameworks with a great variance in the performable actions.

- **The ability to recognize an action starting from few instances of objects**. In this work, the authors propose an instance classifier rather than a sequence classifier. This result is obtained considering each single observation in a sequence as an independent one both in the stage of codebook generation and in that of co-occurrence statistic definition. Once again, the codebook is a key element. If the codebook elements are able to describe in an unambiguous manner each instance, the system will work but, if the operative framework becomes more complex and the number of elements in the codebook becomes insufficient to describe in an unambiguous manner each instance, the system will not work very fine.

A second class of approaches for detecting anomalies in video sequences and/or in single images has been proposed in [25]. This method tries to decompose the normal (or the allowed) video sequences and/or images in small multi-scale portions defining the knowledge base of the system. When the system analyses a query video and/or image, it tries to rebuild the query using the portions stored in the knowledge base. The regions of the query that can be rebuilt using large portions from the database are considered as "normal" while those that can not be rebuilt or that can be rebuilt using small and fragmented portions are regarded as "suspicious".

The single images are decomposed in a set of spatial multi-scale portions. Each portion is described by means of a feature vector considering its absolute position in the image and a synthetic representation of the gradient of each its pixel. A video sequence is decomposed in spatial-temporal multi-scale portions. The feature vector describing each portion reports its absolute position and the absolute values of the temporal derivatives in all pixels of the portion.

In order to speed-up the image/video rebuilding process and thus the whole interpretation task, the authors propose an inference method based on a probabilistic analysis of the portions allowing for small local misalignments in their relative geometric arrangement. This approach allows for a rapid and efficient detection of subtle but important local changes in behaviour.

The main drawbacks of this method are:

- **Computational complexity**: the examples reported in the work are relative to single images and short video sequences. The inference process is quite efficient,

but the database can become very "heavy" when long video sequences are analysed.

- **The feature vector used to describe the portions**: as the same authors say, the used feature vector is quite simple, but the proposed system is modular and so it could be able to use other features. The computer vision research community has elaborated more sophisticated descriptors.

- **Probabilistic analysis**: this is the kernel of the inference process (because the entire task of similarity computation among portions is based on it) and perhaps the most limitative drawback of this system. This analysis is based on an assumption that is almost never applicable in real situations: there are not overlapping areas among the portions. This assumption is necessary to compute the similarity between pairs of portions, but it makes this system unsuitable in each situation where there are partial occlusions among moving objects.

A third class of approaches is based on the study of the trajectories of the moving objects and focuses the attention both on the geometric characteristics of the trajectories and on their cinematic aspects. The idea standing at the base of these works is that in a given place, similar trajectories can be associated to similar activities. In this way, analysing and classifying the trajectories described by a moving object is equivalent to analyse and classify its activities.

[26] proposes a method to distinguish between objects traversing spatially dissimilar paths, or objects traversing spatially proximal paths but having different spatial-temporal characteristics.

The system uses a fixed camera and a tracker to record the trajectories of the moving objects. A recursive min-cut clustering algorithm is used to group similar trajectories. The distance between two trajectories is computed using the Hausdorff distance. This fact allows the system to compute the distance between two trajectories composed of a different number of points. To limit the spatial extent of a path, an envelop is defined using a dynamic time warping algorithm.

The system performs a hierarchical classification using different features at each level. The first level uses only the geometric properties of the trajectory. If a given trajectory is

not geometrically similar to any in the database, it is considered “unusual”. If a given trajectory is similar to one in the database, the second stage classification is performed. The second level works on the cinematic properties of the trajectories thus evaluating the velocities of the moving objects. The third level analyses the discontinuity in the trajectories. The system analyses the discontinuity in speed, acceleration and curvature. This criterion is able to detect irregular motion patterns that can be associated to particular situation (i.e. a drunk man walking).

From a conceptual point of view, this work has good performances. The problems arise when this system analyses crowded scenes. Indeed, in this situation, the tracker is not able to track all the moving objects (due to the mutual occlusions) and so it is difficult to associate a trajectory to each moving object. Actually, this problem is common almost to all the works in the research field on human behaviour analysis because it is the result of the propagation of an error occurring in a lower level of the computational pyramid shown in Figure 4 (object tracking module).

While the ratio behind the association of semantic meaning to the object trajectories is a common point to all the works falling in this framework (scene interpretation), there are sensible differences in the method used to detect, model and handle the object trajectories.

For example, in [27] the authors use non-rigid shapes and a dynamic model that characterizes the variations in the shape structure. This system is specialized in surveillance task where it is important to discriminate between “normal” and “abnormal” behaviours. The authors propose a method to model the shape of group of simultaneously moving objects that in the surveillance framework means to model the activities of group of people.

They use the Dryden and Mardia’s statistical shape theory [28] to represent the shape of the configuration of a group of moving objects and its deformations over time. In other words, they model a trajectory activity as a mean stationary trajectory plus a set of allowed deformations due to slight difference on the paths followed by the various moving objects and/or to the displacement among the various moving objects following the same path.

This method represents a pragmatic solution to the problem of occlusions in crowded scenes. The ratio behind this solution is: since it is not possible (at least till today and in a

perfect way) to detect and track each moving object in a crowded scene, we try to model and characterize the motion of the whole crowd.

This approach has perfect sense in video surveillance systems applied to particular operative context such as transit areas in airports and stations where it is possible to classify the possible paths in common/allowed paths and reserved/forbidden paths.

In [29] the authors propose a system having the same ratio and application framework of the previous one (video surveillance) but a different way to model the shapes. Here they are described as the composition of basis shapes obtained by applying the factorization theorem [30] to the 3D shape that can be recovered from the motion tracks of points in a 2D image sequence.

A different approach is proposed in [31]. Here, the concept of object trajectory is leaved and an activity is represented as a succession of basic events and modelled through interpretation of the temporal and causal correlations among different classes of events. An event is defined as a group of significant changes in pixels in a local image neighbourhood over the time. The events are detected and classified by unsupervised clustering using Gaussian Mixture Model with automatic model selection based on Schwarz's Bayesian Information Criterion [32]. A robust and holistic scene-level behaviour interpretation method is implemented using Dynamic Probabilistic Networks to model the temporal and causal correlations among discrete events.

This approach is quite different from the others presented in this section. It leaves both the concept of object detection and that of trajectory. In this way the problem of occlusions is not considered because the system tries to model the behaviour of the entire group of moving (and self-occluding) objects. According to the experimental results presented by the authors, this system is suitable to model different situations such as: aircraft cargo loading/unloading activities and shopping activities.

The main drawbacks of this work are:

- It requires a large training set to model complex situations
- Since it does not attempt to track each moving object (trying to resolve the problem of partial occlusions occurring in crowded scene) in a scene, it is not able to

evaluate what happens among the elements of the moving group. For this reason it is not suitable in mission critical security systems.

3.3 Human recognition

The works falling in this class try to infer information about the human beings activities analysing the dynamic of their movements. Some of these works try to recognize and study the motion of the individual body parts while others consider the whole body as a unique element.

The works falling in this class produce a motion analysis with a semantic level richer than that obtained by the works belonging to the scene interpretation class. Indeed, while the latter are able to classify a scene/behaviour as known or unknown, the former works try to identify elementary actions such as walk, running, the human gait but also more complex actions when applied in a narrow domain (for example the actions of a tennis match in [33]).

Also these works are characterized by a learning stage and an operative stage and so, also for these works the number of recognizable actions is limited to that of the actions learnt in the learning stage.

Furthermore, since these works try to analyze the single action with a fine level of detail, the execution speed of the various movements and hence the sampling rate of the camera can become critical parameters. A direct consequence of this fact is the difficulty in delimiting the action in the operating stage.

An example of work considering the whole body as a unique element is [34]. Here the authors consider videos where the human beings are tall about 30 pixels (they call this situation “medium field” while they define “far field” the scenes where human beings are tall about 3 pixels and “near field” those where they are tall 300 pixels). They motivate this choice considering that this is the classical resolution of many videos of sport events where people can easily track individual players and recognize actions such as running, kicking, jumping (despite the small dimension of the players). Each moving person is tracked so that the image is stabilized in the middle of a tracking window. This removes the effects of the absolute motion of the tracked person in the scene (it is equivalent to the panning movement by a camera operator who keeps the moving figure in the centre of the field of

view). In this way, any residual motion within the spatio-temporal volume is due to the relative motions of different body parts: limbs, head, torso etc. The analysis of these residual motions is at the base of this method. Given a stabilized figure-centric sequence, the optical flow is computed and decomposed into two scalar fields corresponding to the horizontal and vertical components of the optical flow. These fields are decomposed into four non negative channels. Each channel is blurred with a Gaussian and normalized to obtain the final descriptor for the image sequence.

According to the authors, this system has a good performance in terms of sequences recognition. It works onto scenes recorded in the "medium field" maximizing the information in the blurred areas due to the residual motion. The critical points can be:

- the system is able to recognize similar sequences also if they are recorded with slight different frame rates, but the system is not able to analyse sequences of different lengths.

- the length of the motion descriptors (i.e., the number of frames used to analyse an action) can be a critical parameter. Indeed, it is a constant value fixed at design time. From this point of view, the system is not able to recognize the same action performed at two different speeds.

A hierarchical approach is proposed in [33] where a system for human behaviour analysis in the narrow domain (tennis match) is presented. Here, a given human behaviour is considered as composed of a stochastic sequence of actions. Actions are described by a feature vector comprising both trajectory information (position and velocity), and a set of local motion descriptors. The used local motion descriptors are an advanced version of those used in [34]. Indeed, the coarse optical flow descriptor used in [34] has been endowed with data about the position where a given action is performed. Action recognition is achieved by means of a probabilistic search method applied to the database representing previously seen actions. The possible actions are modelled by means of Hidden Markov Models (HMM), high-level behaviour recognition is achieved by computing the likelihood that a set of predefined Hidden Markov Models explains the current action sequence. Thus, human actions and behaviour are represented using a hierarchy of abstraction: from simple actions, to actions with spatio-temporal context, to

action sequences and finally general behaviours. This system has been used to produce high semantic level video annotations for tennis matches.

The main drawbacks of this work are:

- the high dimensionality of the feature space: according to the authors, "there are 30000 entries in a single local motion feature vector for a 30 × 50 pixel target". Even thought the authors propose a solution to this problem (a database structured as a binary tree via principal component analysis of the data set), it can become a restriction for using this approach in more complex scenarios;

- this system is not able to generalize knowledge about the recorded scenes. Indeed, it is able to recognize and label a given action using a fixed knowledge base created during the learning stage.

Other approaches are based on the concept of "*temporal templates*" (a static vector-image where the vector value at each point is a function of the motion properties at the corresponding spatial location in an image sequence). This idea was proposed in [35] where the authors used a two components version of the templates: the first value is a binary value indicating the presence of motion and the second value is a function of the recency of motion in a sequence. These components are called MEI (Motion-Energy Image) and MHI (Motion-History Image) respectively. MEIs are cumulative binary motion images, namely, binary images where the value of each pixel is set to 0 if its value does not change for each frame of a given sequence (namely, no moving objects have passed over it) while it is set to 1 otherwise. MHIs are grey scale images where pixel intensity is a function of the temporal history of motion at that point. In these images, the more recently moving pixels are brighter. From a certain point of view, considering a given scene, the MEI image describes *where* the motion occurs while the MHI describes *how* it occurs. Matching temporal templates is based on Hu moments [36].

The main limit of this approach is its dependency by the speed of the action and by the frame rate. Indeed, the same action, executed at different speed or recorded with different frame rates, gives different MHI and MEI.

A variant of the concept of MHI is called *timed MHI* (tMHI) and is presented in [37]. This method tries to overcome the limit of the previous approach due to its dependence by

the frame rate and/or execution time. In order to obtain this result, the authors use the timestamp and introduce a limit into the duration of an action (a constant of few seconds). In this way, a given gesture will cover the same MHI area at different capture rates.

Another approach is that of ‘‘*Actions Sketches*’’ or ‘‘*Space-Time Shapes*’’ in the 3D XYT volume. In [38] the authors propose to model an action based on both the shape and the motion of the object performing the action. When the object performs an action in 3D, the points on the outer boundary of the object are projected as 2D (x, y) contour in the image plane. A sequence of such 2D contours with respect to time generates a spatiotemporal volume (STV) in (x, y, t), which can be treated as 3D object in the (x, y, t) space. The differential geometric surface properties of this 3D object (such as peaks, pits, valleys and ridges) are considered specific action descriptors capturing both spatial and temporal properties. For example, a pit surface is generated when the contour first moves in the direction that is normal to the contour, then stops and moves in the opposite direction.

Instead of using spatio-temporal volumes, a large number of papers choose the more classical approach of considering sequences of silhouettes. For example, in [39], the authors present a method for human motion pattern recognition based on Principal Component Analysis (PCA) and neural networks. They extract the silhouettes of the moving objects in the various frames. Each 2D silhouette contour is converted into a one dimensional signal computing the distance between each point of the silhouette contour and its barycentre. The PCA is used to reduce the dimension of the feature space (using 32 components to describe each silhouette, the 96% of variance is retained) and a three layers neural network has been trained to recognize the *walking*, *running* and *other* actions.

In a number of publications, recognition is based on HMMs and Dynamic Bayes Networks (DBNs).

For example, in [40], the authors present an object-based system for video analysis and interpretation. The basic unit for analysis is the video object (VO), a concept introduced by MPEG-4. Instead of considering as basic unit of analysis the whole frame, in this approach the low-level features from individual objects in the frames (VO) are considered. From extracted VOs, they followed a pattern analysis methodology by modelling the VO behaviour using DBNs, which can generate a hierarchical description for the video events.

They showed that the object-based approach is effective for a complete characterization of video sequences, which includes both macro-grained and fine-grained semantics contained in the video sequences.

According to the authors, a serious limitation of this approach is that the description generated by DBNs is very fine-grained. It works the best for simple video events. But for long video sequences involving various video events, this approach is unlikely to be satisfactory.

Unsupervised methods (such as HMMs) can be trained automatically but yield models whose internal structure - the nodes - are difficult to interpret semantically. Manually constructed networks typically have nodes corresponding to sub-events, but the programming and training of these networks is tedious and requires extensive domain expertise. In [41] the authors propose a semi-supervised approach where a manually structured, Propagation Network (a form of a DBN) is initialized from a small amount of fully annotated data, and then refined by an Expectation Maximization based learning method in an unsupervised fashion. During node refinement (the M step) a boosting-based algorithm is employed to train the evidence detectors of individual nodes. The proposed results shown that, starting from few full annotated example accompanied by a small number of positive but non-annotated training examples, the system can achieve good performance in indoor activity analysis and also in other applications.

In literature there are also many works attempting to infer information about human behaviours analysing the dynamics and settings of the individual body parts. Once that they have recognized the position of the various body parts, it is possible to consider some constraints and extract some features that can be indicative of some specific actions/behaviours.

For example, in [42] the authors consider three dynamic regularity features. They are temporal properties and are generally independent of camera position:

- **Cycle Time**: this time is referred to the cycle time of a leg which decreases with increasing walking speed. It is computed by measuring the time interval between two successive minima or maxima in the trajectory of a foot.

- **Stance/Swing Ratio**: *stance time* is the period of time when the foot is in contact with the ground while the *swing time* is the period of time when the foot is not in contact with the ground. The ratio stance/swing decreases when a person walks faster.

- **Double Support Time**: this is the period of time when both feet are in contact with the ground. This occurs twice in the gait cycle, at the beginning and end of the stance phase.

Using these features, the authors are able to distinguish walking examples across multiple speeds from other non-walking actions.

A critical element for this kind of systems is the fact that they are substantially view dependent also if the dynamic regularities features that they use are not view dependent.

An attempting to create a view invariant system has been done in [43]. Here the authors, starting from the findings in [44] (where the authors developed relationships between six-tuple 3D points and their corresponding image coordinates that are satisfied for all views of the 3D points), propose a 3D approach aiming for viewpoint invariance. Each action is represented as a unique curve in a 3D invariance-space, surrounded by an acceptance volume ('action-volume'). The "action-volume" concept derives from the experimental observation of the fact that each person performing more than one time the same task produces curves in the 3D invariance-space that are slightly different among them. To take into account these "slightly" differences, the authors consider an "acceptance-volume" surrounding the mean 3D curve describing the action and they call this volume "action-volume". Given a video sequence, 2D quantities from each frame are calculated and matched against candidate action volumes in a probabilistic framework.

As the results presented by the authors show, this approach has encouraging results, but it is far from the goal of a full view-independent system. Indeed, a constraint of this approach is that at least 4 of the 6 3D points representing the "invariant" must lie on different planes. Once that the designers chooses the six points on a person, it is possible that, for a given view-point, they do not satisfy this constraint (according to the instantaneous pose) and so, for that view-point, the performance decreases.

Another approach to view-independent system has been proposed in [45] where the authors propose an approach to matching human actions that is both fully descriptive in terms of motion and is both invariant to view and execution rate. They use a point-based representation of the human body. In particular, each point represents the spatial coordinate of an anatomical landmark on the human body. A central point of this work is the fact that it uses the statistical results about the proportion among the various human body parts described in [46] to introduce geometric constraints among the points representing a human body. In this way they are able to recognize a given action also if it was carried out by different people. Since it is expected that different people may perform some portions of the same action at different rates, the dynamic time warp was used to make this approach invariant to the different execution rates.

According to the authors, this system can achieve good results but it was tested only in particular conditions where, for example, there were not occlusions.

3.4 Action primitive and grammars

Works falling into this class attempt to decouple actions into action primitives and to interpret actions as a composition on the alphabet of these action primitives.

Some of the works falling in this class still use a learning based approach and so they are able only to recognize the learnt actions but in this class there are also works using a generative approach [47] and works starting without any models [48]. In this way it is possible to overcome the limit intrinsic to the learning stage. Also for this class of works the video sampling rate can be a critical parameter for the same reasons seen for the previous class.

Due to the fine level of observation of the scene, the performances of these works are heavily influenced by the noise and the occlusion problems.

A method employing techniques from the dynamical systems framework is presented in [49] where the authors propose to decompose a human activity into a set of elementary actions. These elementary actions can be seen as symbols of an “alphabet” and so, they can be used to describe human motions similar to the way phonemes are used in speech. They call these primitives of motion “movemes”. By using system identification techniques and pattern recognition techniques they develop an on-line joint segmentation and

classification algorithm and provide analytical error analysis. Once that the primitives are detected, an iterative approach is used to find the sequence of primitives for a novel action.

The authors show the results obtained using the system to describe the movements carried out by five different people that are drawing a set of shapes using a computer mouse.

The ratio behind this approach is very interesting because this system proposes a "generative" approach to the human behaviour analysis. The main drawback of this work is that it is quite difficult to apply it in real live situations where problems of noise and occlusions occur.

Another approach from the system theoretic point of view is presented in [50] where the authors try to segment and represent repetitive movements. They use a two-threshold multidimensional segmentation algorithm to automatically decompose a complex motion into a sequence of simple linear dynamic models (second order AR models). The problem of action segmentation is resolved in terms of model changes. Namely, the motion segmentation problem is solved detecting the times at which the dynamical parameters of the AR model used to describe the current action change significantly. No a priori assumptions were made about the number of models that comprise the full motion or about the duration of the task cycle. A compact motion representation is obtained for each segment using parameters of a damped harmonic dynamic model.

The main drawback of this system is the fact that it is able to segment and recognize variations of motions known to the classifier. This fact diminishes the generality of this approach making it usable only in tasks where repetitive motions are present. Furthermore, according to the authors, this system is not suitable for real-time analysis.

A vision based approach is proposed in [48] where the authors propose to describe human actions in terms of *action units* called "*dynamic instants*" and "*intervals*" which can be computed studying the spatio-temporal curvature of a 2-D trajectory. The *dynamic instants* are due to changes in the forces applied to the object during the activity. They are perceived as a change in the direction and/or speed and can be reliably detected by identifying maxima in the spatio-temporal curvature of the action trajectory. An *interval* is the period of time between two dynamic instants during which the motion characteristics do not change. The authors formally show that the *dynamic instants* are view-invariant, except in the limited cases of accidental alignment. The ratio behind this approach has a

psychological root in works such as [51, 52, 53]. The authors focus their attention on human actions performed by a hand. Examples of such actions are: opening and closing overhead cabinets, picking up and putting down a book, picking up and putting down a phone, erasing a white-board, etc. Starting without a model, they use this representation for recognition and incremental learning of human actions. The system tracks the hand using a skin detector algorithm. According to the authors, the proposed method can discover instances of the same action performed by different people from different view points. In the experimental section are shown results on 47 actions performed by 7 individuals in an environment with no constraints obtaining good performances in terms of action recognition.

