Angular Variation as a Monocular Cue for Spatial Perception

Agustín Alfonso Navarro Toro

PhD dissertation to aim for a Doctor in Science Degree-Biomedical Engineering Specialty awarded by the Universitat Politècnica de Catalunya

Supervisor: Prof. Dr. Joan Aranda
Automatic Control Department (ESAII)
Universitat Politècnica de Catalunya

November 2008

Table of Contents

Abstract

Monocular cues are spatial sensory inputs which are picked up exclusively from one eye. They are in majority static features that provide depth information and are extensively used in graphic art to create realistic representations of a scene. Since the spatial information contained in these cues is picked up from the retinal image, the existence of a link between it and the theory of direct perception can be conveniently assumed. According to this theory, spatial information of an environment is directly contained in the optic array. Thus, this assumption makes possible the modeling of visual perception processes through computational approaches. In this thesis, angular variation is considered as a monocular cue, and the concept of direct perception is adopted by a computer vision approach that considers it as a suitable principle from which innovative techniques to calculate spatial information can be developed.

The expected spatial information to be obtained from this monocular cue is the position and orientation of an object with respect to the observer, which in computer vision is a well known field of research called 2D-3D pose estimation. In this thesis, the attempt to establish the angular variation as a monocular cue and thus the achievement of a computational approach to direct perception is carried out by the development of a set of pose estimation methods. Parting from conventional strategies to solve the pose estimation problem, a first approach imposes constraint equations to relate object and image features. In this sense, two algorithms based on a simple line rotation motion analysis were developed. These algorithms successfully provide pose information; however, they depend strongly on scene data conditions. To overcome this limitation, a second approach inspired in the biological processes performed by the human visual system was developed. It is based in the proper content of the image and defines a computational approach to direct perception.

The set of developed algorithms analyzes the visual properties provided by angular variations. The aim is to gather valuable data from which spatial information can be obtained and used to emulate a visual perception process by establishing a 2D-3D metric relation. Since it is considered fundamental in the visual-motor coordination and consequently essential to interact with the environment, a significant cognitive effect is produced by the application of the developed computational approach in environments mediated by technology. In this work, this cognitive effect is demonstrated by an experimental study where a number of participants were asked to complete an action-perception task. The main purpose of the study was to analyze the visual guided behavior in teleoperation and the cognitive effect caused by the addition of 3D information. The results presented a significant influence of the 3D aid in the skill improvement, which showed an enhancement of the sense of presence.

Chapter 1

Introduction

The interpretation of the visual world by the human visual system is achieved by the performance of a number of complex processes distributed through different neural links of the brain. These processes are dedicated to perceive color, motion, form and depth, which are considered basic cues that together provide the interpretation of the environment and can even give rise to profound emotions. The perception of depth or distance could be considered the most important property of the visual system. It is capable of extracting 3D information from bidimensional retinal images, thus a conception of space is achieved. Effortlessly tasks executed in daily life such as grasping an object are accomplished due to this spatial conception. It is considered a visual perception process, which requires a previous visual analysis by part of the brain to control action and is fundamental to interact with the environment.

An efficient performance of an action-perception application involves a developed visual guided behavior. It integrates body movements as an intentional motor action in response to visual stimulation. To acquire an accurate coordination, the organism seems to depend on experience, where the active interaction with the perceived spatial environment is essential. According to the theory of direct perception, this spatial information is directly contained in the optic array. It holds that through movements caused by the interaction with the environment, there are certain attributes that remain preserved. This implies that determined visual processes are not influenced by a further evaluation and integration of other cues. Instead, the spatial information is contained fully within the retinal image projected in the eyes.

In this thesis, the theory of direct perception is conveniently adopted. It serves as a link between depth cues represented by image features contained in the retina and the biological processes performed by the human visual system to perceive space. This important relation makes possible the modeling of visual perception processes through computational approaches. It is a suitable principle from which computer vision techniques can rely on to develop innovative methods to calculate spatial information. Since it can be considered as an emulation of the complex processes carried out by the human visual system, it can be inferred that the accomplishment of a general and accurate solution to this problem is not trivial. Nevertheless, as a support to the direct perception theory, a first assumption entails that the visual effect created by the 2D distortion of angular variation is an important source of 3D information that might be modeled. This apparent distortion is a result of the linear perspective and thus presents changes depending on the position of the observer. It is therefore the main focus of this thesis the analysis and the extraction of the spatial information given by angular variations, thus a computational approach to direct perception is achieved. The angular variation is consequently referred as a monocular cue from which 3D information is obtained.

Implications of a direct perception approach may be of significant importance in mediated environments. These are environments that are reinforced by technology to overcome sensorial limitations caused by the physical separation between the operator and the workspace. A considerable aid in this type of applications is the recovery and presentation of lost spatial information. In this sense, through the implementation of a computational approach, as it calculates 3D information from the captured images, it is possible to enhance the optical stimulus of the operator by the introduction of more spatial cues. This leads to assume that the estimation of the orientation of teleoperation tools with respect to the camera is capable of providing the egocentric information necessary to effectively link action and perception. An enhanced view of the workspace would be the inclusion of a new computer generated sight of the scene. The knowledge of this orientation permits to select any angled spot of the workspace, acting as a subjective camera. Thus, self-initiated movements are integrated with the perceived environment providing the operator with the sense of presence necessary to interact with the visual world.

This first chapter is an introduction of this dissertation in which an overview of the relation between the 2D-3D pose estimation problem and the spatial perception is presented. It states the proposed strategy and hypotheses formulated to extract 3D information from images and the efficient application of the obtained data as a perceptual aid in mediated environments. A significant importance has been granted to the visual behavior performed by the human visual system. It interprets visual data to perceive space, thus it is possible to interact properly with the environment. The emulation of certain processes related to this visual function is therefore considered useful in the development of this thesis and a model to fulfill its proposed objectives.

1.1. OBJECTIVES

Angles play a fundamental role in the content of this thesis. They are defined as the figure formed by two lines sharing a mutual endpoint. A perspective representation of this geometric configuration on a flat surface presents a distorted appearance that changes depending on the position of the observer. This angular variation is a result derived from the linear perspective, which is a monocular cue that provides a 3D visual effect. An example of this visual effect could be given by the angles conformed at the vertices of a cube. They are a source of 3D information that undoubtedly provides the observer the perception of a real 3D object. Even with only one vertex presented, at least a relative 3D orientation of the cube could be perceived. This spatial effect created by angular variations is the key that gives rise to the assumption that this is a 3D visual property that may be described by a computational approach.

The extraction of spatial information from angular variations depicted in images is thus the main objective of this thesis. These angular variations are a consequence of changes produced by the perspective distortion and can be generated by two intersecting lines or the rotational motion of a single line. This leads to regard angles as monocular cues for spatial perception with the aim to fulfill this general objective, which has been divided in these partial objectives:

- Study visual properties employed by the human visual system to perceive space and establish 2D-3D metric relations by the modeling of visual processes. This computational approach could be therefore considered as an emulation of determined biological vision processes related to direct perception.

- Generate contributions to computer vision techniques dedicated to the spatial referencing of objects in a scene through bidimensional images. This implies the study and evaluation of innovative visual properties and the development of new pose estimation methods with real-time capability and dispensable of 3D scene data knowledge and initial conditions.

- Analyze the projective properties of angles formed by sets of lines or derived by rotational motions. This leads to the identification of the minimum requirements from which a unique solution to the pose estimation problem is achieved.

This set of partial objectives could be summarized as the obtaining of valuable 3D scene information from images. As it can be considered as an emulation of a visual perception process, its applications in the form of a computational approach and the study of its cognitive effect conform an objective of this thesis as well. This objective is focused in the perception enhancement and has been separated in this set of partial objectives:

- Design and develop tools and interfaces dedicated to efficiently enhance the spatial perception in environments mediated by technology. As a proposed aid to improve the mediation between action and perception, the importance of the assistance is not contained in the 3D information extracted: instead, it is given by how it is presented to the operator.

- Integrate the use of these tools in real visually guided applications. There are a variety of applications with a strong dependence on indirect perception to fulfill action-perception tasks. Minimally invasive surgery (MIS) is an example and can be benefited from this aid. The integration of suitable interfaces should therefore augment the sense of presence.

- Evaluate the cognitive effect produced by the addition of the aid. It implies the verification of the perceptual enhancement and its implication in the visual-motor coordination.

1.2. SYNOPSIS OF DISSERTATION

The content of this thesis is organized following a continuous structure in which the whole approach is described as an emulation of a biological visual process. This structure aims to emphasize the complexity of the human visual system, a fundamental sensory factor to interact with the environment, and its importance as a model to develop new methods in order to fulfill the proposed objectives. As a consequence, the structure is organized in three parts. The first part introduces the concept of visual perception and the complexity of its artificial interpretation. The second implements this knowledge establishing 2D-3D metric relations to develop a computational approach to direct perception. And finally, the third part applies this metric relation and studies its implications in environments mediated by technology.

The introductory part starts in **Chapter 2**. It presents the human visual system as the most sophisticated procedure developed to interpret the visual world. The components and processes involved in the biological image formation are described in conjunction with the sensory cues required to perceive space. **Chapter 3**, on the other hand, presents this biological interpretation of sight as a complex however feasible procedure to be emulated through computer vision. The significance of the visual perception is therefore highlighted and described as a process beyond the simple capturing of visual light.

2D-3D metric relations are developed in the next two chapters. These relations are represented by the geometric description of the visual projection. The pose estimation from angular variations is initially performed in **Chapter 4** using a conventional strategy relating scene and image features. This strategy is based in the rotational motion of lines and provides two algorithms. In **Chapter 5**, the angular variation is

formerly considered as a monocular cue in the sense that the pose estimation problem is seen as part of a visual perception process. It develops an algorithm that formulates the problem regarding the angular changes as a consequence of perspective distortion. These two chapters constitute the second part of the structure of this thesis and define a computational approach to direct perception.

The study of the implications of the developed direct perception approach is carried out in **Chapter 6**. It is assumed that the knowledge of the orientation of objects in a scene is capable of enhancing the sense of presence in visually guided applications. An application of particular interest in this thesis is the minimally invasive surgery (MIS). Therefore, the cognitive tools developed to assist the operator are designed and employed in order to effectively improve the performance in this type of application. The evaluation of the cognitive effect caused by the addition of the aid is demonstrated by an experimental study that simulates this action-perception task. Thus, this study combined with concluding remarks presented in **Chapter 7**, constitute the third and final part of this thesis.

Chapter 2

Introduction to the human visual system

2.1. INTRODUCTION

The interpretation of the visual world by the human visual system implies a number of processes concerned with the formation of spatial images and their further cognitive manifestation. In the initial stage of vision, the optical function of the eyes is to capture and focus radiant light to obtain detailed images, thus they can be transformed into neural form and transmitted to the brain for analysis [1]. There, the visual information is recombined with stored visual memories and processed in order to form the conscious visual perception [2]. This allows conceiving the cognitive, emotional and creative insight of the environment and, since one is unaware of the number of processes involved, it can be qualified as a simple body function. This is a completely misconception, which is mended by analyzing the distribution of visual information along the brain and the complexity of its processes: a group of coordinated tasks responsible of representing an environment described by a bewildering array of overlapping textures, colors and contours undergoing a constant change depending on the position of the observer [3].

This chapter introduces the visual system, beginning with a description of the image formation and continuing with the processes involved in the visual perception. This final stage of the visual processes is required to interpret and perceive objects, events and people in a coherent manner. It is of significant importance to the development of the visual behavior and hence, of special interest in the content of this thesis. The more advanced is the understanding of the visual system, the more accurate and efficient new computational approaches dedicated to the artificial interpretation of sight are developed. A particular relevance is dedicated to the cognitive processes involved in

the perception of space, from which a number of depth cues are combined. These cues can be monocular, if they require a single eye for their reception, or binocular if they require both. Each of them possesses depth and distance information, and their interaction is what provides the sense of space.

Two major approaches contend different theories about the interaction of the spatial cues and have been influential in the guiding of the research of perception. The first, according their historical development, is the constructivist approach. This approach assumes that depth perception is based on the conjunction of previous experience and knowledge of the spatial environment with the evaluation of the acquired depth cues. Thereby, the observer integrates this information and constructs the perception [4]. On the contrary, the leading critic to the constructivist position was held by Gibson [5], and according to his theory the visual stimuli acquired in the environment contained sufficient information to perceive the physical world directly. This theory is known as the direct perception approach, and contends that the perceptual system have been designed to detect perceptual invariants directly and without the necessity of past experience [2]. From the computational point of view, the basic idea of the direct perception is accepted since it involves mathematical oriented analysis and modeling of the visual perception processes and, as it is shown through the next chapters, it could be considered convenient.

2.2. ANATOMY OF THE EYE

The first stage of the visual system is the formation of detailed spatial images. It is achieved by the reception of physical stimulus through the eye in the form of light energy. Light rays enter the eye through a curved and translucent membrane called the cornea, which converges them to a focus on the rear surface of the eye and are regulated by the iris (Figure 2.1). The iris is a contractile structure that varies in size as a function of light intensity and also controls the focal length of the light beam [3]. The rear surface, where focused light is received, is photosensitive and is called the retina. There, photoreceptors absorb light energy in order to transform the information into neural activity. This sequence of processes responsible of transforming light energy to retinal images is complex and is carried out through elaborated anatomical structures. In this section these basic anatomical structures and their functions are described in detail. Thus, functions and properties of the human optical system may be further compared to the components and features of a photographic camera, from which a number of similarities are present.

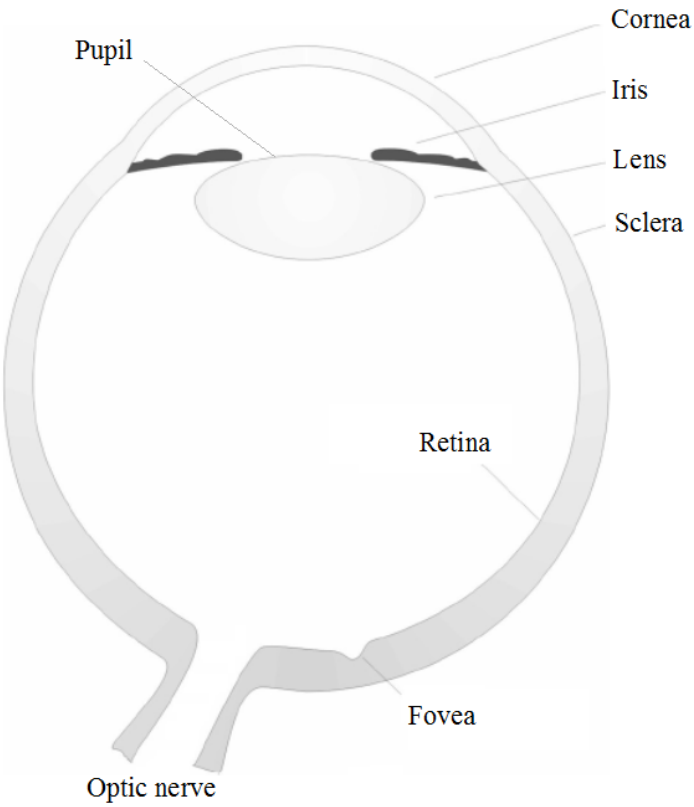

Figure 2.1. Schematic of the cross section of the human eye. The optical system constituted by the anatomical structure of the eye forms detailed spatial images by transforming the incoming light to an array of neural activity, which is considered as the initial stage of the complex process of visual perception (Adapted from [6]).

2.2.1. The cornea

The cornea is the transparent front surface of the eye. It is a powerful refracting surface of about 0.5 to 0.6 mm thick at its center, with a mean refractive index of about 1.376 and a radius of curvature of about 7.7 mm [6]. It is normally clear, there are no blood vessels to interfere with its transparency, and thus it can refract light properly. Its main function, apart of being a protective layer from germs or dust, is the control and focus of light into the eye. The cornea contributes approximately the 75% of the total focusing power of the eye [7]. This optical power is a key contributor to aberrations, which due to its conic shape are reduced about one-tenth of that in the spherical shape lenses with similar power [8].

2.2.2. The iris and the pupil

The colored concentric disk visible through the transparent cornea is called the iris. It is a thin membrane composed of connective tissue and smooth muscle fibers, which is unique for a given individual and is considered a better identification indicator than the fingerprints [9]. The iris controls the amount of light entering the eye. It is a contractile structure that changes in size as a function of light intensity and distance of objects, and regulates the focal length of the light beam [6]. The high pigment serves to block the light and limits it to the pupil. The pupil is the round black center of the iris. Its size determines the amount of light entering the eye, for which a measure between 2 and 3 mm provides a high image quality.

2.2.3. The lens

The crystalline and flexible lens is located behind the iris and the pupil. It is surrounded by a ring of muscular tissue called the ciliary body, which helps to control fine focusing of light to precisely position the visual information. The lens is composed by a sequence of layers with different refracting index. The refraction increases as the lens thickens. With light rays coming from far targets, an easy focus is obtained, as they can be considered parallel. However, with relatively close targets, the light rays are divergent form each other and are focused behind the photoreceptor surface if they are not properly refracted [1]. Therefore, an automatic adjustment process changes the shape of the lens. This process is called accommodation, and increases the lens curvature to be capable to focus on near objects.

2.2.4. The retina

The light sensitive tissue at the inner surface of the eye is called the retina. It is the most photosensitive component of the human central nervous system and converts the captured light rays into neural activity [3]. There are two types of photoreceptors in the retina: rods and cones. The retina contains approximately 125 million rods. They are spread through the periphery of the retina and are most sensitive to light brightness changes, suitable for dim light, shape and movement. While the cones, which are concentrated essentially in the fovea, are most sensitive to one of the three colors: red, green or blue. There are approximately 6 millions cones and, as they are sensitive to color, they are used to appreciate fine details.

As it is shown in Figure 2.2, the neuronal structure of the retina is organized in cellular and fiber layers. It is a network of interconnections between the photoreceptors and the optic nerve. Rods and cones are connected to intermediate cells called bipolar cells, and this cells, are connected to ganglion cells, whose axons are the optic nerve and finalize the vertical neural connection. While two layers of horizontal connections serve to link the adjacent bipolar and ganglion cells [2]. The horizontal cells lay between the

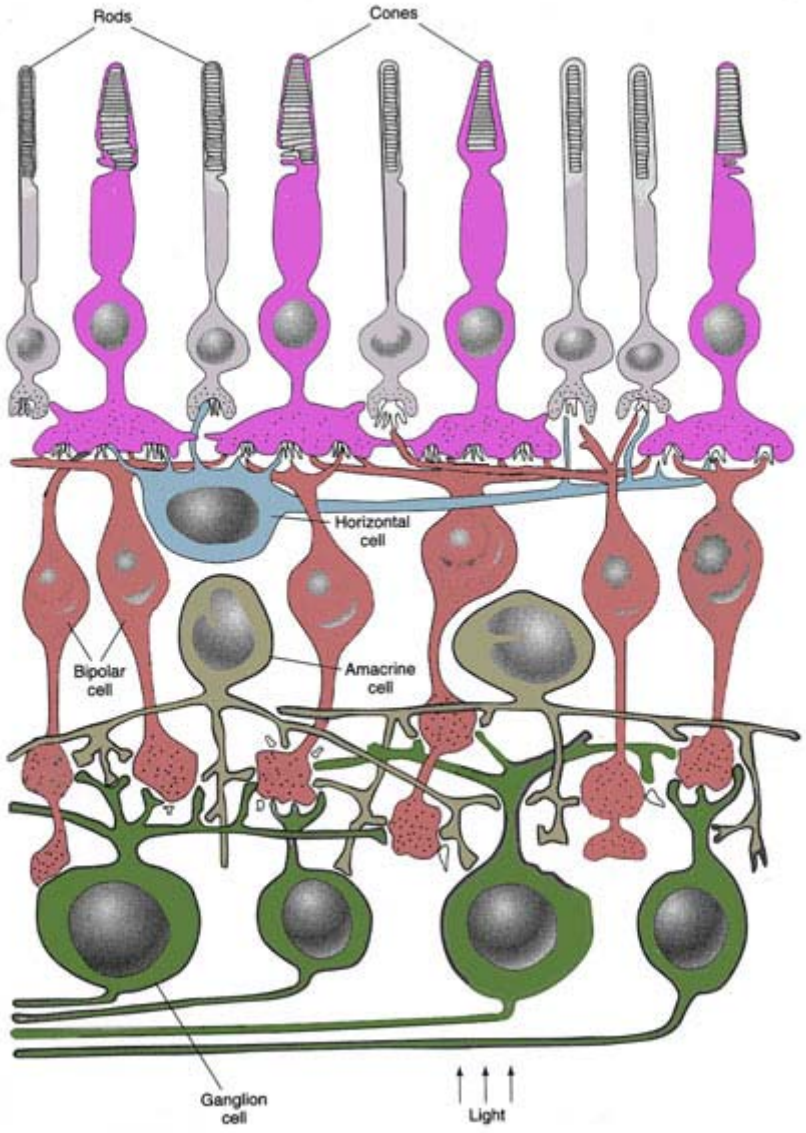

Figure 2.2. Schematic of the neural structure of the retina. The network of interconnections converts the captured light rays to neural activity (Adapted from [10]).

photoreceptors and the bipolar cells, and the amacrine cells lie between the bipolar cells and the ganglion cells. By one side, a high acuity is promoted by the independent connections of cones, while at the other; a high sensitivity is provided by the convergence of rods to intermediate ganglion cells [4].

2.3. VISUAL PATHWAY TO THE BRAIN

The optical functions of the eye involve the formation of spatial images by the transformation of captured energy to an array of neural activity in the retina. Since this visual information at this stage is only a set of neural activity without information processing, a conscious sense of visual perception could not be achieved [3]. Therefore,

a further processing is necessary, which having the resulting retinal images as neural input, follows the visual pathway through a complex network of neural structures to the major center of the visual system, the visual area of the brain [1]. There, the majority of the visual information is sent to separate cortical areas devoted to the primary visual processing, while the remaining is processed for functions such as multiple sensory integration or visually guided motor control [11].

The visual pathway shown in Figure 2.3 starts with the transmission of visual information from the ganglion cells in the retina to the brain through the optic nerve. The majority of the optic fibers in the optic nerve go to the lateral geniculate nucleus (LGN) in the thalamus. These fibers are of several types, from which the 90% of them is

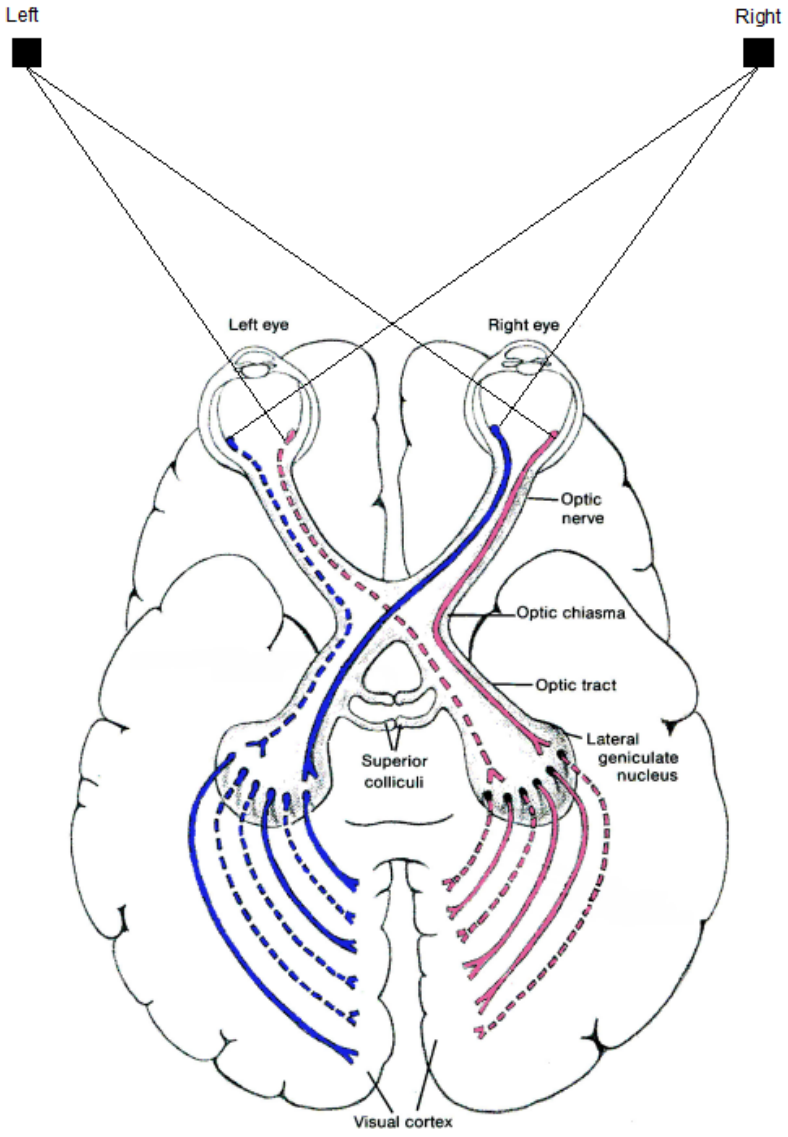

Figure 2.3. Schematic of the visual pathway from the ganglion cells in the retina to the primary visual cortex. The visual information is divided, thus each hemisphere of the brain processes its corresponding visual field. The left hemisphere is related to the right field and the right hemisphere to the left (Adapted from [10]).

constituted by the M-cells and the P-cells. M-cells are sensitive to low spatial frequencies, while P-cells are sensitive for wavelengths and high spatial frequencies [12]. The remaining 10% is constituted by different kinds of cells that together with the M and P cells form a set of data, which is distributed through different routes for parallel processing. This distribution of visual information starts by its division according to the visual field. The right visual field is related to the left hemisphere of the brain, and the left visual field with the right hemisphere. It is a separation achieved by a convergence of the optic nerve at the optic chiasm. There, the optic fibers from the nasal half of the retina cross, while the fibers from the temporal half stay on the same side. Therefore, this is a crossing point that ensures the reception of visual information of a determined side of the visual field to the same hemisphere of the brain, which means that information captured in the same side of both eyes is processed in the same hemisphere.

After the division of the visual information according to its visual field, the majority of axons of the ganglion cells synapse with the corresponding LGN cells on each hemisphere. The pathway where this visual information is sent is called the optic tract. Through this connection, the visual information is processed in a systematic manner, thus a representation of the retinal image is mapped in the LGN [1]. From there, neurons transmit the visual information by optic radiations to the occipital lobe of the cerebral cortex, which is located at the back of the brain, as shown in Figure 2.4, and is referred to as the primary visual cortex or Visual Area 1 (V1). This is the primary route of the visual information flow and is responsible of the conscious visual perception [3]. It is concerned with the basic level processing, primary the detection of features and

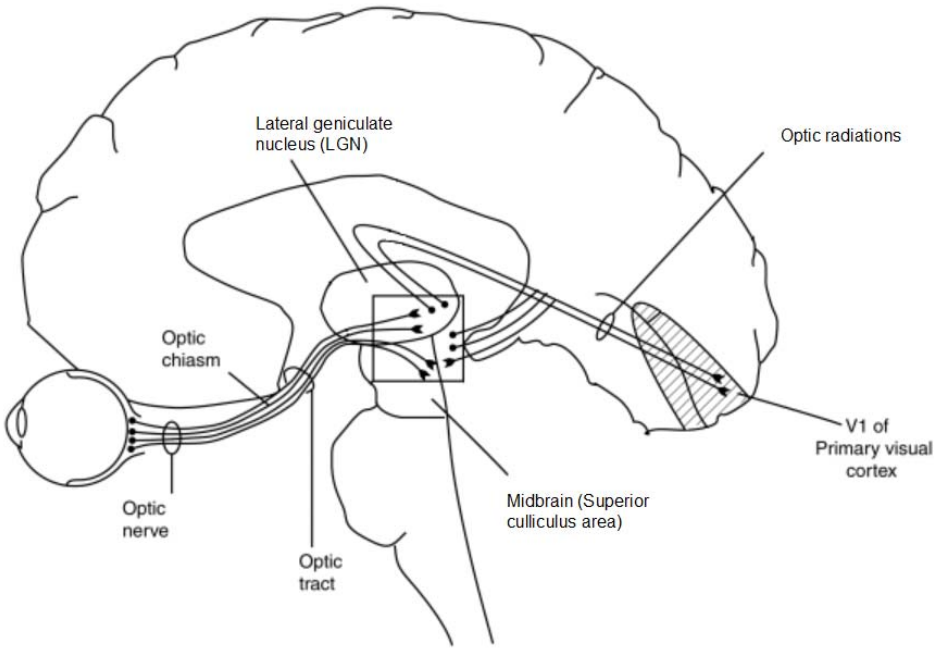

Figure 2.4. Schematic of the major areas of the brain involved in the conscious visual perception (Adapted from [12]).

their orientations. Thereafter, this visual information is distributed to distinct areas of the cortex, which in conjunction form the area called the extrastriate cortex. Each one specialized on a determined function to process specific features.

2.3.1. Functional specialization

The transmission of visual information along different pathways consists in the distribution of specific neural signals through specific areas of the brain, thus its parallel processing provides the conscious visual perception. Since the majority of axons of the ganglion cells project to the LGN, the second primary route synapse with cells at the superior colliculus at the top of the midbrain. It is responsible of controlling the orientation of eye movements and appears to be related to spatial location and visual-motor processing [2]. Others functions such as processing of color, form and analysis of movement appears to be related to the extrastriate cortex, which in conjunction with primary visual cortex and their interactions with the temporal and parietal lobe, form two streams of visual information. These streams are independent and have been proposed to represent the 'what' system, which is related to object identification, and the 'where' system, which is related to the spatial localization [13]. The 'what' system, also called ventral stream, projects to the cortex of the inferior convexity (IC) and the 'where' system, also called the dorsal stream, projects to the dorsolateral prefrontal region (DL) as shown in Figure 2.5 [14].

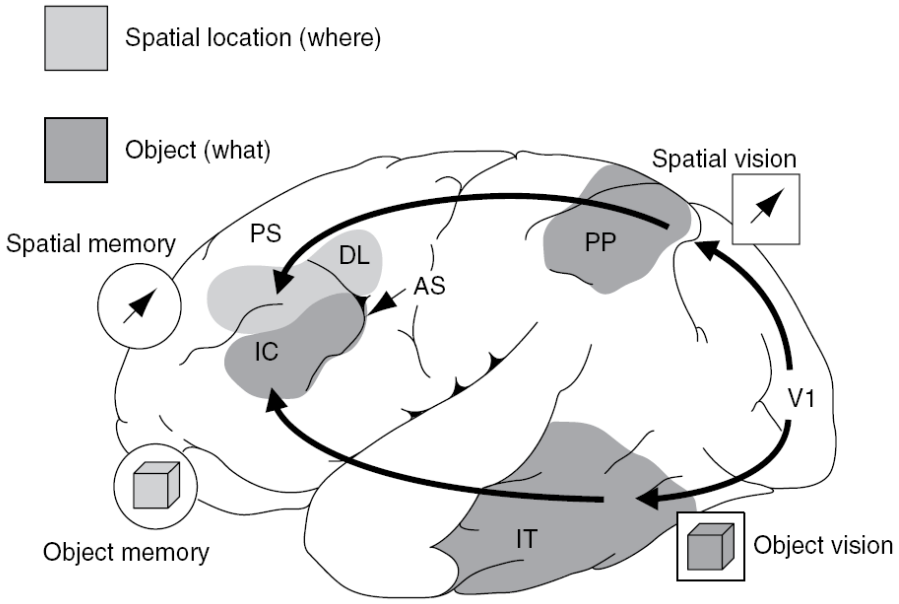

Figure 2.5. Schematic of the two streams of visual information, from which the theory of the two independent pathways, 'what' and 'where', is represented. With PS as principal sulcus, PP as posterior parietal cortex, IT as inferior parietal cortex, DL as dorsolateral frontal cortex and IC as inferior convexity of the frontal cortex (Adapted from [11]).

As a summary, it has been observed that the visual pathway to the brain is a complex structure of neural links between cortical areas. It starts with the physical stimulus in the retina and the subsequent transmission of visual information through two major pathways, the superior colliculus and the LGN. The primary visual information is subsequently transmitted to the visual cortex, situated in the occipital lobe, to be rerouted to parts of the extrastriate cortex, including regions of the parietal and temporal lobes of the cortex. Each subcortical area is related to a determined function and contains a separate map of visual space [15]. They process the visual information differently and from the formation of the ventral and dorsal streams, it has been suggested that IC mediates working memory for objects, while DL mediates spatial working memory [11].

2.3.2. Perception and action

The interpretation of the environment, the conscious visual perception, as has been demonstrated requires a number of complex processes within the visual system. From the objective point of view of the basic functions of the visual system, it is not constrained to the understanding of the environment. The visual stream concerned with the object recognition, which would be the 'what' system or ventral stream, in functional terms works in conjunction with the spatial information provided by the dorsal stream to interact coherently with the environment. Having the two separated streams, an alternative approach has been proposed in which a 'what' system is related to a 'how' system to guide action [16]. In this approach it is supposed that the spatial information from the dorsal visual stream passes directly to the motor guidance system without being consciously perceived. Nevertheless, recent experiments on subjects performing grasping tasks show a significant influence on past experiences and learning behavior from which a conscious calibration of perceived actions occurs [17].

2.4. VISUAL PERCEPTION

This overview about the human visual system would not be complete without mentioning and explaining the functional consequence resultant from the series of neural links. The main function of the visual system is more than the transformation and transmission of an array of light signals; it produces the interpretation of the visual world, and coordinates and executes motor outputs. These functions, apart from the sophisticated anatomy and complex processing dynamics, are what highlight the capacity of the human visual system over any other man-made visual processor. Although the brain works as a unitary system, it distributes visual information thus specific parts are processed separately in order to perceive the environment. Features as color, form, motion and depth constitute the basic cues that together provide the best interpretation and even might create profound emotional effects, as they are perceived.