This approach presents various interesting aspects such as:

- it starts without a predefined model of the actions to be recognized and this, as discussed above, is a desirable characteristic for this kind of systems.
- it is "almost" view invariant because it is able to recognize the same action recorded from different view points. The term "almost" view invariant refers to the fact that it fails in recognizing actions performed on a plane perpendicular to the view plane.

On the other hand, the main drawback of this system is related to the required sample rate. Indeed, since the system recognizes *dynamic instants* studying the spatio-temporal properties of the hand's trajectory (it searches for the maxima in the spatio-temporal curvature of the action trajectory), it requires a detailed representation of this curve. In other words, it must use an adequate sampling rate to sample this curve. This sampling rate varies in a proportional way to the speed at which a given action is performed. Having a low sampling rate, it is possible to loose some *dynamic instants* reducing the performance of the whole system.

In literature it is possible to find also systems attempting to achieve a higher semantic level analysis of the human behaviours in the recorded scenes.

For example, in [54] the authors propose a system performing a hierarchical analysis of a video stream. The lowest level analyses the poses of individual body parts including head, torso, arms and legs are recognized using individual Bayesian networks (BNs), which are then integrated to obtain an overall body pose. The middle level models the activity of a single person using a dynamic Bayesian network (DBN). The higher level of the hierarchy works on the results of the mid-level layer. Here, the descriptions for each person are juxtaposed along a common time line to identify an interaction between two

persons. According to the authors, the following nine interaction types are considered in this paper: the neutral interactions include (1) approaching each other, (2) departing each other, and (3) pointing, and the positive interactions include (4) shaking hands, (5) hugging, and (6) standing hand-in-hand, and the negative interactions include (7) punching, (8) pushing, and (9) kicking.

An interesting aspect of this work is the high semantic level of its output. Indeed, according to the authors, the human action is automatically represented in terms of verbal description according to subject + verb + object syntax, and human interaction is represented in terms of cause + effect semantics between the human actions.

The main drawbacks of this system are:

- the method used to classify the interaction between two people: indeed this task is accomplished using a decision tree. This fact makes this system suitable for recognizing a well defined and fixed set of interactions (the nine listed above);
- as a direct consequence of the previous point, the system is not able to generalize the observed behaviours. Hence, it is not able to recognize interactions that do not belong to the training set and so, for example, it is not able to recognize actions where more than two people are involved.
- occlusions: the whole hierarchy is based on the principle that the system is able to identify the single body parts of the people in the scene. In this way, for example, at the middle level it is possible to infer information about the pose of a man. This is a critical point because in this kind of systems, there are both self-occlusions and mutual occlusions among people.

Working on the same concept of multi level analysis, where at the lower levels the action primitives are recognized and sent at the higher levels to perform a more complex analysis, in [47] the authors propose to use a Stochastic Context Free Grammar (SCFG) to obtain a high semantic level analysis of human behaviour. In this work, the authors propose a probabilistic approach to the analysis of temporally extended actions encompassing also the problem of interactions among moving objects. The system is composed of two levels. The first level detects action primitives using standard independent probabilistic event detectors to propose candidate detections of low-level features. The outputs of these detectors are sent as input stream to the second level. Here a stochastic context-free grammar parsing mechanism is used to analyse the stream and perform a higher semantic level analysis. The main advantages of this approach are that it provides longer range temporal constraints, disambiguates uncertain low-level detections,

and allows the inclusion of a priori knowledge about the structure of temporal events in a given domain.

An interesting aspect of this method is the use of the grammar as a convenient means for encoding the external knowledge about the problem domain, expressing the expected structure of the activity.

The main limit of this approach is the fact that it uses low-level features detectors that are able to model and recognize only a fixed number of action primitives.

An interesting aspect of the human behaviour and of his communicative capacity is the gestural expressiveness. In literature, a large number of works, dealing with the human gesture analysis, are present. This problem can be seen as particular case of the human behaviour analysis.

Many video-based methods have been developed for hand [55], arm [56] and full-body [57] gesture recognition. These systems can be classified in the following classes according to the methodology of analysis that they use:

- landmark based: these systems detect and track some landmarks such as body parts [58], "visual interesting points" [59], "visual cues" [56] or "feature points" [57].
- kinematical based: in these systems, movement kinematical parameters related to the articulated body motion are first recovered as joint-angle vectors or body-centred joint locations. Action recognition is then conducted in such kinematical parameter spaces [47, 60]. It should be noticed as this kind of representation of the human body parts is the same used in many works cited above.
- template based: these systems represent actions using image information such as silhouettes or 3D volumetric reconstruction such as visual hulls. These systems can be further divided into two classes according to the method that they use for feature extraction and action recognition:
    - holistic approaches [61, 62]: these systems model the entire action as a spatio-temporal shape. The recognition task is accomplished comparing this model with a set of learnt models using statistical pattern recognition techniques such as SVM, LDA.
    - sequential approaches [63, 64]: these systems represent an action as a temporal series of key poses. In the training phase, a set of key poses are first selected from the gesture set. Each key pose is described by means of a features vector often called "*pose feature vector*". Action recognition is then

achieved through sequential pattern recognition using methods such as hidden Markov models (HMM) and/or Bayesian networks.

Summary

In this section, an introduction to the problem of human behaviour analysis in video streaming has been presented showing an overview of the main works on this topic present in literature.

These works follow a well defined and accepted processing chain as shown in Figure 4. This fact introduces a hierarchical decomposition of this problem into various modules allowing the researchers for focusing their attention on specific aspects of the problem. Following this principle, the object of this thesis is the definition of an innovative methodology to implement systems for high semantic level analysis of human behaviour in streaming video recorded into the narrow domain. For this reason, this literature overview considers only the works falling in the highest level of the processing chain showed in Figure 4 (namely human behaviour analysis and recognition).

These works can be divided into three main classes according to the aspect of the problem that they consider (whole scene, whole human body and single body parts).

This research topic has attracted many researchers in the last years due to the large number of potential applications in various fields (surveillance, control, analysis). Despite the efforts of scientific community, automatic high semantic level analysis of video streaming still remain a problem far to be solved. This is due to the lack of a comprehensive and universally valid solution to two problems: semantic and sensory gaps. For this reason, in literature it is possible to find many works dealing with specific aspects of the problem and providing satisfying solutions in well defined operating conditions.

One of the most widely accepted assumptions by the works on this topic is the processing chain. Indeed, using this processing chain it is possible to decompose the general problem (semantic human behaviour analysis) into various aspects. In this way, it is possible to face these aspects singularly without considering the whole problem. So, it is possible to build hierarchical systems where the output of the module of a given level becomes the input for the module at the successive level (see Figure 4).

As shown in this literature overview this approach has some drawbacks. Indeed, often the performances of the higher level modules are affected by the errors done to the lower levels. Another open issue to be faced is the definition of an effective representation of the observed scene. This representation should reach an adequate level of detail for all the

semantic processes to be implemented. On the other hands, it should be as compact as possible to avoid problems related to the computational complexity.

Many works attempt to recognize human activity using statistical approaches and searching the query action in a knowledge base composed of a set of recognizable activities. This approach can give good results in some scenarios, but it has an intrinsic limit: it is not able to recognize actions that do not belong to the knowledge base (and hence to the training set used to create it). On the other hand, the idea to define "generative" approaches to the human behaviour analysis has been used in various works. They attempt to recognize a complex activity using the composition of elementary and recognizable actions.

Recognizing activities is an extremely complicated task at which even humans are often less than perfect. The implementation of an automatic system performing this task is an open research field. As shown in this chapter, in literature, some works achieving good results in recognizing specific human activities are present. But the trend is quite clear: the higher is the required semantic level analysis the narrower must be the domain of application.

In this thesis, a methodological approach to implement systems for high semantic analysis of video streaming is proposed. The key issues faced by this methodology are:

- **sensory gap**: since, as shown above, the errors done at the lower level of the processing chain can influence the performance of the semantic analysis, in this thesis an innovative method to reduce the sensory gap is proposed. This method reduces the problem of occlusions among moving objects using a multi-camera approach. The details of this method are shown in chapter 4.
- **model representation**: the proposed methodology represents the actions using string of symbols. In this way it is possible to obtain a compact representation suitable for real time analysis. The details of this method are shown in chapter 5.
- **semantic gap**: the proposed methodology works in narrow domains exploiting some background knowledge about scene and/or the human behaviour, thus narrowing the huge variety of possible behavioural patterns by focusing on a specific narrow domain. A linguistic approach based on the definition of a specific grammar for each domain is used to obtain a high semantic level analysis of human behaviour. The details of this method are shown in chapter 5.

Chapter 4

Sensor data interpretation for symbolic analysis

This chapter describes an original contribute of this thesis: a method to solve the correspondence problem in multi-camera systems without the assumption of epipolar geometry. This method is suitable to reduce the sensory gap and the problem of the presence of mutual occlusions among moving objects inside a scene. Using this method it is possible to improve the performance of the tracking algorithm. The chapter starts with a brief problem overview and then it presents a description of the epipolar geometry. Finally the proposed solution is described and the final considerations are reported.

4.1 Introduction

As seen, in the previous chapter, many approaches to human behaviour analysis work by studying the trajectories of some relevant points. These systems have good performances but they all fail when the motion is perpendicular to the view plane.

In order to overcome this problem many authors propose to use multi-camera systems and in particular binocular systems often in stereoscopic configuration. In these systems two or more cameras are used to record the same scene from different points of view. Using specific algorithms, it is possible to recover 3D data about the scene analysing the various recorded streams. Furthermore this approach helps in reducing the occlusion problem in crowding scenes.

On the other hand, this method requires that the correspondence problem is solved to work properly. This problem refers to locate the match for each pixel of one image with a pixel in the other, and, hence, the name correspondence problem.

Figure 5.a shows a schematic representation of the correspondence problem. From a geometrical point of view, this problem can be solved using the epipolar geometry.

In literature various authors use this epipolar geometry to solve the correspondence problem. This fact means that they introduce strong constraints on the cameras configuration. For example, a widely applied constraint is that the acquisition system

produces stereo images [65, 66]. In this case, the images are taken by two cameras with parallel optic axes and displaced perpendicular to the axes.

Using stereo pairs of images it is possible to assume that: stereo pairs are epipolar and the epipolar lines are horizontally aligned, i.e., the correspondence points in the two images lie along the same scan lines; the objects have continuity in depth; there is a one-to-one mapping of an image element between the two images (uniqueness); and there is an ordering of the matchable points [67].

From a geometric point of view, the corresponding problem can be successfully solved and the methods proposed in literature achieve excellent performance when applied to synthetic images. When these methods are applied to real world images, the main issues to be solved are: noise and illumination changes as a result of which the feature values for the corresponding points in the two images can differ; lack of unique match features in large regions; occlusions, and half occlusions. The wider used methods to solve this problem are: area based [66], feature based [68], Bayesian network [69], neural networks [67, 70], etc.

The stereo-vision approach and, more in general, the multi-camera approach has been applied also to video analysis systems to overcome the occlusion problem and to track people/objects using different cameras [71, 72, 73, 74].

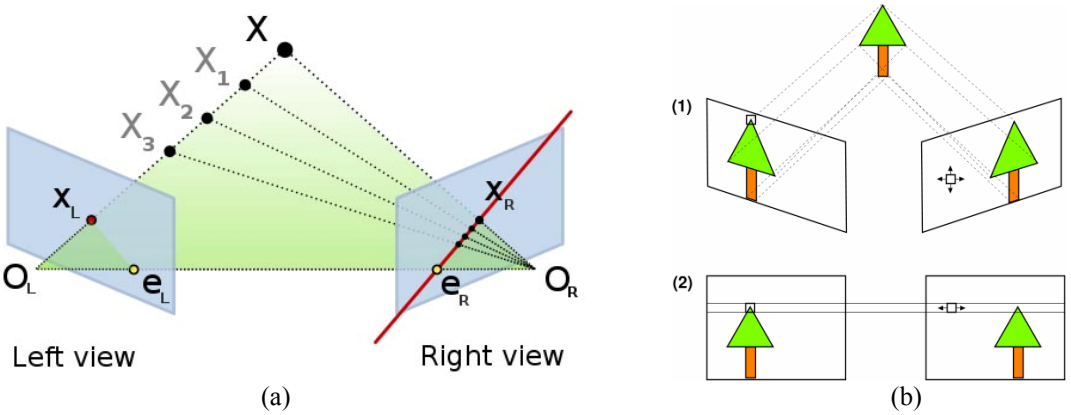
In this thesis, a method to solve the correspondence problem in multi-camera systems based on the merge of two approaches (Self Organizing Map (SOM) and feature based recognition) is proposed. The novelties of this approach are: the proposed method and the ability to work without the assumption of epipolar geometry. Furthermore this method does not require a calibration stage (the initial training of the SOM can not be considered as a calibration stage). This method is not used to fuse two images into one. It is used to find an object into two different images only to handle possible problems of mutual occlusions.

The system must be seen as a stage of the longer processing chain aiming at semantic video analysis (see Figure 4). For this reason, the correspondence problem is not solved for the whole images but only for few relevant points (the barycentres of moving objects).

The remaining part of this chapter is so organized: in section 4.2 a brief introduction to the epipolar geometry is presented and in section 4.3 (and in its sub-sections) the approach to resolve the correspondence problem proposed in this thesis is shown.

4.2 Epipolar geometry

Considering Figure 5.a, let O_l and O_R be the two focal points of the cameras. In real cameras, the image plane is actually behind the focal point, and produces a rotated image.



(a) (b)

Figure 5 – A schematic representation of the correspondence problem (a) and the epipolar constraint (b)

Here, however, the projection problem is simplified by placing a *virtual image plane* in front of the focal point of each camera to produce an un-rotated image.

Since the two focal points of the cameras are distinct points into the 3D world, each focal point is seen by the other and their projection in the respective image planes are called epipoles or epipolar points (e_L and e_R in Figure 5.a). Both epipoles $\mathbf{e}_L$ and $\mathbf{e}_R$ in their respective image planes and both focal points $\mathbf{O}_L$ and $\mathbf{O}_R$ lie on a single 3D line.

The left camera sees the line $\mathbf{O}_L$–$\mathbf{X}$ as a point because it is directly in line with that camera's focal point. On the other hand, the right camera sees this line as a line in its image plane. The line $\mathbf{e}_R$–$\mathbf{x}_R$ in the right camera is called an *epipolar line*. For symmetric reasons, the line $\mathbf{O}_R$–$\mathbf{X}$ is seen as a point by the right camera and as epipolar line $\mathbf{e}_L$–$\mathbf{x}_L$ by the left camera.

The points O_l, O_r and X define a plane called *epipolar plane*. All the epipolar lines lie on this plane. All epipolar planes and epipolar lines intersect the epipole regardless of where $\mathbf{X}$ is located.

If the relative translation and rotation of the two cameras is known, the corresponding epipolar geometry leads to two important observations:

- If the projection point $\mathbf{x}_L$ is known, then the epipolar line $\mathbf{e}_R$–$\mathbf{x}_R$ is known and the point $\mathbf{X}$ projects into the right image, on a point $\mathbf{x}_R$ which must lie on this particular

epipolar line. In other words, for each point observed in one image the same point must be observed in the other image on a known epipolar line. The corresponding image points must satisfy this *epipolar constraint*. This fact can be used as criterion to verify if two points really correspond to the same 3D point. Epipolar constraints can also be described by the *essential matrix* between the two cameras. The essential matrix is a 3x3 matrix which relates corresponding points in stereo images assuming that the cameras satisfy the pinhole camera model.

- Using the triangulation method, it is possible to know the 3D coordinate of X knowing the points $\mathbf{x}_L$ and $\mathbf{x}_R$.

The epipolar geometry is simplified if the two camera image planes coincide (see Figure 5.b). In this case, the search is simplified to one dimension (a horizontal line parallel to the baseline between the cameras O_L-O_R). Furthermore, if the location of a point in the left image is known, it can be searched for in the right image by searching left of this location along the line, and vice versa.

4.3 Proposed System

The proposed system uses two cameras (T_1 and T_2) installed in an arbitrary way. The only constraint is that the scenes recorded by the two cameras must have an overlapping zone. The greater is the overlapping zone, the greater is the area where the correspondence problem can be solved.

As shown in Figure 6, the system has a modular architecture. The stream sampled by each camera follows the chain:

motion detection → object detection → feature extraction → correspondence finder

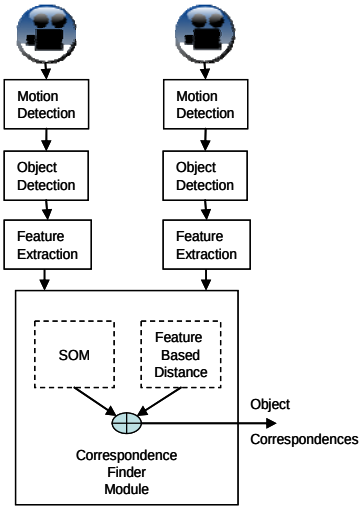


Figure 6 - Block diagram of the proposed system

4.3.1 Motion and object detection module

The models of the target objects and their motions are considered unknown, so as to achieve maximum application independence. In this condition, the most widely adopted approach for moving object detection with fixed camera is based on background subtraction [75]. The background is estimated using a model evolving frame by frame. The moving objects are detected by the difference between the current frame and the background model. A good background model should be as close as possible to the real background and it should be able to reflect as soon as possible the sudden change in the real background.

According to the taxonomy proposed in [76], the proposed system considers the following objects:

- **Moving visual object (MVO)**: a set of connected pixels moving at a given speed
- **Background** is the current model of the real background
- **Ghost**: a set of connected pixel detected as "in motion" by the subtraction algorithm but not corresponding to any real moving object.

The proposed system uses these three elements to implement an object oriented motion detection algorithm. It uses the knowledge about the segmented objects to dynamically improve the background model. In order to classify the object after the blob segmentation the following rules are used:

1) <MVO> ← (foreground blob)Λ (large area) Λ (high speed)

2) <GHOST> ← (foreground blob) Λ (large area) ¬ (high speed)

4.3.2 Features extraction modules

Once that a MVO is detected, it is described using a Content Based Image Retrieval (CBIR) technique. CBIR is the application of computer vision dealing with the problem of retrieving a set of relevant images from an image database. These systems work implementing the following general scheme:

Stimuli → *Signatures* → *Distance*

Stimuli are sought as points in some perceptual space [77] while the notion of similarity between two stimuli is one of the fundamental concepts of the cognitive theories of similarity. For the visualization of the underlying ideas, refer to Figure 7.

Consider two different images from the semantic point of view. Let A_p be the semantic (human centric) space. In this space for each image there is a stimulus. On its basis, it is easy for a human being to assess the similarity between two or more images. CBIR systems try to emulate this cognitive chain. Such systems associate a signature (A_s and B_s in Figure 7) to each image defining a signature space. By endowing the signature space with a distance model it is possible to define an artificial similarity space in which it is possible to measure the distance between two or more images (denoted here by $A_{a\text{-}b}$ and $A_{b\text{-}a}$). In many distance models (such as the one utilized in this work) symmetry property is assumed to be valid, namely make $A_{a\text{-}b}$ equal to $A_{b\text{-}a}$. There could be other constructs such as, e.g., Tversky's "contrast model" [78] in which the symmetry requirement is not used. Indeed, a central assumption of this model is that the similarity between object A and B is a function of the features which are common to A and B ("A and B"), those in A but not in B (symbolized as "A-B") and those in B but not in A (denoted by "B-A"). Based on this concept and several other assumptions, Tversky postulated the following relationship:

$$S(A,B) = xf(A \text{ and } B) - yf(A\text{-}B) - zf(B\text{-}A) \qquad (1)$$

where S is an interval scale of similarity, f is a function of salience of the various features which have been considered, and x, y and z are weights that underline the relations among the features of the objects in A and B.

Following this scheme (in the form of the chain of associations *stimuli→signatures→distance*), several hypotheses have been investigated. In the purely psychological approach based on multidimensional representation, an image is represented as a point in some highly dimensional space [79]. The location of the point is typically determined with the use of some scaling techniques such as e.g. Multidimensional Scaling

(MDS) [80]. The MSD leads to a non-metric space. The goal is to find a projection space in which the inter-point distance is monotonically related to a human panel response about (di-) similarity. In the purely computational approach, a natural scene is represented by means of a collection of values (*signatures*) explicitly derived from the 2-D image containing basic low level features. These could be comparable to such features as retinal-brain sensitivity including shape, colours, and patterns.

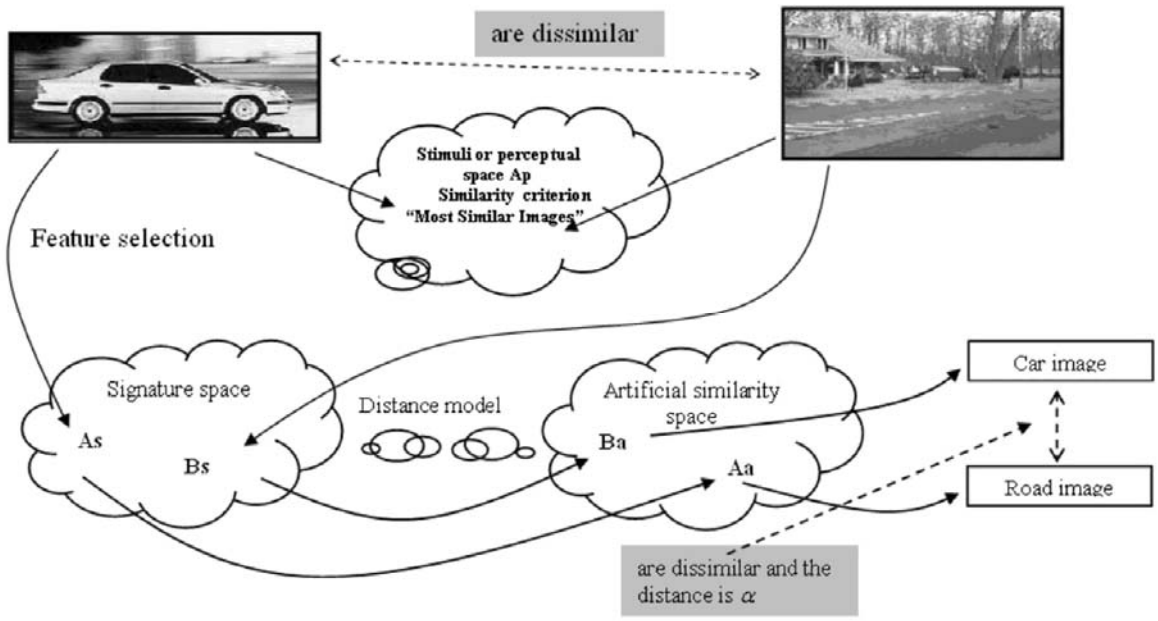


Figure 7 - Relationship diagram between images stimuli-->signatures-->distance (computable similarity)

In this thesis, each MVO is characterized by means of two low level visual features (signatures): colour histogram and texture. These features were used in many Content Based Image Retrieval (CBIR) systems [24, 81, 82].

The colour histogram is computed using the Hue, Saturation and Value (HSV) colour space (Figure 8). This colour space was developed in the late 1970 by computer graphics researchers because they recognized that the geometry of the Red Green Blue (RGB) model (that is the widest used colour space in all the common electronic colour devices) was poorly aligned with the colour-making attributes recognized by human vision.

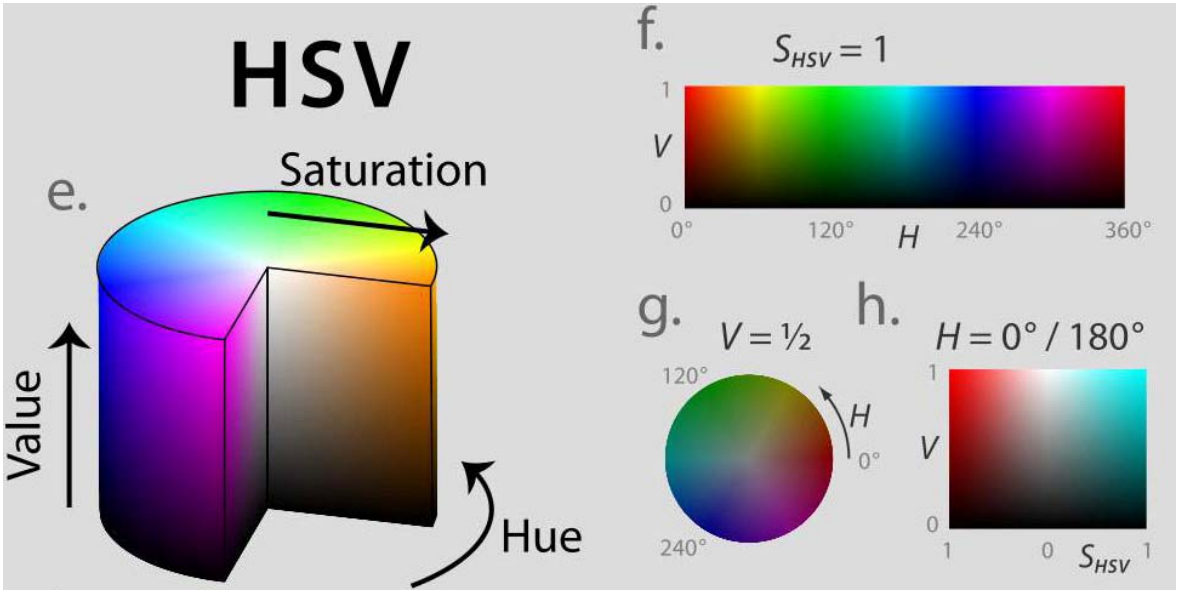


Figure 8 - schematic representation of the HSV colour space

HSV arranges the colours on a cylinder where the *hue* is an angle (0 to 360) representing the pure colour (red, magenta, yellow, etc). The distance to the centre is the *saturation* going from the pure colour (1.0 == fully saturated) to white (0.0 == no saturation). The height within the cylinder represents the *value* or brightness of the colour, going from completely bright (1.0) to no brightness (black, 0.0).

The mathematical transformation from RGB to HSV could be computed using the following equations [83]:

$$H = \begin{cases} 60\left(\frac{G-B}{\delta}\right) \Leftrightarrow MAX = R \\ 60\left(\frac{B-R}{\delta}+2\right) \Leftrightarrow MAX = G \\ 60\left(\frac{R-G}{\delta}+4\right) \Leftrightarrow MAX = B \\ notdefined \Leftrightarrow MAX = 0 \end{cases}$$

$$S = \begin{cases} \frac{\delta}{MAX} \Leftrightarrow MAX \neq 0 \\ 0 \Leftrightarrow MAX = 0 \end{cases} \qquad (2)$$

$$V = MAX$$

Where δ = (MAX – MIN), MAX = max(R, G, B), and MIN = min(R, G, B). Note that the R, G, B values in the equations are scaled to [0, 1]. In order to confine H within the range of [0, 360], H =H + 360, if H < 0.

Using these equations, the original colour space of each MVO (RGB) is converted into HSV colour space and then a histogram composed of 256 bins (16 hue, 4 saturation and 4

value levels) is computed. This colour space was preferred to the RGB colour space because the former is closer than the latter to the human colour perception scheme [84].

Textures can be defined as 'homogeneous patterns or spatial arrangements of pixels that regional intensity or colour alone does not sufficiently describe' [85]. Among contents based features, texture is a fundamental feature which provides significant information for image classification, for this reason it has been used in the proposed system. In order to describe the texture of each MVO, the approach proposed in [86] has been used. This approach is based on using Gabor filters and defining a feature vector composed of 48 components. Using Gabor filters in texture analysis is a well known method presenting good performance as shown in [87, 88].

A two dimensional Gabor filter can be seen as a plane wave restricted by a Gaussian envelope function. A two dimensional Gabor function g(x,y) can be written as:

$$g(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)} e^{i2\pi f_0 x} \qquad (3)$$

Where σ_x and σ_y are the spreads of the Gaussian and f_0 is the spatial frequency of harmonic wave. Gabor functions form a complete but non-orthogonal basis set. Expanding a signal using this basis provides a localized frequency description. Starting from the Gabor function, often called *mother Gabor wavelet*, it is possible to define a self-similar filter dictionary through the appropriate dilatations and rotations of *g(x,y)*. This dictionary is used to decompose the input image into *s*xθ filtered images where *s* represents the number of scales and θ is the number of orientations. In other words, the input image is analysed at *s* different scales to capture different level of details. At each scale, the image is analysed to evaluate the presence of components at θ different orientations. For each filtered image two features are extracted: mean value and standard deviation. The so obtained feature vector is composed of *s*xθx2 components. The experiments reported in this thesis are based on a feature vector composed of 48 components where *s*=4 and θ=6.

Furthermore, for each MVO the bounding box vertex coordinates and the barycentre coordinates are computed.

4.3.3 Correspondence finder module

The main task of this module is to define a mapping among the MVOs present in the various frames extracted simultaneously by the two cameras. It works using two different

methods to find the matching among MVOs: SOM and the distance among their visual features.

The SOM has been chosen because they have the property of effectively creating spatially organized "internal representation" of various features of input signals and their signatures. The training stage is unsupervised. At each training step, the node/neuron closer (in the Euclidean sense) to the input vector is considered the winner. Its weight vector is updated to move it closer to the input vector in the weight space. All the neighbouring nodes/neurons are updated in a weighted way.

The proposed approach is inspired by the work in [67] where the authors propose to use a SOM to obtain a dense disparity map between two images. In this thesis, this approach is used to define a sort of raw region mapping between the images sampled by the two cameras. In this way, the SOM is used to measure the distance between the barycentre of a MVO in a frame and all the barycentres of the MVOs in the frame taken by the other camera. The time correspondence of two frames taken by the two cameras is assured by the acquisition system that labels in real time the frame sampled by each camera with the timestamp. Let $C_1(x',y')$ be the coordinates of the first MVO barycentre in a frame taken by one camera. Using the SOM, it is possible to map C_1 in the frame taken by the other camera obtaining the coordinates $C_1(x'',y'')$. In this way it is possible to measure the distance between $C_1(x'',y'')$ (and hence the MVO_1 represented by C_1) and all the barycentres of the other MVOs in the frame taken by the second camera. The Euclidean distance is used to measure the distance among the various MVOs.

The second method uses the visual features extracted by each MVO to define a mapping among the MVOs in the frames taken by the two cameras. For each MVO present in each couple of frames sampled simultaneously by the two cameras, the extracted features are used to find the correspondence. This module computes the distance among the visual features using a weighted normalized Euclidean distance. This kind of distance is used to consider the similarity contribute of both the features. In the proposed experiments, the weight is equal to 0.5 to give the same relevance to both the features.

The two methods are merged using a weighted distance function (equation 4). Let X and Y be the sets of MVOs in the frame taken by T_1 and T_2 respectively, let $x \in X$ be a MVO in T_1 and $y \in Y$ a MVO in T_2, then the distance between x and y is computed using the following function:

$$d(x,y) = \alpha d_1(x,y) + (1-\alpha)[\beta d_2(x,y) + (1-\beta) d_3(x,y)] \qquad (4)$$

Where d_1 is the Euclidean distance between the barycentre of x and y after the SOM mapping, d_2 and d_3 are Euclidean distances applied respectively to colour feature and texture feature. α and β are the weights used to merge the proposed methods.

Summary

In this chapter a new method to perform images fusion in multi-camera systems, by identifying the corresponding points in the various images without the assumption of epipolar geometry has been presented.

This method must be seen as a stage of the longer processing chain aiming at semantic video analysis. Since it works at the level of the tracking algorithm reducing the problem of mutual occlusions among MVOs, it can improve the results obtained by the proposed methodology for semantic analysis of video streaming but the latter can work also without the former (the semantic analysis module can work also in system using a single camera).

The proposed method is based on the fusion of two different approaches: SOM and CBIRs.

The SOM is used to create a sort of feature based mapping between some relevant points into the two images. It should be noticed that it works only on the background of the scenes. In this way, giving the coordinate of a point in an image, it is possible to find its coordinates into the other image (with a certain level of approximation).

The CBIRs based module describes the detected moving objects present into the two images using two low level visual features (colours and texture). Using this description, this module finds the correspondence among the moving objects present in the two images.

The system uses both this methods weighting their outputs to find the final list of corresponding objects (and their coordinates) into the two frames.

The ratio behind this approach is to enrich the information in the visual features (those used by the CBIR based module) with geometrical information (the output of the SOM).

Using this method, it is possible to reduce the problem of occlusions in crowding scenes and the sensory gap. Indeed, when the system does not find the correspondence between some MVOs into the two frames it means that an occlusion condition has been detected. These conditions are handled at a higher level of the processing chain by the tracking algorithm.

In literature there are other methods that could be used to perform this task. For example, in the framework of CBIR systems an interesting alternative can be the use of

salient points. This kind of technique works on local aspects of the images and it finds a wide application in CBIR systems [89, 90, 91]. An interesting point is characterized by two properties: distinctiveness and invariance. This means that a point should be distinguishable from its immediate neighbours and the position as well as the selection of the interesting point should be invariant with respect to the expected geometric distortions [92].

A comparative evaluation between the proposed method and one based on salient points has not been carried out because the thesis is focused on the high semantic level analysis of video streaming and this module is only functional to the main goal. This approach was introduced to build a solid base for the next processes in the computational chain. It is possible to change this module in order to implement any methods. This fact does not change the methodology used for the semantic analysis of the videos, it can only improve its accuracy.

Chapter 5
Semantic analysis

This chapter describes the second and most important original contribution of this thesis namely a methodology to implement systems suitable for high semantic level analysis of video streaming recorded into a narrow domain. This methodology is independent by the method proposed into the previous chapter. Since the latter improves the performance of the object tracking algorithm, the former works with or without the latter. This methodology can work also using single camera systems. The chapter starts with a brief problem overview and then presents the proposed methodology.

5.1 Introduction

As shown in Figure 4, the semantic analysis of human behaviour is the last stage of a complex processing chain.

This chapter presents the grammar based methodology proposed in this thesis. This methodology allows for a hierarchical analysis of the recorded scene. According to the level of details used in scene recording, this methodology can provide from a semantic analysis of the whole scene till a detailed behaviour analysis of a single person. From this perspective, this methodology can be seen as a unifying approach to the three classes of methods described in the chapter 3.

The grammar based approaches, as shown in chapter 3, allow for human behaviour recognition and classification. The ability to define dynamically, for example using some clustering methods, the classes of the observed scenes is not present in other powerful tools used in literature such as the Hidden Markov models (HMM).

The proposed methodology starts from the idea at the base of the scene interpretation systems that try to interpret the scene studying the trajectories of some relevant points of the moving objects (their barycentres). The idea behind this kind of approaches is the mapping between trajectories and behaviours. The ratio can be synthesized into the observation that in order to accomplish a given task one must follow a prefixed series of movements.

In this thesis, the trajectories are represented by means of string of symbols (the alphabet of the domain specific grammar) each one having a semantic value into the specific domain. The mapping among symbols and the semantic meaning of areas of the scene is done manually at design time. For each domain of application, a grammar is defined in order to specify the recognizable sentences of that domain. In this way it is possible to define a correspondence between the set of the sentences writable with this grammar and the set of allowed actions in the scene.

Since the mapping among portions of scene and symbols/semantic meanings and the grammar are specific for each application, this system belongs to the class of narrow domain systems.

In general, the main drawback of the existing approaches to semantic analysis of the human behaviour, even in narrow domains, is inefficiency due to the high computational complexity related to the complex models representing the dynamics of the moving objects and the patterns of the human behaviours. In this perspective this thesis explores an innovative, original approach to human behaviour analysis and understanding by using the syntactical symbolic analysis of video streaming described by means of strings of symbols.

The remaining part of this chapter is so organised: in section 5.2 the key elements of the proposed context switching from trajectories to word is presented, an overview of the proposed methodology is shown in section 5.3. In 5.4 some relevant aspects about grammars and languages are introduced, in section 5.5 the used grammar is shown in details and in 5.6 the aspects related to the time are discussed.

5.2 Switching domains: from trajectories to words

This approach introduces a domain switching for the problem of trajectories analysis. Indeed, by labelling the environment, it is possible to “translate” the geometric data about the trajectories into words. In this way, studying the characteristic of a word means to study the geometric characteristic of a trajectory. So the geometric analysis becomes a linguistic problem.

From this perspective, the problem changes its appearance. The issue of understanding which behaviours (and so which trajectories) are allowed in a given environment become

the issue of understanding which words one can write using the symbols (labels) defined for that environment.

This problem can be faced defining a specific grammar for each environment. This approach gives a strong flexibility and reliability to the proposed methodology. Indeed, in this context, defining a grammar means to define the utilizable rules to write the words describing the behaviours. In this way, this methodology inherits one of the most interesting characteristics of the language theory: the possibility of defining infinite set of words (behaviours) starting from a finite set of symbols (the labels used to describe the domain of interest).

This aspect is very important and represents a significant advantage of the grammar based approaches in comparison with other statistic methods (for example the HMM). Indeed, also thinking at a method based on a learning process, it is well known that it is able to recognize all the behaviours that belonged to the training set. Of course, there is the process of generalization, but human behaviour analysis is a complex task. Here, this property gives robustness to small changes into the observed behaviours but, for example, it does not give to the system the ability to recognize for example a complex behaviour composed of the concatenation of two successive behaviours (also if both these behaviours belonged to the training set).

Defining the grammar for a given domain of interest (namely the environment from which the scenes are recorded), it is possible to define a set of rules allowing to discriminate between words belonging to that grammar (corresponding to the set of allowed behaviours) and words do not (corresponding to forbidden behaviours).

From this point of view, this methodology can be used to implement systems that are able to recognize a large number of behaviours and to raise alarms when a given behaviour is not recognized.

Figure 9 stresses the concept of mapping among domains showing the successive translations among the three domains under analysis: real world human behaviour, trajectory and linguistic.

Figure 9.a, shows a schematic overview of a real world room. A man is entered from the door (the brown rectangle) and he is walking along the perimeter of the room. He stops his

walk into a certain position. This corresponds to the behaviour of entering into a room and going to a specific position. This representation shows all the positions occupied by the man during his walk into a single frame. To highlight his actual position, his shape has been drawn darker in the last frame than in the previous frames.

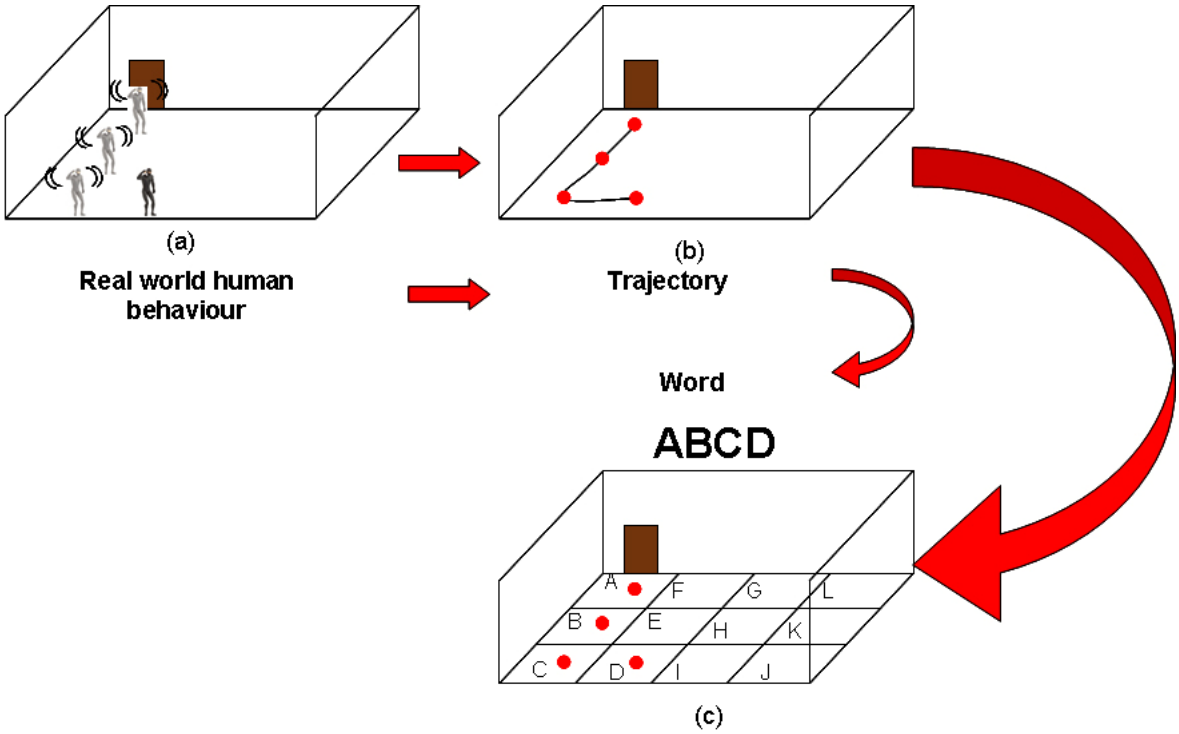


Figure 9 - An example of domain conversion. The real world action (a) is "translated" into a curve in the domain of the trajectories (b). The trajectory is "translated" into a word into the linguistic domain (c)

This behaviour is projected into the domain of the trajectory as shown in Figure 9.b. Here the man is represented by his barycentre (the red points into the figure) and its trajectory is represented by a curve intersecting his barycentre in the successive frames.