2.4.1. Perception of color

An important property of the human visual system is its capacity to distinguish colors. It is an additional source of information, which forms part of the object identification process and provides a solid impression of the visual environment. The perception of color depends more on the wavelength than on the intensity of the light. The visible spectrum for humans ranges from 380 to 760 nm. A light ray referred as red, comes from the range of short wavelengths, while a light ray referred as blue comes from the range of long wavelengths. Color information at the early stage of vision is determined by the wavelength of the reflected light from objects. It is decomposed in three physical attributes of light: hue, brightness and saturation. Hue is related to the wavelength, thus it is the main component and represents the true meaning of color. Brightness is related to light intensity and saturation to its spectral purity [18]. A color perception theory proposed by Young-Helmholtz maintains that only three different ocular sensors, each one with its corresponding spectral sensitivity to red, green and blue, are required to produce the visible spectrum [19]. Therefore, three types of cones in the retina are sensitive to these three wavelengths. On the contrary, an opponent theory was held by Ewald Hering [20]. In his theory he proposed three opponent mechanisms, each one divided in pairs, which are antagonistic to each other: a red-green pair excited by red and inhibited by green or vice versa; a blue-yellow and a white-black [3]. More recent studies by Hurvich and Jameson [21], confirm the two theories dividing the complete process in two stages: the processing of wavelength information in the retina by different types of cones and the three antagonistic processes at neural level. Therefore, the sensation of color is a product of the visual system.

2.4.2. Perception of form

Object recognition is the result of a combined number of separated processes, which depend strongly on the neural activity within the ventral stream. There, visual information is processed hierarchically from the primary visual cortex to the anterior occipitotemporal cortex [22]. Visual cortical neurons are specialized to respond to contours and contrast, while higher up in the ventral stream, cells are specialized to respond to distinguish patterns and shapes [3]. To obtain the meaning of this sensory input, a perceptual organization based on two basic approaches, bottom-up and top-down; simplify the analysis by organizing the complete process. The bottom-up processes begin with simple features, which afterward are combined to construct a defined pattern or shape; while the top-down processes begin with complex sensory input as a global set of information, which aided with previous experience and knowledge creates a first interpretation of the environment, to afterwards emphasize on details. Since the amount of information possibly contained in a captured sight of a scene is beyond the capabilities of the visual system, it filters and selects the relevant information. Details are strongly related to the frequency of contrasting light and dark areas, thus patterns with high spatial frequencies contain fine details, while patterns

with low spatial frequencies are part of broad elements. On the other hand, a selective process, called attention, permits to extract only relevant information of the environment. It is a sophisticated process affected by experience and nonsensory factors such as emotions, intentions or expectations. Thus, the amount of information is reduced and a conscious visual perception is produced in only selected areas [1].

2.4.3. Perception of motion

The interpretation of motion information is a significant property of the visual system since it is essential to interact and navigate through the perceived environment. Detection of location, orientation and rate of movement are basic functions highly developed through different motion-sensitive components along the human visual system. They start at the ganglion level in the retina with the M-cells, which are dedicated to react to moving stimuli, specifically object directions [23]. Thereafter, a convergence of information projects through the LGN to the primary visual cortex, where a detailed processing of moving objects is performed to finally distribute the motion information to higher levels of the visual cortical system for further processing. The medial temporal (MT) lobe of the cortex is one of the most important higher level areas involved in the processing of visual motion and serves to integrate information from large regions of the visual field. The optical stimulation to perceive motion is detected by a succession of neural activity of neighboring retinal elements. Therefore, the perception of real movements is achieved by the relation of motion information directly detected in the retina with the one obtained from head or body movements [1]. As a person moves through the environment the retinal images continually changes, these changes create a pattern called optic flow, which provides information about the direction of the person's movements. In the same way, changes of the relative size of a stationary object produce a retinal expansion, which provides information about the rate of movements and, as can be used as a collision indicator, it can also be considered a source of spatial information. It is generally produced by the distorted perception of movements, which is explained in detail in the next subsection and, as it is shown further in Chapter 5, it is a useful visual property that together with other spatial cues, provide 3D information.

2.4.4. Perception of space

The perception of depth or distance could be considered the most important property of the visual system. It is capable of extracting 3D information from bidimensional retinal images, thus a conception of space is achieved. This reconstruction of the third dimension is the product of a complex processing structure from which different sensory inputs are analyzed. These sensory inputs are called spatial cues and can be picked up from the retinal image formed in the optical array of one eye, in which case are called monocular cues, or from the information provided by the two eyes, which are called binocular cues.

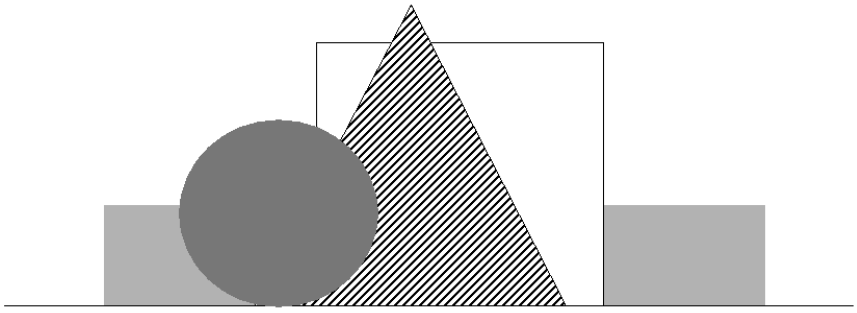

Figure 2.6. Occlusion cues provide a relative perception of depth by the overlapping and interposition of objects. In this example the circle appears to be the nearest object.

Monocular cues

Monocular cues are in majority static features picked up from the scene that provide depth information. They are known as pictorial cues due to their use in graphic art to achieve a realistic representation of depth and consequently an impression of space [24]. Nevertheless, some cues are present only through motions of elements in the scene or self-movements. The monocular cues are:

- **Occlusion**. This monocular cue is also known as interposition and provides relative depth information. It is manifested by the overlapping and covering of elements in the scene. An object is nearer as it can partially or completely conceal those behind it. This cue only indicates which object is nearer from others it occludes (Figure 2.6).

- **Size**. A relative judgment about the depth of an object is given by its size. The larger it is, the closer is the distance from the viewer. This judgment is strongly dependent on past experience and familiarity with similar objects. However, representations of different sizes of the same object within a static retinal image provide a sensation of depth without previous knowledge of the object (Figure 2.7).

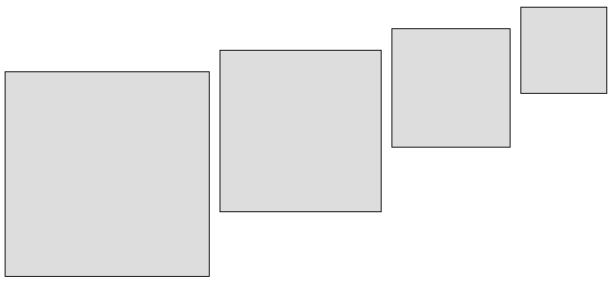

Figure 2.7. The relative size of objects contributes to the perception of depth even without the influence of past experience. In this example different sizes of the same object provide a sensation of depth as they reduce their size (Adapted from [3]).

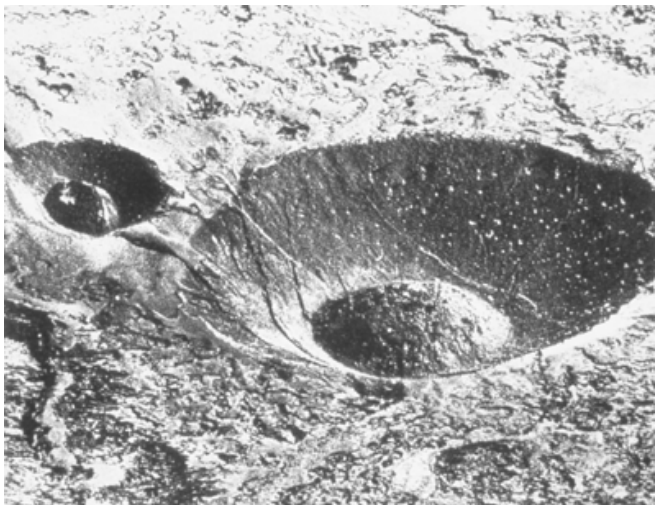

Figure 2.8. Shading cues provide a relative perception of 3D shape. This effect is created by the pattern of bright and dark areas produced by the fall of light (Adapted from [25]).

- **Shading**. The fall of light on objects is a source of depth information. Normally the surface closer to the emitted light is the brightest and provides clues about its orientation with respect to other objects in the environment. In the same way, as an object is illuminated with a single light, the effect given by a created pattern of bright and dark areas produce a perception of its 3D shape (Figure 2.8).

- **Aerial perspective**. This is a monocular cue effective for the relative depth of objects at long range distance. It is due to the fact that the atmosphere causes light to be scattered, thus perceived colors of objects vary depending on their distance and as further they are, their contrast appear to decrease. As textures, the clearer they appear to be, the closer they are (Figure 2.9).

Figure 2.9. Aerial perspective cues are effective for long distance objects where scattering of light at the atmosphere produce distant elements to reduce their contrast (Adapted from [26]).

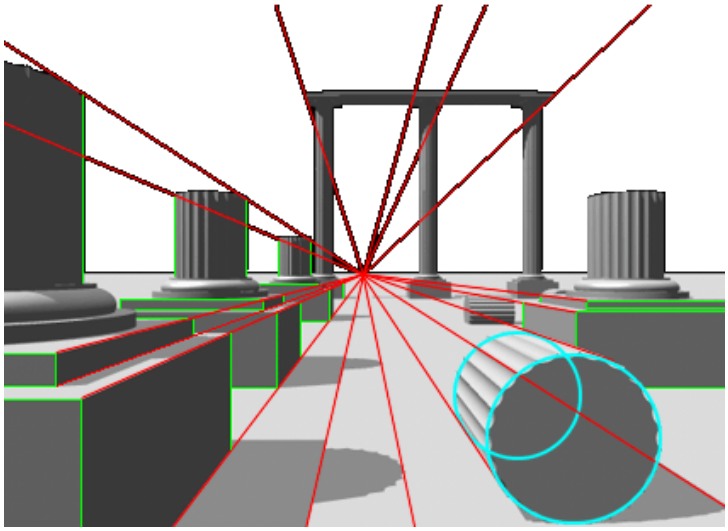

Figure 2.10. The linear perspective is described by a transformation of geometric properties. Parallel lines converge at a single point and the space and size of object reduce as they become distant (Adapted from [27]).

- **Linear perspective**. As a 3D scene is projected to the retinal image, it suffers a systematical reduction of element's size and the space among them according to their distance. It is a transformation of geometric properties that serve to interpret depth. Its potential is of relevant importance in graphic artwork due to its capacity to provide a 3D impression on a flat surface. As it is shown in Figure 2.10, physically parallel lines converge at a single point.

- **Motion parallax**. This is a monocular cue in which motion is required, by part of objects in the scene or by self-movements, to perceive depth. The displacement of the retinal images of objects varies depending on their distance. Near objects appear to move faster than those at far distances (Figure 2.11).

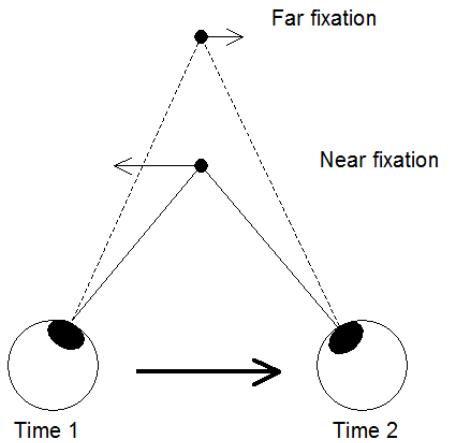

Figure 2.11. The depth effect provided by motion parallax cues derives from the apparent velocity of objects in the retina. As the observer moves, elements move at different velocities depending on their distance.

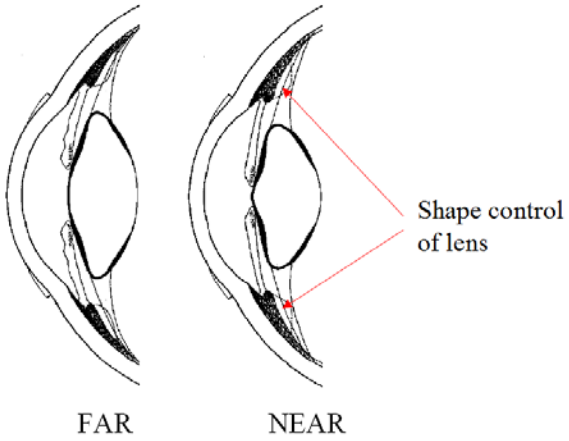

Figure 2.12. Accommodation cues consist in the capacity of the eye to control the lens' shape through the ciliary muscles in order to focus sharply (Adapted from [26]).

- **Accommodation**. As the eye muscles adjust and control the shape of the eye in order to focus the lens and form a sharp retinal image, depth information is obtained. A different response is obtained according to the distance of the object. This monocular cue is effective only at short distances (Figure 2.12).

Binocular cues

Although monocular cues are effective as sources of depth information, a complete perception of the 3D aspect of the environment would not be possible without the information provided by the two eyes. Binocular vision is effective for acute depth perception of short and medium range distances, and are composed by this set of cues:

- **Convergence**. This binocular cue, as the monocular accommodation, is an oculomotor cue for depth perception of nearby objects. It is an automatic coordinated action where the two eyes converge in order to focus a determined object. The closer the object is from the observer, the more the eyes should turn to each other to obtain a sharp focus (Figure 2.13).

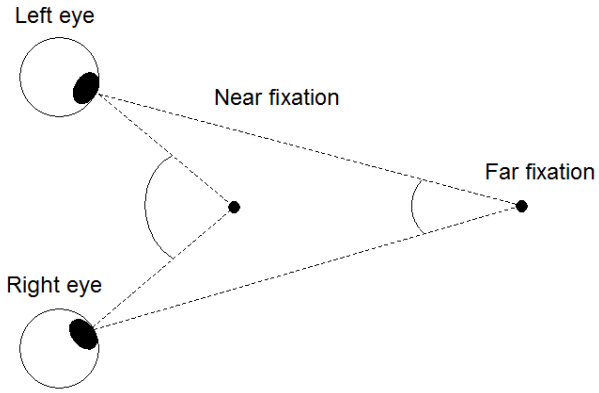

Figure 2.13. The convergence of both eyes is an oculomotor action that contributes to perceive relative depth. The resultant angle represents the relative distance of a determined element.

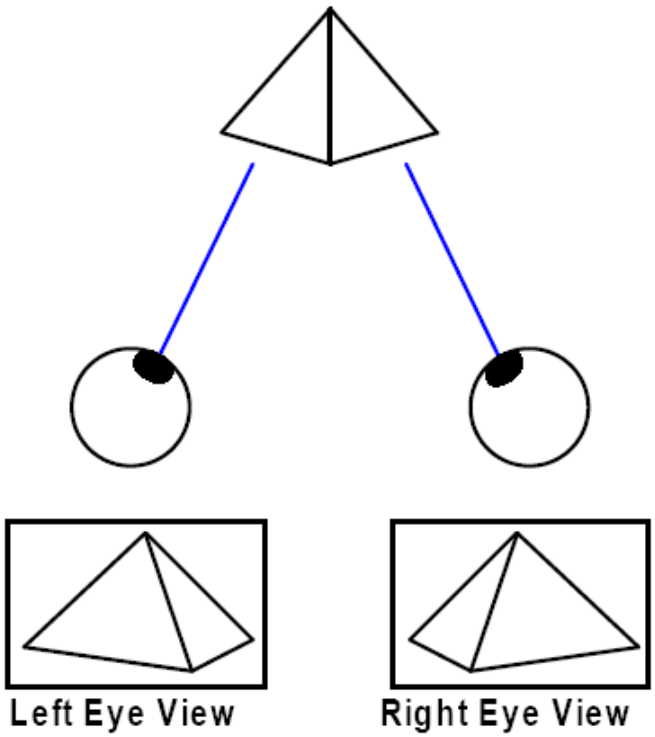

Figure 2.14. Focusing a single object results in slightly different images by the left and right eyes. The binocular disparity of this couple of images provides a powerful source of depth information (Adapted from [26]).

- **Binocular disparity and stereopsis**. Parting from the visual information captured by each of the two eyes, there is a region of the visual field in common that contains depth information. Despite this information represents the same area of a determined scene; the images of the left and right eye are slightly different. This difference or disparity supplies a reliable source of depth information, especially for short range distances. Further elements appear to have a higher disparity than those close to the viewer. The final perception, from the disparity of the two perceived images, is a unique and acute conception of the 3D environment. This solid depth effect is known as stereopsis and plays an important role in spatial vision (Figure 2.14).

Interaction of cues

The complete process of space perception is a combination and evaluation of depth cues. As a natural image is composed by multiple cues, the visual system integrates them in order to reconstruct the 3D environment. The more depth cues are presented; the better is the sense of depth (Figure 2.15) [28]. Some depth cues can be classified as physiological, since they depend on oculomotor mechanisms of adjustment of the eye and are conformed by the accommodation, motion parallax, convergence and binocular disparity. While the rest of the depth cues (interposition, size, shading, aerial perspective, linear perspective and stereopsis), are considered psychological cues, as

Figure 2.15. Pictorial example of the integration of cues. Different monocular cues are used to augment the sensation of depth (Adapted from [28]).

they require a higher level processing in the neurological system of the brain. These psychological cues separated from their coherent nature in a bidimensional image can be ambiguous and, as is shown in Figure 2.16, illusions may be achieved [29]. An approach to spatial perception contends that the way the brain processes and interprets the combination of depth cues is based on previous experience, knowledge of the environment and evaluation of the spatial cues. This approach is called the constructivist and proposes that higher level neural processing is responsible of the reconstruction of the space [3]. In opposition, the direct approach presented by Gibson holds that depth information is contained in the optic array and the interpretation of space is directly perceived without processing [5]. Therefore, according to the direct approach to spatial perception, computational approaches can be developed, since depth information is presented directly in the image. Changes of geometric properties due to perspective transformations, as is explained in Chapter 5, show to be a reliable source of depth information.

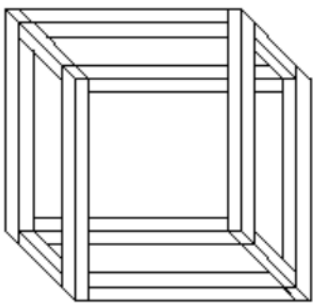

Figure 2.16. The misuse of coherent spatial cues causes a confusing representation of apparent real objects. Impossible figures, as the cube presented in this example defy the laws of geometry (Adapted from [30]).

2.5. REFERENCES

[1] H. R. Schiffman, *Sensation and Perception*, 5th ed. John Wiley & Sons, 2001.

[2] A. Slater, *Perceptual Development: Visual, Auditory and Speech Perception in infancy,* Psychology Press, 1998.

[3] K.R. Huxlin and W.H. Merigan, "Deficits in complex visual perception following unilateral temporal lobectomy," *Journal of Cognitive Neuroscience,* vol. 10, pp. 395-407, 1998.

[4] I. Rock, *Perception,* Scientific American Library, New York, 1995.

[5] J.J. Gibson, *The Ecological Approach to Visual Perception*, Houghton-Mifflin, Boston, 1979.

[6] A. Roorda, "Human visual system - Image formation," in: J.P. Hornak (Eds.), *Encyclopedia of Imaging Science and Technology*, John Wiley & Sons, New York, pp. 539-557, 2002.

[7] National Eye Institute, "Facts about the cornea and corneal disease," US National Institute of Health, www.nei.nih.gov, 2007.

[8] K.M. Charman and J. Cronly-Dillon, *Visual Optics and Instrumentation,* CRC Press, Boca Raton, 1991.

[9] C. Holden, "Eyeball ID," *Science*, 1998.

[10] M.W. Matlin and H.J. Foley, *Sensation and Perception,* 4th ed. Allyn and bacon, 1997.

[11] M.J. Tovée, *An Introduction to the Visual System*, 2nd ed. Cambridge University Press, 2008.

[12] M.M. Gupta and G.K. Knopf, *Neuro-Vision Systems,* IEEE Press, 1994.

[13] M. Mishkin, L.G. Ungerleider and K.A. Macko, "Object vision and spatial vision: two cortical pathways," *Trends in neuroscience*, vol. 6, pp. 414-417, 1983.

[14] F.A.W. Wilson, S.P. O'Scalaidhe and P.S. Goldman-Rakic, "Dissociation of object and spatial processing domains in primate prefrontal cortex," *Science*, vol. 260, pp. 1955-1958, 1993.

[15] T. Allison, A. Begleiter, G. McCarthy, E. Roessler, A.C. Nobre and D.D. Spencer, "Electrophysiological studies of processing in human visual cortex," *Electroencephalography and Clinical Neurophysiology,* vol. 88, pp. 343-355, 1993.

[16] M.A. Goodale and A.D. Milner, "Separate visual pathways for perception and action," *Trends in Neuroscience*, vol. 15, pp. 20-25, 1992.

[17] J.C. Culham, C. Cavina-Pratesi and A. Singhal, "The role of parietal cortex in visuomotor control: what have we learned from neuroimaging?," *Neuropsychologia*, vol. 44, pp. 2668-2684, 2006.

[18] R. Maunsfeld and D. Heyer, *Colour Perception: Mind and the Physical World*, Oxford Press, New York, 2003.

[19] H. von Helmholtz, *Handbook of Physiological Optics*, (Orig. 1850) Translated and reprinted Dover Press, New York, 1962.

[20] E. Hering, *Outlines of a Theory of the Light Sense*, Harvard University Press, Cambridge, 1964.

[21] L.M. Hurvich and D. Jameson, "Opponent processes as a model of neural organization," *America Psychologist,* vol. 29, pp. 88-102, 1974.

[22] K. Grill-Spector, Z. Kourtzy and N. Kanwisher, "The lateral occipital complex and its role in object recognition," *Vision Research*, vol. 41, pp. 1409-1422, 2001.

[23] G. Yang and R.H. Masland, "Direct visualization of the dendritic and receptive fields of directionally selective retinal ganglion cells," *Science*, vol. 258, pp. 1949-1952, 1992.

[24] E.B. Goldstein, *Sensation and Perception*, 6th ed. Pacific Grove, Wadsworth, 2002.

[25] A. Huk, "Seeing in 3D," Lecture Notes: Stanford University Psychology, www.psych.stanford.edu, 1999.

[26] G. Mather, *Foundations of Perception*, Psychology Press, London, 2008.

[27] Perspective drawing, www.artyfactory.com, 2008.

[28] J.D. Pfautz, "Depth perception in computer graphics," Technical Report 546, U. of Cambridge, 2002.

[29] M. Dalton, "Visual perception and the holographic image," www.holographer.org, 2004.

[30] L.S. Penrose and R. Penrose, "Impossible objects: a special type of visual illusion," *British Journal of Psychology*, vol. 49, pp. 31-33, 1958.

Chapter 3

Computer vision: towards an artificial interpretation of sight

3.1. INTRODUCTION

Technology advances and the anxiety to understand and emulate biological vision, have led the emergence of computer-oriented techniques dedicated to artificially interpret the environment through visual stimulus. Initially connected to mathematics and computer science, these techniques were focused on the common objective of making a computer see, giving rise to a new discipline called computer vision. Since the interpretation of the visual world is beyond the capturing and collection of single images, disciplines such as psychology and neuroscience have been added, as they are involved in the visual perception. The perception of motion, forms, color and depth in the human visual system, as is outlined in the previous chapter, are the result of a conjunction of a number of different complex processes from which the suggestion of being emulated by a computer seems to be an unreachable task. Nevertheless, attempts based on image feature analysis and specially the variation of their geometric properties depending on the point of view of the observer, have obtained significant achievements.

The idea of developing new techniques based on biological processes appears to be appropriate to create a reliable artificial vision system. With the better understanding of the biological visual system, the better computational approaches dedicated to model perceptual processes can be developed. These models are useful to simplify the complexity of visual processes, some of them slightly understood, and permit to take practical advantage of the human vision properties in certain applications. Examples of

these applications are the color television or the image compression. The color resolution in the human visual system is lower than the brightness [1], thus a limited bandwidth is required. In the same way, the concept in which image compression is based on, takes advantage of the low spatial frequency property. The human eye is not capable of perceiving detail in high frequency areas, which are regions of the retinal image with high rate of variations in luminance [2]. Thus, a suppression of detail in these areas is not perceived.

In this chapter the potential of computer vision techniques is presented as a capable approach to the artificial interpretation of sight. Parting from the detection of light to the image formation, camera models and projections; the conception of an optic system consistent with its biological counterpart, results in an array of components that constitute sets of contours and shapes that describe a determined scene. From this visual information, biologically inspired techniques have been developed to emulate the visual perception. It could be seen as the translation of the complex neural processes carried out in the brain through mathematic equations and geometric relations, which are implemented in order to fulfill certain visual tasks. Some of these computational approaches are described in this chapter, with a special emphasis in the estimation of spatial attributes of elements. Pose estimation is a relevant field of computer vision implied in the calculation of these spatial attributes and, though its methods are not highly influenced by biological processes, it is an important field to introduce. Since it is developed to model and perform a visual perception function, it can be devised as other methods, taking advantage of the human vision properties, inspired in cognitive and biological processes.

3.2. IMAGE FORMATION

The process of image formation in an electronic camera and the eye is similar. Both optical systems are based on the capturing and filtering of light emitted from elements of a scene to subsequently be focused onto an imaging sensor. There, the visual information is transformed to a video signal, which in the biological case is represented by neural activity, while in a camera is represented by electric charge. As can be seen in Figure 3.1, functions of specific components of the anatomy of the eye are emulated in the mechanic structure of the camera. The amount of light entering to the optical system of the camera is controlled by an adjustable barrier, which acts as the iris of the eye. The diameter of this barrier permits to supply the imaging sensor with the light intensity it can handle. Afterwards, the light rays are focused through the lens to project the image onto a flat surface. An appropriate focus permits to obtain a sharp projection of objects and is achieved by moving the lens toward or away from the imaging sensor, instead of deforming the flexible structure of the eye's lens by the ciliary muscles [3]. These light rays are therefore focused on the imaging sensor, which transforms the visual stimulus to electric charge, acting as the sensitive surface at the rear of the eye, the retina.

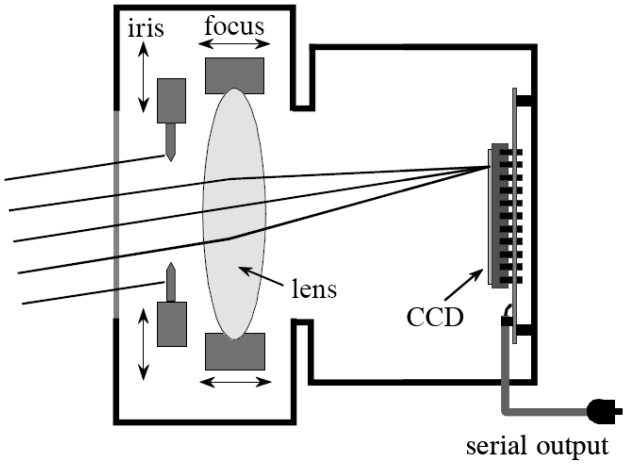

Figure 3.1. The structure of an electronic camera resembles the optical system of the eye. Two major components, the lens and the photoreceptive surface (CCD), present a key role in the image formation (Adapted from [3]).

In an electronic camera, the light sensitive tissue of the retina is emulated by an array of photodetectors sites, or potential wells, in a silicon substrate. This photosensitive array built in an integrated circuit is called charge coupled device (CCD), and serves to sample the light intensity. The charge accumulated in each potential well is proportional to the number of incident light photons and represents a sample of the light pattern [4]. Each sample is called pixel and its intensity value ranges from 0 to 255 after analog to digital conversion (Figure 3.2). These quantization levels display a grayscale image, having a value of 0 for black and 255 for white, and are suitable for data managing as they correspond to a single byte. Color imaging is therefore represented by three values for each pixel (one value for each intensity of the three primary colors: red, green and blue). The more pixels are added to represent an image; the better is its quality. However, this increment of resolution also augments the amount of data to manage.

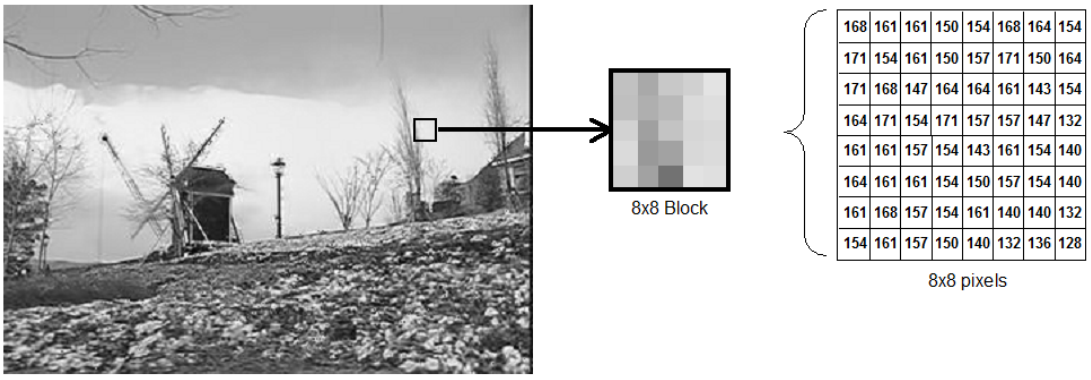

Figure 3.2. Each sample of light intensity is called pixel. The more is the number of pixels; the better is the quality of the image. This example shows a grayscale array of 8x8 pixels with their corresponding values.

3.3. CREATING AN IMAGE-BASED FRAMEWORK

The visual information obtained by the camera results in an array of samples that represents the projected image of a determined scene. At this early stage of vision, the visual world is limited to a set of data encoded in the spatial domain. The information of objects and entities is described by intensity discontinuities or sample contrasts, which require further analysis to be recognized. As their visual interpretation relies on the extraction of basic image features as edges and contours, the artificial perception of the visual environment appears to be a non-trivial task. However, accurate quantitative models of cameras have been defined in order to develop metric measurements from images [5]. This framework describes the nature of perspective projection and the geometric properties of image features. Thus, perceptual functions of the visual system such as object recognition, shape, motion and depth information could be calculated. In this section, this quantitative approach to the image formation and the geometric properties involved are presented. It starts with a brief introduction of the simplest model of the optical system, followed by an introduction of the perspective projection, and finalizes with an important property derived by this kind of projection, which is the geometric invariance.

3.3.1. Optical system model

The simplest camera model that describes the mapping of 3D entities to a bidimensional image is the pinhole camera. It is a suitable model for perspective projection and consists of a center of projection (*o*) and the image or retinal plane. As is shown in Figure 3.3, the projection of a scene point corresponds to the intersection of a line passing through this point and the center of projection with the image plane. Therefore, having the center of projection as the origin of a Euclidean coordinate frame, the mapping of a 3D point $\mathbf{X} = (X, Y, Z)^t$, is an image point $\mathbf{x} = (x, y)^t$, which is the intersection of the line joining *o* and **X** with the image plane at $Z = f$, being *f* the focal length.

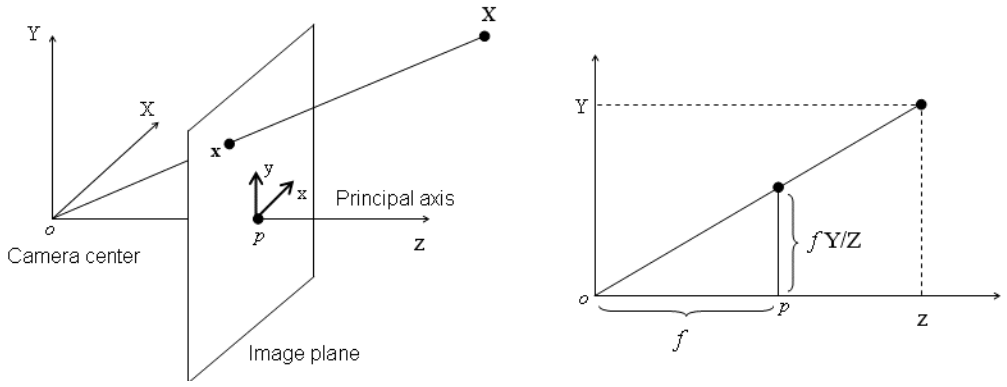

Figure 3.3. Schematic of the pinhole camera model. The mapping of a 3D point is obtained by the intersection of the line uniting the point and the center of projection with the image plane (Adapted from [6]).

Using similar triangles, it can be seen that the Euclidean mapping of the 3D point **X** is described by:

$$x = f\,X/Z.$$
$$y = f\,Y/Z. \tag{3.1}$$

This 3D-2D relation is useful to represent geometrically spatial properties of the image features and expresses the perspective transformation in the image plane. Nevertheless, though most cameras are described by the pinhole model, other specialized models have been developed to describe and take advantage of some properties defined according to their application [6].

3.3.2. Perspective mapping and projective geometry

The central projection described by the pinhole camera model can be specialized to a general projective camera and expressed in a simpler form. As the 3D-2D mapping represented in (3.1) is non-linear, homogeneous coordinates permit to express the projection by the matrix:

$$\begin{pmatrix} fX \\ fY \\ Z \end{pmatrix} = \begin{bmatrix} f & & & 0 \\ & f & & 0 \\ & & 1 & 0 \end{bmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}. \tag{3.2}$$

If the 3D point is now represented by the homogeneous 4-vector **X** and the image point by a homogeneous 3-vector **x**, the perspective transformation expressed in (3.2) could be represented as:

$$x = PX. \tag{3.3}$$

Having P as the transformation matrix, which defines the camera matrix for the model of central projection [5]. It is assumed that the origin of this central projection coincides with the world frame origin. However, in practice it may not be physically possible. Thus, the mapping adds two variables (p_x, p_y), representing the coordinates of the principal point. This point is the intersection of the perpendicular axis of the camera frame to the image plane and its addition is expressed as:

$$(X, Y, Z)^t \rightarrow \left(fX/Z + p_x, fY/Z + p_y \right)^t. \tag{3.4}$$

If the matrix K is:

$$K = \begin{bmatrix} f & & p_x \\ & f & p_y \\ & & f \end{bmatrix} \tag{3.5}$$

the transformation described in (3.3) can be rewritten as $\mathbf{x} = K[I|0]\mathbf{X}$, where [I|0] represents the 3x3 identity matrix plus a column vector of zeros. Thus, now $P = K[I|0]$, with K known as the camera calibration matrix [6]. Since it is assumed that the camera coordinate system coincides with the world coordinate system, the 3D point $\mathbf{X}$ can be called $\mathbf{X}_{cam}$. This point is generally expressed in world coordinates, thus it is necessary to be translated and rotated to the camera coordinate system, which is done by:

$$X_{cam} = \begin{bmatrix} R & -Ro \\ 0 & 1 \end{bmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}. \tag{3.6}$$

with R as the 3x3 rotation matrix and $\boldsymbol{o}$ the 3-vector representing the position of the camera center in the world coordinate system [7]. This leads to have this expression:

$$x = KR\left[I \mid 0\right] X_{cam} \tag{3.7}$$

which represents the mapping described by a pinhole camera, a representation that can be generalized by adding the possibility of having non-square pixels. Thus, by this addition the calibration matrix may be written as:

$$K = \begin{bmatrix} \alpha_x & & x_0 \\ & \alpha_y & y_0 \\ & & 1 \end{bmatrix} \tag{3.8}$$

where, if the pixel dimensions in the x and y directions are represented by m_x and m_y respectively, $\alpha_x = fm_x$ and $\alpha_y = fm_y$, and the principal point coordinates in terms of pixel dimensions is given by $x_0 = m_x p_x$ and $y_0 = m_y p_y$. These parameters form the calibration matrix that defines pixel coordinates of image points with respect to the camera frame, and are called the camera intrinsic parameters.

There is a strong relation between the perspective projection and projective geometry. An important aspect of the projective geometry is the representation of projective

transformations as a matrix multiplication by the inclusion of homogeneous coordinates. A number of computer vision methods take advantage of these properties to easily solve problems, which some of them might be impossible using Euclidean geometry. Some of the major contributions of projective geometry to vision are the described linear method for determining the projective transformation matrix for a camera, the computation of the orientation of planes from vanishing points, camera motion from n matched points and projective invariants [8]. This last contribution is of significant importance in this thesis and serves as the basis for the spatial estimation method proposed in Chapter 5. Nevertheless, approaches based on Euclidean coordinates are introduced previously in Chapter 4, as an initial attempt to extract spatial information from the perspective transformation of angles.

3.3.3. Geometric invariance

The significant importance of the invariance of geometric configurations comes from their property of remaining unchanged under an appropriate class of transformations [9]. Having created an image-based framework for visual perception, where the perspective projection is represented by the collineation of the spatial and image point through the central projection, decisive tasks for the recognition of objects and spatial description of the environment are the extraction of projected geometric configurations and the determination of invariants. An example of invariance is the length under Euclidean transformations. Having the Euclidean distance of two points $D(x_1, x_2)^2$, their displacement can be expressed as $(x_1 - x_2) = R(X_1 - X_2)$, with X_1 and X_2 as their representation after the rotation, for any rotation matrix $R^tR = I$. Therefore, under Euclidean transformations the distance between two points, as well as angles and areas are preserved.

In computer vision, invariants are commonly the property of interest from which many methods are based on [10] [11] [12]. Since the image features that describe an object change depending on the point of view, the invariance of certain geometric configurations under perspective transformations is an important source of information, which can be applied for their recognition. Shape descriptors of 3D objects constructed from the correspondence of specific geometric configurations of two different views also permit to reproduce their 3D projection in any other view without camera calibration [8]. The construction of these invariants comes from their 2D projection in the image plane and therefore, their computation requires more image features than the Euclidean case. Concurrency, collinearity, intersections and the cross-ratio are invariant properties of projective transformations [13]. The cross-ratio, which represents the ratio of ratios, is one of the most important invariant properties with a several number of applications, and as it is explained in Chapter 5, it is strongly related with the angular variation.

3.4. COMPUTATIONAL APPROACHES TO VISUAL PERCEPTION

The interpretation of the environment by the human visual system, as is described in the previous chapter, is achieved by the performance of a diversity of complex processes distributed through brain. The perception of color, motion, form and depth constitute the basic cues that together provide the interpretation of the visual world. An artificial emulation of this perceptual function is the objective of computer vision. It is the fact that would confirm that a machine could really see. Although the accomplishment of this objective is regarded as difficult and even ambitious, several computer vision methods have succeeded in the emulation of determined visual tasks. Some of these methods are inspired in the biological functions of the human visual system. Motion or depth perception is achieved following processes as binocular disparity, accommodation, motion parallax, optical flow or shading patterns. In this section, a further explanation of the principles of some of these methods dedicated to perceive spatial information from monocular cues is presented. These methods have had a relevant interest and formed their own field in computer vision research. Thus, after the accomplishment of image formation and the creation of a special framework from which spatial metrics can be performed, computational approaches have led the ambitious task of visual perception as at least an accessible objective to fulfill.

3.4.1. Optical flow

An important cue used by the human visual system to perceive motion is the analysis of changes in the retinal image caused by head or body movements or by dynamics of the environment. It is expressed in the retina as a succession of neural activity of neighboring elements, which produce optical flow patterns. In the image plane of a perspective camera, these neighboring elements are replaced by intensity samples. Each sample represents the movement of the projection and serves to identify image brightness patterns from which motion information of local regions can be measured from a sequence of images [14]. Therefore, since motion represents spatial changes over time, it is a useful cue for object detection and 3D reconstruction.

A number of computer vision methods have been developed to estimate optical flow vectors. Their initial hypothesis is based on the approximately constancy of intensity structures of local time-varying image regions under motion for at least a short duration [15]. It is assumed that intensity changes are due only to translations and also the objects visualized in the environment are rigid. The computation procedure starts with the measurement of spatio-temporal intensity derivatives, followed by the integration of normal velocities into full velocities [16]. Thus, the problem can be stated by expressing the intensity function as:

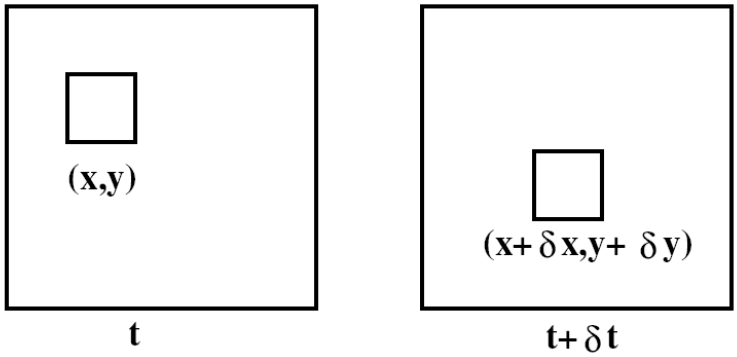

Figure 3.4. The motion constraint equation of differential optical flow assumes that the intensity of regions is approximately constant after short intervals of time for small translations (Adapted from [16]).

$$I(x,t) \approx I(x+\delta_x, t+\delta_t) \tag{3.9}$$

where δ_x represents the displacement at the image region (x, t) after time δ_t, as shown in Figure 3.4. This is a good approximation for small translations, which after being expanded (3.9) in Tailor series is expressed as:

$$I(x,t) = I(x,t) + \nabla I \cdot \delta_x + \delta_t I_t + O^2 \tag{3.10}$$

with $\nabla I = (I_x, I_y)$ and I_t as the first order partial derivatives and O^2 as the second and higher order terms, which are assumed to be small and can be ignored. Thus, subtracting $I(x, t)$ on both sides and ignoring O^2, the resultant equation:

$$\nabla I \cdot v + I_t = 0 \tag{3.11}$$

having $v = (v_x, v_y)$ as the image velocity of pixel (x, y) at time t and representing the motion constraint equation [17]. The computing of full image velocity consists in the addition of more constraints and an example of its results is shown in Figure 3.5.

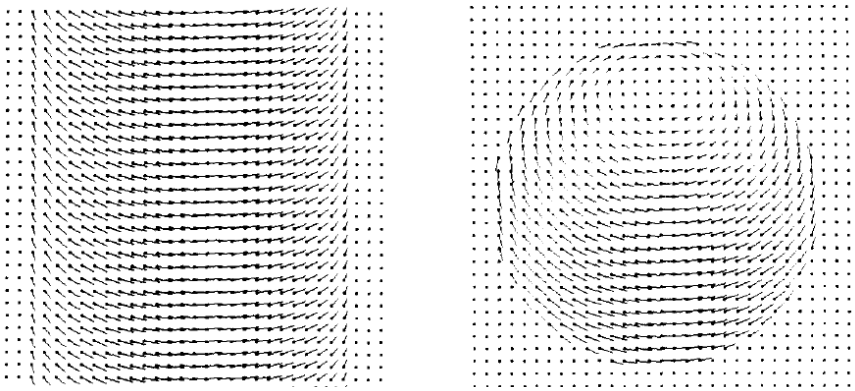

Figure 3.5. Example of the optical flow of a cylinder and a sphere (Adapted from [15]).

3.4.2. Motion parallax

The monocular cue given by the relation between the apparent displacement of elements in the retina and the distance of objects in the scene is also used in computer vision methods. Together with optical flow form the field known as structure from motion, which are based on the extraction of motion information from multiple images to obtain depth. In this case, the optical effect given by the resultant displacement of a determined scene feature in an interval of time is the source of information from which depth can be calculated. As mentioned in the previous chapter, near objects appear to move faster than those at far distances.

Several methods have been proposed to produce a 3D description of a scene based on this monocular cue. An accurate camera calibration is fundamental to obtain good results. This could be considered as the first step of the 3D reconstruction process and can be performed before the application (pre-calibrated) or at run time (online calibrated). Parting from the assumption that determined image features can be identified in a set of multiple projections; it can be considered that the obvious and simpler form to compute depth is by triangulation. However, the presence of noise complicates the problem, generating back projection rays that do not meet in 3D space [7]. Algorithms based on image feature reconstruction have been developed in order to tackle this problem by finding a suitable point of intersection [18]. In the case of feature-based reconstruction, performance is dependent on the accuracy of the feature detection process and its respective correspondence in other frames [19] [20]. While in the case of densely matched pixels, the assumption of dense temporal sampling leads to optical flow as an efficient approximation of sparse feature displacements [21].

Alternative approaches developed for 3D reconstruction are based on the volumetric representation of scene structure by discretizing the scene space into a set of voxels [22] [23] or recover the surface description of objects using variational principles to deform an initial surface [24]. This last approach tackles the 3D reconstruction as a surface evolution problem and can be solved by choosing a surface that minimizes this energy function:

$$E(S) = \int_S \Phi(X)\, dA \tag{3.12}$$

with $\Phi(X)$ small in good matching locations, which can be selected by the summed square errors of the matching pixels:

$$\Phi(X) = \frac{1}{n} \sum_{i \neq j} \Phi_{ij}(X) \tag{3.13}$$

$$\Phi_{ij}(X) = (I_i(x) - I_j(x'))^2. \tag{3.14}$$

Thus, minimizing (3.12), a surface S can be calculated using gradient descent [7].

3.4.3. Depth from focus

A different approach to those methods based on motion is the extraction of spatial information from changes in the optical system of the camera that form a sharp projection of scene elements. As the monocular depth cue used by the eye in which its shape is adjusted in order to focus the lens, the optic system of the camera finds the correct focus of an object by changing the distance between the lens center and the image plane [25]. Computer vision methods emulate this accommodation process by the formulation of a correctly focused expression. It is the representation of finding a sharp image, which is the presence of high spatial frequency content, without blurring, at a determined position [26] [27].

With a first assumption that the focal length f is constant, the correctly focused expression of a defocused spatial point P is given by the finite circular blurry region resultant of its projection in the image plane:

$$d = \frac{Af}{Z - f}\left(1 - \frac{Z}{Z_0}\right). \tag{3.15}$$

Having d as the diameter of the blur circle, in an optical system where A is the lens diameter, Z is the depth of the spatial point P in focus and Z_0 the depth of the spatial point P' out of focus, as seen in Figure 3.6. Thereafter, the intensity distribution is approximated to a two dimensional Gaussian function:

$$g(x, y) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{1}{2}\frac{x^2 + y^2}{\sigma^2}\right) \tag{3.16}$$

where σ is the standard deviation given by the product of the diameter d of the blur circle and a constant k given by the camera.

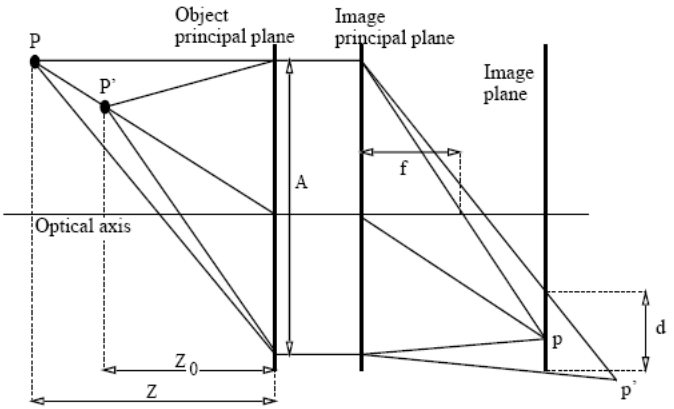

Figure 3.6 Schematic of the optical system properties used to compute depth from focus. A correctly focused expression provides depth information from the blur circle created by a defocused spatial point P' (Adapted from [26]).

3.4.4. Shape from shading

The visual effect given by projected patterns of bright and dark areas in the retina as a consequence of the fall of light is a monocular cue used by the human visual system to perceive depth. This shading cue has also been emulated by computer vision methods to calculate depth and is the last example of computational approaches to visual perception presented in this section. As the human visual system takes advantage of the information provided by the shading depth cue, computational approaches have been developed based on gradations of reflected light intensity to determine the shape of objects as well. It relates the angles of scattered reflected rays with the incident light on a determined surface. As is shown in Figure 3.7, there are three angles of importance: The incident angle, formed between the incident ray and the surface local normal; the emittance angle, formed between the local normal and the emitted ray; and the phase angle, between the incident and emitted rays. Thus, it is possible to obtain the surface shape if the reflectivity and the position of the light sources are known [28].

That initial formulation of the shape from shading defined it as a simple problem where a nonlinear first-order partial differential equation was enough to calculate the solution [29]. This equation related the brightness image I with the reflectance map R, and is expressed as:

$$I(x_1, x_2) = R(n(x_1, x_2)) \tag{3.17}$$

where (x_1, x_2) are image coordinates of a point x and $\boldsymbol{n}$ is the normal vector to the surface. Nowadays, it is known that the solution of this problem is not unique [30] [31]. There is a concave/convex ambiguity and a change of the estimation of the lighting parameters. Thus, assuming that the scene is Lambertian, the reflectance map is the cosine of the incident angle. It leads to this expression: $R = cos(\boldsymbol{l}, \boldsymbol{n})$, where the incident light vector $\boldsymbol{l}$, together with R and $\boldsymbol{n}$ depend on (x_1, x_2) [32].

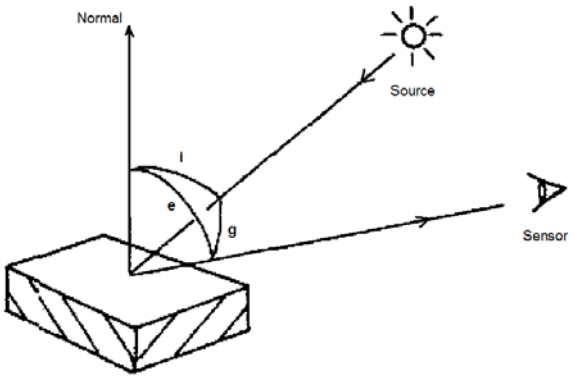

Figure 3.7 Schematic of the three angles involved in the formulation of the shape from shading problem. If the incident (i), emittance (e) and phase (g) angles are known the shape of the surface can be determined (Adapted from [27]).

3.5. POSE ESTIMATION

Although there have been developed a number of vision techniques specialized on the extraction of depth information inspired by biological processes, a different approach to visual perception takes advantage of indirect geometric properties of projected features to estimate the position and orientation of scene objects. This approach is a well known problem in computer vision. It is known as the 2D-3D pose estimation and, as the biological inspired approaches presented in the previous section, it has created its own research field. The pose problem can be defined as the estimation of the position and orientation of a 3D object with respect to a reference camera frame, as shown in Figure 3.8. Therefore, it can be considered as the estimation of the extrinsic parameters of the camera or a solution to the exterior orientation problem. It provides valuable 3D data and leads to the assumption that this spatial information may benefit significantly the cognitive mediation between action and perception in artificial vision applications.

The indirectness from which the pose estimation problem is formulated is what differentiates this approach from biologically inspired methods. While computer vision techniques based on the emulation of determined functions of the human visual

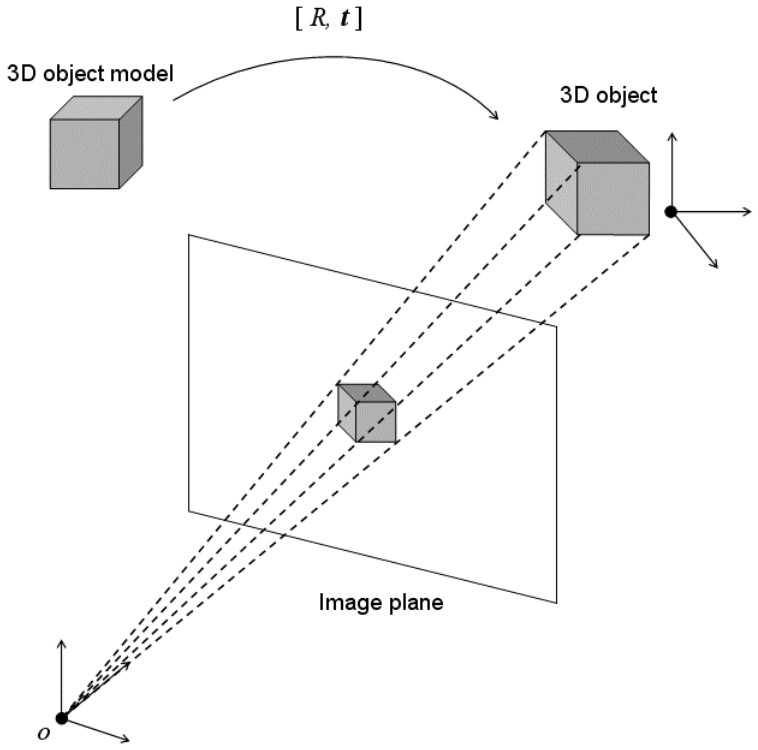

Figure 3.8 Description of the 2D-3D pose estimation problem. The objective is to calculate the rigid body motion that achieves the best fit between object and image features. It represents the position and orientation of the object with respect to the camera.

system take advantage of spatial cues developing computational models for shape and depth calculations, the pose estimation problem generally formulates the problem indirectly imposing constraints equations to relate object and image features [33]. The resultant pose is estimated by minimizing an error measure through constraint equations, instead of applying closed form solutions or using invariants. Most of the developed techniques use the perspective camera model to work with a real representation of scene objects. They are normally separated in categories depending on the type of scene features from which the pose is calculated. There is therefore, a strong dependence on the correspondence of features, which in most cases is assumed to be known by the pose estimation methods.

Pose estimation methods based on points were the first attempts to deal with this problem and thus, a number of algorithms have been proposed. The spatial information obtained from these methods comes from the identification and location of feature points in the image plane [34]. They start with the assumption that the scene is composed by rigid objects. However, their geometric models are not explicitly provided. In general, the problem is seen as the fitting of a set of N object points with its corresponding projection in the image plane [35]. The minimum number of correspondences that provides a finite number of solutions is three [36]. In the case of four coplanar and noncollinear points, a unique solution is given. As N increases, the accuracy of the estimation also increases [37]. Early works focused with a small number of correspondences applied iterative numerical techniques [38] [39]. Applying Newton-Raphson minimization, it was shown that the use of numerical optimization techniques was capable of real-time performance. Other methods apply direct linear transform (DLT) for a larger number of points [40], or reduce the problem to close-form solutions applying orthogonal decomposition [41]. Dual quaternions have also been used to estimate the pose by representing the rotation and translation through the real and dual part of the dual quaternion respectively [42].

An extension of the point configuration concept for pose estimation is the use of higher order geometric entities. It is a generalization of the used features that first introduces correspondences between a 2D line and a 3D point or 2D lines and 3D lines [33]. This different approach is considered a separate category and deals with lines, curves, planar surfaces or quadric surfaces as the features that should be identified in the image plane. Thus, as the 3D object is described by a series of more complex features, it is necessary to know its 3D model to calculate its pose [43]. Some of the methods based on this type of geometric entities are explained in more detail in the next two chapters. Lines are the feature of interest in this thesis and as it is first introduced in Chapter 4, there is a solution of the pose problem using numerical techniques and external angle information. Nevertheless, it is shown in Chapter 5 that an improved solution inspired in the biological processing is possible using geometric invariants, without requiring external angular information.

3.6. REFERENCES

[1] R. Maunsfeld and D. Heyer, *Colour Perception: Mind and the Physical World*, Oxford Press, New York, 2003.

[2] H.R. Schiffman, *Sensation and Perception*, 5th ed. John Wiley & Sons, Inc. 2001.

[3] S.W. Smith, *The Scientist and Engineer's Guide to Digital Signal Processing*, California Technical Publishing, San Diego, 1997.

[4] O. Hainaut, "Basic image processing," www.sc.eso.org, 1996.

[5] O. Faugeras, *Three-Dimensional Computer Vision,* 4th ed. The MIT Press, Cambridge, Mass. 2001.

[6] R. Hartley and A. Zisserman, *Multiple View Geometry*, 2nd ed. Cambridge University Press, 2004.

[7] Y. Lu, J.Z. Zhang, Q.M. Wu and Z. Li, "A survey of motion-parallax-based 3-D reconstruction algorithms," *IEEE Trans. System, Man and Cybernetics - Part C*, vol. 34, no. 4, pp. 532-548, 2004.

[8] J.L. Mundy and A. Zisserman, *Geometric Invariance in Computer Vision*, The MIT Press, Cambridge, Mass. 1992.

[9] D. Mumford, J. Fogarty and F. Kirwan, *Geometric Invariant Theory*, 3rd ed. Springer, 2002.

[10] T. Yuang, S. Yan and X. Tang, "Perspective symmetry invariant and its applications," *Int. Conf. Pattern Recognition*, 2006.

[11] S. Lu and L. Tan, "Camera text recognition based on perspective invariants," *Int. conf. Pattern Recognition*, 2006.

[12] S. Savarece and L. Fei-Fei, "3D generic object categorization, localization and pose estimation," *Int. Conf. Computer Vision*, 2007.

[13] M.A. Rodrigues, *Invariants for Pattern Recognition and Classification*, World Scientific, 2001.

[14] E.P. Simoncelli, "Design of multi-dimensional derivative filters," *IEEE Int. Conf. Image Processing*, vol. 1, pp. 790-793, 1994.

[15] B. Horn and B. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, pp. 185-204, 1981.

[16] J.L. Barron and N.A. Thacker, "Tutorial: computing 2D and 3D optical flow," Tina Technical Report no. 2004-012, 2005.

[17] B.D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," *DARPA Image Understanding Workshop*, pp. 121-130, 1981.

[18] R. Hartley and P. Sturm, "Triangulation," *Computer Vision and Image Understanding*, vol. 68, no. 2, pp. 146-157, 1997.

[19] C. Tomasi and T. Kanade, "Detection and tracking of point features," Technical Report CMU-CS-91-I32, 1991.

[20] Z. Zhang, R. Deriche, O. Faugeras and Q.T. Luong, "A robust technique for matching two uncalibrated images through the recovery of unknown epipolar geometry," *Artificial Intelligence*, vol. 78, pp. 87-119, 1995.

[21] M. Zucchelli, "Optical flow based structure from motion," Ph.D. Thesis, Royal Institute of Technology, 2002.

[22] V. Kolmogorov and R. Zabih, "Multi-camera scene reconstruction via graph cuts," *Proc. European Conf. Computer Vision*, 2002.

[23] G. Slabaugh, B. Culbertson, T. Malzbender and R. Schafer, "A survey of methods for volumetric scene reconstruction from photographs," *Int. Workshop Volume Graphics*, 2001.

[24] O. Faugeras and R. Keriven, "Variational principles, surface evolution, PDE's, level set methods, and the stereo problem," *IEEE Trans. Image Processing*, vol. 7, pp. 336-344, 1998.

[25] J. Ens and P. Lawrence, "An investigation of methods for determining depth from focus," *Trans. Pattern Analysis and Machine Intelligence*, vol. 15, no. 2, pp. 97-108, 1993.

[26] C.S Andersen, J.J. Srensen and H.I. Christensen, "An analysis of five depth recovery techniques," *Proc. Scandinavian Conf. Image Analysis*, 1991.

[27] G. Alenyà, "Localització de robots mitjançant contorns actius," Ph.D. Thesis, Technical University of Catalonia, 2007.

[28] B. Horn, "Obtaining shape from shading information," In P. Winston, ed. *The Psychology of Computer Vision*, McGraw-Hill, New York, 1975.

[29] B. Horn and M. Brooks, *Shape from Shading*, The MIT Press, Cambridge, Mass. 1989.

[30] M. Brooks, "Two results concerning ambiguity in shape from shading," *Proc. AAAI*, pp. 36-39, 1983.

[31] E. Prados and O. Faugeras, "Perspective shape from shading and viscosity solutions," *Proc. Int. Conf. Computer Vision*, vol. 2, pp. 826-831, 2003.

[32] E. Prados and O. Faugeras, "Shape from shading: a well-posed problem," *IEEE Int. Conf. Computer Vision and Pattern Recognition*, pp. 158-164, 2005.

[33] B. Rosenhahn, "Pose estimation revisited," Ph.D. Thesis, University Kiel, 2003.

[34] R.J. Holt and A.N. Netravali, "Camera calibration problem: some new results," *CVGIP Image Understanding*, vol. 54, no. 3, pp. 368-383, 1991.

[35] M.A. Fishler and R.C Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381-395, 1981.

[36] Y. Hao, F. Zhu, J. Ou, W.J. Zhou and S. Fu, "Robust analysis of P3P pose estimation," *Proc. IEEE Int. Conf. Robotics and Biomimetics*, pp. 222-226, 2007.

[37] R.Y. Tsai, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses," *IEEE J. of Robotics and Automation*, vol. RA-3, no. 4, pp. 323-344, 1987.

[38] D.G. Lowe, "Three-dimensional object recognition from single two-dimensional images," *Artificial Intelligence,* vol. 31, No. 3, pp. 355-395, 1987.

[39] R.M. Haralick, C. Lee, K. Ottenberg and M. Nölle, "Analysis and solutions of the three point perspective pose estimation problem," *IEEE Proc. Computer Vision and Pattern Recognition,* 1991.

[40] R.I. Hartley, "Minimizing algebraic error in geometric estimation problems," *Proc. Int. Conf. Computer Vision*, pp. 469-476, 1998.

[41] P.D. Fiore, "Efficient linear solution of exterior orientation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, No. 2, Feb. 2001.

[42] M.W. Walker and L. Shao, "Estimating 3-d location parameters using dual number quaternions," *CVGIP Image Understanding*, vol. 54, no. 3, pp. 358-367, 1991.

[43] J.S. Goddard, "Pose and motion estimation from vision using dual quaternion-based extended kalman filtering," Ph.D. Thesis, University of Tennessee, 1997.

Chapter 4

Line-based rotational motion analysis

4.1. INTRODUCTION

Computer vision methods dedicated to the analysis and understanding of a 3D scene have experienced a notable development over the last decades. Most of the increasing interest in this research field is due to the necessity of interpreting the 3D space from its projection on a flat surface. It could be seen as the recovery of 3D physical attributes of objects represented in images. These attributes can be expressed by its position and orientation with respect to a reference frame, and consequently, spatial information of a 3D scene is provided. To accomplish this task specific geometric configurations are extracted from the captured camera view in the form of projected image features. The angular variation, as a monocular cue used by the human visual system, could be capable of supplying the required data to calculate this 3D information mathematically as well. In this chapter, this geometric configuration is introduced as the responsible to achieve this task. It could be considered as a first attempt to fulfill the objective proposed in this thesis of obtaining spatial information relating variations between 3D angles and their 2D projections.

The previous chapter describes several methods proposed to estimate the orientation of a rigid object. The first step of their algorithms consists on the identification and location of some kind of features that represent an object in the image plane [1] [2]. Most of them rely on feature points and apply closed-form or numerical solutions depending on the number of objects and image feature correspondences [3] [4]. Lines, however, are the features of interest in this chapter. As higher-order geometric primitives, describe objects where part of its geometry is previously known. These kinds of features have been incorporated to take advantage of its inherent stability and

robustness to solve pose estimation problems [5] [6] [7]. A diversity of methods has been proposed using line correspondences, parting from representing them as Plücker lines [8] [9], to their combination with points [10] [11], or sets of lines [12] [13].

In the case of sequence of images, motion and structure parameters of a scene can be determined. These motion parameters are calculated by establishing correspondences between selected features in successive images. The specific field of computer vision which studies feature tracking and correspondence is called dynamic vision [14] [15]. Using line correspondences, it takes advantage of its robustness. Nevertheless, as it is benefited from its stability, it also has to confront some disadvantages: more computationally intensive tracking algorithms, low sampling frequency and mathematic complexity [16]. Therefore, some early works have chosen solutions based on set of nonlinear equations [17], or iterated Kalman filters through three perspective views [18]. Recently, pose estimation algorithms have combined sets of lines and points for a linear estimation [12], or used dynamic vision and inertial sensors [16]. The uniqueness of the structure and motion was discussed for combinations of lines and points correspondences, and their result was that three views with a set of correspondent features, two lines and one point, or two points and one line give a unique solution [19].

This chapter is focused in the analysis of changes in the image plane determined by selected feature correspondences. These changes are expressed as angular variations, which are represented differently depending on their orientation with respect to the camera. They are induced applying a sequence of specific transformations to an object line. Some properties of these motions are useful to estimate the pose of an object addressing questions as the number of movements or motion patterns required which give a unique solution. Making use of a monocular view of a perspective camera, some of these questions are answered in this chapter by the development of two proposed methods. The first of them requires 3D angular information and introduces the line to plane correspondence problem, while the second improves its dependence on 3D information by the addition of new specific transformations of the object. Both of them are based on the accomplishment of a certain 3D structure produced by selected rotations of the object represented by a 3D line, thus determined constraints are applied in order to obtain a unique solution. It could be seen as an exterior orientation problem where objects in the scene are moved to calculate their pose. Therefore, it is shown how knowing the transformations applied, it is possible to estimate the relative orientation of the 3D line with a number of different rotations and the relative position if the length information of the line segment is provided.

4.2. PROJECTIVE NATURE OF LINES

A line can be described as the shortest connection between two points or the intersection of two planes. In Euclidean 3-space a straight line is infinitely extended in both directions. It is usually defined in terms of a point and an orientation. Thus, having the Cartesian coordinates of the position vector of a point $\boldsymbol{p}$ in the line and its orientation vector $\boldsymbol{b}$, the line L can be parameterized in the real variable t with the resulting expression:

$$L = \{ p + tb, -\infty < t < \infty \}. \tag{4.1}$$

Although this representation of a line, defined as a set of points in 3-space, is commonly used, a number of different representations have been proposed [20]. They differ in their mathematical complexity and are applied by pose estimation methods depending on their suitability for calculus operations [21]. In the image plane, their projections are extracted in order to determine displacement information through their correspondences. It is a robust feature, since edges are normally captured and can be reliably extracted. However, the complexity of the problem increases, since the motion information needs more constraints to be calculated than point-based methods [22].

The projective nature of lines presented in this section serves as an introduction of the utility of this geometric configuration in computer vision applications, particularly to solve pose estimation problems. The methods proposed in this chapter are focused in the angular relations between a rotated 3D line and its derived projections. Therefore, a description of different line representations in 3-space is presented. It is followed by an introduction to the line to plane problem, which is a technique commonly used by pose estimation applications. It is strongly related to the line based rotational motion analysis, as they are based on similar imposed constraints.

4.2.1. Representation of lines

The representation of a line in Euclidean 3-space given in (4.1) is a basic definition which presents several disadvantages, all of them related to the lack of uniqueness. It is a point and orientation representation where infinity of points $\boldsymbol{p}$ and orientation vectors $\boldsymbol{b}$ serve to describe the same line. Having also an ambiguity introduced by the sign of the orientation vector, where vectors $\boldsymbol{b}$ and $\boldsymbol{-b}$ describe the same line as well. There have been a number of attempts to remove extra parameters and improve the representation based in some desired properties as the 1-1 correspondence between the set of lines and the set of representations; and a continuous mapping from lines into representations, or from representations to lines [20].