This curve in the domain of the trajectory is translated into a word in the linguistic domain. This process is represented in Figure 9.c. The floor of the room is virtually divided into twelve areas each one labelled using a letter of the English alphabet. The portion of the curve falling into a given area is coded using the label of that area. In this way, a trajectory is transformed into a word by means of the concatenation of all symbols labelling the areas on which the curve lies. **Following this processing chain, it is possible to obtain a mapping between human behaviours and words**.

On the other hand, this methodology allows for an improvement in the semantic level of the human behaviour analysis. Indeed, at design time, it is possible to create a mapping between some symbols and some relevant semantic concepts.

For example, it is possible to think that the room represented in Figure 9 is a laboratory of information science and that a printer is installed in the area labelled with the letter "***D***". **In this perspective, this methodology allows for implementing systems exploiting a higher level of semantic analysis of the human behaviours than systems that are only able to classify a behaviour as belonging to a known class or not.** Indeed, the word "***ABCD***" that is written by the system in the example of Figure 9, it is not only a "*correct*" word (because it is a word that belong to the language defined on the grammar written at design time) and thus an allowed behaviour, but it describes the action of "*a man who enters into the room and goes to the printer*".

Another aspect to be considered of the proposed methodology is the used model of the real world and in particular of the human behaviours. Indeed, this model represents each behaviour as a string of symbols (letters of English alphabet). In this way, the recorded scenes can be represented by means of strings having a high semantic content. As it is well known, the strings are variables easily handled by the modern computers. This approach is not affected by the inefficiency due to the high computational complexity related to the complex models representing the dynamics of the moving objects and the patterns of the human behaviours that are typical of other approaches present in literature.

Furthermore, since this methodology realizes a mapping between the domains of human behaviours and words, storing strings into a database is equivalent to logging human actions. So, it is possible to create databases handling simple variables and using all the powerful research features of the modern database management systems.

From this perspective, the proposed methodology allows for the implementation of advanced system for semantic video indexing.

5.3 Overview of the proposed methodology

Figure 10 shows a schematic overview of the proposed method for human behaviour analysis.

In the bottom left part of the figure, a schematic example of an indoor environment is presented. At design time, the environment is virtually divided into a certain number of areas and each area is labelled with a symbol (in this thesis, letters of the English alphabet have been used). It is possible to define areas of whatever shapes and dimensions.

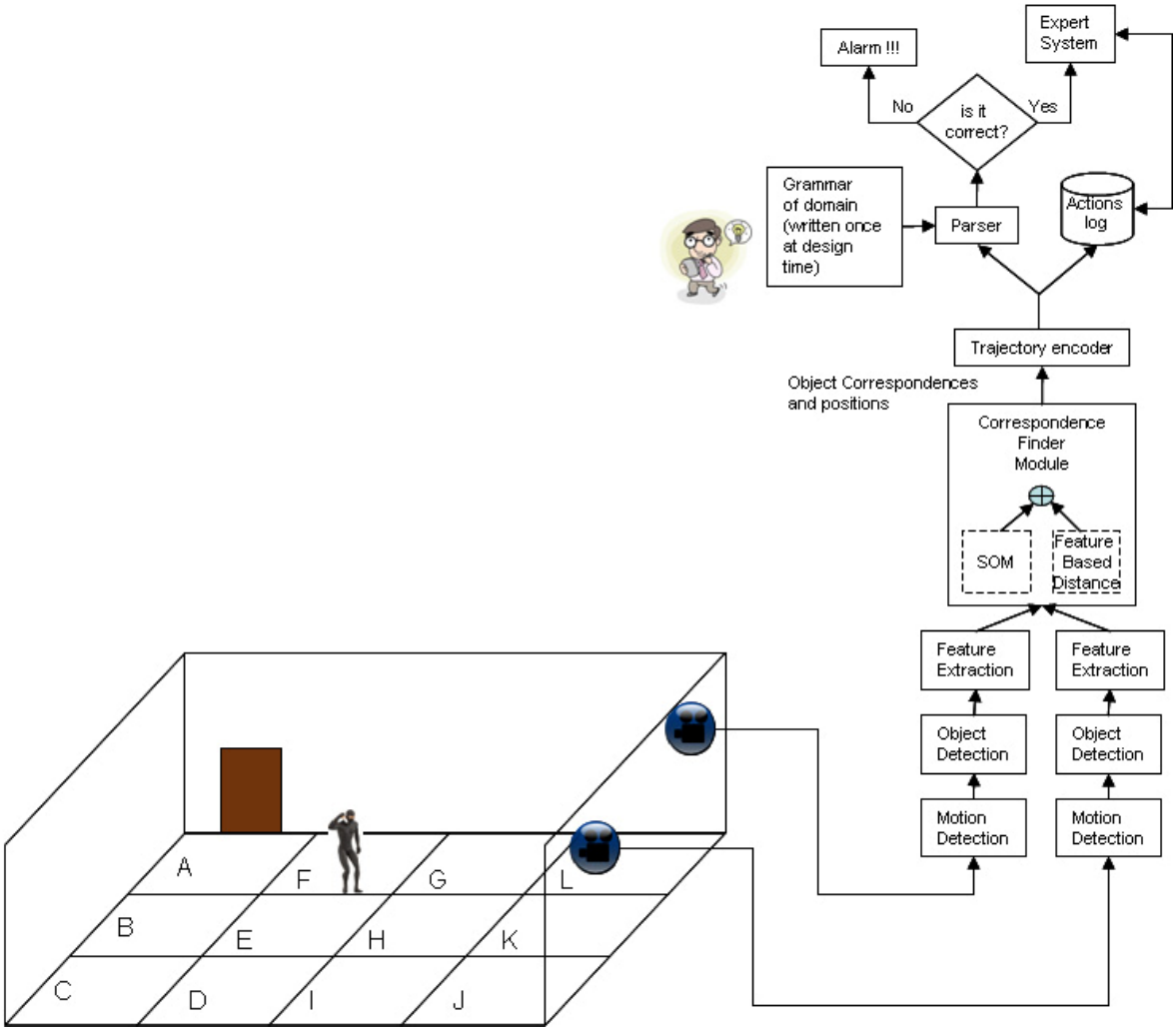


Figure 10 - a schematic overview of the proposed methodology

It is also possible to define areas of different dimensions among them. The only constraints are:

- There are no overlapping areas
- Each portion of the environment belongs to an area (namely, the partition must completely cover the area under analysis)

This fact allows for a more accurate definition of the areas around some particularly relevant points, namely the areas where it is possible to attribute specific semantic meanings (see the above example of the printer).

The dimension of the areas should be coherent with: the scale of observation of the environment (hence with the resolution of the video streaming) and with the typical speed of the MVOs in the environment under analysis. In other words, the partition should be defined in order to obtain a good resolution in the successive stage of string generation. This is a critical parameter because if the areas are too small in comparison to the speed of

the MVOs, it is possible that a given MVO goes from an area to another without passing among all the adjacent areas. In this condition, the parser will not recognize this word and it will consider this one as a forbidden behaviour.

The choice of the correct dimension of each area can be straightaway solved in an automatic way, implementing an early learning stage where the system analyzes the mean speed of the MVOs.

Each trajectory should be represented by a string of symbols with at least a symbol for each semantic area that it crosses.

This great flexibility into the partition of the environment and, from a certain point of view, the coarse grain at which this methodology analyses the trajectories make the system very robust to all the typical problem related to the presence of noise into the input data.

This consideration derives directly from the process at the base of the double context switching from human behaviour to trajectory and from trajectory and word described above. Indeed, a human (or whatever portion of interest according to the scale of observation of the problem) is represented analysing only a point (his barycentre). The position of this point into the room can also be determined with a low level of precision because the useful data are not its coordinates but the label of the area that contains it.

Starting from these considerations, **this methodology does not require techniques to handle the uncertainty due to the noise into input data**.

As shown in Figure 10, this methodology can take advantage of the multi-camera system described in the chapter 4 (but the methodology is applicable also to single camera systems). Indeed, having a binocular vision, it is possible to:

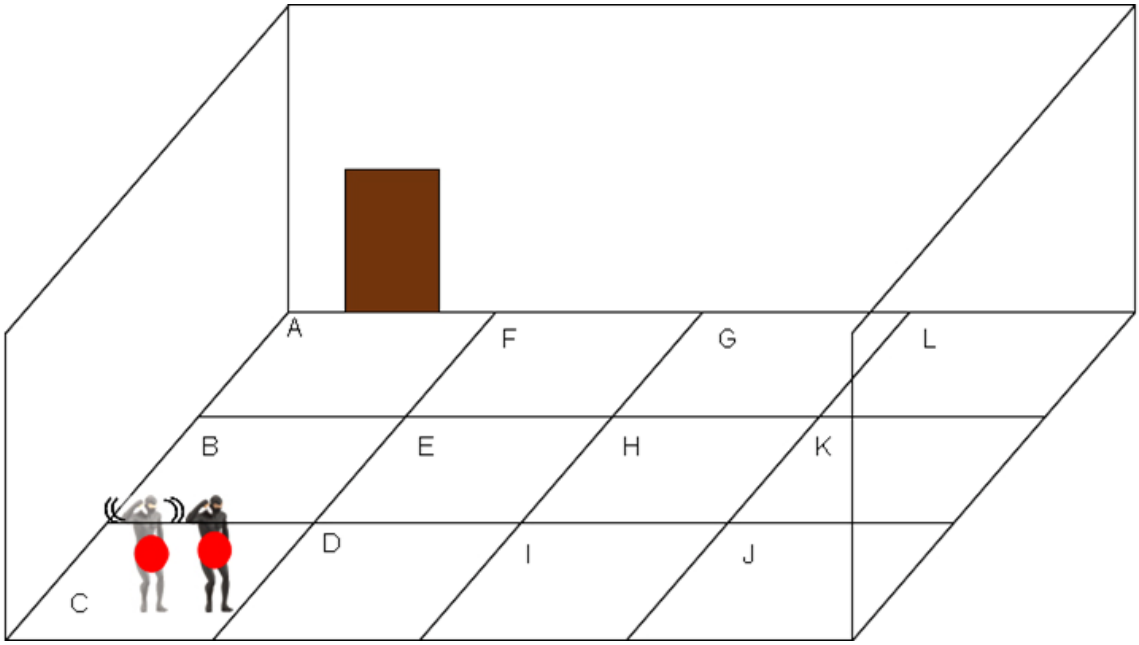
- Improve the precision of the coordinate of the barycentres of moving objects;
- Alleviate the problem of partial occlusions in crowded scenes.

The stream sampled by each camera follows the chain:

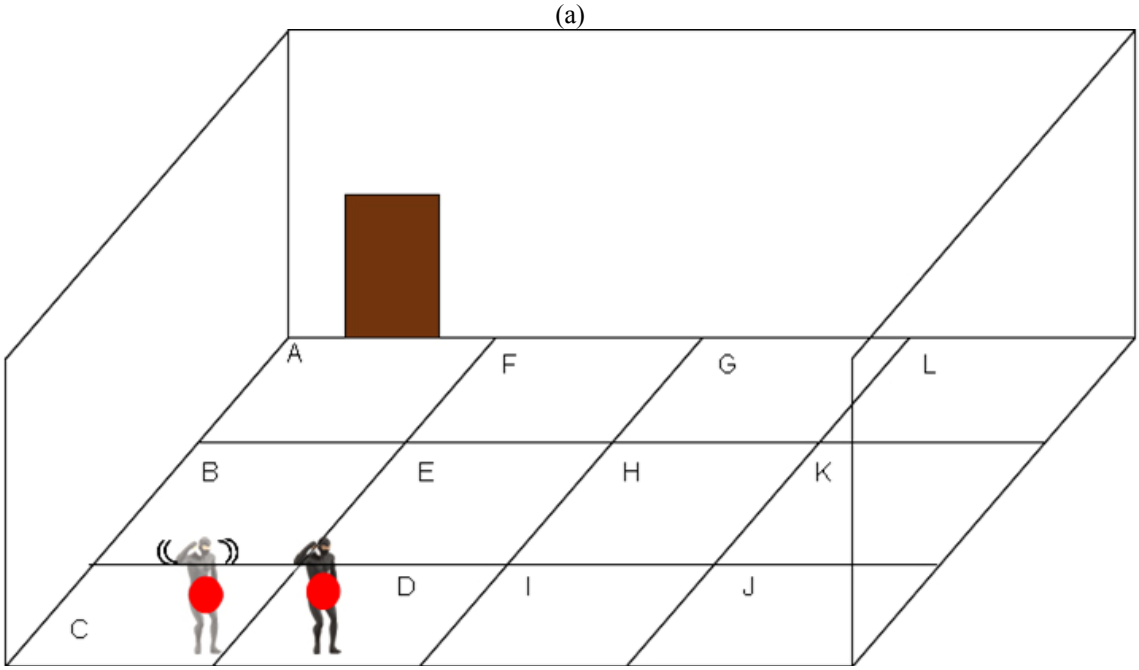
motion detection → object detection → feature extraction → correspondence finder

The output of the correspondence finder module consists of the list of object correspondences and their coordinates (Figure 10).

The tracking algorithm uses a string for each detected moving object, hence, the output of the correspondence finder module is used to update these strings (see the module "trajectory encoder" in Figure 10) appending the symbols with which are labelled the relative areas.



(a)



(b)

Figure 11 – Schematic representation of the process of string generation

The process of strings generation and update is described in the Figure 11. This figure shows two possible events that can occur during a scene recording. In Figure 11.a, an example of a man moving inside the same area is shown. In this case, no one symbol is appended to the end of the string used to track his movements. Figure 11.b shows an example of a man moving from one area to another. In this case, the label of the new area is added to the string ("*D*" in the proposed example).

From a theoretical point of view, it should be noticed that, using this procedure, it is possible to generate any kind of string. On the other hand, the real world phenomenon under analysis, (i.e., the motion of a man into a room) has its physical constraints. For example, looking at Figure 11.b, for the principle of continuity of motion, when the man is in the area "*D*", he can not go directly into the area "*J*". He should follow a path through the area "*I*" or a longer path through the other neighbourhood areas. Furthermore, it is possible that in a real world case there are other constraints. For example, it is possible that a desk is positioned in the area "*E*". In this case, no one can walk on this area. Other possible constraints can be introduced according to the analysed scene.

Using the proposed methodology, the respect of these constraints has a straightforward implementation. Indeed, thanks to the double context switching from real world motion to trajectory and from trajectory to string, this methodology allows to face this problem as a linguistic one.

In this context, there is a reliable methodology derived from the language theory to evaluate if a given *string* belongs to a given language (and so it is correct) or not.

5.4 Grammars and languages

Formal languages are defined with respect to a given alphabet. The alphabet is a finite set of symbols, each of which is called a *letter*. It should be noticed that the terms "letter" does not refer to the "ordinary" letters but it refers to any symbols like numbers, digits, and words. Each finite sequence of letter is defined string or word.

Given an alphabet Σ, the set of all strings over Σ is denoted by Σ^* (where * is the Kleene operator). Notice that no matter what the alphabet is, Σ^* is always infinite. Indeed, even for an alphabet composed of a single letter (for example the letter *a*), Σ^* contains all the combination of this symbol (*a*, *aa*, *aaa*, *aaa*…).

A formal language over an alphabet Σ is a subset of Σ^*. A language is defined as a subset of Σ^*. It can be, finite or infinite. Since Σ^* is always infinite, given any alphabet Σ, the number of formal languages over Σ is infinite.

In order to specify a language, it is possible to use a generative approach by means of the concept of **grammar**. A grammar could be seen as set of rules which manipulate symbols. There are two kinds of symbols: **terminal** ones, which should be thought of as elements of the target language, and **non-terminal** ones, which are auxiliary symbols that facilitate the specification. The non-terminal symbols can be considered as syntactic

categories. Similarly, terminal symbols might correspond to letters of some natural language, or to words.

Rules are used to express the internal structure of "phrases", which should not necessarily be viewed as natural language phrases. Rather, they induce an internal structure on strings of the language, but this structure can be arbitrary, and should be motivated only by the convenience of expressing the required language. A rule is a non-empty sequence of symbols, a mixture of terminals and non-terminals, with the only requirement that the first element in the sequence be a non-terminal one.

A grammar is a finite set of rules. Formally, a grammar is defined by a four-tuple G=(V, Σ, P, S) , where V is a finite set of non-terminal symbols, Σ is an alphabet of terminal symbols, P is a set of rules and S is the start symbol, a distinguished member of V. The rules (members of P) are sequences of terminals and non-terminals with a distinguished first element which is a non-terminal.

A well accepted method to represent the rules is the use of expressions like:

S→A

S→AB | a

In these examples of expressions, the following elements are present:

- capital letters of the English alphabet: they represent the non-terminal symbols
- lower case letter of the English alphabet: they represent the terminal symbols
- The symbol '→' that means 'produce', it represent the relation that exists between various strings of non-terminals and terminals.
- The symbol '|' that means 'or', namely, in this rule, the non-terminal symbol 'S' can produce or two non-terminal symbols 'AB' or the terminal symbol 'a'.

A language L over a grammar G is represented by the symbol L(G) and can be informally defined as the set of all the possible strings that can be generated by G.

Noam Chomsky classified grammars into four types now known as the *Chomsky hierarchy*. The difference between these types is that they have increasingly stricter production rules and can express fewer formal languages.

The Chomsky hierarchy defines the following levels:

- A type-0 grammar (unrestricted grammars) is the set composed of all formal grammars. They are able to generate all the languages that can be recognized by a Turing machine. These languages are also known as the recursively enumerable languages.

- Type-1 grammars (context-sensitive grammars): these grammars are able to generate the context-sensitive languages. The rules of these grammar are expressed in the form $\alpha A\beta \rightarrow \alpha\beta\gamma$ with A a non-terminal and α, β and γ strings of terminals and non-terminals. The strings α and β may be empty, but γ must be nonempty. The machine that is able to recognize these languages is the linear bounded automaton (a nondeterministic Turing machine whose tape is bounded by a constant times the length of the input.)
- Type-2 grammars (context-free grammars): these grammars generate the context-free languages. These are defined by rules of the form $A \rightarrow \gamma$ with A a non-terminal and γ a string of terminals and non-terminals. The machine that can recognize these languages is the non-deterministic pushdown automaton. Context-free languages are the theoretical basis for the syntax of most programming languages.
- Type-3 grammars (regular grammars) generate the regular languages. Such a grammar restricts its rules to a single non-terminal on the left-hand side and a right-hand side consisting of a single terminal, possibly followed (or preceded, but not both in the same grammar) by a single non-terminal. The machine that is able to recognize these languages is the finite state automaton. Additionally, this family of formal languages can be obtained by regular expressions. Regular languages are commonly used to define search patterns and the lexical structure of programming languages.