Plücker originally presented a method to represent a line as the intersection of two planes [23]. It can be expressed as:

$$L = \left\{ (x, y, z) \;\middle|\; x = \kappa z + \lambda,\;\; y = \mu z + \nu \right\}. \tag{4.2}$$

Where the four parameters, κ, λ, μ and ν, are used to describe the line. However, this representation presents a singularity with lines perpendicular to the z axis. Therefore, a new approach was presented based in the use of homogeneous coordinates. Lines in the 3-space have four degrees of freedom, thus it is a difficult geometric configuration to represent. It would be a homogeneous 5-vector, which cannot be freely used in mathematical expressions since points and planes are represented by 4-vectors [21]. To overcome this problem an elegant representation was proposed:

$$\left(B,\;\; M,\;\; N,\;\; P,\;\; Q,\;\; R \right) = \left(b, m \right). \tag{4.3}$$

Where $\boldsymbol{m} = \boldsymbol{p} \times \boldsymbol{b}$ is the moment of the line about the origin, which is a six parameter representation called Plücker coordinate. It has interesting properties as the intersection of two lines, ($\boldsymbol{b}_1$, $\boldsymbol{m}_1$) and ($\boldsymbol{b}_2$, $\boldsymbol{m}_2$), if $\boldsymbol{b}_1.\boldsymbol{m}_2 + \boldsymbol{b}_2.\boldsymbol{m}_1 = 0$.

In the same way, the previous definition for a line presented in (4.1) may obtain a unique representation by the imposition of some constraints. This is a point and orientation representation where the equivalence of different values of vectors $\boldsymbol{p}$ and $\boldsymbol{b}$ to describe the same line is reduced in order to obtain a single instance. It is possible by selecting $\boldsymbol{p}$ as the point of the line L nearest to the origin of the reference frame and constraining $\boldsymbol{b}$ to be a unit vector. Thus, it is achieved by these non-trivial constraints equations:

$$p.b = 0 \tag{4.4}$$

$$b.b = 1. \tag{4.5}$$

Even though these two constraints define a unique physical line in 3-space, there is still an ambiguity determined by the sign of $\boldsymbol{b}$. Therefore, this orientation vector must be limited to a single half-space to get uniqueness [20], which is accomplished by choosing the sign of the z axis component of $\boldsymbol{b}$ positive: $\boldsymbol{b}_z \geq 0$. In the case $\boldsymbol{b}_z = 0$, the y axis component of $\boldsymbol{b}$ should be chosen so that $\boldsymbol{b}_y \geq 0$. And if both $\boldsymbol{b}_z = \boldsymbol{b}_y = 0$, the x axis component of $\boldsymbol{b}$ should be set to 1.

The outlined representations differ in their mathematical complexity. They are selected depending on their suitability on a required application. Plücker coordinates provide an elegant representation of lines and are useful in algebraic derivations. However, the representation through point and orientation describes physical attributes of lines in a natural form. Thus, it is used in this chapter, as it provides an easier handling of this geometric configuration to derive desired orientations.

4.2.2. Pose estimation from lines

The representation of lines in 3-space with respect to a camera reference frame is the purpose of the pose estimation from line correspondences. Since the input data are the projection of segments of lines in the image plane, a 3D-2D relation must be established in order to obtain the desired representation. Having *X*, *Y* and *Z* as the camera coordinates of a space point *P*, which is projected to the image at *p*, the first relation given in the previous chapter leads to:

$$\begin{aligned} x &= f\,{}^{X}\!/_{Z} \\ y &= f\,{}^{Y}\!/_{Z} \end{aligned} \tag{4.6}$$

where *x* and *y* are the 2D image plane coordinates of *p* and *f* is the focal length. Thus, with this point constrained to lie on a projected line with equation:

$$ax + by + c = 0. \tag{4.7}$$

A relation between the 3D line and its projection in the image plane can be expressed as:

$$aX + bY + cZ = n.\overrightarrow{oP} = 0. \tag{4.8}$$

which is the equation of the projection plane that pass through the center of projection *o*, the 3D line and the image line, with a normal vector ***n*** as seen in Figure 4.1.

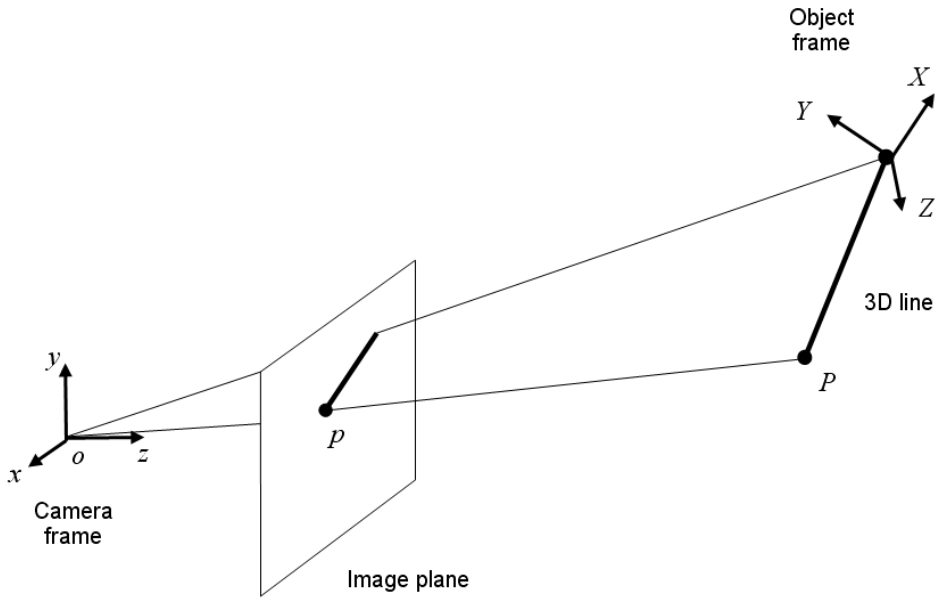

Figure 4.1. Projection of a 3D line in the image plane. A projection plane is created and passes through the center of projection, the projected line and the object line.

The pose of the 3D line with respect to the camera reference frame is given by a 3×3 rotation matrix R and a translation vector $\boldsymbol{t}$. If the 3D line is represented by a point vector $\boldsymbol{m}$ and an orientation vector $\boldsymbol{b}$, the rigid transformation from the object frame to the camera frame can be expressed as:

$$m' = Rm + t \tag{4.9}$$

$$b' = Rb. \tag{4.10}$$

Where $\boldsymbol{m'}$ and $\boldsymbol{b'}$ represent the 3D line in point and orientation respectively in the camera reference frame. Therefore, to estimates the coefficients of the rotation matrix and the translation vector, some constraints must be applied. Considering the fact that the object line, the projected line and the center of projection are coplanar, a plane constraint can be imposed by:

$$\begin{aligned} n.m' &= 0 \\ n.b' &= 0. \end{aligned} \tag{4.11}$$

With a point and orientation representation, the number of unknown parameters to describe a 3D line is six, three for rotation and the other three for translation. This implies that the two constraints provided by a single line are not enough to calculate its pose. Furthermore, it demonstrates the necessity of at least three line correspondences. Therefore, since three lines can be built from a variety of combinations from points, a solution to this problem can be taken from the general solution for point and line correspondences for any number of correspondences [24]. The minimum of input data sets is three points or three lines. This general solution is given by:

$$f(R,t) = \sum_{i=1}^{N} (n_i.(Rb_i))^2 + \sum_{i=1}^{N} (n_i.(Rm_i + t))^2 \tag{4.12}$$

It solves the problem of N line correspondences as is solved the problem of a set of $2N$ nonlinear constraints, which is equivalent to solving the problem of minimizing the error function. This is a general solution for $N \geq 3$, where from $2N$ point correspondences can be built N line correspondences.

4.2.3. Line to plane correspondences

Normally reference features of the same form are extracted from images to estimate the position and orientation of an object with respect to a determined coordinate frame. As presented above, correspondences between sets of points or lines are commonly used and provide the required information. A different approach to fulfill this task is the selection of lines as the sensory features and planes as the matched reference feature [25]. This is generally known as the line to plane correspondence problem and is based on

determining the transformation in which the set of lines corresponds to their respective projection planes. It is strongly related to the theory involved in the pose estimation methods proposed in this chapter, as a 3D line is rotated in order to obtain its respective projection plane until the complete set of concurrent lines satisfy the plane constraint.

The general representation point and orientation of a line requires six parameters. Knowing that each pair of corresponding features provides two equations, it is evident that at least three feature pairs are required to solve the problem. If $L_i = \boldsymbol{p_i} + t\boldsymbol{b_i}$ is the representation of a 3D line, where $\boldsymbol{p_i}$ is a point in the line and $\boldsymbol{b_i}$ is its direction, and l_i its 2D projection in the image plane, which define a projection plane S_i with normal vector $\boldsymbol{n_i}$, the orientation problem is formally stated as:

$$n_i^t R b_i = 0. \tag{4.13}$$

Where $\boldsymbol{n_i}$ and $\boldsymbol{b_i}$ are unit vectors and i = 1, 2, 3. If the proper rotation R is calculated, the rotated line direction vector will be perpendicular to the projection plane normal, as shown in Figure 4.2. Once the rotation has been calculated, the translation can be easily computed knowing that:

$$n_i^t (RP_i + t) = p_i. \tag{4.14}$$

Where $\boldsymbol{t}$ is the translation vector and P_i is the position of a point in L_i, with p_i representing the distance from the center of projection to S_i, as it is a point and orientation representation of a line. Thus, when the proper translation is calculated, the transformed point will lie in the plane.

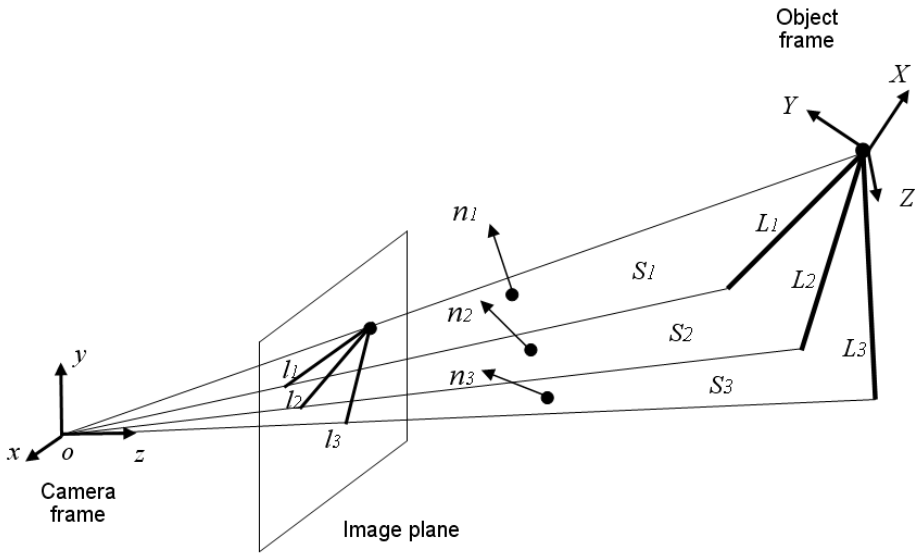

Figure 4.2. Projection of a set of three concurrent lines. The line to plane correspondence first transforms the lines L_i to a configuration in which each one is made perpendicular to its corresponding plane S_i.

A number of degenerate configurations of lines and planes have been addressed in [25], where no solution can be determined. Nevertheless, for valid configurations a closed-form solution was established in order to provide a solution for three lines to plane correspondences. It results in an eight degree polynomial with one unknown. Thus, having the model lines, the method first rotates it in a way the orientation of the fist line $\boldsymbol{b_1}$ lies in its corresponding plane S_1, achieving a perpendicularity with its normal vector $\boldsymbol{n_1}$. This rotation is performed through an axis $\boldsymbol{e} = \boldsymbol{n_1} \times \boldsymbol{b_1}$, which forms a coordinate system with $\boldsymbol{n_1}$ and $\boldsymbol{b_1}' = \boldsymbol{e} \times \boldsymbol{n_1}$. It can be seen as the coordinate frame x, y and z axes, where the normals and the rotated model lines are represented. In this configuration, L_1 has only two degrees of freedom in rotation. It implies that there are only two unknowns left; hence the problem has been simplified. This resulting configuration is called canonical configuration [26], and can be expressed as:

$$R(n_1,\theta)R(b'_1,\rho) = \begin{pmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{pmatrix} \begin{pmatrix} \cos\rho & -\sin\rho & 0 \\ \sin\rho & \cos\rho & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{4.15}$$

where θ and ρ are the two unknowns to be determined. Consequently, substituting equation (4.13) in (4.15), a fourth degree equation is obtained:

$$\sum_{i=0}^{4} \lambda_i \cos^i \rho + \left(\sum_{i=0}^{3} \mu_i \cos^i \rho \right) \sin\rho = 0. \tag{4.16}$$

And taking the squares of both terms of the equation, an eighth degree equation with one unknown results:

$$P(\rho) = \sum_{i=0}^{8} \sigma_i \cos^i \rho = 0 \tag{4.17}$$

where the coefficients σ_i are functions of the components of $\boldsymbol{b_2}$, $\boldsymbol{n_2}$, $\boldsymbol{b_3}$ and $\boldsymbol{n_3}$. The roots of $P(\rho)$ can be calculated by numerical operations with no established initial conditions. However, there are certain special feature configurations that a closed-form solution can be obtained. As an example, there is the case of a coplanar configuration, where the three line orientation vectors have a common origin and lie in a plane. This coplanarity is $\boldsymbol{b_1}.\boldsymbol{b_2} \times \boldsymbol{b_3} = 0$. If it is assumed that this plane is $x = 0$, the expression of $P(\rho)$ becomes:

$$P(\rho) = \sigma_8 \cos^8 \rho + \sigma_6 \cos^6 \rho + \sigma_4 \cos^4 \rho + \sigma_2 \cos^2 \rho + \sigma_0. \tag{4.18}$$

It could be seen as a fourth degree polynomial with $\cos^2(\rho)$ as the unknown, which can be solved analytically. Orthogonal and parallel configurations are the two other cases from which a closed-form solution can be obtained and are based on polynomial formulations as well.

4.3. POSE FROM LINE-BASED ROTATIONAL MOTION ANALYSIS

The selection of line features as the required geometric configuration to determine the pose of an object has demonstrated to be a suitable and robust option. As said before, techniques based on this type of feature are normally focused on the establishment of correspondences with lines or planes. This provides a set of constraints that leads to obtain the 3D representation of the object line. As the relation between 3D and 2D angular variations is of significant importance in this research, motion analysis of feature lines is the base of the pose estimation algorithm developed in this section. It serves as a first approach, being a reconstruction of a set of 3D concurrent lines. In this case known 3D rotations of a line and its subsequent projections in the image plane are related to compute its relative position and orientation with respect to a perspective camera. Vision problems as feature extraction and line correspondences are not discussed and we suppose the focal distance *f* as known. The main goal is, having this image and motion information, estimate the pose of an object represented by feature lines with the minimum number of movements and identify angular patterns that permit to compute a unique solution without defined initial conditions.

4.3.1. Mathematical analysis

A 3D plane is the result of the projection of a line in the image plane. It is called the projection plane and passes through the projection center of the camera and the 3D line. This 3D line, in this case, is the representation of an object. With three views after two different rotations of the object, three lines are projected in the image plane. Thus three projection planes can be calculated. These planes are S_a, S_b and S_c, and their intersection is a 3D line that passes through the projection center and the centroid of the rotated object; being the centroid the point of rotation. Across this line a unit director vector v_{dd} can be determined easily by knowing *f* and the intersection point of the projected lines in the image plane. The intention is to use this 2D information to formulate angular relations with the 3D motion data.

Working in the 3D space permits to take advantage of the motion data. In this case where a 3D line represents the object, the problem could be seen as a unit vector across its direction that is rotated two times. In each position of the three views this unit vector lies in one of the projection planes as seen in Figure 4.3. It is first located in S_a, then it rotates an angle α_{ab} to lie in S_b and ends in S_c after the second rotation by an angle α_{bc}. To estimate the relative orientation of the object the location of the three unit vectors, v_a, v_b and v_c, must be obtained. They should coincide with the 3D motion data and lie on their respective planes. To do this, it is known that the scalar product of:

$$v_a . v_b = \cos \alpha_{ab} \tag{4.19}$$

$$v_b . v_c = \cos \alpha_{bc} \tag{4.20}$$

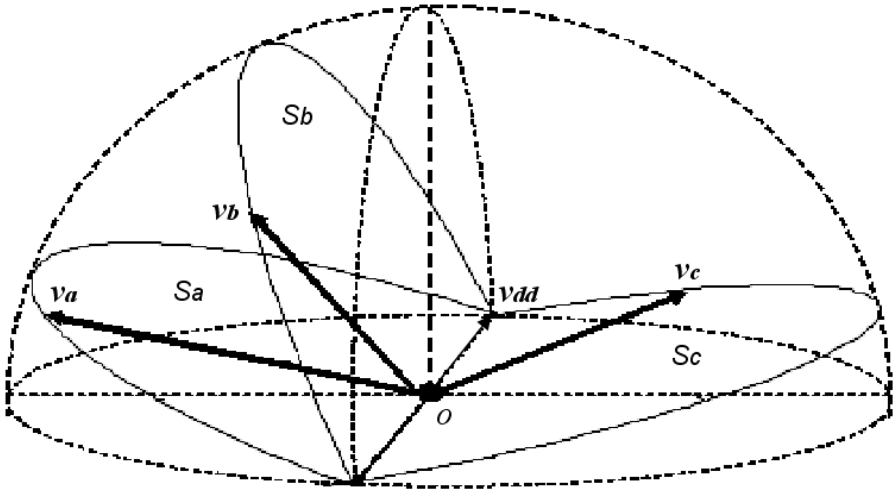

Figure 4.3. Unit vectors v_a, v_b and v_c are constrained to lie on planes S_a, S_b and S_c respectively. Their estimation can be seen as a semi sphere where their combination must satisfy the angle variation condition.

And calculating the angle γ between the planes formed by $v_a v_b$ and $v_b v_c$ from the motion information a third equation is introduced:

$$\left(v_a \times v_b\right)\left(v_b \times v_c\right) = \cos\gamma \tag{4.21}$$

which after applying vector identities, it can be expressed as:

$$v_a . v_c = \cos\alpha_{ab} \cos\alpha_{bc} - \cos\gamma \tag{4.22}$$

With the set of equations conformed by (4.19), (4.20) and (4.22), the three unit vectors are related. However, although there are more unknowns than equations, there is not a unique solution, thus constraints must be applied.

4.3.2. Projection planes constraint

There are many possible locations where the three unit vectors can satisfy the equations in the 3D space. To obtain a unique solution unit vectors v_a, v_b and v_c must be constrained to lie in their respective planes. Unit vector v_a could be seen as any unit vector in the plane S_a rotated through an axis and an angle. Using unit quaternions to express v_a results in:

$$v_a = q_a v q_a^* \tag{4.23}$$

where q_a is the unit quaternion applied to v_a, q_a^* is its conjugate and v is any unit vector in the plane. For every rotation about an axis e, of unit length, and angle Ω, a corresponding unit quaternion q = *(cos* Ω/2, *sin* Ω/2 e*)* exists [27] [28]. Thus v_a is expressed as a rotation of v, about an axis and an angle by unit quaternions multiplications. In this case e must be normal to the plane S_a if both unit vectors v_a and v are restricted to be in the plane.

Applying the plane constraints and expressing v_a, v_b and v_c as mapped vectors through unit quaternions, equations (4.19), (4.20) and (4.22) can be expressed as a set of three nonlinear equations with three unknowns:

$$q_a v_d q_a^* q_b v_d q_b^* = \cos\alpha_1 \tag{4.24}$$

$$q_b v_d q_b^* q_c v_d q_c^* = \cos\alpha_2 \tag{4.25}$$

$$q_a v_d q_a^* q_c v_d q_c^* = \cos\alpha_1 \cos\alpha_2 - \cos\gamma \tag{4.26}$$

The vector to be rotated is v_{dd}, which is common to the three planes, and their respective normal vectors are the axes of rotation. Therefore, extending the set of equations (4.24), (4.25) and (4.26), multiplying vectors and quaternions, it can be appreciated that there are only three unknowns that are the angles of rotation Ω_a, Ω_b and Ω_c.

4.3.3. Relative position and orientation estimation

Applying iterative numerical methods to solve the set of nonlinear equations, the location of v_a, v_b and v_c with respect to the camera frame in the 3D space is calculated. Now there is a simple 3D orientation problem that can be solved easily by a variety of methods as least square based techniques. However, in the case where motions could be controlled and selected movements applied, this last step to estimate the relative orientation would be eliminated. Rotation information would be obtained directly from the numerical solution. If we assume one of the coordinate axes of the object frame coincide with the moving unit vector and apply selected motions, as one component rotations, a unique solution is provided faster and easier.

The estimation of the relative orientation serves now to compute the relative position. It is necessary to track the projection of the end point of the 3D line that represents the object. The relation between the 3D position of this end point, (X, Y, Z), and its projection in the image plane, (x, y), can be expressed as $x=fX/Z$ and $y=fY/Z$. In our case where we know the 3D transformations applied to the object and its image projections, the relative position can be easily calculated rotating them to camera frame coordinates and by relating point position differences ΔX, ΔY and ΔZ and their projections Δx and

Δy. Thus two views are sufficient to obtain the position of the centroid if the length to its end point is provided.

4.3.4. Uniqueness of solution and motion patterns analysis

The pose estimation algorithm explained above guarantee a unique solution. With two views, there would be many combinations of unit vectors constrained to their planes that satisfy the angular variations condition. The third view constrains the set of combinations, thus the solution is unique. It is estimated from two rotations and is considered as a robust orientation method due to the use of angle between lines. A pattern of motions analysis is useful in the case movements are controlled and can be selected to simplify calculations.

In the form the orientation estimation is tackled by this method, where small rotations in different axes provide a unique solution and the object is represented by a 3D line, some selected rotations provide faster computations. If it is assumed that one of the axes of the coordinate frame of the object is the unit vector that describes the 3D line direction, with two different one component rotations, the relative orientation result comes directly from the numerical solution. With normal axes of rotations the angle between the planes formed by $v_a v_b$ and $v_b v_c$ is $\pi/2$, as seen in Figure 4.4. It also serves to reduce calculations. This is relevant when applications require real time performance, providing fast results that in many cases should be updated frequently.

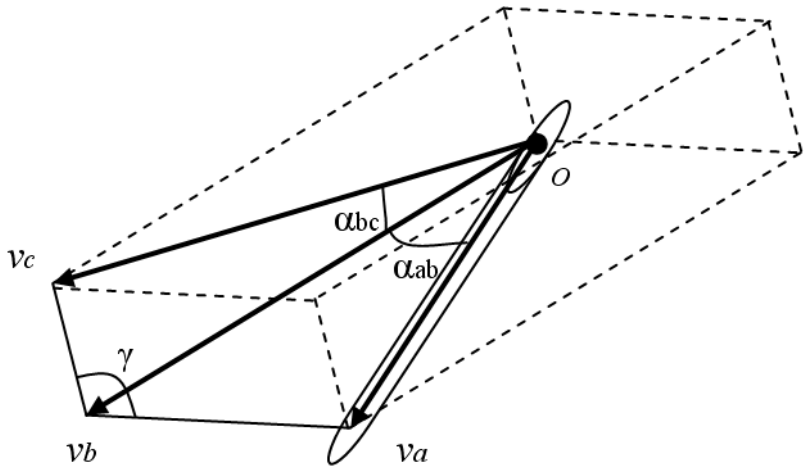

Figure 4.4. Two selected rotations through normal axes simplify computations. Having a determined aperture of planes γ, the angular variation could generate a different pattern even when α_{ab} and α_{bc} are equal.

4.4. POSE FROM CONSTRAINED ROTATIONS

The line-based rotational motion analysis described on the previous section served as an introduction to the visualization of projective properties of the angular variation. It demonstrates the possibility to estimate the relative orientation of an object line through defined rotations. However, as this algorithm is based on exterior or 3D constraints by the imposition of lying in projection planes, the 3D rotation angles are required as an input. In this section an extension of this algorithm is proposed. It improves its dependence on 3D information by the addition of new specific rotations of the object. Therefore a relation of the 2D-3D angles is applied, in which geometric invariants of projected features serve as the base to determine Euclidean magnitudes. First the orientation of the object is estimated through simple determined motions. Subsequently, according to its orientation, specific registered positions are identified to fulfill some geometric conditions. This permits to estimate the 3D angles and consequently the orientation of the object with respect to the camera, as well as its position.

4.4.1. Projected geometric invariants

Projected features of the rotated object are represented in the image plane by lines. A perspective camera maps these projections where certain geometric properties are preserved [29]. Straight lines map to straight lines. Presenting the line-based rotational motion problem as in the previous section, where three projection planes S_a, S_o and S_c correspond to three views of the object after two rotations around a centroid, it is possible to approximate the 2D-3D angular relation if small rotations are performed about the same axis. This is due 3D angles, which are Euclidean invariants, are not preserved under perspective mapping. Therefore, if the object line is rotated about the same axis by a determined angle, according to this property, the angles between projection planes are approximately equal. This is:

$$n_a . n_o \approx n_o . n_c \tag{4.27}$$

where $\boldsymbol{n}_a$, $\boldsymbol{n}_o$ and $\boldsymbol{n}_c$ are the normal unit vectors to planes S_a, S_o and S_c respectively. Thus projective features provide relevant information through Euclidean invariants. The smaller are the 3D angles of rotation, the more approximated are the projection planes separated.

Applying the line-based rotational motion analysis, this set of equations is obtained:

$$v_a . v_o = \cos\alpha \tag{4.28}$$

$$v_o . v_c = \cos\alpha \tag{4.29}$$

$$\left(v_a \times v_o\right)\left(v_o \times v_c\right) = -1 \tag{4.30}$$

having v_a as a unit vector representing the initial orientation of the object, and v_o and v_c representing the orientation after two subsequent rotations. Since the 3D angles α_{ao} and α_{oc} are equal, they are expressed by α. And according to the principle of this three unit vector configuration, *cos(γ)* of equation (4.21) is -1, as the rotations are performed about the same axis. With this set of equations the three unit vectors can be calculated only if the 3D angles are provided. Therefore, constraints must be applied considering the properties of projective mapping of feature lines.

4.4.2. Projection planes constraint

Applying the same concept presented in the previous section, the infinity of possible orientations which satisfy the set of equations (4.28), (4.29) and (4.30), is constrained through the imposition of unit vectors v_a, v_o and v_c to lie in their respective projection planes. Thus, using unit quaternions to represent the unit vectors as a result of a rigid transformation, equation (4.28) can be expressed as:

$$q_a v_{dd} q_a^* q_o v_{dd} q_o^* = \cos\alpha \tag{4.31}$$

which is an example of the representation of one of the equations of the set. Where q_a is the unit quaternion applied to v_a, q_a^* is its conjugate and v_{dd} is the unit director vector that connects the center of projection and the centroid. The vector to be rotated is v_{dd}, which is common to the three projection planes, and its normal vector n_a is the axis of rotation. Thus, applying the constraint to (4.29) and (4.30) as well, with n_o and n_c as their respective axes of rotation, a set of three nonlinear equations is obtained. This constraint simplifies calculations resulting in three unknowns which are the angles of rotation through their respective projection planes: Ω_a, Ω_b and Ω_c.

Although the projection plane constraint simplifies the equations, it is not sufficient to estimate the 3D angles. Therefore, it is necessary to include 3D information and take advantage of the projective invariant approximation. It can be accomplished by rotating the instrument two consecutive times about the same axis a determined angle ($\alpha_{ao} = \alpha_{oc} = \alpha$). This permits to reduce the unknowns, Ω_a, Ω_b and Ω_c, to only one Ω due to the approximation of equation (4.27). If the unit vectors v_a, v_o and v_c lie on a plane, with angles α_{ao} and α_{oc} equal, and are constrained to their projection planes, there are two possible combinations to satisfy the equations. The first is centering v_o, it is $\Omega_b = \pi/2$, and v_a and v_c rotated in opposite directions, $\Omega_a = \Omega - \pi/2$ and $\Omega_c = \pi/2 - \Omega$, as seen in Figure 4.5. And the second Ω_a, Ω_o and Ω_c equal to Ω.

The projective invariant approximation reduces the three unknowns to only one in the case there are three unit vectors lying in a plane. It is possible by the establishment of a rule, to apply plane constrained rotations. The first combination mentioned above is

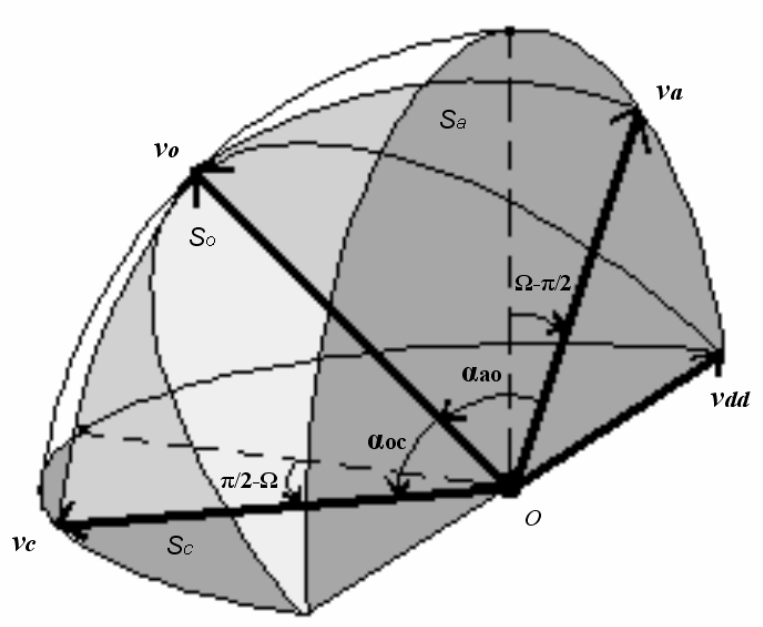

Figure 4.5. Unit vectors v_a, v_o and v_c are constrained to lie on projections planes S_a, S_o and S_c respectively. There is the same angle between the planes. Having the condition that $\alpha_{ao} = \alpha_{oc}$, the constrained unit vectors can be estimated by rotating v_{dd} in a determined relation of angles.

the rule implemented in order to obtain a unique solution. The second is not useful in this application: it is a singularity where the plane formed by the three vectors coincides with the projection plane. However, this case can be avoided applying different motions, or simply by augmenting or decreasing the produced angle.

The reduction of unknowns facilitates the estimation of the 3D angles; however there are many possible locations to satisfy equations (4.28), (4.29) and (4.30). In order to obtain a unique solution, 3D information is included. As seen in Figure 4.6 a second plane is added formed by three unit vectors. This plane is perpendicular to the formed by v_a, v_o and v_c, with equal angular rotations. It forms a right circular cone shape with the intersection of the two planes as its axis. Thus, only two unit vectors are added, v_b and v_d, with v_o as the cone axis.

With the inclusion of a second plane forming a right circular cone a new set of nonlinear equations is produced. To satisfy the angular condition of unit vectors around the central axis, as $\alpha_{ao} = \alpha_{oc} = \alpha_{bo} = \alpha_{od}$, equations (4.28) and (4.29) lead to:

$$v_a.v_c = v_b.v_d \tag{4.32}$$

and as the two planes are perpendicular:

$$(v_a \times v_d).(v_b \times v_d) = 0 \tag{4.33}$$

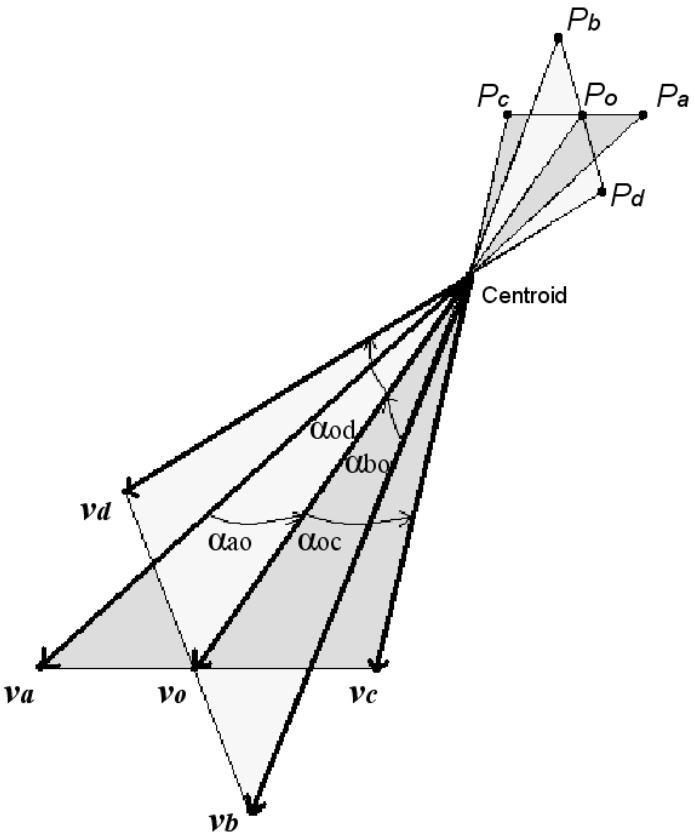

Figure 4.6. Determined transformations are applied to the object to obtain a unique solution. Unit vectors are estimated in order to calculate the 3D angles produced by the rotations, forming in this case a right circular cone with the intersection of the two planes as its axis.