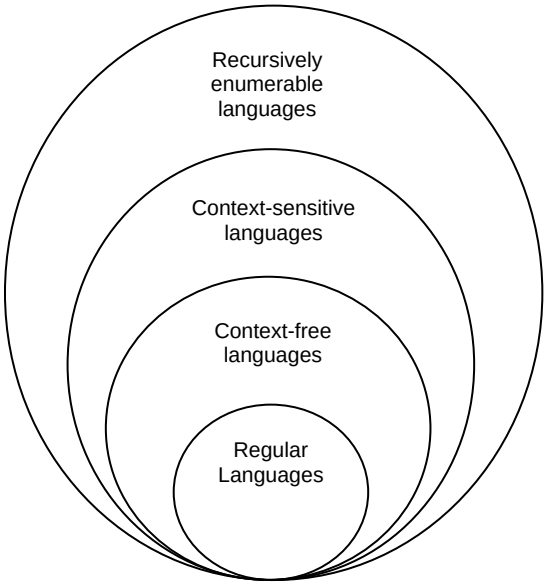


Figure 12 - a graphical representation of the Chomsky's hierarchy

5.5 The grammar used in the proposed methodology

Using a grammar becomes a cardinal point of the proposed methodology because, as shown above, the grammars can be used to decide if a given string belongs to a given language or not.

Thanks to the proposed double context switch, it is possible to represent behaviours using strings. Hence, it is possible to model the behaviours allowed into a given environment by defining a grammar on the symbols used to label it. This task must be done once at design time. At run time, when the system records and interprets a scene, it translates the observed trajectories into strings and tests if they belong to the defined language or not. If a string does not belong to the defined language, it means that the corresponding behaviour does not belong to the set of behaviours considered compatible (or acceptable) in that environment.

The principles used to define a grammar for a given environment can be summarized into the following points:

- Define a virtual partition of the environment under test
- Attach a label to each element of the partition. These labels will be the non-terminal symbols of the grammar
- Assume that the entry point of the environment is the start symbol of the grammar
- For each labelled area, write the production rules that allow to a moving object to go from that area to each allowed adjacent area.

According to the grammar classification proposed by Chomsky, the grammar used in this thesis belongs to the grammar of type 3 namely it is a regular grammar. Indeed, the production rules satisfy the conditions of such kind of grammar (a single non-terminal on the left-hand side and a right-hand side consisting of a single terminal, possibly followed (or preceded, but not both in the same grammar) by a single non-terminal) and overall because all finite languages are regular. The proof of this theorem is carried out using the principle of induction as shown in [93].

The language defined by the proposed grammar can be theoretically composed by a large number of words but this number is finite because the time spent by people in a giving environment is finite.

In the following, an example of definition of such a grammar is proposed.

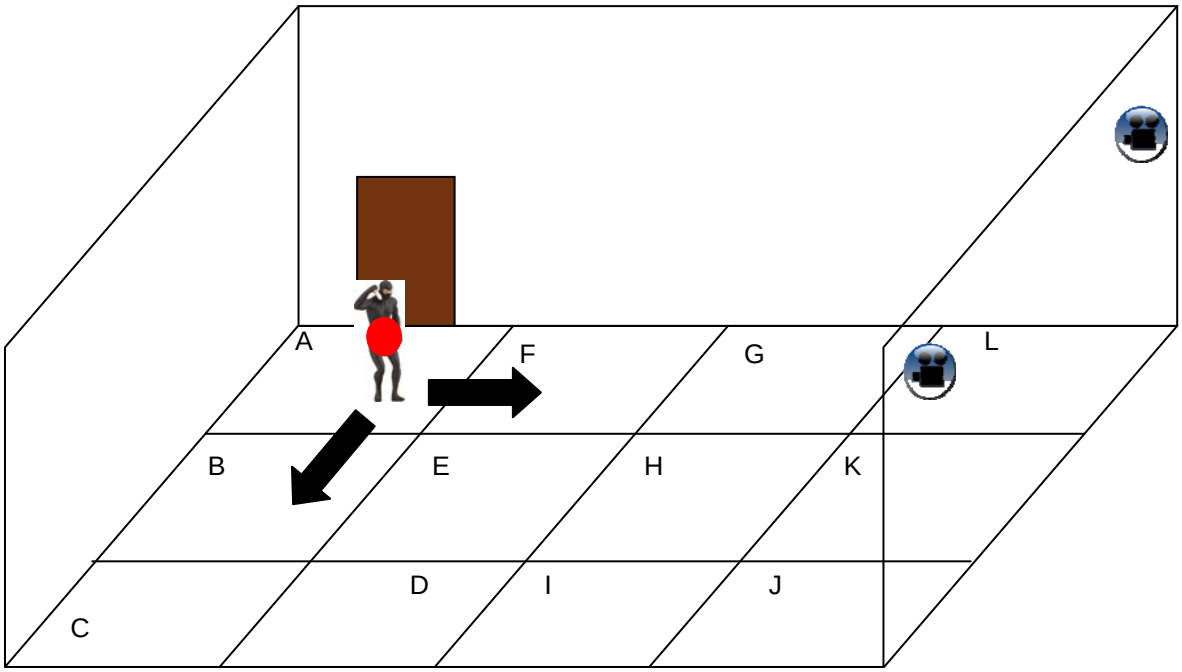


Figure 13 - An example of labelled indoor environment with an entering moving object

Figure 13 shows a schematic view of an indoor environment. In the right side, two cameras are represented in order to show the compatibility of this methodology with the method to solve the correspondence problem presented in the previous chapter. The environment has been partitioned into twelve areas. Each area has been labelled with a capital letter of the English alphabet.

As said above, formally, a grammar is defined by a four-tuple G=(V, Σ, P, S) , where V is a finite set of non-terminal symbols, Σ is an alphabet of terminal symbols, P is a set of rules and S is the start symbol, a distinguished member of V.

In this example, the set of non-terminal symbols is:

V = {A, B, C, D, E, F, G, H, I, J, K, L}

The set of terminal symbols is:

Σ = {a, b, c, d, e, f, g, h, i, j, k, l}

The start symbol is S={A}

In order to complete the formal definition of the grammar, the set P, namely the set of production rules, must be defined.

For each area, the production rules to be written are those allowing for the moving object to move from that area to each allowed adjacent areas. Hence, in this example, supposing to implement a quad-connection schema, the rules for the start element “A” are:

S → A

A → aB | aF | a

In this context, using the "|" symbol it is possible to obtain a more compact representation of the rules.

This rule (A → aB | aF) means that a walking man that enters in this environment can go or in the area labelled with the symbol "B" or in that labelled with the symbol "F". The lower case "a", namely the terminal symbol "a", is written in the string describing the motion to record the fact that the man went through the area "A".

Just starting from this first rule, it is possible to show a concrete application of the proposed methodology: if at run-time, the recording system generates a string containing the combination of symbols "ae", then the moving object is acting a forbidden behaviour because this string does not belong to this grammar.

Figure 14 shows another possible configuration: the man is entered into the area labelled with the symbol "B".

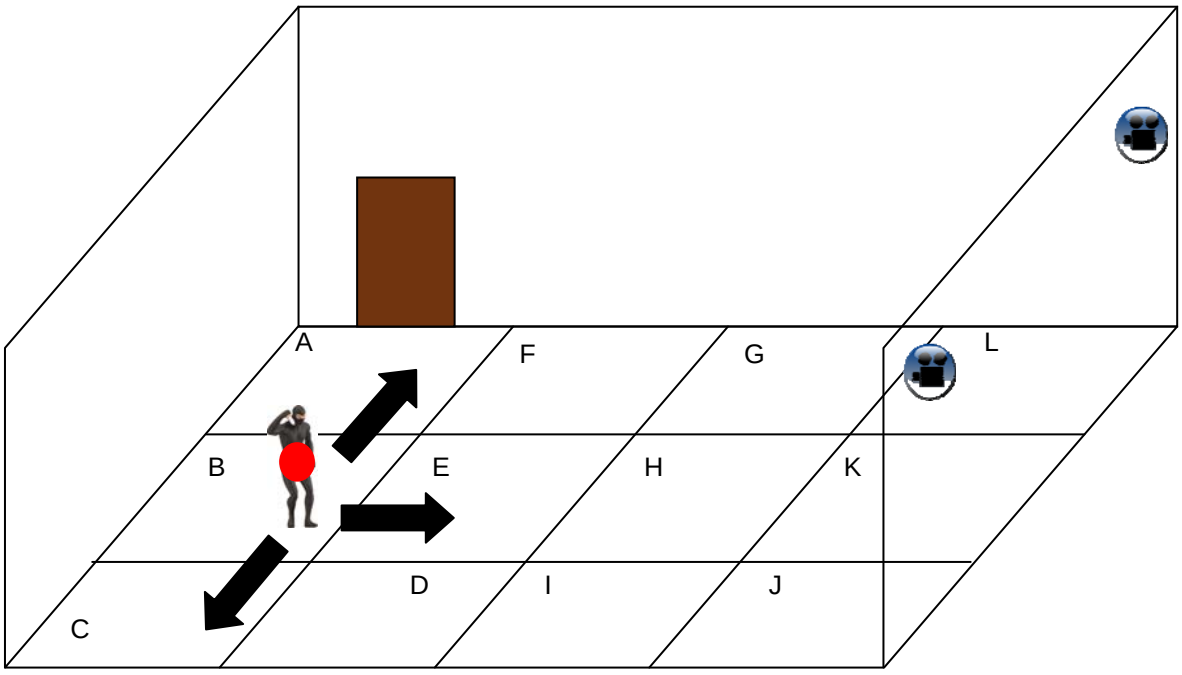


Figure 14 - the walking man is in the area "B"

Starting from this position, in the real world (with the constraint of the quad-connection schema), the man has three possible chooses of movements: he can go back into the position "A" or he can go forward in position "C" or he can turn left and go in position "E".

The formal rule of the grammar under definition encompassing these three options is:

B → bA | bC | bE | b

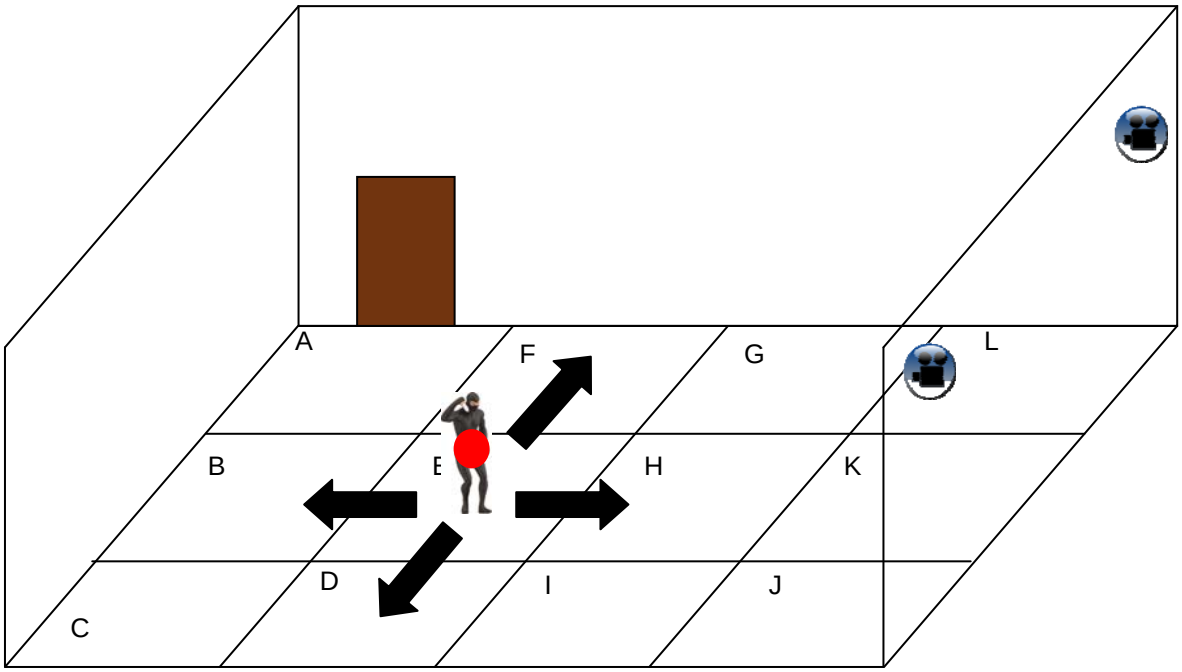


Figure 15 - the walking man is in the area "E"

One of the areas with the greatest number of possible movements is that labelled with the symbol “E”. As shown in Figure 15, starting from the area “E”, the man can go in “F” or in “H” or in “D” or in “B”.

The formal rule of the grammar under definition encompassing these four options is:

E → eH | eD | eB | eF | e

The complete grammar G=(V, Σ, P, S) for this example can be defined in the following way:

V = {A, B, C, D, E, F, G, H, I, J, K, L}

Σ = {a, b, c, d, e, f, g, h, i, j, k, l}

S = {A}

P is composed of the following rules:

S → A

A → aB | aF | a

B → bA | bC | bE | b

C → cB | cD | c

D → dC | dI | dE | d

E → eH | eD | eB | eF | e

F → fA | fE | fG | f

G → gF | gH | gL | g

H → hG | hI | hE | hK | h

I → iD | iH | iJ | i

J → jI | jK | j

K → kJ | kH | kL | k

L → lK | lG | l

Analysing the grammar written for this example, it is possible to make the following considerations:

- On this grammar, it is possible to define the language L(G). This language is composed of the set of strings describing all the continuous paths that are possible into that environment. Hence, all the continuous trajectories will generate strings belonging to L(G). Hence, a system implementing this grammar can be used to find discontinuous trajectories. In real world applications, a discontinuous trajectory exists when the tracker loses the moving object for an interleave of time. **From this point of view, this grammar can be used as a system to recover the trajectories of moving objects in crowding scenes**.
- Each rule produces either a couple (terminal, non-terminal) symbols or a single terminal symbol. This means that this language contains strings ending with any symbol $x \in \Sigma$. In the real world, this means that this language contains all the strings describing trajectories starting from the area "A" and ending everywhere into the room. In order to build a more realistic model of the room in Figure 15 that has only a gateway (the door in area "A"), the previous rules can be rewritten in the following way:

  S → A

  A → aB | aF | a

  B → bA | bC | bE

  C → cB | cD

  D → dC | dI | dE

  E → eH | eD | eB | eF

  F → fA | fE | fG

  G → gF | gH | gL

  H → hG | hI | hE | hK

  I → iD | iH | iJ

  J → jI | jK

  K → kJ | kH | kL

L → lK | lG

This set of rules describes all the closed trajectories starting from the area “A”. Indeed, the only non-terminal symbol that can produce only a terminal symbol is “A”. This is a realistic condition in an indoor environment with a single gateway as that shown in Figure 15.

In Figure 16 a more complex indoor scenario is presented. Also in this scenario the environment has a single gateway, i.e., the door in the area “A”. The environment has been partitioned into twelve areas. Each area has been labelled with a capital letter of the English alphabet. In the areas labelled with the symbols “E” and “H” is positioned a desk. In this scenario, a walking man can go everywhere except for the areas “E” and “H” due to the presence of the desk.

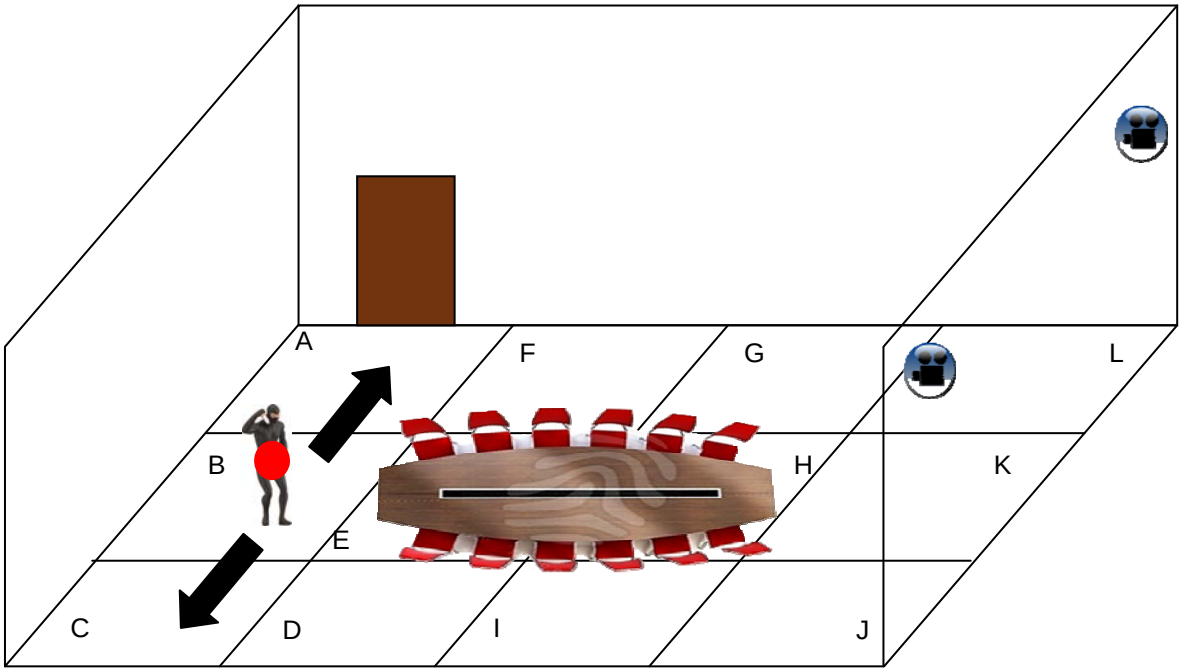


Figure 16 - a schematic representation of an indoor environment with a desk in the areas "E" and "H"

Also this scenario can be described with a grammar G=(V, Σ, P, S) where:

V = {A, B, C, D, E, F, G, H, I, J, K, L}

Σ = {a, b, c, d, e, f, g, h, i, j, k, l}

S = {A}

The set P must contain the production rules allowing for the definition of a language L(G) that is able to describe all the possible trajectories into this environment. The start symbol and the rule for the entering area “A” are the same of the previous example, namely:

S → A

A → aB | aF | a

Figure 16 shows another possible configuration: the man is entered into the area labelled with the symbol “B”.

This time, starting from this position, in the real world (with the constraint of the quad-connection schema), the man has only two possible choices of movements: he can go back into the position “A” or he can go forward in position “C”. He can not turn left and go in position “E” due to the presence of the desk.

The formal rule of the grammar under definition encompassing these three options is:

B → bA | bC

In this way, if at run-time, the recording system generates a string containing the combination of symbols “be”, then the moving object is acting a forbidden behaviour (he is walking on the desk !!!) because this string does not belong to this grammar. The complete set P is composed of the following rules:

S → A

A → abs | aft | a

B → bad | be

C → cob | cod

D → do | did

F → far | fig

G → go | gal

I → ad | in

J → jig | joke

K → kJ | kill

L → elk | lag

This set of rules takes in account the presence of a single gateway in the area “A”, so L (G) is composed of all the strings describing closed continuous trajectories starting from the area “A”. L (G) does not contain strings with the symbols “E” and “F”. **From this point of view, this grammar can be used both as a system to recover the trajectories of moving objects in crowding scenes and as a system that is able to raise alarms when forbidden behaviours are recognized.**

5.6 The proposed methodology and the time

As shown above, the proposed methodology allows the implementation of systems for high level semantic analysis of human behaviour in a given scenario.

Figure 17 shows an example of a more complex scenario than that analysed in the previous examples. This time, the room contains various elements: a printer in area "L", a PC station in "K", a professor's desk in area "J" and a plotter in area "C".

In this scenario, it is possible to say that the string "alfalfa" means "a man is entered into the room, he has taken a print and then he has gone out of the room". In the same way, it is possible to say that the string "afghijihgfa" means "a man is entered into the room, he has gone to the professor desk and then he has come back".

The proposed methodology allows for attributing a high semantic level description to both the events. Nevertheless, a certain level of ambiguity still remains: who was the man that went to the printer? Was he a user that took a print or a technician that repaired it? In the same way, who was the man that went to professor's desk? Was he somebody that went to the desk to take a document or the professor that went to his desk to work?