Applying the projective invariant approximation and the relation of angles established as rules in each of the two perpendicular planes, the result is a set of two nonlinear equations with two unknowns: Ω_1 and Ω_2. Having a plane conformed by v_a, v_o and v_c in which the projective constraint is implemented with their respective unit quaternions with the common rotated unit vector v_{dd} and unknown Ω_1. And the perpendicular plane conformed by v_b, v_o and v_d, with unknown Ω_2. Solving the set of equation conformed by (4.32) and (4.33) a unique solution is obtained. Thus, as the 3D angles have been estimated, the orientation of the unit vectors, and consequently the relative orientation of the object can be calculated. Moreover, once having calculated the 3D angle, the position of the centroid can also be estimated knowing the transformations applied.

4.5. EXPERIMENTAL RESULTS

In order to analyze the angular variation produced by 3D transformations, and to validate the two proposed algorithms developed in this chapter, a series of simulations and real world data tests were carried out. First, simulations were performed to obtain a profile of the effects of different 3D angular amplitudes over the perspective angular variation. And subsequently, validations of the line-based rotational motion analysis and the constrained planes algorithms were performed using real data, from which interesting properties and drawbacks were found and are presented in this section.

4.5.1. Angular variation analysis through simulations

Simulations were performed through a graphics interface application where 3D transformations could be handled. The rigid object to be transformed was a line segment where one of its end points was the centroid. The position in the 3D space of the viewpoint was known and represented the center of projection of the camera. The input parameters were the angles of rotation of the object, α_1 and α_2, and the angle between the planes generated by these rotations, γ. Line features were identified in the image plane by straight-line detection algorithms. It served to measure the projected angles, β_1 and β_2 as the slope of its respective lines, and the unit vectors that described their projection planes to compute the relative orientation.

In Figure 4.7 can be observed the angular variation from the projected angles, β_1 and β_2, as a result of changes of γ. It is an example to depict the changes of the projected angles through specific rotations in a determined pose of the centroid. With static 3D angles, α_1 and α_2, and only variations of γ, it is shown that not only the projected angle differs from the 3D angle, but it also varies depending on its point of view. This is represented by the contrast between the constancy of the measured angle β_1, resultant from motions over an unchanged orientation, and the variations of β_2, which changes its orientation due to increments of γ. There are values of γ where the angular variation is higher with small variations of α. Thus γ can be employed to determine the minimum number of movements to estimate the orientation easier and decrease errors. In the case of 3D transformations with high angular variations, the 2D difference decreases constantly. This case is not useful. Motions with such a magnitude are not always applicable. Figure 4.8 is shows the performance of the orientation estimation through the rotational motion analysis. It can be observed the percentage error and the computation time with variations of α_1 and α_2. The error decreases with increments of α and provide good results for small and middle angles. On the other hand, the computation time increases exponentially with increments of α, due to the greater number of iterations required to solve the nonlinear set of equations. This example in particular results from calculations performed over a computer equipped with a Pentium IV@2GHz processor.

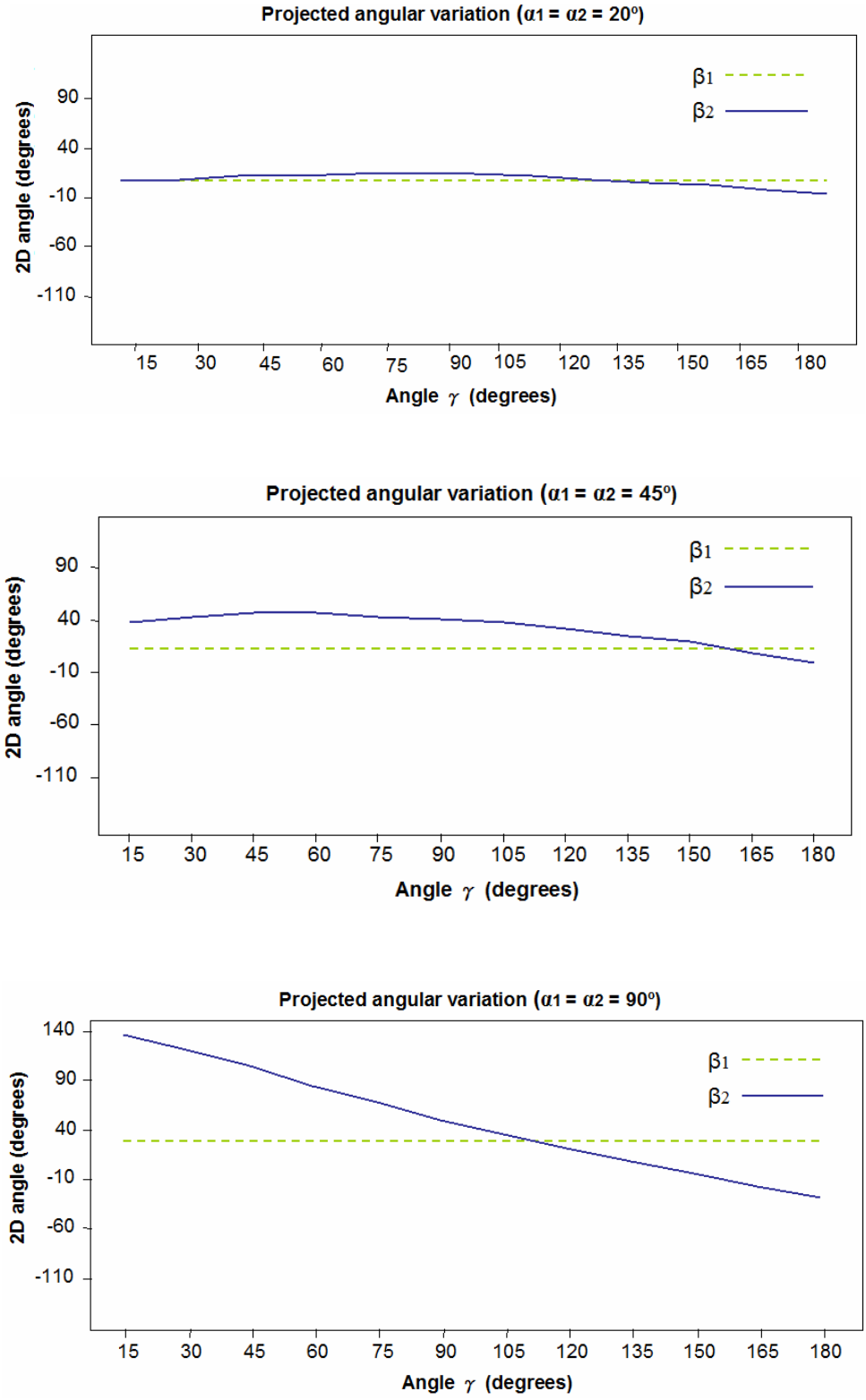

Figure 4.7. Example of the 2D angular variation induced by 3D rotations. The pose of the object was selected in such a form that variations of γ could only affect the orientation of the second motion. As a result uniform values of β_1 were obtained, while values of β_2 changed. It implies that there are angular variation patterns determined by the point of view of the observer.

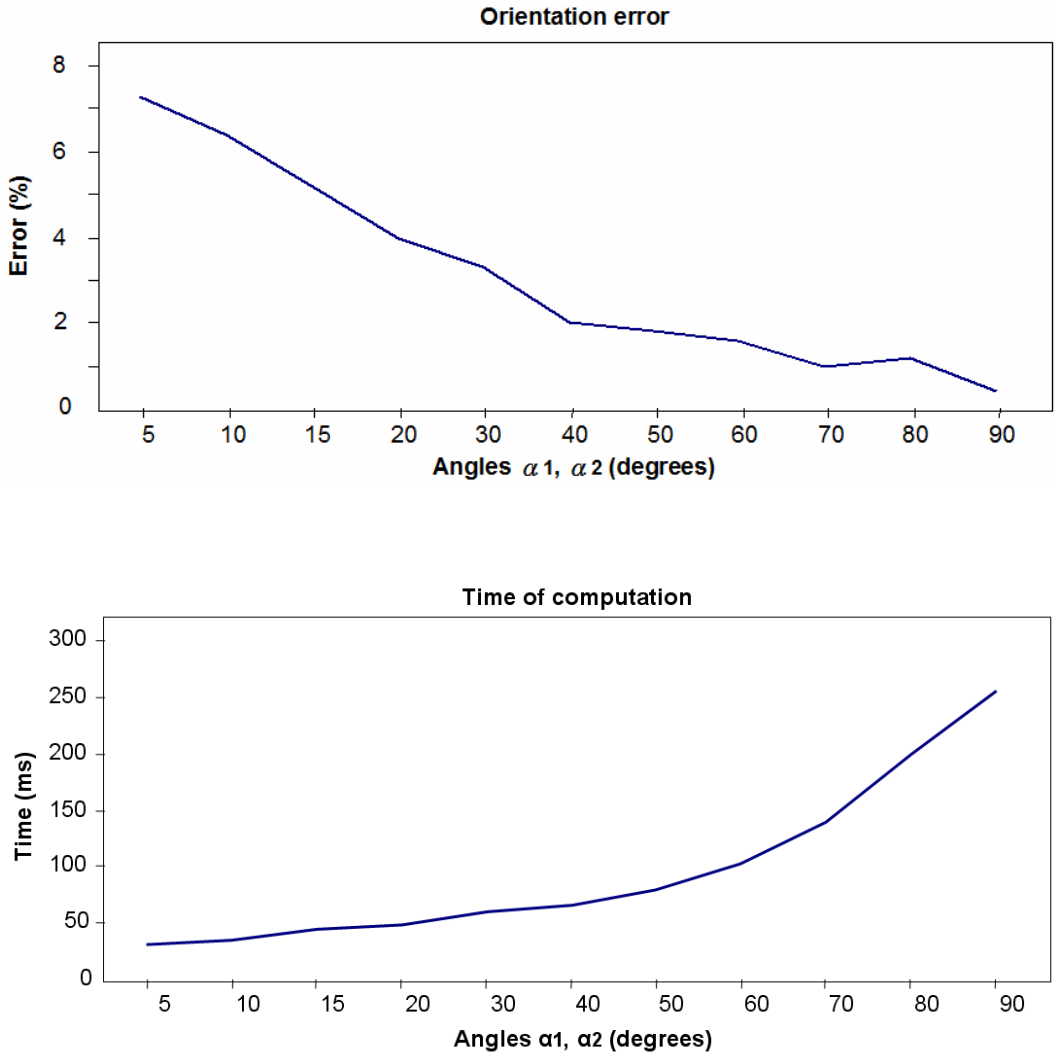

Figure 4.8. Relative error and computation time of the orientation using the rotational motion analysis algorithm.

4.5.2. Line-based rotational motion analysis results

Real world data was used to validate the algorithm. Experiments were carried out through a robotic test bed that was developed in order to get high repeatability. It consisted on an articulated robotic arm with a calibrated tool frame equipped with a tool, easily described by a line, which was presented in different precisely known orientations to a camera. The camera field of view remained fixed during the image acquisition sequence. The tool center of rotation was programmed to be out of the field of view, as it would be presented in an action-perception application.

With this premises, a standard analog B/W camera equipped with known focal length optics was used. This generated a wide field of view that was sampled at 768x576

pixels resolution. After image edge detection, Hough Transform was used in order to obtain the tool contour and the straight line in the image plane associated to it. Tool contour was supposed to have the longest number of aligned pixel edges in the image.

Having this setup, feature lines were identified and located in a sequence of images. These images captured selected positions of the rotated tool. Once the equations of the lines projected in the image plane were acquired, unit vectors normal to the constraint planes and v_{dd} could be calculated. This unit vectors and the motion angles α_1 and α_2 served as the input to the proposed algorithm. The intersection of the lines was needed to calculate v_{dd}. This calculation is prone to errors due to be located out of the field of view. It means the intersection of a different number of lines is not usually a single concurrent point. Table 1 shows the standard deviation in pixels of the intersection points calculated through different motion angles. The intersections converge to a single point when the angles between lines are higher.

Table 1. Standard deviation of intersecting lines (in pixels) through different motion angles, being the intersection point defined by coordinates *Xint* and *Yint* of the image.

Degrees	*Xint*	*Yint*	Standard. Dev.
5	895.81	264.11	62.31
10	928.28	278.91	35.65
15	861.41	250.71	6.58
20	876.67	256.52	5.83

The orientation of the object line attached to the robotic arm was calculated by the algorithm performing a sequence of determined rotations. It related this 3D angular information to the projected lines detected by the vision system. Tests for motions with an aperture angle γ between 5 and 20 degrees were carried out. Figure 4.9 shows the percentage error for different angles α_1 and α_2. Which were selected to be equal ($\alpha_1 = \alpha_2$), thus the three required views were given by each position of the object. There can be seen a common feature to the entire test results, it is the considerable high error at small values of angles α_1 and α_2. It implies that at small rotation angles, there is not an enough projected angular difference to be appreciated. This identification of angular difference is a crucial factor to obtain an optimum solution and is determined by the resolution of the line feature extraction method applied. This effect can also be seen through the error decreasing tendency as the rotation angles increase. Therefore, small aperture angles γ influence as well in the solution of the algorithm and explain the stability of a high error even when the motion angles increase.

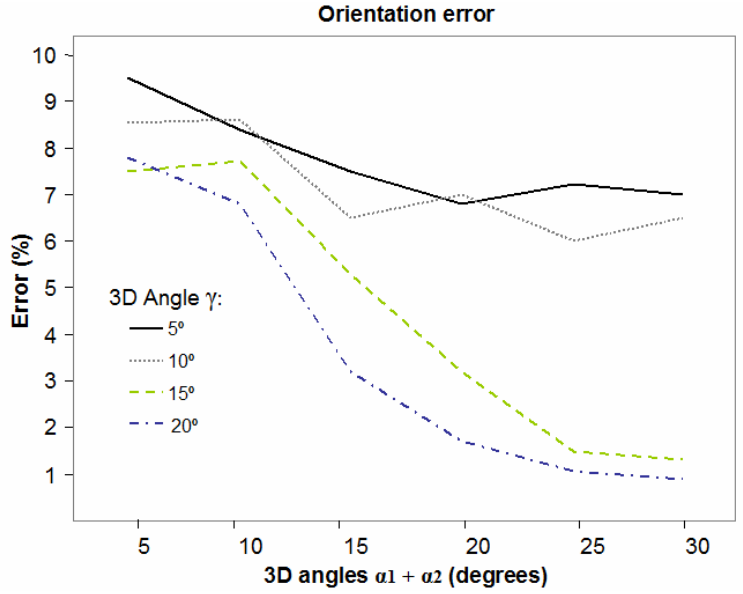

Figure 4.9. Orientation error using the rotational motion analysis algorithm. There is an important source of errors at small angular differences.

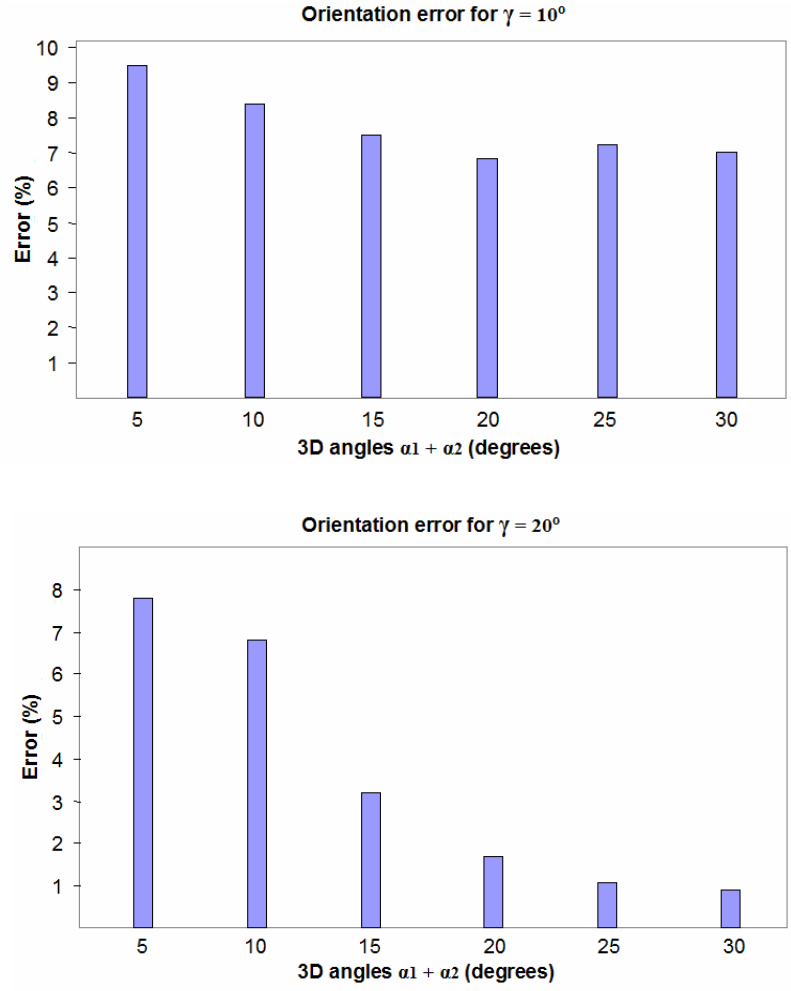

Figure 4.10. Algorithm performance comparison between 10 and 20 degrees, with a first rotation α_1 followed by a second α_2 of the same magnitude.

Figure 4.10 compares the algorithm performance for 10 and 20 degrees aperture angles. There the error varies differently. It can be seen as two tests that are below and above a threshold determined by the capacity of the line extraction method to detect precisely angular measures from the image plane. In the case of 10 degrees, the test is below the threshold, thus the error remain stable even with increments of α_1 and α_2. On the other hand, in the case of 20 degrees, the test is above the threshold. Hence the error is reduced as the angular measures become appreciable. It is also notable the effect of the intersection point, the calculation of $\boldsymbol{v}_{dd}$ has a great impact.

4.5.3. Constrained rotations results

Having the setup conformed by the object line attached to a robotic arm, feature lines were identified and located in a sequence of images. These images captured selected positions of the rotated object. Once the equations of the lines projected in the image plane were acquired, unit vectors normal to the constraint planes and $\boldsymbol{v}_{dd}$ were calculated online. Therefore, having the facilities given by the calibrated configuration, the 3D points P_0, P_a, P_b, P_c and P_d needed to build the right circular cone were located. $\boldsymbol{v}_{dd}$ was calculated by the intersection of feature lines. This calculation is prone to errors due to be located out of the field of view.

Having the right circular cone formed by determining the 3D points P_0, P_a, P_b, P_c and P_d, the resulting angles between the projection planes can be observed in Figure 4.11. There, two couples of projective planes are easily distinguished. The projective planes S_a and S_c containing $\boldsymbol{v}_a$ and $\boldsymbol{v}_c$ which lie on a plane, and S_b and S_d with $\boldsymbol{v}_b$ and $\boldsymbol{v}_d$ which lie on its perpendicular. As the 3D angle of rotation (α) increases, the angle between couples of planes tends to separate. This demonstrates that in this case, the projective invariant property can be implemented with small angles. However, as shown in Figure 4.12, the percentage error grows exponentially with increments of α, which on contrary to the previous algorithm tests, was estimated from its angular projection. There the length of the object line was estimated in order to determine the position of the centroid. Therefore the applicability of the algorithm is strongly influenced by the motions performed, especially by the magnitude of the angles and the reliability of the approximation. As in the previous tests, there is an error increment when angular magnitudes are small. It is consequence of the resolution of the line extraction method, where angular differences cannot be appreciated. In the same way, the reliability of the approximation is not easy to be achieved. Thus, although it is possible to estimate 3D information directly from image data, this method requires a set of conditions that are difficult to fulfill.

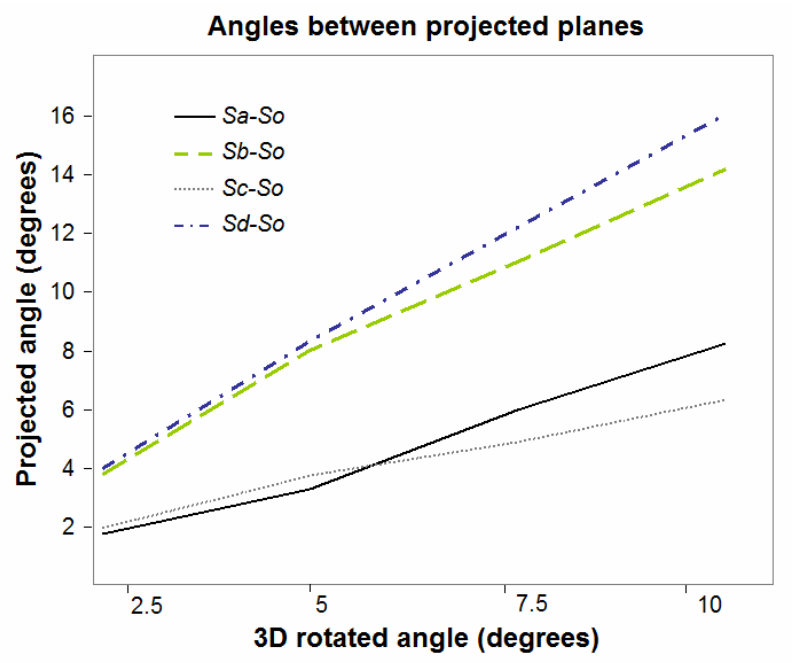

Figure 4.11. Angular variation between projection planes in an arbitrary view. Two couples of projective planes are distinguished. *Sa-So-Sc* as v_a, v_o and v_c lie on a plane, and *Sb-So-Sd*, as v_b, v_o and v_d lie on its perpendicular. As the 3D angle of rotation of the instrument (α) increases, the angle between couples of planes tends to separate.

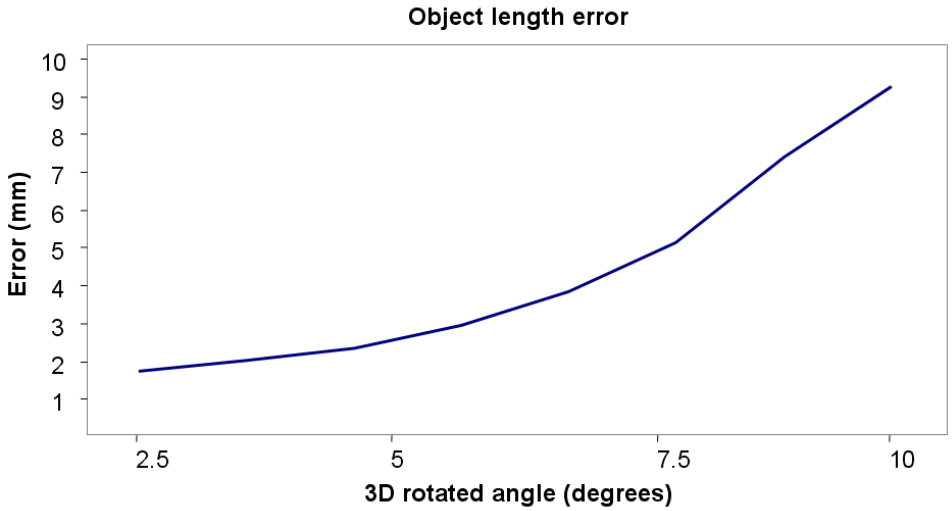

Figure 4.12. Object length error estimated with the constrained planes motion analysis algorithm in an arbitrary view. The error increases as the projected lines approximation deviates.

4.6. DISCUSSION

The angular variation produced by projections of an object line under 3D rotations presents interesting properties from which spatial information can be obtained. A simple line feature resultant from the perspective mapping of an object line can provide relevant 3D information through angles formed by its own motion or by a set of concurrent lines. It has been seen in this chapter that 3D angles are mapped differently in an image. Their magnitudes vary depending on the point of view of the observer. This fact suggests that there is a relation between the angular variation performed by a certain object line and its orientation with respect to the camera. Therefore, two methods are proposed to estimate the pose of an object through the analysis of the angular variation caused by 3D rotations.

A common aspect of the two proposed algorithms is the imposition of 3D constraints. Since the information provided by a projected line is not enough to obtain a unique solution, the application of constraints over the lines to lie on their respective projection planes reduces the range of possible solutions. Thus, applying this constraint to a set of projected lines, a unique solution can be obtained by the accomplishment of a certain 3D structure. For the first algorithm, the line-based rotational motion analysis, the set of concurrent lines must be at least of three. However, the 3D angles between them must be known. It shows that with only two rotations the angular variation between lines provides sufficient information to estimate the relative orientation of the object. This motion analysis produced answers to addressed questions as the uniqueness of solution for the minimum number of movements and possible motion patterns to solve it directly.

The second proposed algorithm, the constrained planes motion analysis, adds new movements or lines to be projected in the image plane to calculate the spatial information. On contrary to the previous algorithm, it first estimates the angular magnitudes between the 3D lines. It requires an increased number of constraints due to the lack of supplied 3D information. Thus, a cone-shaped structure should be described by the rotated object line. This is a notable drawback of the method, which in order to accomplish this type of 3D structure specific rotations must be performed. This is not a common model; therefore it should be manipulated, as in the experiment tests, by a robotic arm. This algorithm also makes use of an angular approximation, which aside of being difficult to fulfill, is based on a determined orientation of the cone. As a right circular cone, its base is conditioned to remain tangent to a sphere centered at the origin of the camera frame. Thus, the approximation can be considered reliable.

4.7. REFERENCES

[1] T.S. Huang and A.B. Netravali, "Motion and structure from feature correspondences: A review," *Proc. IEEE*, vol. 82, pp. 252-268, 1994.

[2] J. Olensis, "A critique of structure-from-motion algorithms," *Computer Vision and Image Understanding,* vol. 80, pp. 172-214, 2000.

[3] C. Harris, "Geometry from visual motion," In: A. Blake and A. Yuille (Eds.), *Active Vision*, Chapter 16, MIT Press, Cambridge, Mass, 1992.

[4] B. Wrobel, "Minimum solutions for orientation," In: A. Gruen and T. Huang (Eds.), *Calibration and Orientation of Cameras in Computer Vision*," Chapter 2, Springer-Verlag, 2001.

[5] M. Spetsakis, "Structure from motion using line correspondences," *Int. J. Computer Vision,* vol. 4, pp. 171-183, 1990.

[6] S. Christy and R. Horaud, "Fast and reliable object pose estimation from line correspondences," *Proc. Int. Conf. Computer Analysis Images Patterns,* pp. 432-439, 1997.

[7] F. Dornaika and C. Garcia, "Pose estimation using point and line correspondences," *Real-Time Imag.*, vol. 5, pp. 215-230, 1999.

[8] J.M. Selig, "Some remarks on the statistics of pose estimation," Technical report SBU-CISM-00-25, South Bank University London, 2000.

[9] F. Shevlin, "Analysis of orientation problems using plücker lines," *Int. Conf. Pattern recognition*, vol. 1, pp. 685-689, 1998.

[10] Y. Liu, T.S. Huang and O. Faugeras, "Determination of camera location from 2-D to 3-D line and point correspondences," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 12, pp. 28-37, 1990.

[11] R. Horaud, T.Q. Phong and P.D. Tao, "Object pose from 2-d to 3-d point and line correspondences," *Int. J. Computer Vision*, vol. 15, pp. 225-243, 1995.

[12] J.S. Park, "Interactive 3D reconstruction from multiple images: a primitive-based approach," *Pattern Recognition Letters*, vol. 26, no. 16, pp. 2558-2571, 2005.

[13] A. Ansar and K. Daniilidis, "Linear pose estimation from points and lines," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, pp. 578-589, 2003.

[14] A. Matveev, X. Hu, R. Frezza and H. Rehbinder, "Observers for systems with implicit output," *IEEE Trans. Automatic Control*, vol. 45, pp. 168-173, 2000.

[15] E.D. Dickmanns, "Dynamic vision-based intelligence," *AI Magazine*, vol. 25, no. 2, pp. 10-30, 2004.

[16] H. Rehbinder and B.K Ghosh, "Pose estimation using line-based dynamic vision and inertial sensors," *IEEE Trans. Automatic Control,* vol. 48, no. 2, pp. 186-199, 2003.

[17] B.L Yen and T.S Huang, "Determining 3-D motion and structure of a rigid body using straight line correspondences," *Image Sequence Processing and Dynamic Scene Analysis,* Springer-Verlag, 1983.

[18] O. Faugeras, F. Lustran and G. Toscani, "Motion and structure from point and line matches," *Proc. Int. Conf. Computer Vision,* 1987.

[19] J.R. Holt and A.N. Netravali, "Uniqueness of solution to structure and motion from combinations of point and line correspondences," *Journal of Visual Communication and Image Representation,* vol. 7, no. 2, pp. 126-136, 1996.

[20] K.S. Roberts, "A new representation for a line," *IEEE Proc. Computer Vision and Pattern Recognition,* pp. 635-640, 1988.

[21] R. Hartley and A. Zisserman, *Multiple View Geometry*, 2nd ed. Cambridge University Press, 2004.

[22] O. Faugeras, *Three-Dimensional Computer Vision*, 4th ed. The MIT Press, Cambridge, Mass. 2001.

[23] M. Hohmeyer and S. Teller, "Determining the lines through four lines," *J. of Graphic Tools*, vol. 4, no. 3, pp. 11-22, 1999.

[24] T.Q. Phong, R. Horaud, A. Yassine and P.D. Tao, "Object pose from 2-D to 3-D point and line correspondences," *Int. Journal of Computer Vision,* vol. 15, no. 3, pp. 225-243, 1995.

[25] H. H. Chen, "Pose determination from line-to-plane correspondences: Existence condition and closed-form solutions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, pp. 530-541, 1991.

[26] A. Rhijn and J. D. Mulder, "Optical tracking using line pencil fiducials," *Eurographics Symposium on Virtual Environments,* 2004.

[27] K. Daniilidis, "Hand-eye calibration using dual quaternions," *Int. J. Robotics Research*, vol. 18, pp. 286-298, 1999.

[28] J.M. Kuang and M. Liu, "A novel explicit pose estimation algorithm based on euclidean geometry," *Proc. Int. Conf. Information Acquisition*, 2007.

[29] D. Mumford, J. Fogarty and F. Kirwan, *Geometric Invariant Theory*, 3rd ed. Springer, 2002.

Chapter 5

Perspective distortion model as a function of angular variation

5.1. INTRODUCTION

The angular variation has demonstrated to be a stable and reliable feature which provides 3D information. Previous approaches described in Chapter 4 were developed as an attempt to introduce this feature as a monocular cue for spatial perception. Although both proposed algorithms, from the simple line-based rotational motion analysis [1], to the more sophisticate cone-shape line rotation structure [2], successfully provide pose information, they are strongly dependent on certain scene data conditions. They are based on the application of 3D constraints on 2D input data. It implies that angular information extracted from images must satisfy 3D conditions based on a specified scene structure. This is a relevant limitation when 3D information is not available or the required structure cannot be accomplished. An alternative approach is described in this chapter. Inspired in the monocular cues used by the human visual system for spatial perception, it is based on the proper content of the image. It is focused on the distortion of geometric configurations caused by the perspective projection. Thus, the necessary information is completely contained within the captured camera view.

The monocular cue for spatial perception used in this approach is the linear perspective. It is related with the appearance of objects under perspective transformations on a flat surface. This bidimensional view generated by the representation of a 3D object is the result of a planar projection. It is obtained by mapping each point of the object onto a plane through passing lines emanated from a center of projection. Depending on the desired visual effect derived by the representation, this mapping may be different. It

could show the general appearance of an object or depict its metric properties. When the center of projection is finite, a perspective projection is obtained. It presents an object as it is seen by the eye and is generally used in computer vision applications.

A perspective projection provides a realistic representation of an object. An impression of depth is created on a 2D surface and its 3D shape can be visualized. However, to provide this impression the geometry of the object is strongly distorted. Different parts are represented at different scales and parallel lines converge at a single point. It implies that such a projection is not considered by the principles of Euclidean geometry [3]. Under perspective transformations distances and angles, which are Euclidean invariants, are not preserved. Nevertheless, different properties of geometric configurations remain unchanged. For instance, a straight line is mapped to a straight line.

The invariance of geometric configurations under perspective transformations plays an important role in computer vision. It is a geometrical description of objects from perspective images which prevail unchanged, a useful property in object recognition and navigation [4]. In many cases projective invariants are the only properties of interest on a perspective image. However, the changes on geometric configurations derived by perspective transformations can provide useful 3D information as well.

In this chapter the particular interest is focused on the analysis of the perspective distortion. The visual representation of an object depends on the point of view of the observer. This implies that 3D information, as position and orientation of an object, could be estimated from changes on geometric configurations. In this case the geometric property to analyze is the angular variation. This simple property is deeply influenced by perspective transformations and serves to model the perspective distortion.

The angle between two projected straight lines varies depending on their orientation with respect to the camera. In a series of coplanar incident 3D lines separated by a constant angle, a perspective projection affects its appearance changing each angular magnitude differently. The only geometric property that remains invariant is its cross-ratio. Since the locus of points that forms the pencil of lines is a conic curve, the invariance of the cross-ratio is an important property to take into account. It serves as the base of the perspective distortion model developed in this chapter to establish the nature of the projection. It describes the distribution of angular magnitudes of the projected pencil of lines, and consequently, the orientation of the 3D plane where they lie can be estimated.

Numerous techniques have established conics as the ideal features to estimate the pose of an object. Particularly circles are considered as compact geometric primitives with enough 3D information [5]. The perspective projection of a circle in any arbitrary orientation is always an exact ellipse [6]. Thus, the first step adopted by previous approaches to solve the model analysis of circular features has been the estimation of the

elliptical parameters by least-square fitting techniques [7] [8] [9]. In our case, straight lines are the features to be extracted and the angular magnitudes are the geometric properties from which the elliptical parameters are obtained.