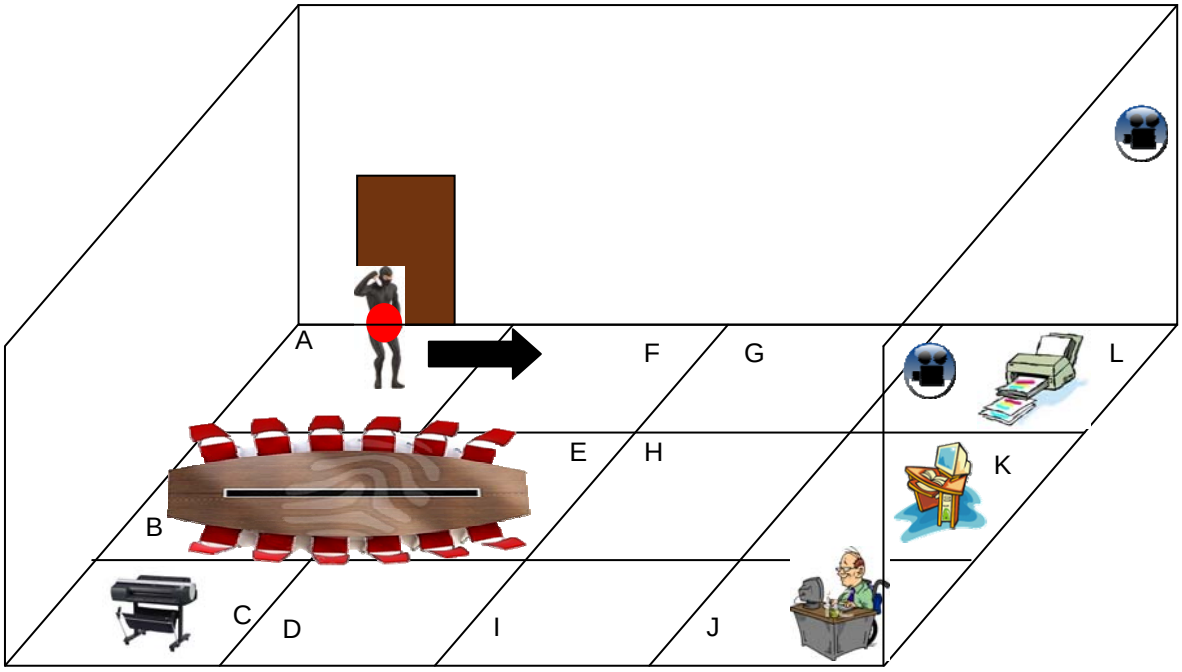


Figure 17 - an example of complex indoor scenario

Since this methodology is suitable for systems working in the narrow domain, a possible answer to these questions could result from the analysis of another parameter: **the time**.

Supposing that the room in Figure 17 is a laboratory of a University, it is possible to add more knowledge to the model. For example, in the first example, if somebody goes to the printer to take a print, there is a high probability that this task is executed in few time. Vice versa, if a technician does a work of maintenance of the printer, there is a high probability that this task requires more time than the previous one.

The time plays a fundamental role also in the second scenario. If the man arrives in "J" and stays there for a while, it is very probable that he is the professor. Vice versa, it is probable that he is an assistant taking a document from the professor's desk.

The proposed methodology uses a system for string generation (described in Figure 11) that does not take into account the time of permanence of a moving object in a given area. Indeed, as stated above, in the scenario represented in Figure 11.a, where the walking man is moving inside the same area, the system does not produce any symbol. Nevertheless, as further proof of the flexibility of the proposed methodology, in the following **a methodology to handle the time is proposed**.

Thanks to the modularity of the proposed approach, the only module to modify in order to take into account the time is the *trajectory encoder* (see Figure 10). Indeed this module evaluates the coordinates of a given moving object. If these coordinates belong to the same area of the previous recorded point, the system does not generate any symbol. Vice versa, if the coordinates belong to a new area, the system generates the symbol corresponding to its label.

From this perspective, it is possible to say that the *trajectory encoder* has an **event driven** behaviour because it generates symbols when a new event (namely a change of area) is recognized.

In order to consider the time, the *trajectory encoder* must have a **time driven** behaviour. In this configuration, it generates a symbol ever T seconds. The value of T is a constant for the system and it must be chosen at design time according to the dynamic of the analysed environment.

A little value of T makes the system more reactive to the rapid changes. In other words, using a small T, it is possible to analyse the behaviour of human beings that are moving in a fast way. An example of real world environment where these kinds of behaviours are common can be the corridor of an airport or of a subway station. In these scenarios, typically the people go from a point to another of the scene as quick as possible.

Using a big value of T, the system updates the positions of the tracked people more slowly. In this way, it is possible to analyse the behaviour of people moving very slowly or

staying in a fixed position for a while. Some examples of real world environments where these kinds of behaviours are common can be: a waiting room, the queue in some offices, the stands in the football stadium.

A small value of *T* forces the *trajectory encoder* to generate a long word to describe a behaviour while a big value of *T* gives short words. The choice of the correct value of *T* for a given environment must be carried out analysing at design time the specificity of the environment where the system will work. On the other hand, this parameter is not too critical from a computational point of view, because, thanks to the kind of grammar used to generate the language describing the behaviours, the string analysis process has a linear time complexity.

Thanks to this switching of the *trajectory encoder* from event driven to time driven, it is possible to resolve, at certain extent, the examples of ambiguities highlighted in the environment of Figure 17.

Indeed, the behaviour of "a man is entered into the room, he has taken a print and then he has gone out of the room" will generate a string similar to: "afglgfa" while the behaviour of a technician repairing the printer can be described by: "afgllllllllllllgfa".

In the same way, the assistant that goes and takes a document from the desk of the professor generates the following word: "afghijihgfa" while the professor that goes and works to his desk can generate a string like "afghijjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjihgfa".

Hence, introducing the concept of succession of the same symbol into a string and **generating new symbols in time driven modality, it is possible to consider the time using the proposed methodology**.

The grammar used to generate the language describing the allowed behaviour in the environment shown in Figure 17 can be the following:

G=(V, Σ, P, S) where:

V = {A, B, C, D, E, F, G, H, I, J, K, L}

Σ = {a, b, c, d, e, f, g, h, i, j, k, l}

S = {A}

The terminal and the non-terminal symbols sets and the start symbol remain the same of the previous example because the environment has been partitioned and labelled in the same way. The production rule set becomes:

1. S → A
2. A → aF | aA | a
3. C → cD | cC

4. D → dC | dI | dD
5. F → fA | fG | fF
6. G → gF | gL | gH | gG
7. H → hK | hG | hI | hH
8. I → iD | iJ | iH | iI
9. J → jI | jK | jJ
10. K → kJ | kL | kH | kK
11. L → lK | lG | lL

These rules take into account the fact that the *trajectory encoder* can generate two or more consecutive times the same symbol. Figure 18 shows an example of trajectory for a man who goes from the area "H" to "D" staying for 2x*T* in the area "I". The movements are analysed into the following instants of time:

T_0 the man is in the area "H".

$T_1 = T_0 + T$ the man has gone from "H" to "I"

$T_2 = T_0 + 2*T$ the man has walked between two points of the area "I"

$T_3 = T_0 + 3*T$ the man has gone from the area "I" to that "D"

As the figure shows, at the time T_0 the man is in the area "H". At T_1 the man is arrived in the area "I" starting from "H". This event is recognized and, using the third option of the rule number 7 of the grammar (see the above list), the trajectory encoder writes the symbols "hI" into the string that describes the motion of this man. It should be noticed that "h" is a terminal symbol while "I" is a non terminal one. At T_2 the man is in a position different from that in T_1 but belonging to the same area "I". Also this event is recognized and the trajectory encoder writes the symbols "iI" using the fourth option of the rule 8. Finally, at T_2 he arrives in "D". This time the trajectory encoder writes the symbols "iD" using the first option of the rule 8. At this time the string describing this piece of trajectory is "hiiD" where "h" and "ii" are terminal symbols and "D" is a non terminal one.

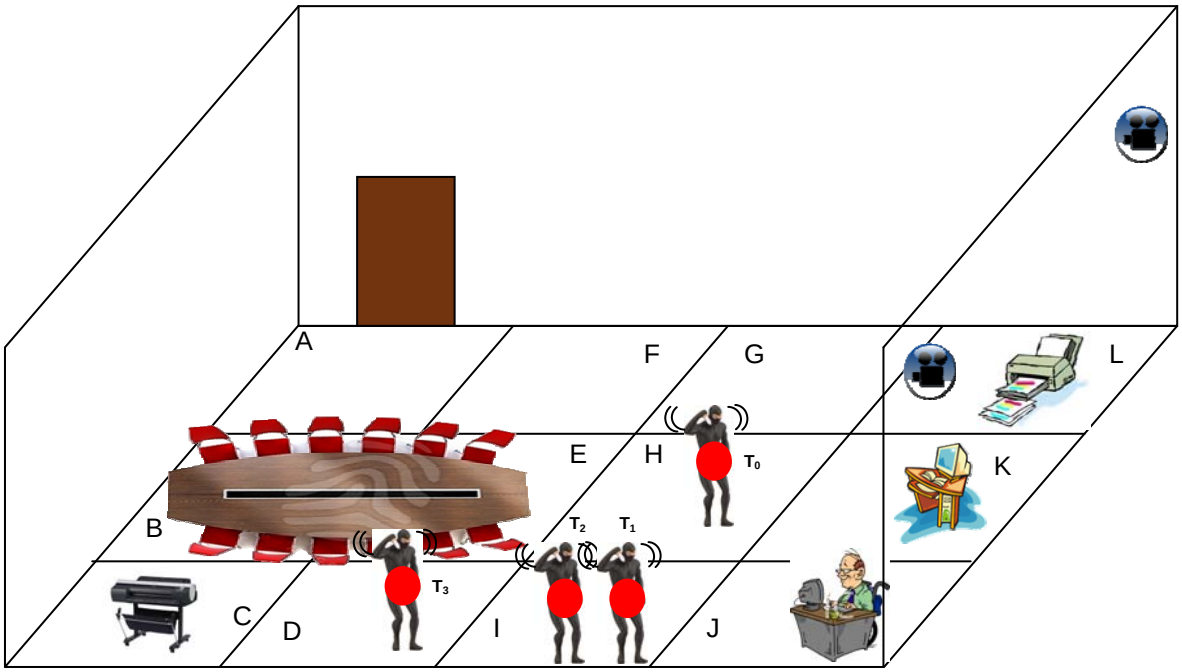


Figure 18 - An example of man moving into the room and staying in the same area for 2x*T* seconds

Using this grammar, it is possible to write strings describing all the allowed behaviours in this environment. Since forbidden behaviours (such as "going on the desk) are not described by any rules, this grammar can be used to generate alarms when it recognizes a string that does not belong to the language L(G). Furthermore, the strings written by the *trajectory encoder* module are stored into a database that can become the knowledge base of an *expert system* that is able to perform various tasks such as semantic behaviour classification and video indexing.

5.7 Hierarchical scene analysis

The proposed methodology allows for creating complex systems to perform a hierarchical analysis of the scenes. Indeed, it allows for performing a hierarchical analysis of the recorded scenes defining a different grammar for each level. At the higher level, the behaviour of the human beings moving in the scene is analysed studying their motion parameters (the trajectory that they are following). This process is shown in the left part of Figure 19 and has been described in the previous sections. According to the classification proposed in the related works section, this approach can be considered as belonging to the "scene analysis" class. In this way it is possible to have an analysis of scenario considering the moving objects. At the lowest level, the system can produce a detailed analysis of the action taken by each single human being in the scene.

As shown in Figure 19, some areas of particular semantic interest can be partitioned in a more detailed way. Defining a specific grammar for this area, it is possible to obtain a more detailed semantic analysis of the actions performed in it. At the finest level of detail, this approach can be used in the context of gesture analysis (see the right part of Figure 19). From this perspective this system can be considered as belonging to the class of human recognition systems.

According to the complexity of the task, between these two levels, it is possible to define as many intermediate levels as they are necessary. This hierarchical analysis can exploit the full potentiality of the modern video surveillance systems where there is a fixed camera of scenario and one or more moving cameras that can focus their attention on some areas of interest. In this application, the first level of the proposed hierarchy is applied to the camera of scenario and the second to the moving cameras.

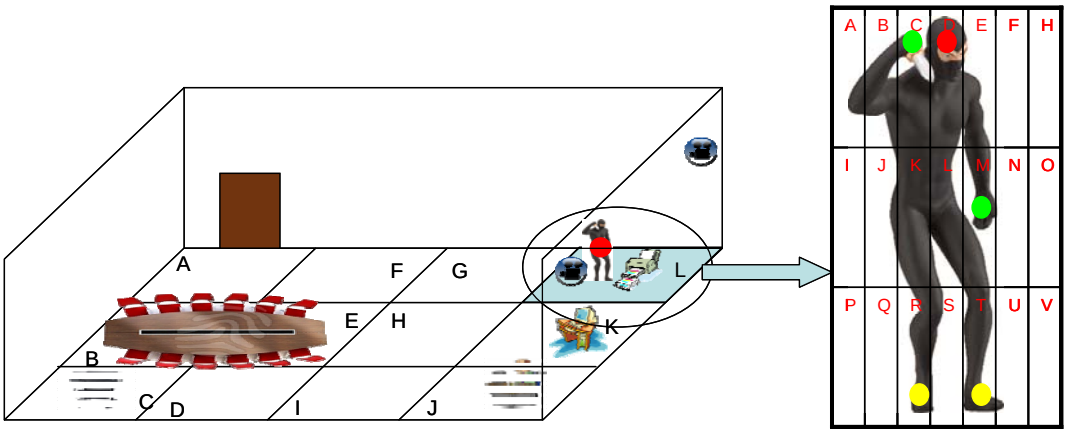


Figure 19 - An example of hierarchical decomposition of the scene

Summary

In this chapter, the proposed methodology for semantic analysis of video streaming has been presented. This methodology can be seen as a unifying approach encompassing the three main approaches to human behaviour analysis existing in literature (scene interpretation, human recognition and action primitive and grammars). Indeed, this methodology allows for a hierarchical analysis of the recorded scene. According to the used level of detail in scene recording, this methodology can provide from a semantic analysis of the whole scene till a detailed behaviour analysis of a single person.

This methodology is designed for systems belonging to a narrow domain. The knowledge about the domain is expressed by means of the labelling process. This is a key process enabling the double context-switch from human behaviours to trajectories and from trajectories to strings. In this way, this methodology faces the problem of semantic analysis of human behaviour as a linguistic problem.

The language composed of all the words corresponding to allowed human behaviours is obtained using a generative process defining a specific grammar for each domain. This is an interesting aspect of the proposed methodology that overcomes the issues related to the heavy learning process used by other methods.

The trajectory encoder describes all the possible trajectories using the label of the various areas. In this way it is possible to describe complex trajectories such for example zigzag and backward motions. It will be the expert system (Figure 10) that will be able to answer to query about the presence of such kind of trajectories.

The proposed methodology can be used in the implementation of advanced systems for: video surveillance, semantic video indexing, control applications, etc.

Chapter 6

Evaluation of the proposed methodology

6.1 Introduction

In this chapter a discussion about the possible applications of this methodology is proposed. Furthermore, the experimental results obtained applying this methodology to a surveillance systems are shown. Also the proposed solution to the correspondence problem has been tested and the obtained results are shown in this chapter.

6.2 Possible applications of the proposed methodology

The proposed methodology is suitable for the design and implementation of systems for semantic analysis of human behaviour in streaming video. Such kinds of systems have a great number of possible applications in various fields:

- **Control systems**. In this framework, the proposed methodology allows for the

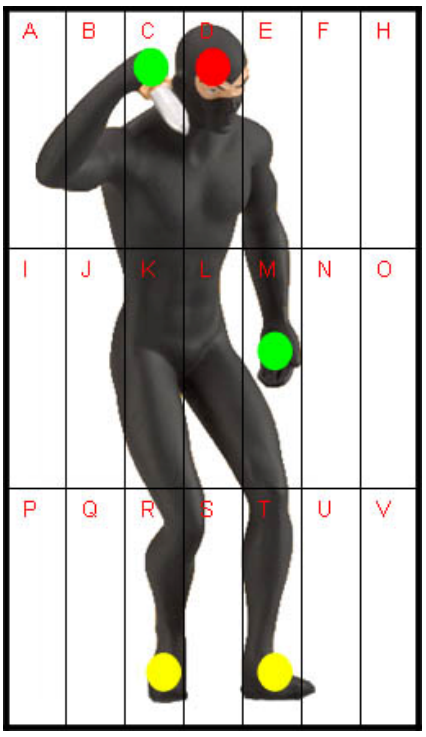


Figure 20 - application to control systems

implementation of advanced human-computer interfaces. According to the specific application, the system will focus on the tracking of some relevant body parts (e.g., hands, head, feet). The recorded scene can be partitioned and labelled as shown in Figure 20. Since the trajectory encoder is time driven, it is possible to write a word for the trajectory of each body part without the problem of synchronization. Using a grammar it is possible to define the possible trajectories for each body part. Analysing the produced strings it is possible to infer the pose of the subject (in the example of Figure 20 "man with raised right hand".

- **Analysis systems**. Applying the same method shown for the control system, it is possible to study also the pose of athletes during their performance. Using a specific grammar, it is possible to write the words corresponding to perfect exercises. In this way, it is possible to implement systems suitable for the processes of training and evaluation of the athletes' performances.
- **Surveillance systems**. As shown in the previous chapter, using the proposed methodology it is possible to design and implement systems for high semantic level

Figure 21 - a screenshot of the monitored room

analysis of human behaviours. Hence, it is possible to implement video surveillance systems that are able to recognize the behaviours of the monitored people and raise alarms when forbidden behaviours are recognized.

6.3 An example of application: video surveillance system

To show the effectiveness of the proposed methodology, a demonstrative video surveillance system has been implemented and applied to a real indoor environment with valuable results. This system has been used also to evaluate the performance of the proposed solution to the correspondence problem in multi-camera systems.

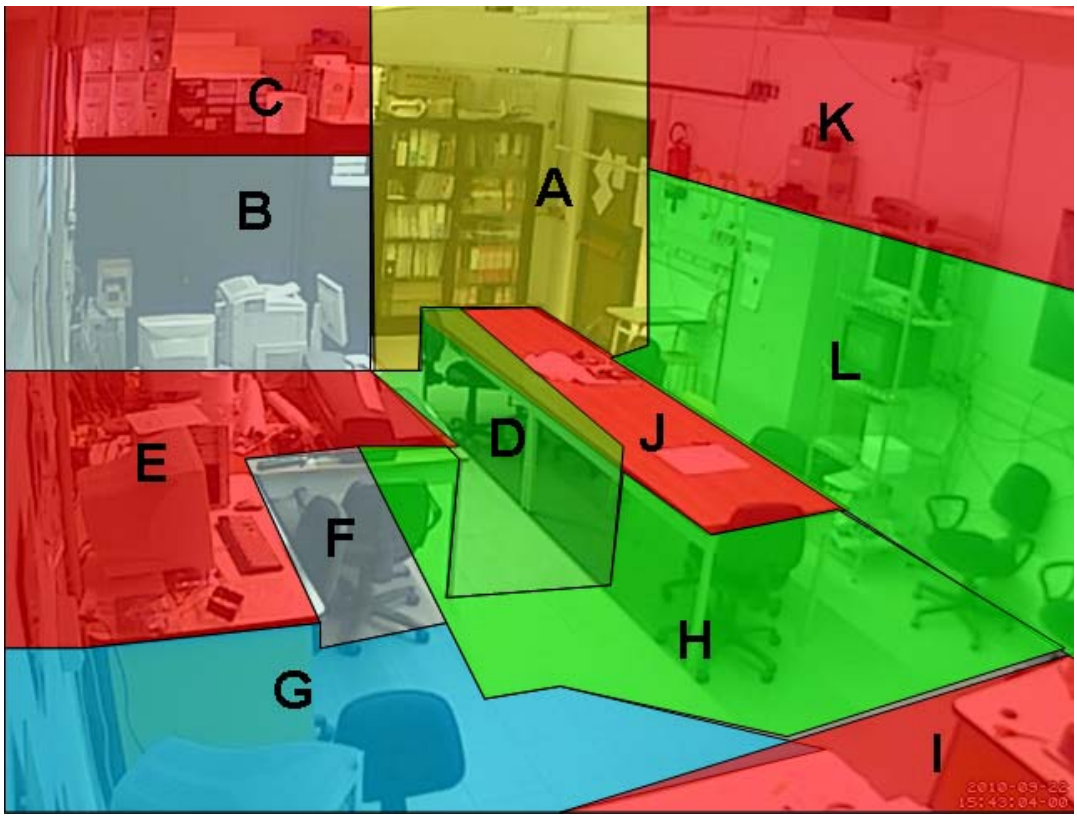


Figure 22 – an example of the area partition and mapping

The system was tested carrying out numerous experiments in a set of indoor environments of our Faculty. In this thesis, the tests carried out in a single room of our Faculty are reported to show how the system works and its potentiality. A comparative evaluation with other systems present in literature is quite difficult due to the lack of a common accepted testbed database. The proposed system can work also in outdoor environments. In this case, only the used motion detection algorithm should be changed (for example using one modelling the background as a Gaussian mixture [18]).

A screenshot of the monitored room is shown in Figure 21.

The room was virtually divided into a partition as shown in Figure 22. In this example, the red areas (namely the area labelled “C”, “E”, “I”, “K”, “J”) are forbidden. In the areas “B” and “F” there are two PC stations. The area “G” is in front of the professor’s desk. The area “A” is the gateway of the room while areas “D”, “H” and “L” are areas where it is possible to walk.

Using these labels, it is possible to associate semantic meaning to each of them and also to their combinations. For example, if the trajectory encoder produces a string with a large number of contiguous “G”, there is a high probability that someone is speaking with the professor. In the same way, strings with large numbers of contiguous “F” or “B” mean that somebody is working on the PC in the area “B” of “F”.

The grammar used to generate the language describing the allowed behaviour in the environment shown in Figure 22 can be the following:

G=(V, Σ, P, S) where:

V = {A, B, C, D, E, F, G, H, I, J, K, L}

Σ = {a, b, c, d, e, f, g, h, i, j, k, l}

S = {A}

The production rules are:

1. S → A
2. A → aB | aD | aL | aA | a
3. B → bA | bB
4. D → dA | dH |dD
5. F → fH | fG | fF
6. G → gF | gH | gG
7. H → hD | hF | hG | hL | hH
8. L → lH | lA | lL

The language L(G) is composed of all the words describing human behaviours considered “normal” at design time. These rules allow behaviours like: walking in the room with whatever trajectory, going to some working stations and staying there for a while and do on.

By the other hands, they do not allow other behaviour such as: going on the desks, or

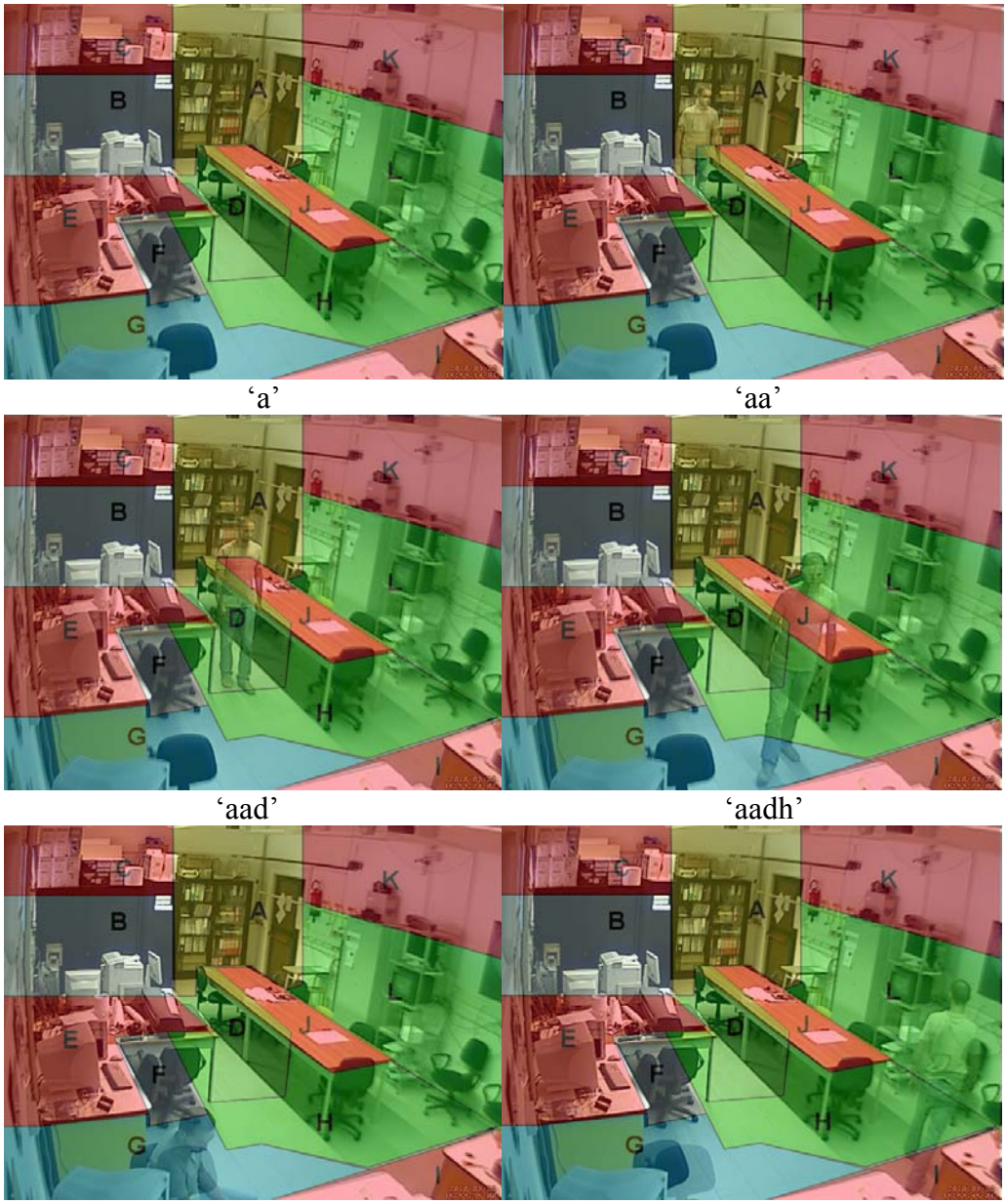


Figure 23 - An example of output of the trajectory encoder. For each analysed frame, the label of the area containing the barycentre of the man is appended to the string. All these strings are recognized by the grammar G and so no alarms are raised.

climbing on walls. Particular attention requires the area “C”. This area comprises a set of PC positioned on a cabinet. These PCs belong to a cluster of workstation. Typically these PCs are managed by remote stations. The only rare operations of ordinary administration can be: power on, insert a DVD, etc. To avoid that the system raises alarms during these operations, the area “C” was sized in a specific manner. In particular, its size was defined in such a way that during the normal operations, only the user’s arms enter in this area. In this way, the barycentre of the user still remains in the area “B” without alarms.

Figure 23 shows an example of the output of the trajectory encoder. In this figure, six frames of a video of a walking man are shown. In this sequence, the man enters the door (area “A”), walks into the room (areas “DH”) and arrives to the area “G”. Then he comes back. Under each frame the output of the trajectory encoder module is shown.

Since in this sequence the trajectory encoder writes a word belonging to L(G), the system does not raise alarms.

Figure 24 shows a situation where a man has a forbidden behaviour (he goes over a desk). In the first frame, he enters in the area “A”, so the trajectory encoder creates a void string and appends the first symbol “a”. In the second frame, he is moving in the room but he is still in area “A”, so again the encoder appends a symbol “a” to the string describing the trajectory of this man.

Till now, the system does not raise alarms because the string “aa”, describing the trajectory of the man, belongs to the language L(G). Indeed, it can be produced by applying two consecutive times the production rule number 2.

In the third frame, the man jumps over the desk and his barycentre is in the area “J”. This time the trajectory encoder appends the symbol “j” to the string. Now the string becomes “aaj”. This string does not belong to the language L(G) because there is no one combination of production rules allowing for writing it.

‘a’ ‘aa’

‘aaj’

Figure 24 - An example of output of the trajectory encoder. For each analysed frame, the label of the area containing the barycentre of the man is appended to the string. Some of these strings do not belong to L(G) and so alarms are raised.

6.4 Test of the proposed solution to the correspondence problem

Some experiments have been carried out to evaluate the effectiveness of the proposed solution to the correspondence problem in multi-camera systems. They have been carried out in the same room of the previous ones using an additional camera as shown in the schema in Figure 25.

T_1 and T_2 are the two used cameras. They are installed at a distance of about three meters from the ground plane. In these experiments two Axis 210 network cameras were used. These cameras are able to sample up to 30 frames per second but in these experiments only 15 frames per second were used. This because a higher frame rate does not introduce more details in the correspondence problem process. Indeed, using more than

15 frames per second the differences between two successive frames sampled by the same camera are negligible.

The SOM is trained using various manually selected relevant points of objects present in the frames sampled by both the cameras. The system was developed using the Matlab® environment. The proposed results are obtained using a SOM with 300 nodes.

In all the experiments both α and β are set to 0.5. In this way both colour and texture

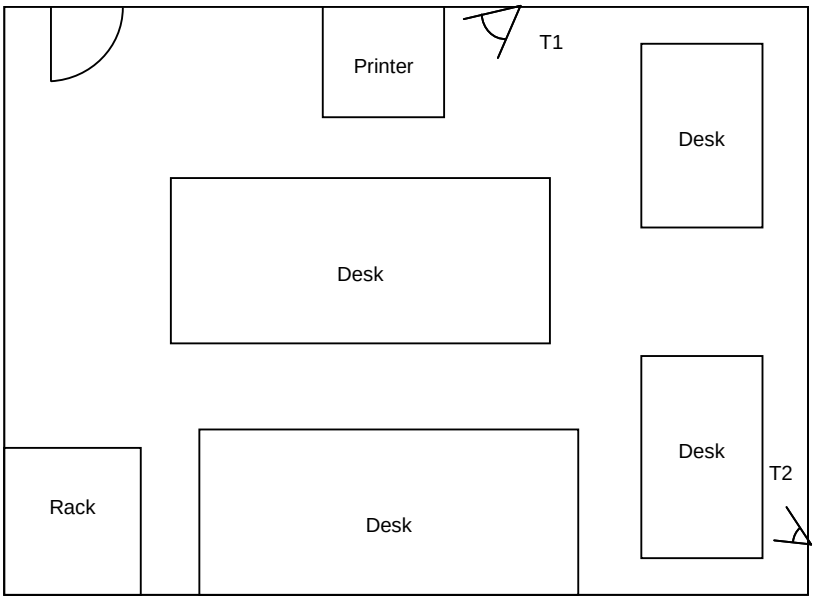


Figure 25 - a schematic view of the experiment setup. T_1 and T_2 are the

features and SOM and feature based method are considered with the same relevance. An in-depth research of optimal values for α and β will be matter of future works.

In order to test the system, various sequences were sampled where one or more people are moving in the room. Figure 26 shows an example of the obtained results. In the first column there are the frames as they are sampled by the cameras. The recorded scenes are quite complex because both cameras see partial occlusions. Indeed, there is a desk occluding the low part of the walking people. Furthermore there are various "dynamic" occlusions namely occlusions due to the overlap of walking people (this kind of occlusions is typical of crowded scenes). In particular, the first column of figure 3 shows a frame where two people are walking without "dynamic" occlusions.

The second column shows the output of the proposed object detection algorithm applied to both the cameras.

The third column shows the results obtained by the proposed solution to the correspondence problem applied to these frames. The corresponding moving objects into the two frames are coloured using the same colour.

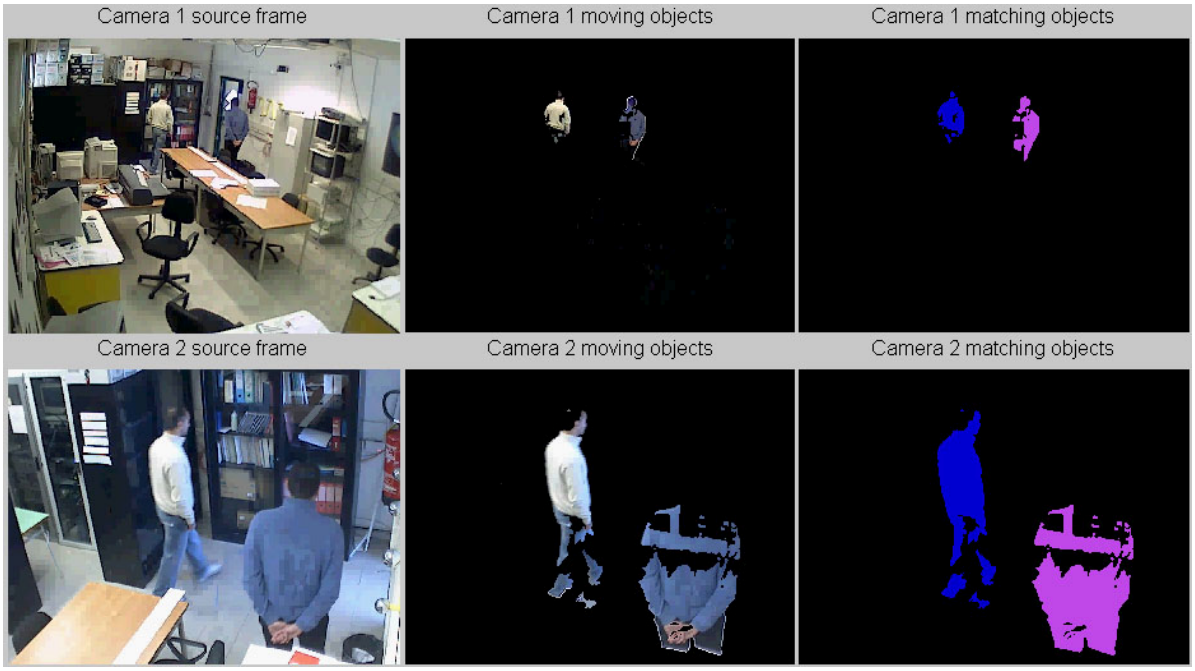


Figure 26 - An example of the obtained results. Starting from the left column: source frames, object detection algorithm output, found object correspondences. The corresponding moving objects into the two frames are colored using the same colour.

Table I shows some statistics about the mean results obtained by analysing ten videos where various people are walking into the room. This table shows the obtained recall and the percentage of detected false matching for videos where there were respectively one person, two and three people walking. The obtained results highlight the effects of occlusion on the proposed system. Indeed, when there is a person walking in the room, the recall is 94% and the false matches are 4%. These values are due to the static partial occlusions due to the presence of the desk in the scene. When there are two and three people walking in the room, dynamic occlusions have a clear effect on the performance of the proposed system. Indeed the recall decreases to 72% for two people and 77% for three people while the percentage of false matching increases to 24% and 28% for two and three people walking respectively. These results highlight the robustness of the system to this kind of occlusions. Indeed the recall and false matching remain almost constant in the experiments with two and three people walking.

Table 1 - results obtained analysing ten different videos

Number of people	**1**	**2**	**3**
Recall (%)	94	72	77
% false match	4	24	28

Summary

In this chapter a discussion about the possible applications of this methodology and the proposed experiments have been reported.

The proposed methodology has a great flexibility. It can be used to describe the actions at various levels of detail. In this way it is possible to implement specific applications in the field of: control, analysis and surveillance systems.

For example, in the proposed test application, the methodology has been applied to implement a surveillance system. It can describe the actions of a human being specifying where he went and how long he was in each position. It should be noticed the high semantic level of the action analysis, indeed the system consider positions as "semantic places". So, considering Figure 22, when the system recognizes that a human being stays for a given period of time in the position "G", it is possible to say that automatically that "a human being stays for a given period in front of professor desk"

This level of details is more than sufficient for the most parts of the video surveillance applications but thanks to the hierarchical approach proposed in this methodology, it is possible to implement a further level of detail where the actions performed in some relevant areas are described. For example, focusing the attention on the area "G" of Figure 22 (the area in front of the professor's desk), it is possible to describe the gesture of the people in that area applying the same methodology. In this way it is possible to implement hierarchical systems having a different level of detail for each level of the hierarchy.

In this chapter has been reported also the results of some experiments carried out to evaluate the performance of the proposed solution to the correspondence problem.

It is difficult to compare the obtained performance with other works in literature. Indeed, in literature there are two main classes of works dealing with the correspondence problem:

1. works aiming to solve the problem for all the points of the images. These works try to obtain a dense disparity map and they evaluate the performance analyzing the percentage of correctly mapped pixels.
2. works using the multi-camera vision in video analysis applications (i.e. video surveillance systems). These works do not give too emphasis to the correspondence problem in the performance evaluation sections.

Starting from these considerations, in Table 1 the obtained results are presented in terms of recall and percentage of false matching because these parameters can give to the reader an objective idea of the system performance.

Conclusions

In this thesis an innovative methodology to implement systems for high semantic level analysis in the narrow domain has been presented.

The methodology proposes a double context switch. The first one is from human motion to barycentre trajectory and is a well known and accepted method in literature. The second one is an original contribute of this thesis and is from trajectory to word. It is based on the domain of application of this methodology, namely the narrow domain. The external knowledge is introduced into the system by labelling with a set of symbols the various areas in which the scene is partitioned. The *trajectory encoder* produces a word for each trajectory. The portion of the trajectory falling into a given area is coded using the label of that area. In this way, a trajectory is transformed into a word by means of the concatenation of all symbols labelling the areas on which the curve lies. Furthermore, since the *trajectory encoder* has a **time driven** behaviour, this methodology is able to handle the issues related to the different execution times of the actions.

Since in the narrow domain it is possible to attrib a semantic value at each area and thus at each symbol, using this methodology it is possible to achieve a high semantic level description of the recorded scene.

This methodology uses a robust approach to the problem of strings/words recognition. Once that the environment has been labelled, a grammar on the set of symbols used to label the scene is defined. The production rules set of this grammar contains the rules to describe all the allowed behaviours in a given scenario. Using this grammar it is possible to define a language composed of the set of recognizable words (and thus the allowed behaviours).

This methodology allows for implementing hierarchical analysis of the scenes defining a different grammar for each level. At the higher level, the behaviour of the human beings moving in the scene are analysed studying their motion parameters (the trajectory that they are following). In this way it is possible to have an analysis of scenario considering the moving objects and their interactions. At the lowest level, the system can produce a detailed analysis of the action taken by each single human being in the scene. According to the complexity of the task, between these two levels, it is possible to define as many intermediate levels as they are necessary. This hierarchical analysis can exploit the full potentiality of the modern video surveillance systems where there is a fixed camera of

scenario and one or more moving cameras that can focus their attention on some areas of interest. In this application, the first level of the proposed hierarchy is applied to the camera of scenario and the second to the moving cameras.

From this point of view, this methodology can be seen as a unifying approach encompassing the three main approaches to human behaviour analysis existing in literature (scene interpretation, human recognition and action primitive and grammars).

Another element of innovation of this methodology is the method used for human behaviour description and recognition. Indeed, this methodology proposes a generative approach to human behaviour recognition. Defining a specific grammar *G* for a given domain, the system uses its rules to describe a human behaviour writing a word. The system is able to recognize all the behaviours that can be described using a word belonging to the language defined on the grammar *G*. This is a strong improvement in comparison to many works in literature that are able to recognize only a finite set of actions learnt in a training stage.

This thesis gives also an original contribute to two relevant issues in this research field: sensory gap and partial occlusions proposing a novel solution to the correspondence problem in multi camera systems. This solution is based on the fusion of two approaches: SOM and CBIRs. The SOM is used to create a sort of feature based mapping between some relevant points into the two images while the CBIRs based module describes the detected moving objects present into the two images using two low level visual features (colours and texture). Using this description, this module finds the correspondence among the moving objects present in the two images.

Future developments

The proposed methodology allows for a hierarchical analysis of the recorded scene. According to the abstraction level of details used in scene recording, this methodology can provide from a semantic analysis of the whole scene till a detailed behaviour analysis of a single person. This architecture has great potentialities but much work is to be done in the definition of the various levels.

In particular, in future works the aspects related to the interactions among human beings can be further analysed. Since the *trajectory encoder* is time driven, it is possible to maintain a good level of synchronization among the strings describing the behaviours of the various people in the scene. This fact can be the basis for the design of grammars on which to define languages describing various interactions among human beings.

Another aspect to be further investigated is the use of this methodology to implement systems for high semantic level video indexing. The strings produced by the *trajectory encoder* are stored into a database (see Figure 10) called "*action log*". Searching for a given string into this database means searching for a given behaviour. From this point of view, it is possible to apply techniques of clustering of words on this database to produce a more in depth comprehension of the observed behaviours. This database contains data characterized by a high semantic value, but it requires an advanced human interface that should be able to exploit all the system potentiality.