The position and orientation of a circular feature can be completely specified by the coordinates of its center and the surface normal vector [10]. The projection of this center does not correspond to the ellipse center [11]. An analysis of this position distortion model of the projected center was proposed in [12]. It studies the visual effect caused by perspective projection. However, it requires projections of the circular feature and its center as well. Through the perspective distortion model developed in this chapter, the angular variation specifies the axis and angle on which the constructed circular surface is oriented. This novel approach is accurate and simple. We show its feasibility in static images, using a pencil of lines, as in a sequence with a moving line feature.

5.2. PROJECTIVE NATURE OF CONICS

A conic is a plane curve obtained by the intersection of a right circular cone with a plane. Since this intersection does not pass through the vertex of the cone, the resulting curve can also be called a non-degenerate conic section or proper conic [13]. In Euclidean geometry proper conics can be differentiated as parabolas, ellipses and hyperbolas. However, in 2D projective geometry these types of proper conics are equivalent as they describe the same properties under projective transformations [14]. An ellipse, including a circle as a special case, is a conic in which its intersecting plane cuts all the elements of one half of the cone. This geometric configuration in particular plays an important role as a common feature to extract from images due to its regular presence and the spatial information it contains. It is a curve strongly related to the cross-ratio and defined in terms of its invariant nature.

The hypothesis of modeling the perspective distortion, further developed in this chapter, is based on the identification of variation patterns which describe the spatial relation between the object and the observer. It could be surprising to find out that these variation patterns are related, as will be shown later, with the projective invariance of specific geometric configurations. In this section the concept of the conic as a projective invariant is introduced. It is presented as a useful geometric configuration and explains its importance in a diversity of computer vision applications. Especially, it is presented as a key feature to provide visual cues to space perception through pose estimation algorithms or, as in this case, describing the perspective distortion derived by a projective transformation and its relation with angles.

5.2.1. Conics and the cross-ratio

A notable fact to emphasize in terms of projective invariant theory over the projective nature of conics is that it is a curve defined by the cross-ratio. It is a result of Chasles' Theorem:

A proper conic is the result of the locus of four coplanar points, no three collinear, which forms a pencil of lines with point of incidence in the same plane preserving a constant cross-ratio.

This projective invariant property can be compared to the Euclidean construction of the circle. It is formed by the locus of points with a fixed distance from a determined point. In that case distance is the invariant under Euclidean transformations [4].

The strong relation between the cross-ratio, a fundamental projective invariant, and conics provides answers to general ambiguities raised by the analysis of uniqueness of projections which involves planar conics. As stated by the Chasles' Theorem, a cross-ratio is also defined for a set of four points on a proper conic. It can be seen in Figure 5.1 that having a set of points x_i, $1 \leq i \leq 4$, on the proper conic Γ, and another point p also on Γ, the lines from p to x_i define a cross-ratio which is by definition independent on the choice of point p [15].

The order in which the points are taken on a line or a conic affects the value of the resultant cross-ratio (Cr). There are 4! = 24 possible permutations, nevertheless due to the symmetries there are only six distinct values and come in reciprocal pairs. These permutations can be produced by selecting all the point combinations. However, there exists a rotational function called the j-invariant [16], which is independent of the order of the points and is defined by this equation:

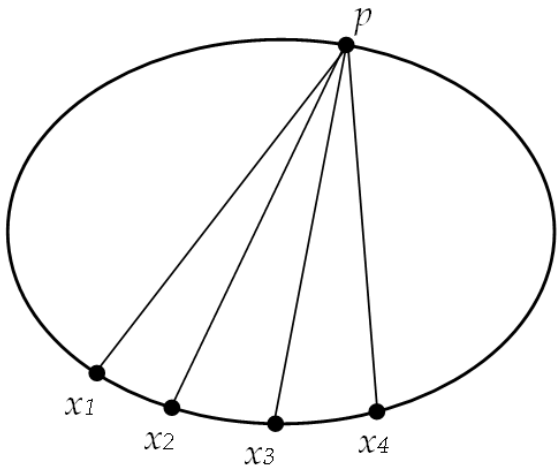

Figure 5.1. Definition of the cross-ratio by a set of four points on a proper conic. The value of the cross-ratio is independent of the choice of p.

$$j(Cr) = \frac{(Cr^2 - Cr + 1)^3}{Cr^2 (Cr - 1)^2} \tag{5.1}$$

It is a relevant property in computer vision applications, particularly in object recognition through image feature invariance, the possibility to acquire a unique value of the cross-ratio independent of the order in which they were taken. Nevertheless, as is the interest of this approach, the invariant property of the cross-ratio also provides an extent source of projective cues to visual perception of space as will be shown in next sections.

5.2.2. The projection of a circle

The realistic impression of depth represented on scene painting or artistic drawing is achieved following certain rules based on the laws of perspective. These rules were first introduced by Greek painters and geometers during classical antiquity [3]. The fact that the appearance of objects under projective transformations is not preserved and, in particular, the necessity to understand how these shape changes could be represented, led geometers to formalize the effect of perspective on geometric properties. A notable feature they took into account about the effect caused by perspective was the dependence of the distortion of a geometric configuration on the point of view of the observer. This dependence has been studied as a source of space information in computer vision applications, especially in conics, where the effect of perspective distortion is pronounced.

A circle, a special case of ellipse, is a geometric configuration strongly distorted under projective transformations. As is shown in Figure 5.2, the nonparallel projection of a circle is an exact ellipse in which its projected center and the center of the ellipse do not coincide. This is an important feature extensively used in computer vision applications. Having a distorted circle and its center, the exterior orientation can be determined by a diversity of methods. In contrast, it is important to mention another projective property of the projected circle, which is invariant to Euclidean transformations. This is the absolute conic, a particular conic contained in the plane at infinity [17]. As some applications take advantage of the perspective distortion, others make use of this invariance, especially in camera self-calibration algorithms. In general the absolute conic demonstrates that there are four positions of a circle with known radius, which corresponds to a given image conic [4].

The projected circle, as a conic, is strongly related with pencils of lines. As it is stated in the Chasles' Theorem, conics can be constructed by the locus of points that conform a pencil of lines with a fixed cross-ratio. The projective property of this geometric configuration, represented by a set of concurrent lines, is the invariance of the cross-ratio defined by the angles between its lines. Since a conic is distorted by projective transformations, the angular distribution between these lines must satisfy the constant

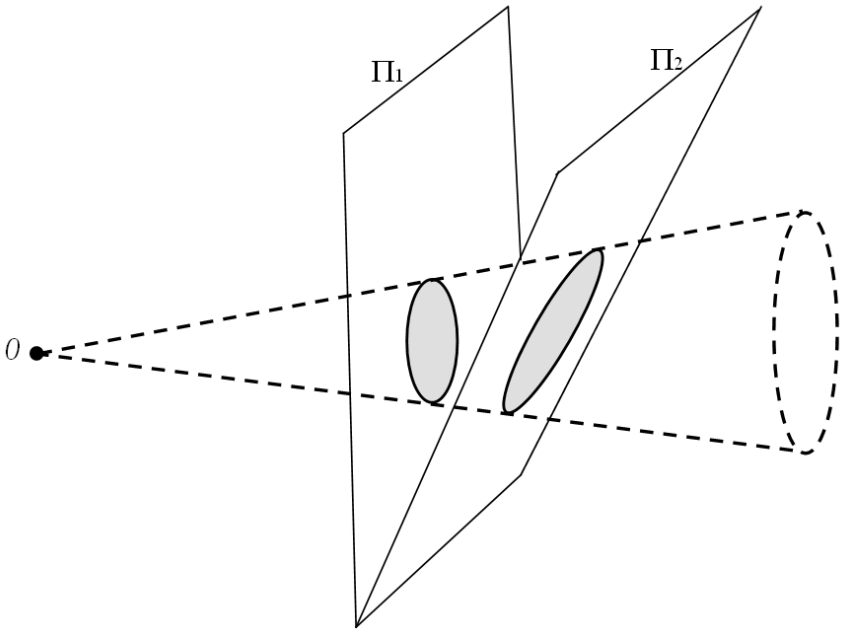

Figure 5.2. Example of the projection of a circle. As a conic is the result of the intersection of a right circular cone with a plane, the projection of a circle is an ellipse.

cross-ratio property. Nevertheless, while the ratio of ratios represented by the angles is invariant, the angular distribution varies in a pattern that depends on the orientation of the conic with respect to the observer. This angular variation pattern serves as the base of this approach and through it; the perspective distortion caused by projective transformations can be modeled.

5.2.3. Pose estimation from conics

Numerous computer vision applications have used conic features for 3D pose estimation. Particularly the circle is considered as a compact geometric primitive in which the pose information of an object is contained. This feature is common in nature or man-made objects. Applications as the estimation of the pose through structured light in cylindrical workpieces [12], or camera calibration algorithms [18], have used the regular presence of this feature.

A closed-form solution was reported in [19] for determining the pose from a single image. It is an analytical solution based on a reduction of the general equation of conicoids. However, it has the exception of a two possible orientations solution. Since an elliptical shape is the result of the perspective projection of a circle in any arbitrary orientation, a second image was suggested in [5]. It uniquely determines the 3D motion parameters based on the eccentricity change of the circle. Thus, based on the elliptical parameters of the projection, the 3D orientation can be estimated.

The estimation of the projected elliptical parameters is considered as the first step to determine the 3D pose of a circular feature with known radius. Subsequently, the problem is defined as the estimation of the pose of the plane that intersects a given cone and generates a circular curve. This given 3D conic surface is described by a base, which is the projection of the circular feature in the image plane, and a vertex, which is the center of projection.

The general form of a quadratic curve can be expressed as follows:

$$au^2 + buv + cv^2 + du + ev + g = 0. \tag{5.2}$$

It represents Γ_i, the circular base of the cone. With image axes u and v parallel to x_c and y_c respectively. In a camera coordinate system $\Omega_c(x_c,y_c,z_c)$, with center of projection o and z_c axis perpendicular to the image plane Πi, as shown in Figure 5.3.

In a perspective camera with focal length f, these axes are related by

$$u = f.x_c / z_c \tag{5.3}$$

$$v = f.y_c / z_c. \tag{5.4}$$

It permits to represent the conic surface, with vertex o, in the camera coordinate system by the following equation:

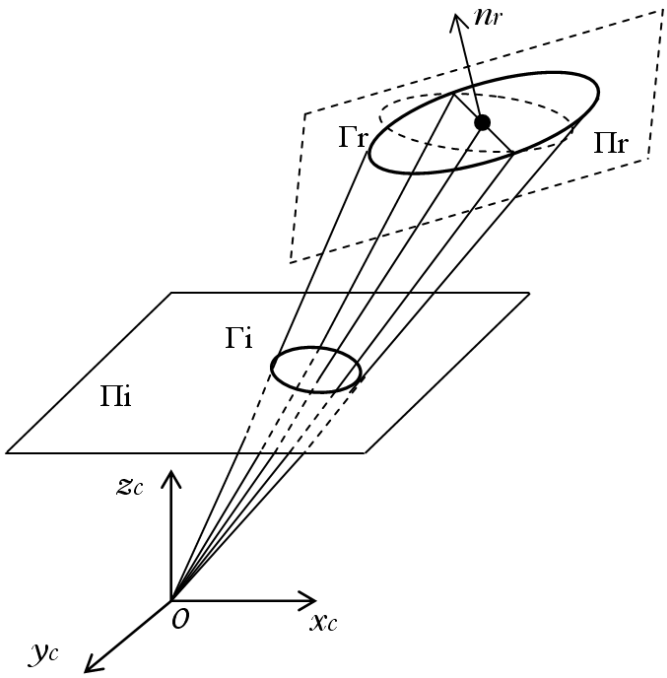

Figure 5.3. The transformation of a conic under perspective projection is a conic. A circular object projected in any arbitrary orientation is mapped to an ellipse in the image plane.

$$Ax_c^2 + Bx_c y_c + Cy_c^2 + Dx_c z_c + Ey_c z_c + Fz_c^2 = 0 \tag{5.5}$$

which can be expressed in matrix form:

$$\Omega_c^t C \Omega_c = 0 \tag{5.6}$$

with the conic coefficient matrix C given by

$$C = \begin{pmatrix} A & B/2 & D/2 \\ B/2 & C & E/2 \\ D/2 & E/2 & F \end{pmatrix}. \tag{5.7}$$

An orthonormal transformation of the reference frame Ω_c must be applied to obtain a new cartesian coordinate frame centered on the same origin and aligned with the principal axis of the cone. This new reference frame $\Omega_e(x_e,y_e,z_e)$ results from a transformation of Ω_c by a rotation matrix R. Thus $\Omega_e = R\ \Omega_c$, with z_e as the axis of the cone.

The expression of (5.6) after the transformation is:

$$\Omega_e^t R^t C R \Omega_e = 0 \tag{5.8}$$

which is the equation of the cone in its central form and can be represented in a more compact form as:

$$\lambda_1 x_e^2 + \lambda_2 y_e^2 + \lambda_3 z_e^2 = 0 \tag{5.9}$$

where λ_1, λ_2 and λ_3 are eigenvalues of R^tCR. The parameters of matrix R can be calculated since it is a diagonalizing matrix for C, which is $R^tCR=Diag(\lambda_1, \lambda_2, \lambda_3)$ [4].

The pose of the circular feature can be specified by the coordinates of its center and the surface normal vector. Therefore, having the equation of the cone, the required vector is defined by the normal to the plane

$$lx_e + my_e + nz_e = p \tag{5.10}$$

whose intersection with the cone is a circle. Thus the solution can be expressed as the determination of the coefficients l, m and n, where $l^2+m^2+n^2=1$.

Methods focused on the pose estimation from conics have been commonly used and are based on the intersection of the cone, obtained from the image projection of the circle, with a determined plane. Geometric constraints, such as the change of eccentricity in a

second image or error compensation of the circle center, have been applied to obtain a unique solution.

In this approach conics are not the features of interest. Nevertheless, since it is a curve that can be constructed from the cross-ratio, it is strongly related with the projective properties of angular variation and consequently with the perspective distortion model developed in this chapter.

5.3. PERSPECTIVE DISTORTION MODEL

As mentioned above, the base of the perspective distortion model developed in this approach is the variation pattern of geometric configurations under perspective projection. A notable property of this type of projection is the natural representation of objects in the image plane. It gives a real appearance, as if they are seen by the eye. However, their shape is not preserved and their geometry is strongly distorted. Since these changes of a geometric configuration can be classified in a pattern, determined by the point of view of the observer, a distortion model exists from which 3D pose information of an object can be calculated.

A perspective projection is characterized by the convergence of lines, the diminution of size and nonuniform foreshortening [3]. Parallel lines converge to a single point if they are not parallel to the projection plane. This point is called vanishing point, and is the intersection of the projection plane and a line through the center of projection parallel to the set of parallel lines. This convergence of parallel lines results, as shown in Figure 5.4, in the diminution of size and nonuniform foreshortening of objects. As the distance from an observer increases, objects of equal size appear smaller. Likewise, depending on the position of the observer, parallel lines are unequally foreshortened.

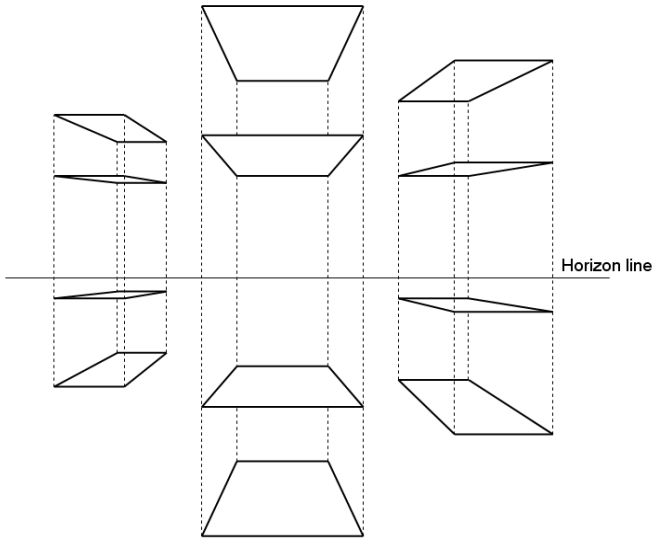

Figure 5.4. The diminution of size and nonuniform foreshortening are a consequence of the convergence of lines under a perspective projection.

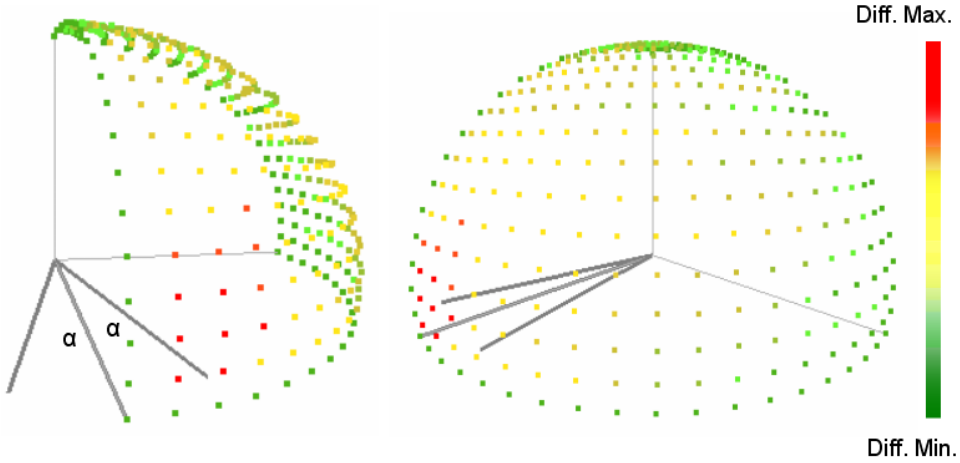

Figure 5.5. Angular difference of a pencil of lines from radial perspective views. The unequal foreshortening of lines depends on the position of the observer. It produces an increment of the angular difference as the parallelism between the projection plane and the pencil of lines deviates.

Under perspective projection geometric configurations of objects are strongly distorted. Only parts parallel to the projection plane preserve their shape. In the case of a pencil of lines separated by a constant angle, the concurrent lines are unequally foreshortened. The angular difference, as can be seen in Figure 5.5, varies depending on the point of view of the observer. This variation increases as the parallelism between the projection plane and the pencil of lines deviates. From radial perspective views this increment is progressive. Nevertheless, an abrupt growth is produced when the pencil is angled towards the observer. This is an area of extreme perspective, where distortion is notably pronounced. Therefore, the angular variation is a simple geometric configuration deeply influenced by perspective transformations and serves to model the perspective distortion.

5.3.1. Projective properties of lines

The change of geometric configurations under perspective transformations, particularly the angle between lines, plays an important role in modeling the perspective distortion. However, certain configurations invariant to these transformations present essential properties necessary to describe the nature of the projection. A qualitative property of this invariance is that the mapping of a straight line is a straight line. Likewise, as the ratio of distances is not preserved, the ratio of ratios of distances is invariant. This quantitative invariant is called cross-ratio, which in the case of four lines is given by:

$$Cr(U_1,U_2,U_3,U_4)=\frac{\sin(\alpha_{12})\sin(\alpha_{34})}{\sin(\alpha_{12}+\alpha_{23})\sin(\alpha_{23}+\alpha_{34})}. \tag{5.11}$$

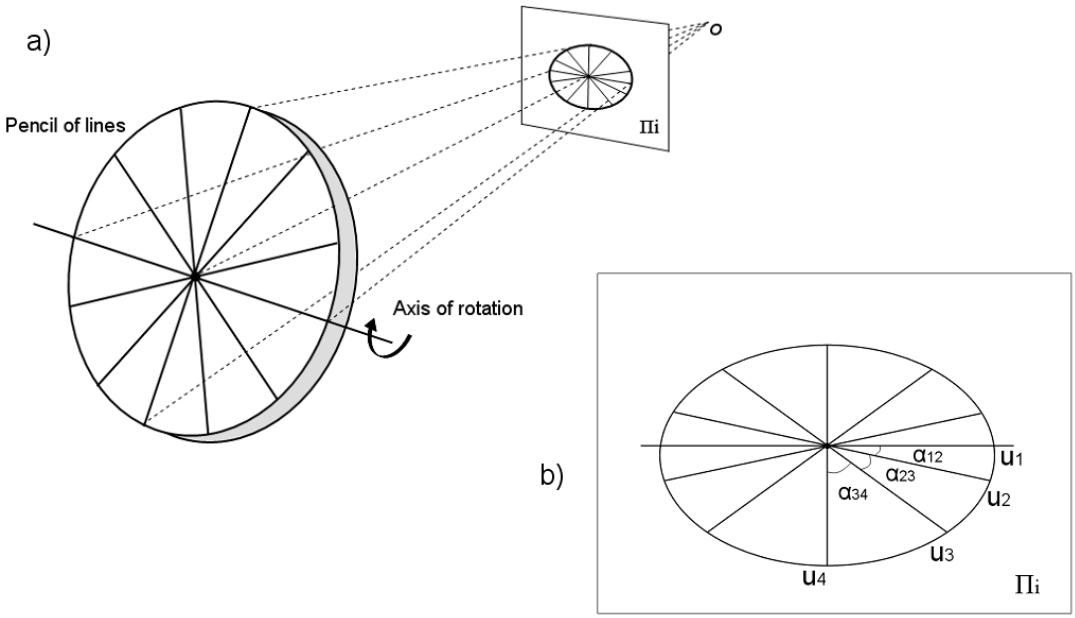

Figure 5.6. Perspective projection of a pencil of lines: a) The rotated circle in which the pencil is contained is projected to an ellipse in the image plane; b) An angular variation pattern is produced bringing the lines closer to the major axis of the ellipse.

Where the four lines (U_1, U_2, U_3, U_4), are incident at a single point, forming a pencil of lines in which its cross-ratio is defined in terms of the angles between them.

The invariance of the ratio of ratios defines the angular distribution of the projected pencil, in this case, the distribution of uniformly separated coplanar lines. Figure 5.6 shows first the 3D rotation of the circle in which the pencil is contained and afterward its resulting projection. There can be seen a tendency of the projected lines to concentrate closer to the axis of rotation with progressive increments of the angular difference between them. This tendency grows with the applied angle of rotation, having the property of maintaining a constant cross-ratio.

An exact ellipse is the result of a projected circle if it is not parallel to the projection plane (Πi). It is the locus of points that forms the pencil of lines, confirming the geometric property of the construction of conics from the cross-ratio. The invariance of this property is strongly related with the angular distribution of the projected pencil of lines, and consequently, with the eccentricity of the ellipse. It provides information about the orientation of the object, knowing that the eccentricity (e) of the ellipse can be expressed as a function of the angle of rotation (γ) in the form $e = sin\ (\gamma)$ [20] [21].

The unequal angular variation of the projected pencil of lines depends on the eccentricity of the formed ellipse. As shown in Figure 5.7, as the eccentricity grows, the set of

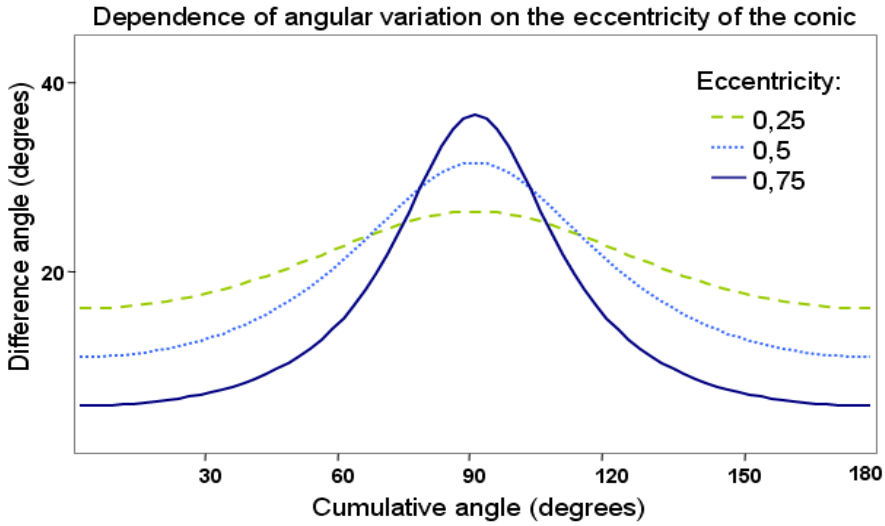

Figure 5.7. Example of the unequal angular variation of a 20 degrees projected pencil of lines. It depends on the eccentricity of the conic, which is function of the angle of rotation. The cumulative angle, parting from the major axis, shows a higher concentration of lines on the extremes and an abrupt change at the minor axis generated by eccentricity increments.

coplanar lines deviates from the parallelism with the projection plane. It is shown by the higher concentration of lines with a low angular difference at the extremes and an abrupt growth at the center as a result of eccentricity increments. This is an example of angular difference measures carried out between the lines of a pencil with an angular separation of 20 degrees under different rotations, being the parting point of the cumulative angle the axis of rotation (U_1). It implies an enhancement of the unequal foreshortening, and consequently, more projected lines concentrated at the axis of rotation. This axis is not the major axis of the ellipse. However, they are parallel if the center of projection and the incident point of the pencil are aligned. Likewise, the point of incidence of the pencil and the center of the ellipse are not coincident. Therefore, the distortion pattern represented by the angular variation appears to be described by a normal like function, where from any pose of the pencil of lines, a constant cross-ratio is obtained.

5.3.2. Aligned center model

The angular variation pattern of a projected pencil of lines provides sufficient information to model the perspective distortion. It describes the resulting projection of a determined angle depending on the position of the point of incidence of the pencil and the orientation of the circle it forms. It could be seen as a circle rotated about an axis. In the case the center of projection is aligned with the point of incidence of the pencil, this axis is coplanar to the circle and parallel to the major axis of the resulting ellipse. Therefore, the pose of the pencil with respect to the center of projection is defined by the angle of rotation of the circle (γ), which is given by the eccentricity of its own projection.

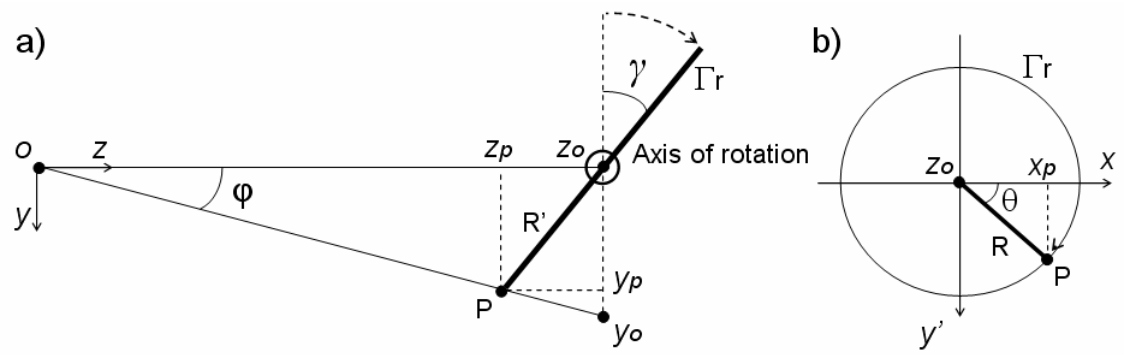

Figure 5.8. The alignment of the center of projection with the point of incidence of the pencil of lines describes the distortion model as the rotation of the pencil about an axis: a) Angle of rotation γ, from lateral view; b) Cumulative angle θ of the rotated line with length R, from frontal view.

Aligning the centers is the simpler case to model. Nevertheless, serves as starting point to analyze the projective properties of a rotated circle. If a sphere is centered at the center of projection (*o*), shown in Figure 5.8, with radius *Zo*, which is the position of the point of incidence, the axis of rotation is coplanar to the circle Γr, and tangent to the sphere at the point of incidence. This implies the possibility to estimate the pose of the pencil of lines by an axis and an angle of rotation defined by the angular variation model. The axis is given by the tangent line (where the projected lines concentrate), which is the point of minimum angle difference between the lines of the pencil. The angle of rotation (λ), however, as the eccentricity of the ellipse (*e*), must be calculated from the angular variation model.

The cross-ratio defines the conic constructed by the series of concurrent projected lines. To be calculated, a minimum of four lines is needed. If they are part of a pencil or form part of a sequence of a rotated line, it is necessary a constant angle between them. The rest of projected angles are calculated knowing the cross-ratio. Thus, parting from the axis of rotation, line U_1, the angular variation pattern of a projected pencil of lines can be determined. It can be expressed by the relation between the 3D applied angular changes and its projected angular data.

The angular variation model developed in this chapter describes the relation between a 3D angle and its own mapping in the projection plane. It is based on the rotation of a circle Γr, formed by the pencil of lines, about a determined axis. In the case of aligned centers, the axis of rotation is perpendicular to the *z* axis, as shown in Figure 5.8. Having the rotated circle, with radius *R*, centered at *Zo*; an angle α is the projection of the 3D angle θ at *Zp*, which is the position of the projection plane Πi. This 3D angle θ is the angle of rotation around the circle of a line with length *R*, centered at *Zo*. Therefore, if the angle θ leads to a point *P* along the circle, the projected angle α is formed by the angle between the projection of *P* at *Zp* and the axis of rotation of the circle. It can be expressed as:

$$\tan(\alpha) = y_p / x_p = y_0 / x_0 . \tag{5.12}$$

Where, having the x axis as the axis of rotation, the slope of the projected line in Πi describes the projected angle through x_p and y_p, which are the respective components of P in the plane xy. Similarly, the same projection in the base plane preserves the angle α through x_0 and y_0 as the respective components of the projection of P at Zo. Thus, by the geometry of the configuration, under a rotation γ, in the yz plane:

$$\tan(\varphi) = R'\cos(\gamma)/(Z_0 - R'\sin(\gamma)). \tag{5.13}$$

Having R' as the component of R in the y axis, it is:

$$R' = R\sin(\theta). \tag{5.14}$$

This leads to the expression of y_0 by:

$$y_0 = \frac{Z_0 R\sin(\theta)\cos(\gamma)}{Z_0 - R\sin(\theta)\sin(\gamma)}. \tag{5.15}$$

Similarly, in the plane xz at Zo, x_0 can be expressed as:

$$x_0 = \frac{Z_0 R\cos(\theta)}{Z_0 - R\sin(\theta)\sin(\gamma)}. \tag{5.16}$$

It implies the function which describes the angular variation in an aligned center configuration, knowing the relation $e = sin(\gamma)$, is defined as:

$$\tan(\alpha) = \tan(\theta)\sqrt{1-e^2}. \tag{5.17}$$

This model satisfies the fact that length and scale are not influential in the angular variation pattern of a projected pencil of lines. It is function of the applied 3D angle and the eccentricity of the ellipse. Figure 5.9 shows the influence of the angle of rotation γ in the projection of a rotated line around a circle. It parts from the axis of rotation, which is the point of minimum angular difference, to the minor axis of the ellipse, which is the point of maximum angular variation. This effect, in conjunction with the angular pattern extracted from the projected lines, make possible the calculation of the eccentricity of the ellipse. It is carried out by fitting the angular data to the model. Thus, the orientation of the circle is calculated.

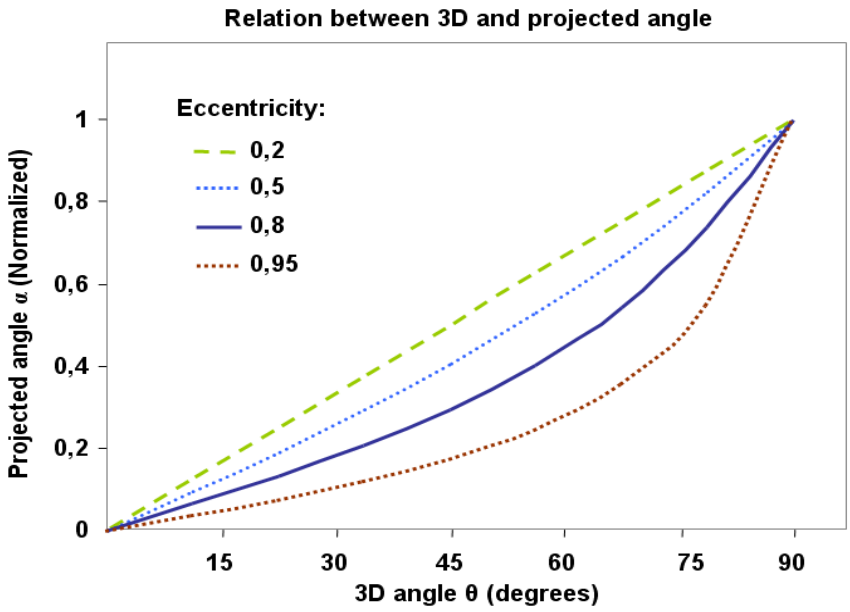

Figure 5.9. Effect of the eccentricity on the projection of a 3D angle. It parts from the axis of rotation to the point of maximum angular difference (minor axis of the ellipse).

5.3.3. General case model

Generally the point of incidence of the pencil of lines is not aligned with the center of projection along the z axis. Its projection can be located at any position in the image plane Πi and represented, in this case, by a unit director vector v_d from the center of projection. Using the concept of the sphere centered at o, presented in Figure 5.10, a tangent circle Γp is not parallel to the image plane, as would be in the aligned center case and consequently, the axis of rotation of the circle Γr is not parallel to the major axis of the projected ellipse. It implies an enhancement on the complexity of the configuration. However, the projective properties are equally satisfied by the angular variation pattern.

The methodology used in this general case approach is based, as in the aligned center model, on the calculation of the axis and angle of rotation of the circle Γr formed by the pencil of lines. Having the angular relation in (5.17) of two circles representing the rotation of a tangent circle about an angle, the general case configuration can be divided in two simpler aligned center models as shown in Figure 5.11. This is due to fact that in (5.17) it is assumed the image plane Πi is tangent to the sphere, and therefore, perpendicular to the axis of projection defined by v_d. Thus, the first part is conformed by the rotated circle Γr and a tangent circle Γp; and the second part by Γp and a circle Γi coplanar with Πi. Both parts are individual aligned center models. It can be seen as the result of the projection of the rotated circle in a tangent plane, and its consecutive projection in the image plane.

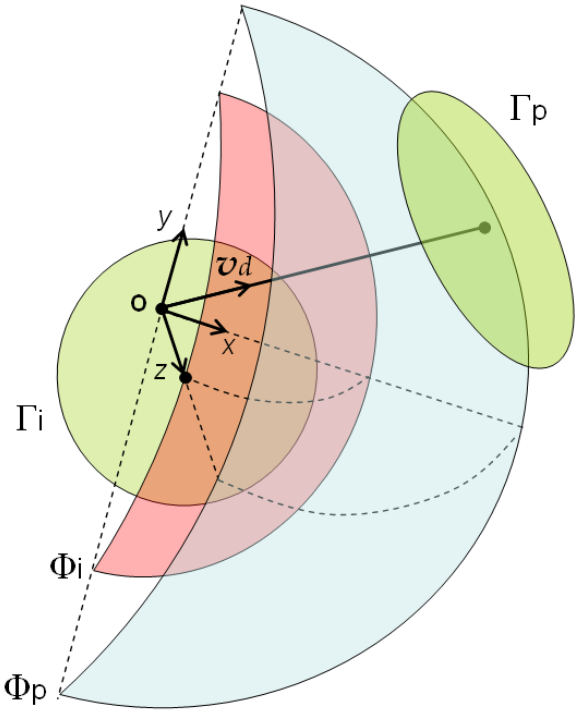

Figure 5.10. Using the concept of the sphere centered at *o*, the aligned center case is modeled referencing the orientation of the rotated circle to the tangent circle Γi. In the case the centers are not aligned, two concentric spheres, Φi and Φp, serve to establish a relation between the new reference tangent circle Γp and the image plane.

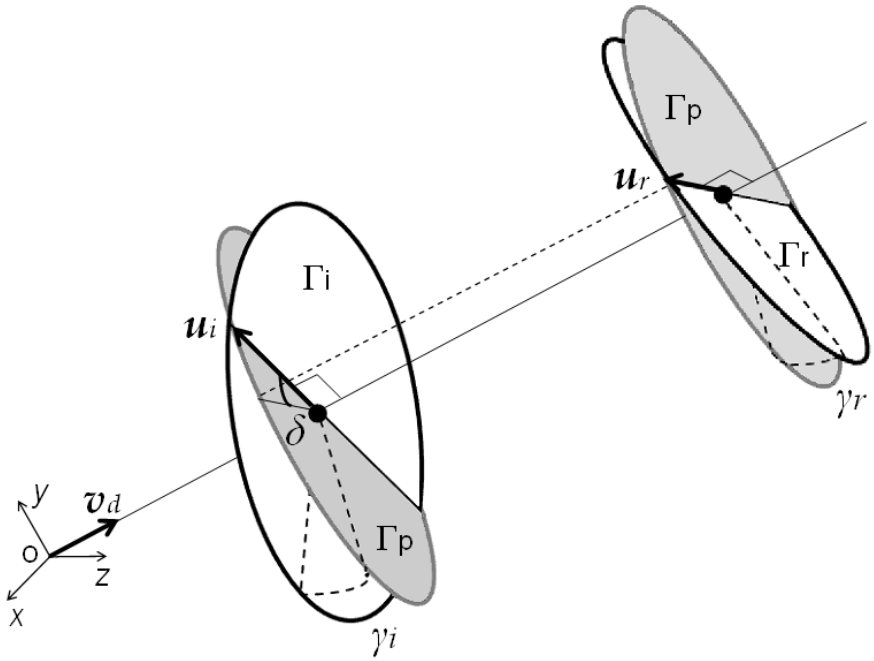

Figure 5.11. Division of the general case configuration in two simpler aligned center models. The rotated circle Γr is projected in a tangent plane to the sphere, and its consecutive projection in the image plane.

The projection of the tangent circle Γp in the image plane could be considered as the first step to define the model, since the initial data is contained in the image. The extraction of the feature lines of the pencil defines the angular pattern of the projection; and the orientation of Γp with respect to Πi through $\boldsymbol{v}_d$ provides the axis of rotation $\boldsymbol{u}_i$. This permits to describe the relation between the angle of rotation θ_p, around Γp, and the angle θ_i, around Γi, which is the result of the rotation of Γp about a given angle γ_i. Applying the center aligned model, this relation can be expressed as:

$$\tan(\theta_p) = \tan(\theta_i)\cos(\gamma_i). \tag{5.18}$$

Equally, the angular relation between the angle of rotation 3D θ_r, around Γr, and its projection in the tangent plane θ_p can be expressed by the aligned center model having γ_r as the angle of rotation of Γr and $\boldsymbol{u}_r$ as the the axis determined from (5.18).

The two parts of the divided configuration are related by the tangent circle Γp. However, the alignment of the two axes of rotation $\boldsymbol{u}_i$ and $\boldsymbol{u}_r$ is only a particular case of the model. Therefore, to relate the 3D angle θ_r with the angle projected in the image θ_i, the difference between these two axes must be taken into account. If this difference is defined by the angle δ, the angular variation pattern of Γr in the tangent plane is given by:

$$\tan(\theta_p + \delta) = \tan(\theta_r)\cos(\gamma_r) \tag{5.19}$$

which in conjunction with (5.18) express the angular variation of the 3D angle in the image plane as:

$$\tan(\theta_i) = \frac{\tan(\theta_r)\cos(\gamma_r) - \tan(\delta)}{\cos(\gamma_i)(1 + \tan(\theta_r)\cos(\gamma_r)\tan(\delta))}. \tag{5.20}$$

The projected angular variation depends on the angles of rotation with the respective tangent and image plane, and the difference between their axes of rotation δ. If there is no difference between these axes and the image plane is tangent to the sphere, the equation is reduced to the aligned center model. In any other case, γ_i and δ are determined from the image data. Thus, through the fitting of the angular data to the model, γ_r can be estimated and consequently, the orientation of the circle with respect to the center of projection can be calculated.

5.4. EXTERIOR ORIENTATION ESTIMATION

The perspective distortion model developed in the previous section describes the effect of the perspective projection on a rotated 3D line. It is based on the angular variation pattern depicted by the line features projected in the image plane. Since this pattern is affected by the position of the observer, spatial information of the 3D line with respect

to the camera can be calculated. This spatial information can be considered the orientation of the 3D line as a first approach given by the output parameters of the model. There are three output parameters, γ_i, γ_r and δ, which describe the angular variation pattern. Once they have been calculated, there is enough information to estimate the orientation of the 3D line. Therefore, parting from the input data provided by the image extracted features, spatial information is obtained.

The input features acquired from the captured images represent the Euclidean transformations of the 3D line. These transformations can be originated by 3D rigid rotations or directly constructed by a set of concurrent lines. The extraction of their image projection in the form of line features, despite of being considered implicit to the orientation estimation process, could be regarded as its first step. Thus, after having this 2D angular information, the perspective distortion model is applied in order to estimate the orientation of the line. This is a process that leads to follow a sequence of steps. Each of these steps performs a task with the general objective of calculating the output of the model through the fitting of the input data [22].

Normally regression analysis methods are used to fit sample data to a determined model. To achieve a close agreement between values of the regression model and the collection of sample data, a number of model parameters are adjusted. Thus, a smooth continuous function is obtained. It describes the sample data at certain values of the independent variable [23]. This independent variable, also called predictor variable, represents the input to the model, while the dependent variable represents its response. In this case, if these variables are x and y, respectively, the perspective distortion model could be expressed as:

$$y = Arc\tan\left(\frac{\tan(x)\cos(b1)-\tan(b2)}{\cos(b3)(1+\tan(x)\cos(b1)\tan(b2))}\right) \tag{5.21}$$

where $b1$,$b2$ and $b3$ are the parameters to be adjusted. Certainly, the perspective distortion model describes the relation between 3D and 2D angular variations. Therefore, the response of the model is the projected angular variation caused by a 3D rotation. Nevertheless, the important values taken into account to obtain spatial information are the adjustable parameters: $b1=\gamma_r$, $b2=\delta$ and $b3=\gamma_i$. Which are the required parameters to calculate the exterior orientation and hence the output of this fitting process. This process is shown by a block diagram in Figure 5.12.

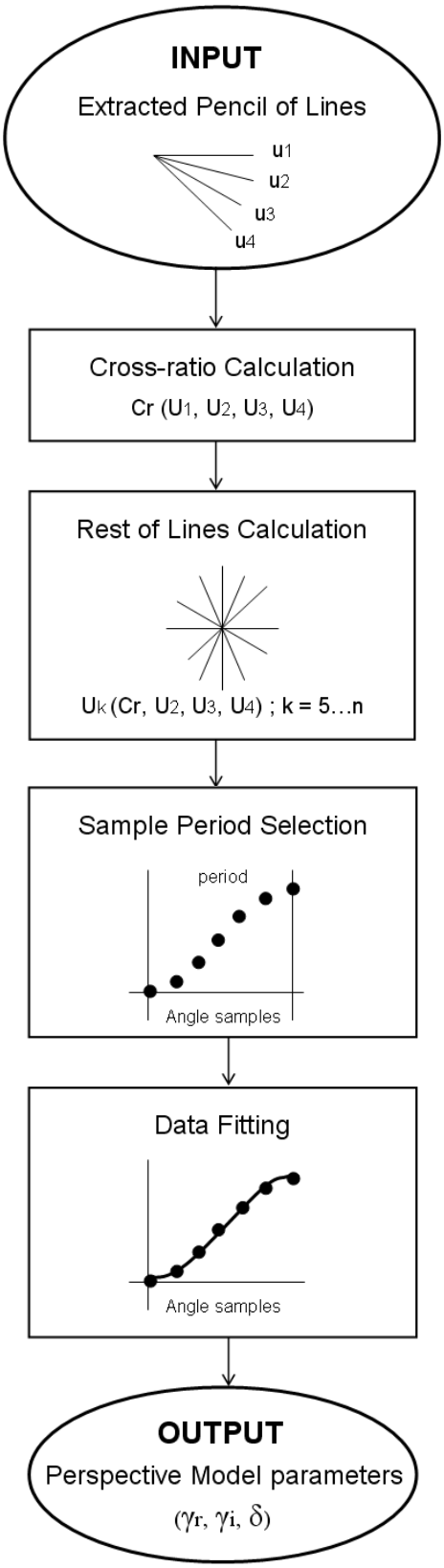

Figure 5.12. Block diagram of the process to calculate the required orientation parameters. The necessary information is provided by the adjustable parameters of the regression.

The regression analysis is based on the agreement between the sample data and the regression model. The measurement of this agreement is called the merit function. The idea is to adjust the model parameters to obtain the smallest value resultant from the merit function. It represents a close agreement between the collected data and the regression model. Therefore, the parameters are adjusted iteratively in order to obtain the best fit [24]. In this case the model has a nonlinear dependence on the adjustable parameters. Thus, the iterative process starts with some initial conditions, which become starting point for the next iteration and continue until the merit function achieves its minimum value [25] [26].

In the block diagram shown in Figure 5.12, the parameter calculation process of the perspective model could be divided in three parts. The first is the extraction of the line features from the image plane. The second is the preparation of the angular data as an ordered collection of samples. And third, the calculation of the output parameters from the nonlinear regression analysis. The preparation of the angular data, though it seems to be a simple arrangement of data, is an important part where a specified amount of determined samples are essential to obtain a good fit. In this particular case, the response of the model is not continuous. It is shown in Figure 5.13. As it is defined by trigonometric tangent functions, smooth continuous periods are clearly identified. Therefore, the collection of sample data is chosen in a way the regression analysis is performed in a single period.

The selection of the samples which conform a single period provides an optimum collection of data to achieve a good fit. It is considered a good fit a result with a merit

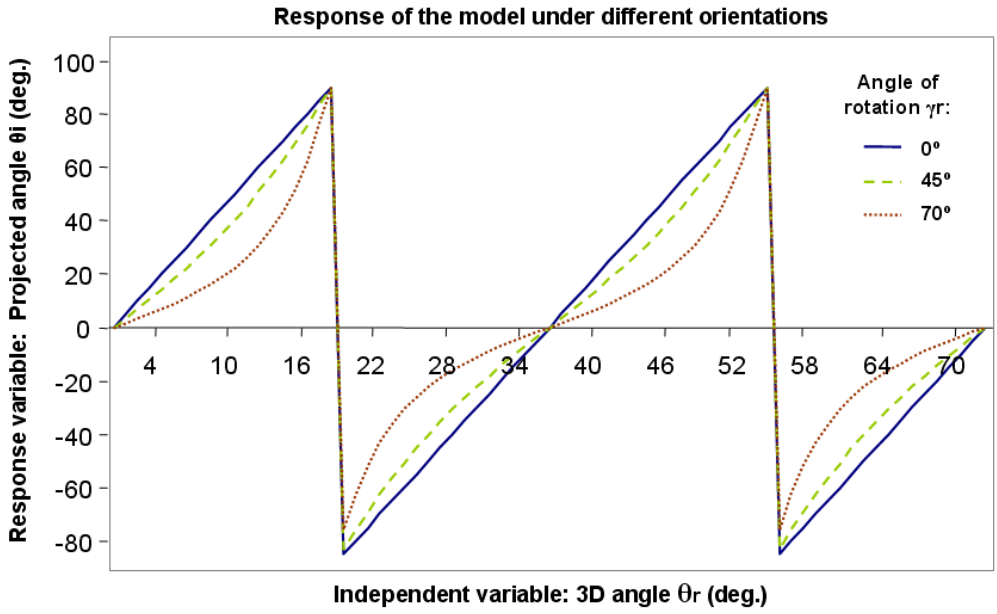

Figure 5.13. Response of the model to rotations of a pencil plane with a separation of five degrees between lines. The projected angle (θi), which is the response of the model, describes a non-continuous function. However, smooth periods are clearly identified.

function below an error threshold. It guarantees the agreement between the sample data and the regression model. By this means, there is enough data to perform a good fit choosing the samples of an entire period. The more measurements are given to be fitted; the better is the agreement with the model. However, despite the collection of samples represents a continuous part of the model function, the angular shift from where the measurements were taken with respect to the axis of rotation is unknown. This fact causes the regression analysis to fail, as it is unable to perform a good match when the collection of measurements does not correspond with the real angular shift.

An option to prevent the disparity between the real measured data and the model is adding a new variable. This new adjustable parameter (*b4*) serves to add an angular shift to the independent variable and consequently an offset. Therefore the collection of samples is adjusted to coincide with the model through the regression analysis. This addition of a new variable, though it solves the sample correspondence problem, implies more calculations and increases the difficulty of the regression algorithm to obtain a good fit. However, having the projection of the 3D lines intersection in the image plane, the unit vector $\boldsymbol{v}_d$ can be calculated. Thus, the angle γ_i, which is one of the model parameters, can be considered as known ($\boldsymbol{v}_d.k = cos(\gamma_i)$, having k as the unit vector normal to the image through the z axis). It changes the equation (5.21), except for the number of parameters to adjust that remain in three, and is expressed as:

$$y = Arc\tan\left(\frac{\tan(x+b4)\cos(b1)-\tan(b2)}{\cos(\gamma_i)(1+\tan(x+b4)\cos(b1)\tan(b2))}\right) \tag{5.22}$$

Finally, the exterior orientation is estimated from the model parameters. They can be seen as the path, through angular shifts, to move from the camera reference frame to the circle normal vector. This is accomplished by a sequence of rigid body transformations, which represent the orientation of the object. First, parting from the z axis unit vector k, to the unit director vector $\boldsymbol{v}_d$. And afterward, from $\boldsymbol{v}_d$ to the circle normal unit vector $\boldsymbol{n}_r$, as seen in Figure 5.14.

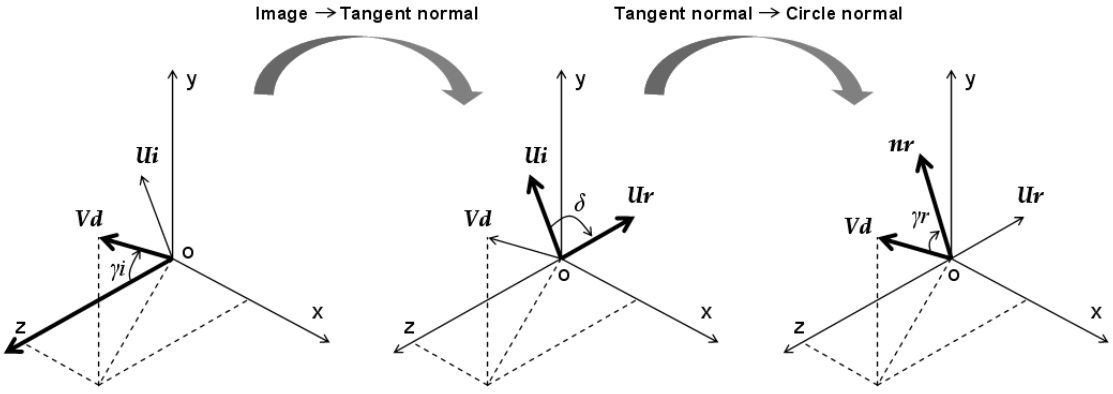

Figure 5.14. The exterior orientation is estimated from a sequence of rotations given by the model parameters: γ_i, γ_r and δ.

5.5. EXPERIMENTAL RESULTS

The perspective distortion model developed in this approach was validated by using real world data. The experimental setup consisted on a fixed standard analog B/W camera equipped with known focal length optics and a mobile frontal frame able to perform a desired orientation. The calculation of the best-fit parameters to the model was determined by the DataFitX$_{TM}$ nonlinear regression software. It utilized a merit function that minimized the sum of the squares of the distances between the actual data points and the regression line. The Levenberg-Marquardt method was the algorithm used to adjust the variables, having the parallel and center aligned position as initial condition and an error threshold of 0,001. The tests carried out were aimed to analyze the response of the model to different 3D angle inputs under a range of rotations of the frontal frame. Two sets of tests were employed to analyze the angular variation and its response to the model. The first set was focused on a single rotated 3D angle, while the second on a pencil of lines.

5.5.1. Single angle tests

A single 3D angle is formed by the intersection of two lines. It can be seen as the difference of the slope of each line on a plane, which is how the projected angle in the image plane was calculated. This angle served as input to the model, where its angular shift with respect to the axis of rotation, as a required parameter, was known. It is a drawback of the method using single angles. It is solved, as will be shown further by the use of a sequence of concurrent lines. According to the model, the angular variation of a single angle under rotation is depicted in Figure 5.15. It can be seen, parting from the parallelism between the image and the frontal frame, the reduction of the projected angle under rotations. Likewise, as the resolution of the response depends on the detected angular variation, it does not only increase with higher 3D angles, it also augments as the angle of rotation increases. As the area of extreme perspective provides enough angular variation with the more accurate response, the area of minor rotations provides negligible variations and consequently a misread response.

The sensitivity of the model is highly affected by the line feature error and the resolution of its acquisition method. Since the model response depends on the projected angle, its extraction using real world data is prone to inaccurate detections, particularly with small angles. The resulting effect of this line feature error is an unstable and poor performance of the model below a sensitivity threshold. Equally, the negligible variation at minor rotations produces a misread response below a threshold caused by the low perspective distortion.

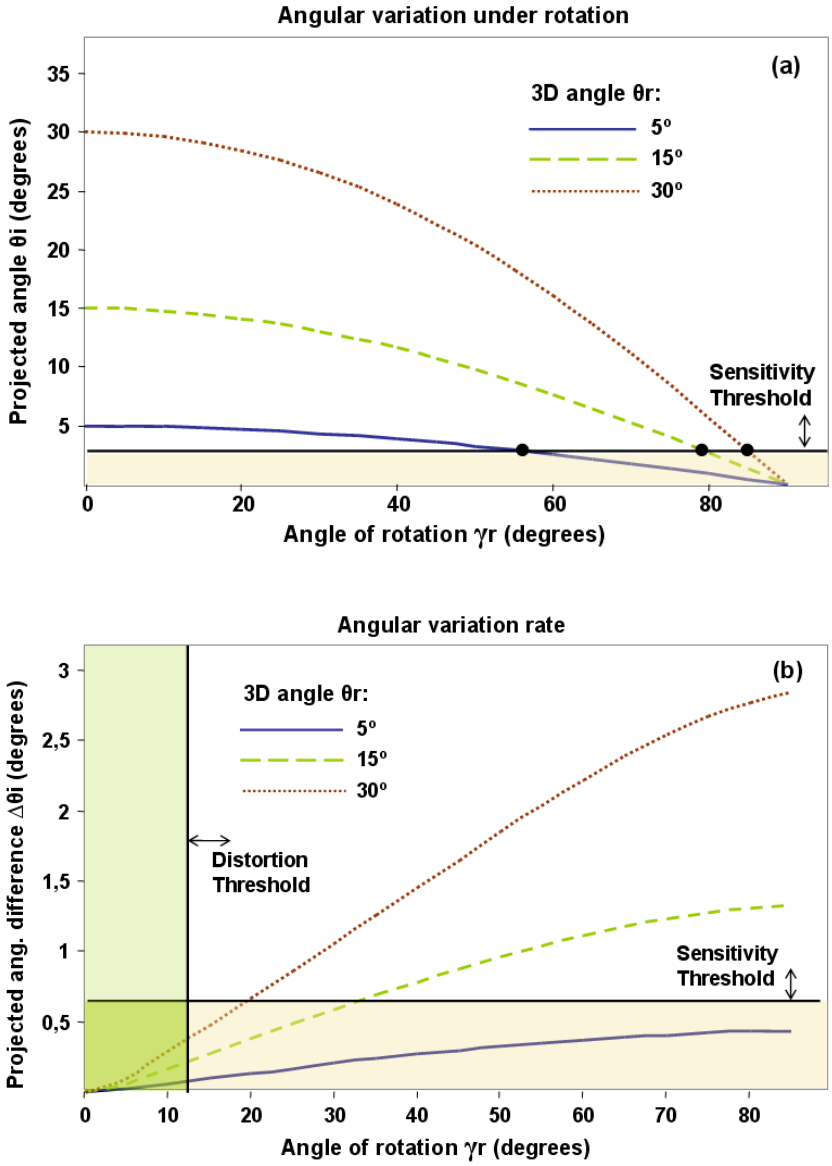

Figure 5.15. Angular variation of a single projected angle under rotation. The response of the perspective distortion model is highly affected by line feature errors and the resolution of its acquisition method: a) reduced projected angles caused by long rotations produce unstable performance under a sensitivity threshold; b) the response resolution, dependent on the angular variation rate, misreads data under a distortion threshold.

Figure 5.16 shows the performance error of the model in the single angle test having real world data as input. Two areas of notable high error are differentiated along the applied rotations. The first is located below the distortion threshold, where minor rotations were applied. This is an area in which the error is uniformly high independently of the 3D angle used. It is caused by the resolution of the line feature extraction method and the low perspective distortion. The second area, in contrast, is located at long rotations. It decreases as the 3D angle applied augments, caused by the sensitivity of the model with reduced angles at long rotations. In general, as can be expected, the error decreases with increments of the 3D angle applied.

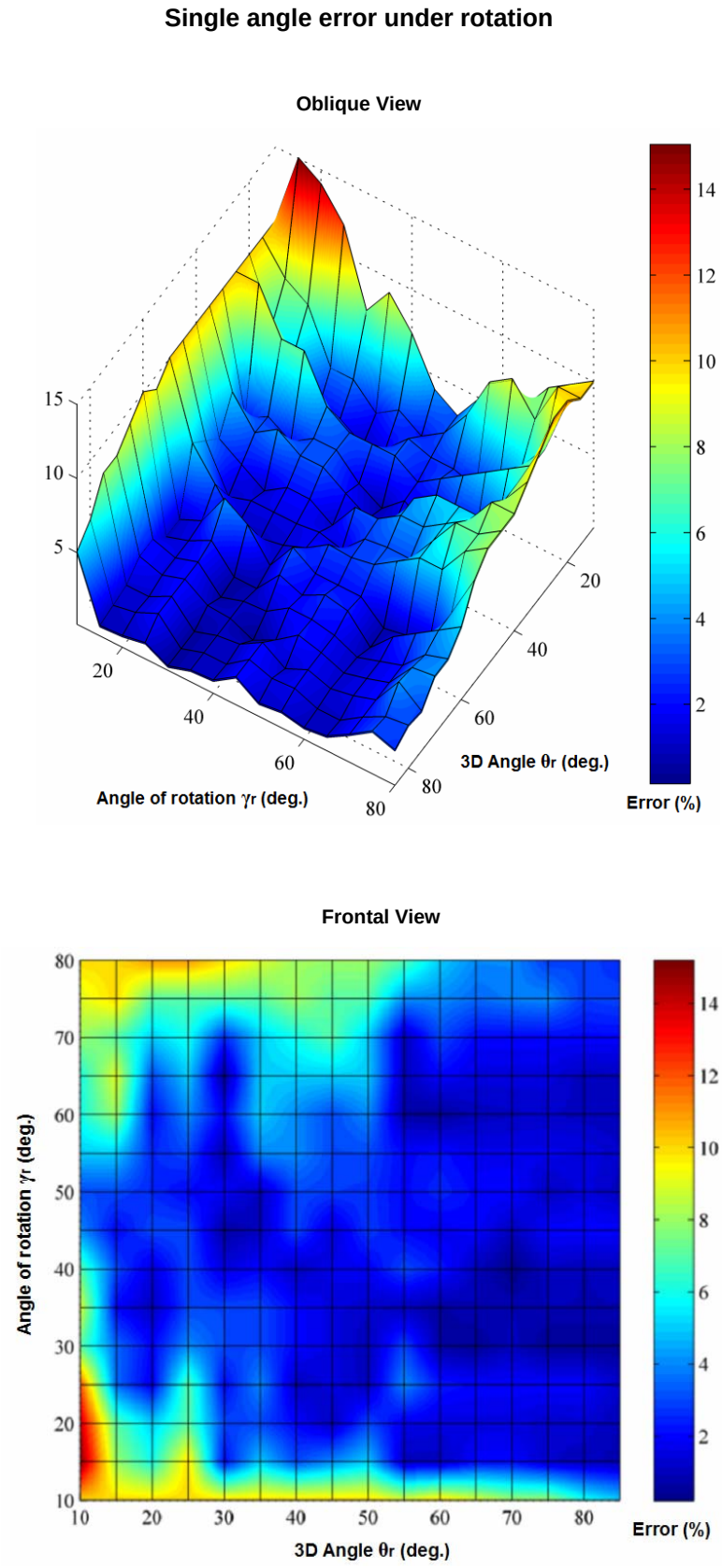

Figure 5.16. Oblique and frontal view of the performance error of the single angle test. The response of the model is prone to errors with negligible variations at minor rotations and small projected angles. This improves as the 3D angles augments.

5.5.2. Pencil of lines tests

The second set of tests employs pencils of lines of different constant angle separations. Figure 5.17 shows an example of the perspective distortion of a 15 degrees pencil at four distinct rotations. The method used to estimate the model parameters is based on the fitting of sample data, which are the projected measured angles, thus the model is satisfied. It requires at least four line samples to calculate the cross-ratio or three in the case it is known. It permits to calculate the next angles of the sequence of concurrent lines and consequently more samples to obtain an improved fit.

According to the invariant property of the pencil of lines, the distribution of the angles is unequal, while the cross-ratio is constant. Figure 5.18 shows the tendency of this angular distribution along rotations of the pencil plane. It presents the performance sensitivity to line feature error. As in the previous set of tests, the model response is highly affected by the resolution and error of the line feature extraction method at

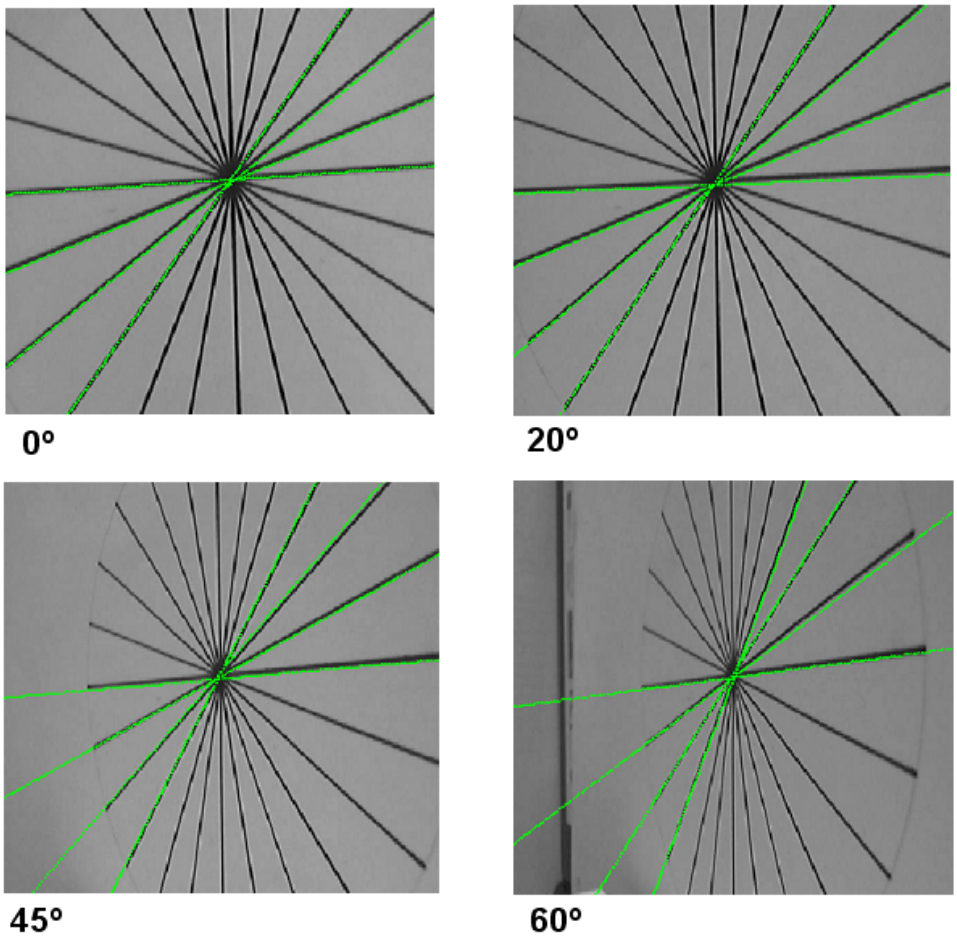

Figure 5.17. Example of the perspective distortion in a rotated pencil of lines separated a constant angle of 15º. Four extracted feature lines are enough to apply the model if the cross-ratio is unknown.

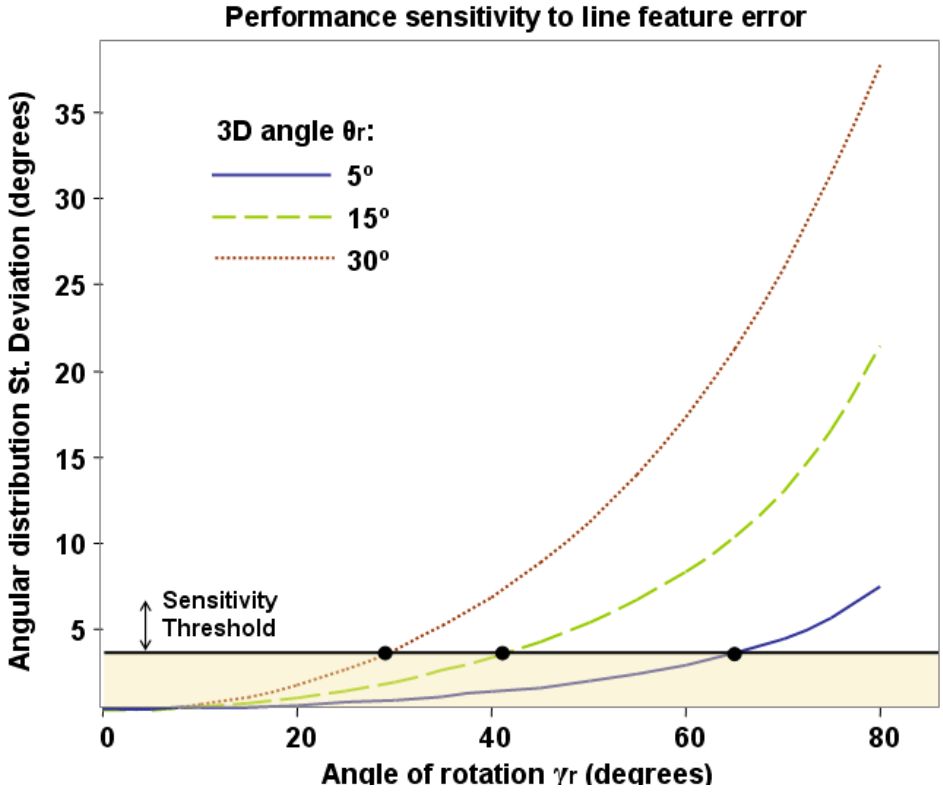

Figure 5.18. Angular distribution variation in a rotated pencil of lines. The model response is highly affected under a sensitivity threshold caused by line feature errors at minor rotations.

minor rotations. The standard deviation of the angular distribution of the pencil indicates a superior performance of the model with pencils of higher 3D angle separation. It implies that the model response is prone to errors below a sensitivity threshold. This is depicted in Figure 5.19, where real world data was used. Only one area presents a notable high error. It is located where minor rotations were applied. In contrast to the previous set of tests, this error is caused by the low angular distribution variation. The change at this area is not only negligible, which depends on the resolution of the line feature extraction method, it is also highly sensitive to line feature errors since the pencil angles are similarly distributed.