A limit of the systems implementing the proposed methodology is the fact that they are view dependent. Indeed, the partitioning and labelling of the scene is referred to the camera view. In future works, a solution to this problem can be searched in the framework of the computer vision field and in particular of the CBIR systems.

References

[1] Lymberopoulos, D.; Teixeira, T.; Savvides, A.; , "Macroscopic Human Behavior Interpretation Using Distributed Imager and Other Sensors," Proceedings of the IEEE , vol.96, no.10, pp.1657-1677, Oct. 2008 doi: 10.1109/JPROC.2008.928761

[2] Ko, T.; , "A survey on behavior analysis in video surveillance for homeland security applications," Applied Imagery Pattern Recognition Workshop, 2008. AIPR '08. 37th IEEE , vol., no., pp.1-8, 15-17 Oct. 2008 doi: 10.1109/AIPR.2008.4906450

[3] Heckenberg, D.; , "Performance Evaluation of Vision-Based High DOF Human Movement Tracking: A Survey And Human Computer Interaction Perspective," Computer Vision and Pattern Recognition Workshop, 2006. CVPRW '06. Conference on , vol., no., pp. 156, 17-22 June 2006 doi: 10.1109/CVPRW.2006.157

[4] Saito, H.; Inamoto, N.; Iwase, S.; , "Sports scene analysis and visualization from multiple-view video," Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on , vol.2, no., pp.1395-1398 Vol.2, 30-30 June 2004

[5] Mizuno, H.; Nagai, H.; Sasaki, K.; Hosaka, H.; Sugimoto, C.; Khalil, K.; Tatsuta, S., "Wearable Sensor System for Human Behavior Recognition (First Report: Basic Architecture and Behavior Prediction Method)", in International Conference on Solid-State Sensors, Actuators and Microsystems Conference, 2007. TRANSDUCERS 2007. Publication Year: 2007 , Page(s): 435 - 438

[6] Isoda, Y., Kurakake, S., and Imai, K. 2008. Ubiquitous sensor-based human behaviour recognition using the spatio-temporal representation of user states. Int. J. Wire. Mob. Comput. 3, 1/2 (Jul. 2008), pp. 46-55. DOI= http://dx.doi.org/10.1504/IJWMC.2008.019717

[7] Smeulders, A.W.M.; Worring, M.; Santini, S.; Gupta, A.; Jain, R.; , "Content-based image retrieval at the end of the early years," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.22, no.12, pp.1349-1380, Dec 2000 doi: 10.1109/34.895972

[8] Santini, S.; Jain, R.; , "Beyond query by example," Multimedia Signal Processing, 1998 IEEE Second Workshop on , vol., no., pp.3-8, 7-9 Dec 1998 doi: 10.1109/MMSP.1998.738904

[9] Philipose, M., Fishkin, K., Patterson, D., Perkowitz, M., Hahnel, D., Fox, D., and Kautz, H. Inferring activities from interactions with objects. IEEE Pervasive Computing Magazine 3, 4 (Oct.–Dec. 2004), pp. 50–57.

[10] Sugimoto, C.; Tsuji, M.; Lopez, G.; Hosaka, H.; Sasaki, K.; Hirota, T.; Tatsuta, S.; , "Development of a behavior recognition system using wireless wearable information devices," Wireless Pervasive Computing, 2006 1st International Symposium on , vol., no., pp. 5 pp., 16-18 Jan. 2006 doi: 10.1109/ISWPC.2006.1613624

[11] Sato, T.; Otani, S.; Itoh, S.; Harada, T.; Mori, T.; , "Human behavior logging support system utilizing fused pose/position sensor and behavior target sensor information," Multisensor Fusion and Integration for Intelligent Systems, MFI2003. Proceedings of IEEE International Conference on , vol., no., pp. 305-310, 30 July-1 Aug. 2003. doi: 10.1109/MFI-2003.2003.1232675

[12] L. Bao and S. S. Intille. (2004) Activity recognition from user-annotated acceleration data. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.68.5133

[13] http://digital.pho.to/ (accessed August 2010)

[14] Zhenke Yang, "Multimodal Datafusion for Aggression Detection in Train Compartments (Whitepaper)", Delft University of Technology, February 2006, http://www.zheek.com/downloads/zhenke_whitepaper.pdf, (accessed August 2010)

[15] Abdullah, L.N.; Noah, S.; , "Integrating Audio Visual Data for Human Action Detection," Computer Graphics, Imaging and Visualisation, 2008. CGIV '08. Fifth International Conference on , vol., no., pp.242-246, 26-28 Aug. 2008 doi: 10.1109/CGIV.2008.65

[16] Zhihong Zeng; Jilin Tu; Pianfetti, B.M.; Huang, T.S.; , "Audio–Visual Affective Expression Recognition Through Multistream Fused HMM," Multimedia, IEEE Transactions on , vol.10, no.4, pp.570-577, June 2008. doi: 10.1109/TMM.2008.921737

[17] Ros, J.; Mekhnacha, K.; , "Multi-sensor human tracking with the Bayesian Occupancy Filter," Digital Signal Processing, 2009 16th International Conference on , vol., no., pp.1-8, 5-7 July 2009 doi: 10.1109/ICDSP.2009.5201201

[18] C. Stauffer, W.E.L. Grimson, Adaptive background mixture models for real-time tracking, in: Computer Vision and Pattern Recognition, Santa Barbara, CA, June 1998, pp. 246-252.
[19] S.J. McKenna, S. Jabri, Z. Duric, H. Wechsler, Tracking Interacting People, in: International Conference on Automatic Face and Gesture Recognition, Grenoble, France, March 2000, pp. 348 - 353.
[20] R. Cucchiara, C. Grana, M. Piccardi, A. Prati, Detecting moving objects, ghosts, and shadows in video streams, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (10) (2003), pp. 1337–1342.
[21] Yilmaz, A., Javed, O., and Shah, M. 2006. Object tracking: A survey. ACM Comput. Surv. 38, 4, Article 13 (Dec. 2006), 45 pages.
DOI = 10.1145/1177352.1177355 http://doi.acm.org/10.1145/1177352.1177355
[22] C. Stauffer, W.E.L. Grimson, Learning patterns of activity using real-time tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000), pp. 747–757.
[23] N. Johnson and D.C. Hogg, "Learning the Distribution of Object Trajectories for Event Recognition", Proc. British Machine Vision Conf., D. Pycock, ed., pp. 583-592, Sept. 1995.
[24] W.Pedrycz, A. Amato, V. Di Lecce, V. Piuri, Fuzzy Clustering with Partial Supervision in Organization and Classification of Digital Image, IEEE Trans. On Fuzzy System, Vol.16, no.4,pp. 1008-1026, August 2008. ISSN: 1063-6706, Digital Object Identifier 10.1109/TFUZZ.2008.917287
[25] O. Boiman, M. Irani, Detecting irregularities in images and in video, in: Internatinal Conference on Computer Vision, Beijing, China, Oct. 15–21, 2005, pp. 462 - 469 Vol. 1.
[26] I.N. Junejo, O. Javed, M. Shah, Multi feature path modeling for video surveillance, in: International Conference on Pattern Recognition, Cambridge, UK, Aug 2004, pp. 716 - 719 Vol.2.
[27] N. Vasvani, A. Roy Chowdhury, R. Chellappa, Activity recognition using the dynamics of the configuration of interacting objects, in: Computer Vision and Pattern Recognition, Madison, Wisconsin, USA, vol.2, no., pp. 633-640, 18-20 June 2003.
[28] I.L.Dryden and K.V. Mardia, Statistical Shape Analysis, John Wiley and Sons, 1998. ISBN: 978-0-471-95816-1
[29] Chowdhury, Amit K. Roy; Chellappa, Rama; , "A Factorization Approach for Activity Recognition," Computer Vision and Pattern Recognition Workshop, 2003. CVPRW '03. Conference on , vol.4, pp.41-48, 16-22 June 2003, doi: 10.1109/CVPRW.2003.10040.
[30] C. Tomasi, T. Kanade, Shape and motion from image streams under orthography: a factorization method, International Journal of Computer Vision 9 (1992), pp. 137–154.
[31] T. Xiang, S. Gong, Beyond tracking: modelling action and understanding behavior, International Journal of Computer Vision 67 (1) (2006), pp. 21–51.
[32] Schwarz, Gideon E. (1978). "Estimating the dimension of a model". Annals of Statistics 6 (2): pp. 461–464. doi:10.1214/aos/1176344136.
[33] N. Robertson, I. Reid, Behaviour understanding in video: a combined method, in: Internatinal Conference on Computer Vision, Beijing, China, Oct 15–21, 2005, pp. 808-815.
[34] A.A. Efros, A.C. Berg, G. Mori, J. Malik, Recognizing action at a distance, in: International Conference on Computer Vision, Nice, France, Oct 13–16, 2003, pp. 726 - 733.
[35] A. Bobick, J. Davis, The recognition of human movement using temporal templates, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (3) (2001), pp. 257–267.
[36] M. Hu, Visual Pattern Recognition by Moment Invariants, IRE Trans. Information Theory, vol. 8, no. 2, pp. 179-187, 1962.
[37] G.R. Bradski, J.W. Davis, Motion segmentation and pose recognition with motion history gradients, Machine Vision and Applications, 13 (3) (2002), pp. 174–184.
[38] A. Yilmaz, M. Shah, Actions sketch: a novel action representation, in: Computer Vision and Pattern Recognition, San Diego, California, USA, June 20–25, 2005, pp. 984-989, vol. 1.
[39] H. Yu, G.-M. Sun, W.-X. Song, X. Li, Human motion recognition based on neural networks, in: International Conference on Communications, Circuits and Systems, Hong Kong, China, May 2005, pp. 979-982.
[40] Y. Luo, T.-W. Wu, J.-N. Hwang, Object-based analysis and interpretation of human motion in sports video sequences by dynamic Bayesian networks, Computer Vision and Image Understanding 92 (2003), pp. 196–216.
[41] Y. Shi, A. Bobick, I. Essa, Learning temporal sequence model from partially labeled data, in: Computer Vision and Pattern Recognition, New York City, New York, USA, June 17–22, 2006, pp. 1631 - 1638.
[42] J.W. Davis, S.R. Taylor, Analysis and recognition of walking movements, in: International Conference on Pattern Recognition, Quebec, Canada, Aug 11–15, 2002, pp. 315 – 318 vol. 1.

[43] V. Parameswaran, R. Chellappa, View invariance for human action recognition, International Journal of Computer Vision 66 (1) (2006), pp. 83–101.
[44] I. Weiss and M. Ray. Model-based recognition of 3d objects from single images. IEEE Trans. on Pattern Analysis and Machine Intelligence, 23, February 2001, pp. 116 - 128.
[45] A. Gritai, Y. Sheikh, M. Shah, On the use of anthropometry in the invariant analysis of human actions, in: International Conference on Pattern Recognition, Cambridge, UK, Aug 23–26, 2004, 923 - 926 Vol.2.
[46] R. Bridger. Human Performance Engineering: A Guide For System Designers. Prentice-Hall, 1982
[47] Y. Ivanov, A. Bobick, Recognition of visual activities and interactions by stochastic parsing, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000), pp. 852–872.
[48] C. Rao, A. Yilmaz, M. Shah, View-invariant representation and recognition of actions, Journal of Computer Vision 50 (2) (2002), pp. 203–226.
[49] D.D. Vecchio, R.M. Murray, P. Perona, Decomposition of human motion into dynamics-based primitives with application to drawing tasks, Automatica 39 (12) (2003), pp. 2085–2098.
[50] C. Lu, N. Ferrier, Repetitive motion analysis: segmentation and event classification, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (2) (2004), pp. 258–263.
[51] Jagacinski, R.J., Johnson, W.W., and Miller, R.A. 1983. Quantifying the cognitive trajectories of extrapolated movements. Journal of Exp. Psychology: Human Perception and Performance, pp. 43–57.
[52] Rubin, J.M. and Richards, W.A. 1985. Boundaries of visual motion Tech. Rep. AIM-835, Massachusetts Institute of Technology, Artificial Intelligence Laboratory.
[53] Zacks, J. and Tversky, B. 2001. Event structure in perception and cognition. Psychological Bulletin, 127(1): pp. 3–21.
[54] S. Park, J.K. Aggarwal, Semantic-level understanding of human actions and interactions using event hierarchy, in: CVPR Workshop on Articulated and Non-Rigid Motion, Washington DC, USA, June 2004. DOI: 10.1109/CVPR.2004.160
[55] H. Francke, J. R. del Solar, , and R. Verschae, "Real-time hand gesture detection and recognition using boosted classifiers and active learning," in Advances in Image and Video Technology. Berlin/Heidelberg: Springer, 2007, pp. 533–547.
[56] M. Holte and T. Moeslund, "View invariant gesture recognition using 3D motion primitives," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2008, pp. 797 – 800.
[57] S.-W. Lee, "Automatic gesture recognition for intelligent human-robot interaction," in Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2006, pp. 645–650.
[58] H. S. Park, D. J. Jung, and H. J. Kim, "Vision-based game interface using human gesture," in Advances in Image and Video Technology. Berlin/Heidelberg: Springer, 2006, pp. 662–671.
[59] G. Ye, J. J. Corso, D. Burschka, and G. D. Hager, "Vics: A modular hci framework using spatiotemporal dynamics," Machine Vision and Applications, vol. 16, no. 1, pp. 13–20, 2004.
[60] A. Yilmaz, "Recognizing human actions in videos acquired by uncalibrated moving cameras," in Procedings of IEEE International Conference on Computer Vision, 2005, pp. 150–157.
[61] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," Computer Vision and Image Understanding, vol. 104, no. 2-3, pp. 249–257, 2006.
[62] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in Procedings of IEEE International Conference on Computer Vision, 2005, pp. 1395–1402.
[63] F. Lv and R. Nevatia, "Single view human action recognition using key pose matching and viterbi path searching," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.
[64] B. Peng, G. Qian, and S. Rajko, "View-invariant full-body gesture recognition from video," in Proceedings of the International Conference on Pattern Recognition, 2008, pp. 1–5.
[65] C. L. Zitnick and T. Kanade, "A cooperative algorithm for stereo matching and occlusion detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 7, pp. 675–684, Jul. 2000.
[66] S. Yoon, S. K. Park, S. Kang, and Y. K. Kwak, "Fast correlationbased stereo-matching with the reduction of systematic errors," Pattern Recognit. Lett., vol. 26, no. 14, pp. 2221–2231, Nov. 2005.
[67] Y. V. Venkatesh, S. Kumar Raja, and A. Jaya Kumar, "On the Application of a Modified Self-Organizing Neural Network to Estimate Stereo Disparity", in IEEE Transactions On Image Processing, Vol. 16, No. 11, November 2007, pp. 2822-2829.
[68] Lu Yang, Rongben Wang, Pingshu Ge, Fengping Cao, "Research on Area-Matching Algorithm Based on Feature-Matching Constraints", in 2009 Fifth International Conference on Natural Computation, pp. 208-213

[69] Jian Sun; Nan-Ning Zheng; Heung-Yeung Shum; , "Stereo matching using belief propagation," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.25, no.7, pp. 787- 800, July 2003
[70] Ruichek, Y.; , "A hierarchical neural stereo matching approach for real-time obstacle detection using linear cameras," Intelligent Transportation Systems, 2003. Proceedings. 2003 IEEE , vol.1, no., pp. 299- 304 vol.1, 12-15 Oct. 2003
[71] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer, "Multi-camera multi-person tracking for easyliving," in IEEE International Workshop on Visual Surveillance, 2000, pp. 3–10.
[72] A. Mittal and L. Davis, "M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene using region-based stereo," in European Conference on Computer Vision, 2002, pp. 18–36.
[73] W. Hu, M. Hu, X. Zhou, T. Tan, J. Lou, and S. Maybank, "Principal axis based correspondence between multiple cameras for people tracking," IEEE Trans. Pattern Anal. Machine Intell., vol. 28, no. 4, pp. 663–671, 2006
[74] S. Khan and M. Shah, "Consistent labeling of tracked objects in multiple cameras with overlapping fields of view," IEEE Trans. Pattern Anal. Machine Intell., vol. 25, pp. 1355–1360, 2003.
[75] C. Wren, A. Azarbayejani, T. Darrell, and A.P. Pentland, "Pfinder: Real- Time Tracking of the Human Body," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 780-785, July 1997.
[76] Rita Cucchiara, Massimo Piccardi, Andrea Prati, Detecting Moving Objects, Ghosts, and Shadows in Video Streams, in IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 25, no. 10, October 2003, pp. 1337 – 1342.
[77] S. Santini and R. Jain, "The 'El Niño' image database system", in IEEE Int. Conf. on Multimedia Computing and Systems, Florence Italy, Vol. 1, pp. 524-529, 1999
[78] A. Tversky. "Features of similarity", Psychological Review, 84(4): pp. 327--352, 1977.
[79] A. Amato, T. Delvecchio and V. Di Lecce, "Silhouettes based evaluation of the effectiveness in image retrieval", CRETA 6th WSEAS International Multiconference CSCC, Rethymno, Crete Island, Greece. July 7-14 , 2002, pp. 169-176.
[80] J. B. Kruskal, "Multi-Dimensional Scaling by Optimizing Goodness-of-Fit to a Non-Metric Hypothesis", Psychometrika, 29, pp. 1—27, 1964
[81] A. Amato, V. Di Lecce, A knowledge based approach for a fast image retrieval system, Image and Vision Computing (2008), Volume 26 , Issue 11 (November 2008), pp. 1466-1480, ISSN:0262-8856
[82] Q. Iqbal, K. Aggarwal, Feature integration, multi-image queries and relevance feedback in image retrieval, in: Invited Paper, 6th International Conference on Visual Information Systems (VISUAL 2003), Miami, Florida, September 24–26, 2003, pp. 467–474
[83] A. R. Smith, "Color gamut transform pairs," Comput. Graph. 12(3) (1978), pp. 12-19.
[84] Ojala T, Rautiainen M, Matinmikko E & Aittola M, Semantic image retrieval with HSV correlograms, in Proc. 12th Scandinavian Conference on Image Analysis, Bergen, Norway, pp. 621 - 627
[85] J. R. Smith, "Integrated Spatial Feature Image Systems: Retrieval, Analysis and Compression", Ph.D. thesis, Graduate School of Arts and Sciences, Columbia University, 1997
[86] Grgic, M.; Ghanbari, M.; Grgic, S.; Texture-based image retrieval in MPEG-7 multimedia system, in EUROCON'2001, Trends in Communications, International Conference on. Volume 2, 4-7 July 2001 pp. 365 - 368 vol.2
[87] Jianguo Zhang, Tieniu Tan, Brief review of invariant texture analysis methods, Pattern Recognition, Volume 35, Issue 3, March 2002, Pages 735-747, ISSN 0031-3203
[88] Grigorescu, S.E.; Petkov, N.; Kruizinga, P.; , "A comparative study of filter based texture operators using Mahalanobis distance," Pattern Recognition, 2000. Proceedings. 15th International Conference on , vol.3, no., pp.885-888 vol.3, 2000
[89] Stottinger, J.; Hanbury, A.; Gevers, T.; Sebe, N.; , "Lonely but attractive: Sparse color salient points for object retrieval and categorization," Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on , vol., no., pp.1-8, 20-25 June 2009
[90] Muwei Jian; Peng Ma; Shi Chen; , "Content-Based Image Retrieval Using Salient Points and Spatial Distribution," Information Science and Engineering, 2008. ISISE '08. International Symposium on , vol.1, no., pp.687-690, 20-22 Dec. 2008
[91] Patras, I.; Andreopoulos, Y.; , "Incremental salient point detection," Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on , vol., no., pp.1337-1340, March 31 2008-April 4 2008, doi: 10.1109/ICASSP.2008.4517865
[92] N. Sebe, Q. Tian, E. Loupias, M. S. Lew, T. S. Huang, Evaluation of salient point techniques, Image and Vision Computing, Volume 21, Issues 13-14, British Machine Vision Computing 2001, 1 December 2003, Pages 1087-1095, ISSN 0262-8856, DOI: 10.1016/j.imavis.2003.08.012

[93] http://www.cs.odu.edu/~toida/nerzic/390teched/regular/reg-lang/properties1.html (accessed January 2011)