In general the use of pencil of lines improves the performance of the model. It does not require the angular shift from the axis of rotation and provides a robust response at high 3D angle separations. Nevertheless, as the estimation is based on the fitting of sample data, this 3D angle separation is limited due to the lack of measured samples. It is also suitable for real time applications, where a moving line feature forms the pencil of lines and its angular variation is modeled by its perspective distortion.

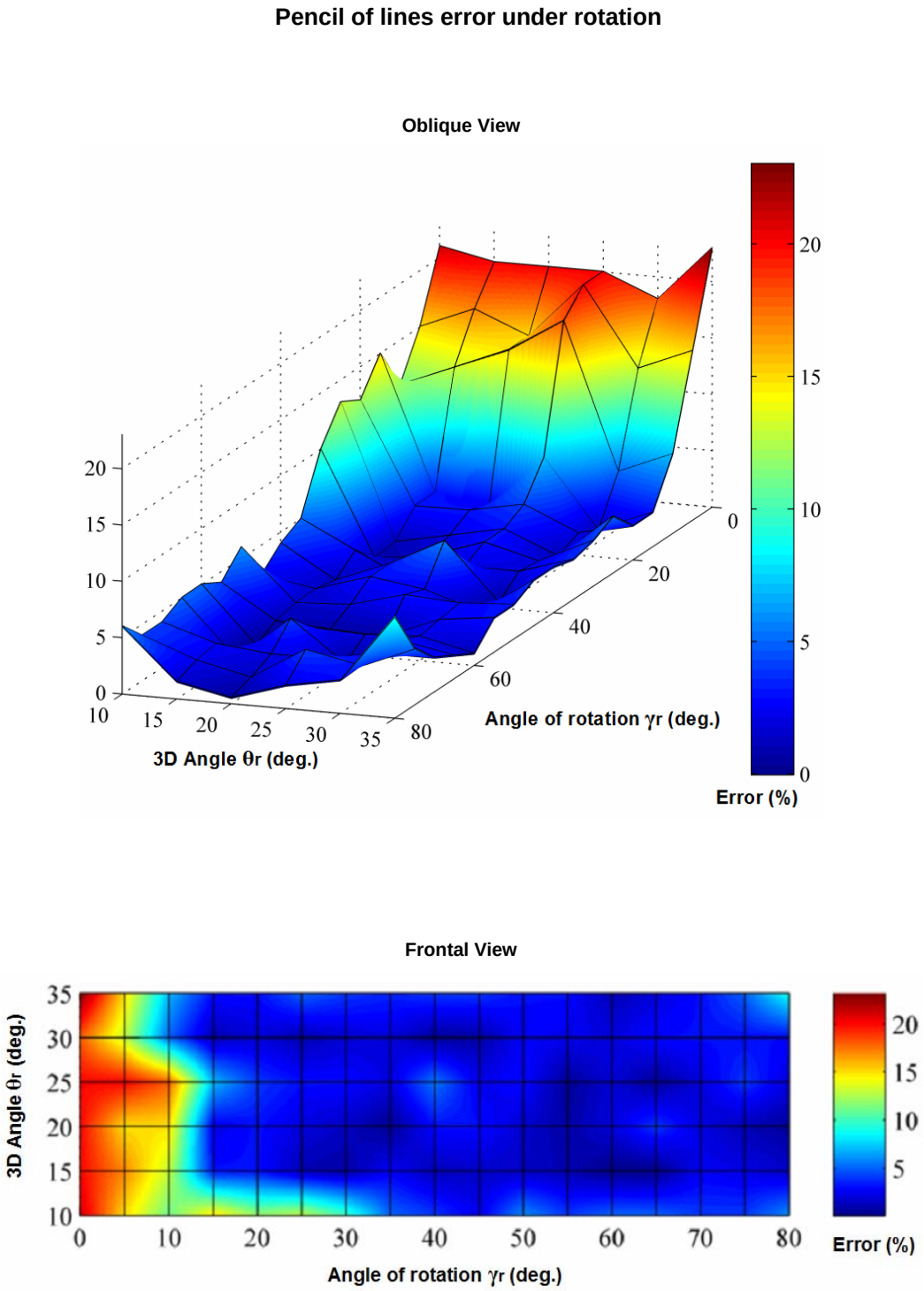

Figure 5.19. Performance error of the pencil of lines test. The poor performance of the model at long rotations is improved, while the 3D separation angle of the pencil is limited.

5.6. DISCUSSION

The capacity of the human visual system to perceive 3D space from a monocular view is described by a set of stimulus and processes. As its perception of depth on a flat surface is enhanced by the use of linear perspective, a set of geometric properties from the projective transformation provide enough 3D information to effectively calculate the pose of an object from 2D images. In this chapter a model of the variance of a geometric configuration under perspective transformation has been developed. Particularly, this geometric configuration is the angular variation caused by the perspective distortion. Since the resulting projected variation depends on the position of the observer, the exterior orientation of the camera can be determined by the model.

In contrast to the previous approaches developed in Chapter 4, 3D information is not required with the use of the perspective distortion analysis. **There is no necessity to measure angular samples from a determined 3D structure**. Thus, this new approach could be considered as an improvement or evolution of the aforementioned algorithms. All of them are based on the angular variation. However, this last approach owns the particularity of containing the complete required information in the image plane.

The projective nature of conics served as the base of this approach. Since angular variation is the geometric configuration analyzed, the rigid rotations applied to a 3D line provide the Euclidean transformations that describe a circle. The perspective projection of this circle is an ellipse in which its eccentricity depends on its orientation with respect to the center of projection. Applying the concept of a tangent plane to a sphere centered at the center of projection, it has been shown that the unit vector directed to the point of tangency is perpendicular to the vanishing point of the projected line. If the plane in which this tangent line is contained is not tangent to the sphere, this line is unique. Additionally, in the case its point of rotation is aligned to the center of the image plane and the center of projection, this 3D line is the axis of rotation and consequently the mayor axis of the projected ellipse.

The projective properties described by perspective transformations effectively supply a real physical interpretation of a rigid body motion. It has been modeled by the geometric distortion where the 2D angular variation is the response to 3D rotations. Experimental results show the efficiency and robustness of the method, in this case using a single angle and a set of concurrent lines. However, its accuracy is highly affected by the reliability of the line feature extraction technique applied.

5.6. REFERENCES

[1] A. Navarro, E. Villarraga and J. Aranda, "Relative pose estimation of surgical tools in assisted minimally invasive surgery," *Iberian Conf. Pattern Recognition and Image Analysis,* 2007.

[2] A. Navarro, A. Hernansanz, E. Villarraga, X. Giralt and J. Aranda, "Automatic positioning of surgical instruments in minimally invasive robotic surgery through vision-based motion analysis," *IEEE Proc. Int. Conf. Intelligent Robots and Systems,* 2007.

[3] D. Mumford, J. Fogarty and F. Kirwan, *Geometric Invariant Theory*, 3rd ed. Springer, 2002.

[4] J.L Mundy and A. Zisserman, *Geometric Invariance in Computer Vision*, The MIT Press, Cambridge, Mass. 1992.

[5] D. He and B. Benhabib, "Solving the orientation-duality problem for a circular feature in motion," *IEEE Trans. System, Man and Cybernetics*, vol. 28, no. 4, pp. 506-515, 1998.

[6] L. Li, Z. Feng and Q. Peng, "Detection and model analysis of circular feature for robot vision," *IEEE Proc. Third Int. Conf. Machine Learning and Cybernetics*, pp. 3943-3948, 2004.

[7] W. Gander, G.H. Golub and R. Strebel, "Least-squares fitting of circles and ellipses," *Journal BIT Numerical Mathematics,* vol. 34, no. 4, pp. 558-578, 1994.

[8] A. Sung, W. Rauh and M. Recknagel, "Ellipse fitting and parameter assessment of circular object targets for robot vision," *IEEE/RSJ Proc. Int. Conf. Intelligent Robots and Systems*, pp. 525-530, 1999.

[9] A. Fitzgibbon, M. Pilu and R.B. Fisher, "Direct least square fitting of ellipses," *IEEE Trans. Pattern Analysis and Machine Intelligence*, pp. 476-480, 1999.

[10] Y.C. Shiu and S. Ahmad, "3D location of circular and spherical features by monocular model-based vision," *IEEE Proc. Int. Conf. Man and Cybernetics*, pp. 576-581, 1989.

[11] J. Heikkilä, "A four-step camera calibration procedure with implicit image correction," *IEEE Proc. Int. Conf. Computer Vision and Pattern Recognition*, pp. 1106-1112, 1997.

[12] G. Zhang and Z. Wei, A position-distortion model of ellipse centre for perspective projection," *Journal of Measurement Science and Technology*, vol. 14, pp. 1420-1426, 2003.

[13] A.V. Akopyan and A.A. Zaslavsky, *Geometry of Conics*, American Mathematical Society, 2007.

[14] R. Hartley and A. Zisserman, *Multiple View Geometry*, 2nd ed. Cambridge University Press, 2004.

[15] O. Faugeras, *Three-Dimensional Computer Vision*, 4th ed. The MIT Press, Cambridge, Mass. 2001.

[16] B.C. Berndt and H.H. Chan, "Ramanujan and the modular j-invariant," *Canadian Mathematical Bulletin*, vol. 42, no. 4, pp. 427-440, 1999.

[17] Y. Ma, S. Soatto, J. Kosecka and S.S. Sastry, *An Invitation to 3-D vision*, Springer, 2003.

[18] J. Heikkilä, "Geometric camera calibration using circular control points," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1066-1076, 2000.

[19] R. Safaee-Rad, I. Tchoukanov, K.C. Smith and B. Benhabib, "Three-dimensional location estimation of circular features for machine vision," *IEEE Trans. Robotics and Automation*, vol. 8, pp. 624-640, 1992.

[20] L.A. Kenna, "Eccentricity in ellipses," *Mathematics Magazine*, vol. 32, No. 3, pp. 133-135, 1959.

[21] D. B. Lloyd, "Some old slants and a new twist to the cone," *Mathematics Magazine*, vol. 34, no. 5, pp. 293-296, 1961.

[22] A. Agarwal and B. Triggs, "Recovering 3D human pose from monocular images," *Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 1, pp. 44-58, 2006.

[23] Oakdale Engineering, *DataFit Curve Fitting Software Manual*, 2007.

[24] W.H. Press, S.A. Teukolsky, W.T. Vetterling and B.P. Flannery, *Numerical Recipes in C++,* 2nd ed. Cambridge University Press, 2002.

[25] N.R. Draper and H. Smith, *Applied Regression Analysis,* Wiley Series in Probability and Statistics, 1998.

[26] S. Kotsiantis and P. Pintelas, "Selective averaging of regression models," *Annals of Mathematics, Computing and Teleinformatics*, vol. 1, no. 3, pp. 66-75, 2005.

Chapter 6

Perception enhancement in visually guided applications

6.1. INTRODUCTION

The spatial information calculated from the proposed methods presented in the previous chapters is expressed by the estimation of the position and orientation of a determined object with respect to a camera. This is the recovery of the 3D physical attributes of the object that were lost in its 2D projection. To accomplish this task, line features are extracted from the image plane. The angular variation of these features presents properties that provide spatial information. Thus, as it is used by the human visual system, certain 3D attributes of the object can be calculated by a computer vision system as well. As this angular variation is used as a monocular cue for spatial perception, it would be reasonable to think that interactive applications, where sensory input is limited to 2D visual information, could be benefited by an enhancement of perceptual cues through this calculated data. In this chapter, the effect of this perceptual enhancement due to the introduction of spatial information is evaluated. It is carried out by the study of task performance on a teleoperation application; specifically a Minimally Invasive Surgery (MIS) based application, where the normal function of the surgeon's perceptual-motor system is highly affected.

The perceptual-motor coordination is the ability to link the organism's perception with its self-initiated movement in the performance of visually guided tasks [1]. These tasks can be simple and normal activities executed in daily life. Effortlessly tasks such as grasping a pencil or bounce a ball, involve an intentional motor movement in response to optical stimulation. This associates the visual perception and hand-eye coordination,

and requires the ability to translate this visual perception into motor functioning; which involves motor control, motor accuracy, motor coordination and psychomotor speed [2]. Therefore, visually guided applications such as teleoperations are notably affected since there is a physical separation between the working space and the operator. In Minimally Invasive Surgery (MIS) a 2D window on a 3D world is imposed. The surgeon is limited to work physically separated from the operation site and must rely heavily on indirect perception [3]. To effectively link action to perception it is necessary to integrate body movements with other senses such as vision. Thus, in this chapter is suggested that locating the instruments with respect to the surgeon through computer vision techniques serves to enhance the cognitive mediation between action and perception and can be considered as an important assistance in this type of surgery.

It is the main purpose of this chapter the study of the effect derived by the introduction and integration of visual cues in action-perception applications. The suggested hypothesis states that the spatial perception enhancement over the 2D visual input provided by a camera view, improves the visual-motor link through the inclusion of orientation depth cues. This provides a sense of presence due to the understanding of the operative site, and relating this with the external environment, the orientation disengagement is reduced by an improved sense of place. To demonstrate this, the 3D attributes of an object line are estimated using the perspective distortion model, which supplies the orientation information included in the application. In this case, the application is based in MIS, due to its compliancy with the rotational motion analysis, which in fact was devised as assistance motivated in this type of operations.

6.2. MEDIATING ACTION AND PERCEPTION IN MIS

The introduction of minimally invasive surgery (MIS) as a common procedure in daily surgery practice is due to a number of advantages over some open surgery interventions. In MIS the patient body is accessed by inserting special instruments through small incisions. As a result tissue trauma is reduced and patients are able to recover faster. However, the nature of this technique limits the surgeon to work physically separated from the operation site. This implies a significant reduction of manipulation capabilities and a loss of direct perception. Therefore, robotic and computer-assisted systems have been developed as a solution to these restrictions to help the surgeon.

There are some solutions currently proposed to overcome the limitations concerning the constrained workspace and the reduced manipulability restrictions. Approaches dedicated to assist the surgeon are basically aimed to provide an environment similar to conventional procedures. In this sense, robotic surgery developments are especially focused on the enhancement of dexterity, designing special hand-like tools or adding force-feedback by direct telerobotic systems [4] [5]. Other systems aid the surgeon through auxiliary robotic assistants, as is the case of a laparoscopic camera handler [6]

[7]. Nevertheless, though the limitation of the visual sense has been tackled by robotic vision systems capable of guiding the laparoscopic camera to a desired view [8] [9], the 3D perception and hand-eye coordination reduction in terms of cognitive mediation have not been extensively developed.

6.2.1. Computer assistance in MIS

The visual sense in the MIS environment is limited. It imposes a 2D window of the operative site. Therefore, approaches focused to assist the surgeon are fundamentally based on image content recognition and presentation. As an example of this computer assistance, there are a number of approaches focused on the experimental verification of surgical tool tracking to be presented in the center of the image [10], the study of the distribution of markers to accurately track the instruments [11], the establishment of models for the lens distortion [12]. These examples are emergent techniques to assist the surgeon by the enhancement of the image content. The work in which this chapter is focused, however, is based on the integration of visual and motion information to perceptually locate the instruments with respect to the surgeon.

Healey in [3] describes the mediation between action and perception in the MIS environment and states the necessity to effectively link action to perception in egocentric coordinates. In this approach it is suggested that the integration of egocentric information, as visual and limb movements can be provided by the capacity to locate surgery instruments at a desired position in the operative site and the knowledge of their orientation with respect to the laparoscopic camera. As a result, the surgeon perception is enhanced by a sense of presence. Thus, computer vision issues such as the 2D-3D pose estimation and exterior orientation, deal with this problem and can be applied to aid the surgeon in this kind of procedures.

It can be seen in Figure 6.1 the schematic of an application where exterior orientation is used and presented through enhanced visual information to assist the surgeon. This presentation is commonly performed by augmented reality. There have been early approaches in which this type of resource is used in different kinds of applications [13], others more specialized in surgery, recognize objects seen by the endoscope in cardiac MIS [14], or design a system for surgical guidance [15], being a visual enhancement which has served as a human-machine interface. In this chapter, the position and orientation of surgery instruments is the information to be imposed in the visual information. It serves to integrate egocentric information, as vision and limb movements, to provide a sense of presence and relate it with the external environment to give a sense of place. Nevertheless, the camera-tool calibration must be calculated. This problem can be tackled by computer vision techniques, as the perspective distortion model presented in the previous chapter. Thus, the computer assisted system can be expressed as a process as shown in Figure 6.2.

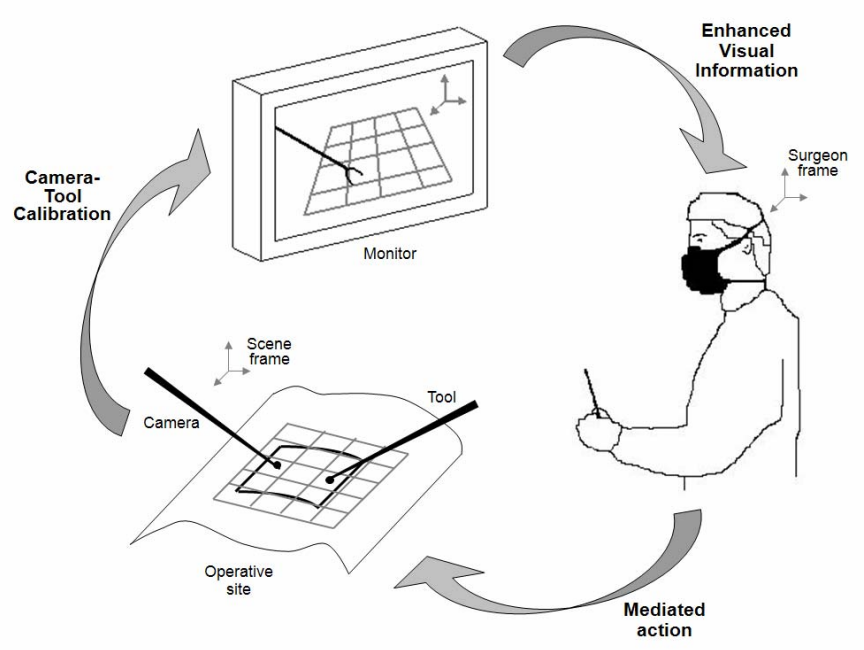

Figure 6.1. Schematic of an application of the exterior orientation as a tool to assist the surgeon in MIS. It permits to control action through the perception enhancement.

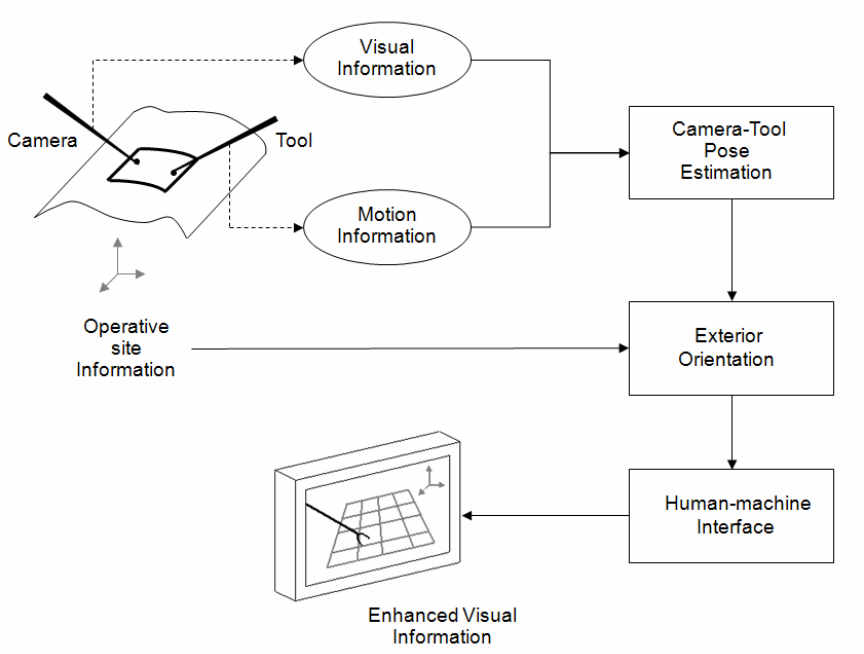

Figure 6.2. Visual enhancement process to assist the surgeon. Motion and visual information is related to calibrate the surgical instrument with respect to the camera.

6.2.2. Robotic assistance in MIS

Robotic assistants, as specialized handlers of surgical instruments, have been developed to facilitate the surgeon performance in MIS. Since the patient body is accessed by inserting surgical instruments through small incisions, passive wrists have been incorporated for free compliance with the port of entry. With this wrist configuration it is only possible to locate accurately an instrument tip if its port of entry or fulcrum point is known. This is a 3D point external to the robotic system and though it has a significant influence on the passive wrist robot kinematics, its absolute position is uncertain.

A number of approaches are focused on the control and manipulability of surgical instruments in MIS through robotic assistants. As can be seen in Figure 6.3, 3D transformations are applied to produce pivoting motions through the fulcrum point. The more accurately has been estimated this port of entry, the more accurately the instrument tip is positioned at a desired location. Some of the approaches evade this difficulty by the incorporation of special mechanisms with actuated wrists [16]. Others based on passive two joint wrists tackle the problem by control strategies. Thus, error minimization methods are applied to determine the outside penetration [17] [18], and compensate the fulcrum location imprecision [7].

A reasonable alternative approach to tackle the passive robot wrist problem is by adding computer vision methods. Since the laparoscopic camera captures the workspace sight, specialized vision algorithms are capable of estimating 3D geometrical features of the scene. The 2D-3D pose estimation problem serves to map these geometric relations estimating the transformation between coordinate frames of the instruments and the camera. There are several methods proposed to estimate the pose of a rigid object. The first step of their algorithms, as described in Chapter 3, consists on the identification and location of some kind of features that represent an object in the image plane.

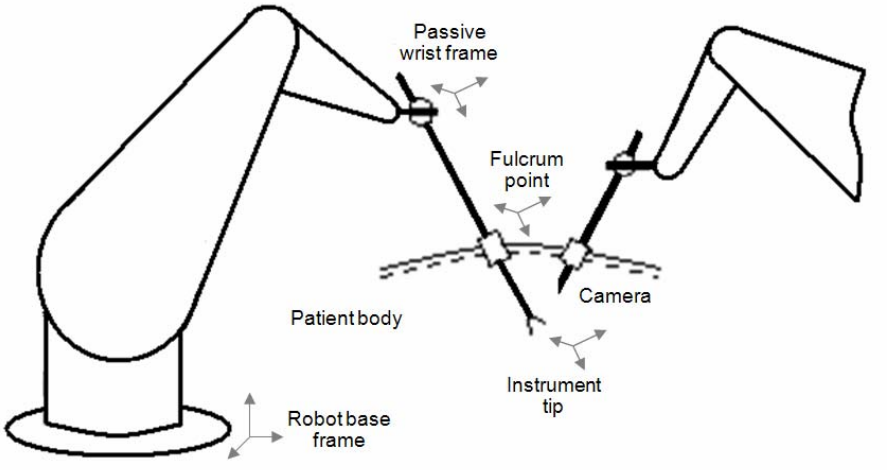

Figure 6.3. Minimally invasive robotic surgery systems with passive wrist robotic assistants. The location of the instrument tip depends on the knowledge of the fulcrum point.

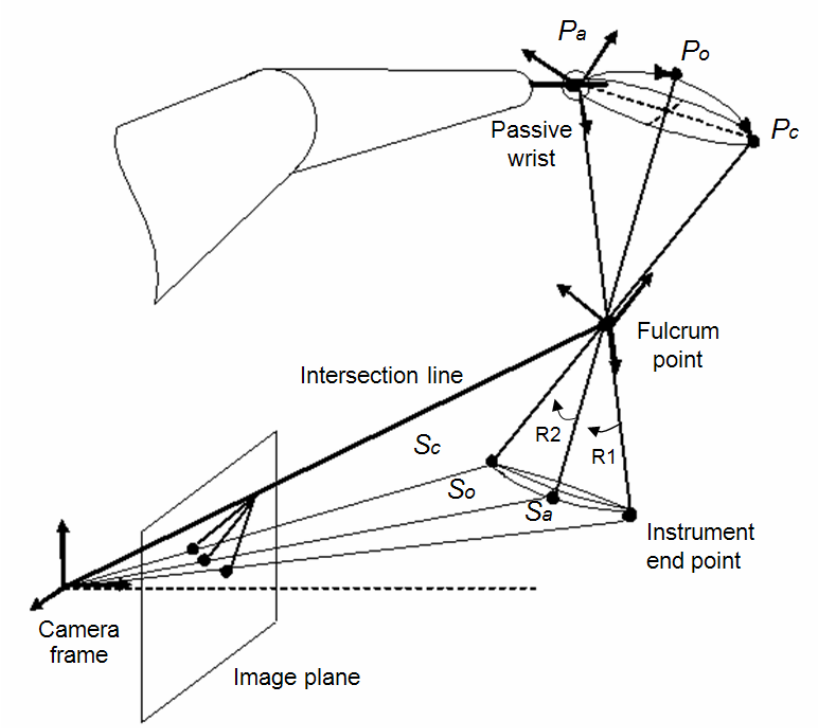

Figure 6.4. 3D transformations of the instrument manipulated by a robotic assistant produce a pivoting motion through the fulcrum point. It results in 3D rotations at the workspace captured by the camera through projection planes.

In the case of image sequences motion and structure parameters of the 3D scene can be determined. Correspondences between selected features in successive images must be established. This provides motion information and can be used, as is in this case, to estimate the pose of a moving object. Figure 6.4 shows the pivoting motion of the surgery instrument through the fulcrum point. As can be seen, instruments are represented by feature lines and are visualized by the laparoscopic camera. In this example, three views after two different rotations generate three lines in the image plane. Each of them defines a 3D plane called the projection plane of the line. These planes pass through the projection center and their respective lines. Their intersection is a 3D line from the origin of the perspective camera frame to the fulcrum point.

6.3. APPLICATION AND COGNITIVE EFFECTS ASSESSMENT

The recovery and presentation of at least the basic sensorial cues to perceive elemental attributes of a workspace is the main purpose of an environment mediated by technology. Commonly, it is constrained by imposing a 2D window of the operative site. Thus, approaches focused to enhance the sensorial cues and assist the operator are fundamentally based on image content recognition and presentation. This presentation is usually performed by augmented reality, where computer generated virtual objects are accurately registered with real attributes of the scene [19]. Applications of this

technology in mediated environments range from fields as telerobotics by improving the remote view by wireframe drawings of structures [13], entertainment by creating virtual studio environments [20], or training simulators [21], to medical 3D model visualization [22] [23], or image guided surgery [14] [15], which is a field highly benefited from this aid and is viewed as one of the most important to develop.

The imposition of virtual arrangements on a bidimensional view of a real scene has demonstrated to exert a strong influence over the perceptual conception of space [24]. In mediated environments, the concept of presence has become a representative property from which a sense of immersion is obtained. It is described as the experience of being there, which becomes more convincing as media content is more perceptually realistic and interactive [25]. The level of presence is augmented, as changes in the environment are perceived in response of one's own movements. Healey in [3] describes the mediation between action and perception in teleoperation. There, it is stated that it is necessary to effectively link action to perception in egocentric coordinates to overcome the indirect cognitive mediation. Therefore, the orientation estimation of the instruments with respect to the camera in an action-perception application may be considered capable of providing such information [26].

In this work, a computational approach to direct perception of space is proposed as useful information to augment the sense of presence in mediated environments. It calculates 3D information from the captured images. Therefore, more spatial cues may be introduced in order to enhance the optical stimulus of the operator. In this case, a new computer-generated sight of the scene represents the additional information provided. This is possible due to the previous calculation of orientation data, which permits to select and visualize any angled spot of the workspace to perform action-perception tasks. Thus, the cognitive effect resultant from the implementation of this aid serves to improve the visual-motor coordination in mediated environments.

An experimental study carried out simulating a real action-perception application with previous knowledge of the tool's exterior orientation is described in detail in the next subsection and demonstrates the utility of this perceptual information.

6.3.1. Material and Methods

Participants. Twenty individuals with ages between 20-40 years of both genders and without any experience related to teleoperation, remote control or robotics participated in the study. All of them presented normal or corrected-to-normal vision.

Apparatus. The experimental setup illustrated in Figure 6.5 was composed by two workstations, a workspace and an operator's control center, divided by a mobile occluding screen. The workspace represented the remote environment where the action-perception task took place. Therefore, it was equipped with a robot arm to perform

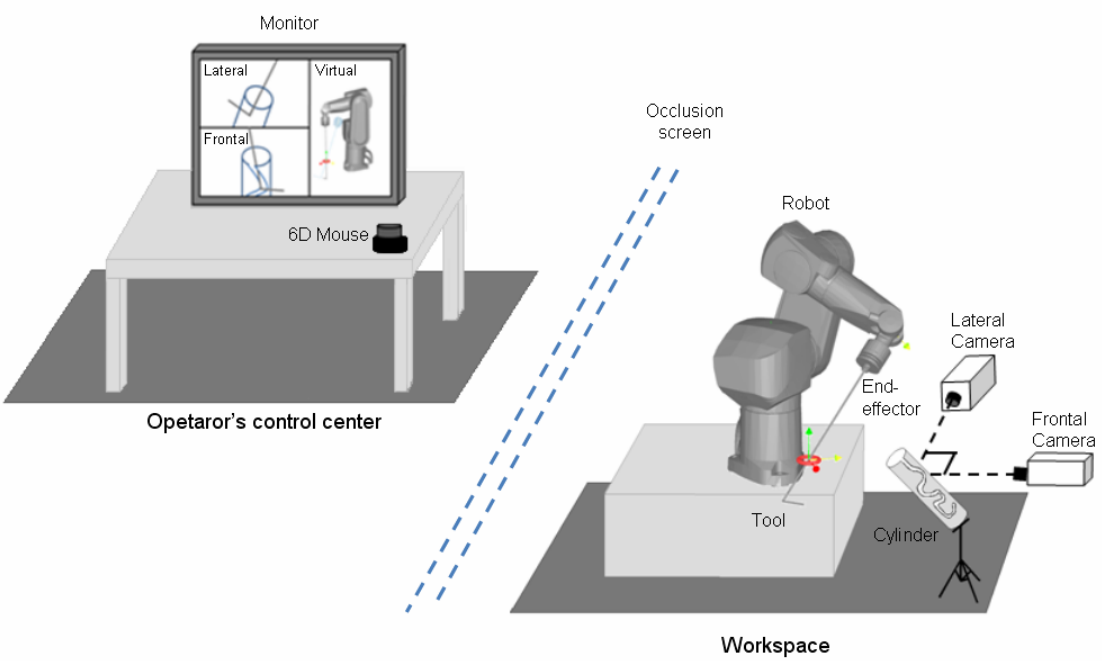

Figure 6.5. The experimental setup conformed by an operator's control center and a workspace simulated a teleoperation system. It created a mediated environment and demanded efficient visual-motor coordination.

actions with a long line-form object as the tool attached to its end-effector ending with an angled tip. The operative site was complemented with a cylinder located in a determined position with respect to the robot arm. This cylinder was shaped with a carved curved route and represented the objective to be traced by the tool (Figure 6.6). In addition, two analog RGB cameras equipped with known focal length optics were located orthogonally directed to the operative site. One was directed to the frontal face of the route of the cylinder and the other to the lateral face. The captured images were sampled at 768x576 pixel resolution and represented the visual feedback of the workspace.

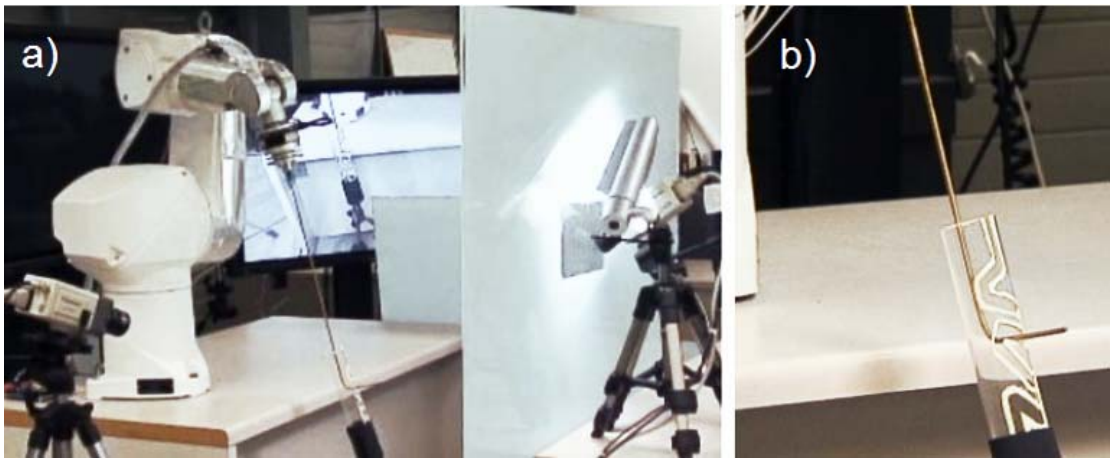

Figure 6.6. The workspace of the experimental setup simulated the remote environment which was composed by: a) A robot arm with a line-form object attached to its end-effector and two orthogonally located cameras; b) A cylinder with a carved curved route, which represented the operative site.

The operator's control center, representing the local environment, was the location where orders were transmitted to the robot by a 6D mouse and the visual feedback of the two cameras was received. Captured images were presented in a wide monitor in conjunction with a virtual enhanced view of the robot. This virtual representation of the robot and its tool was an accurate triangle-based model developed to simulate robotic proximity queries and collision detection [27], and served to depict the robot actions from any angled spot (Figure 6.7). Thus, a desired view could be selected and changed by the operator in real-time. Furthermore, it was possible to save a selected position of the robot during the performance and visualize it as a fixed semi-transparent tool, which guided the operator by providing a spatial reference.

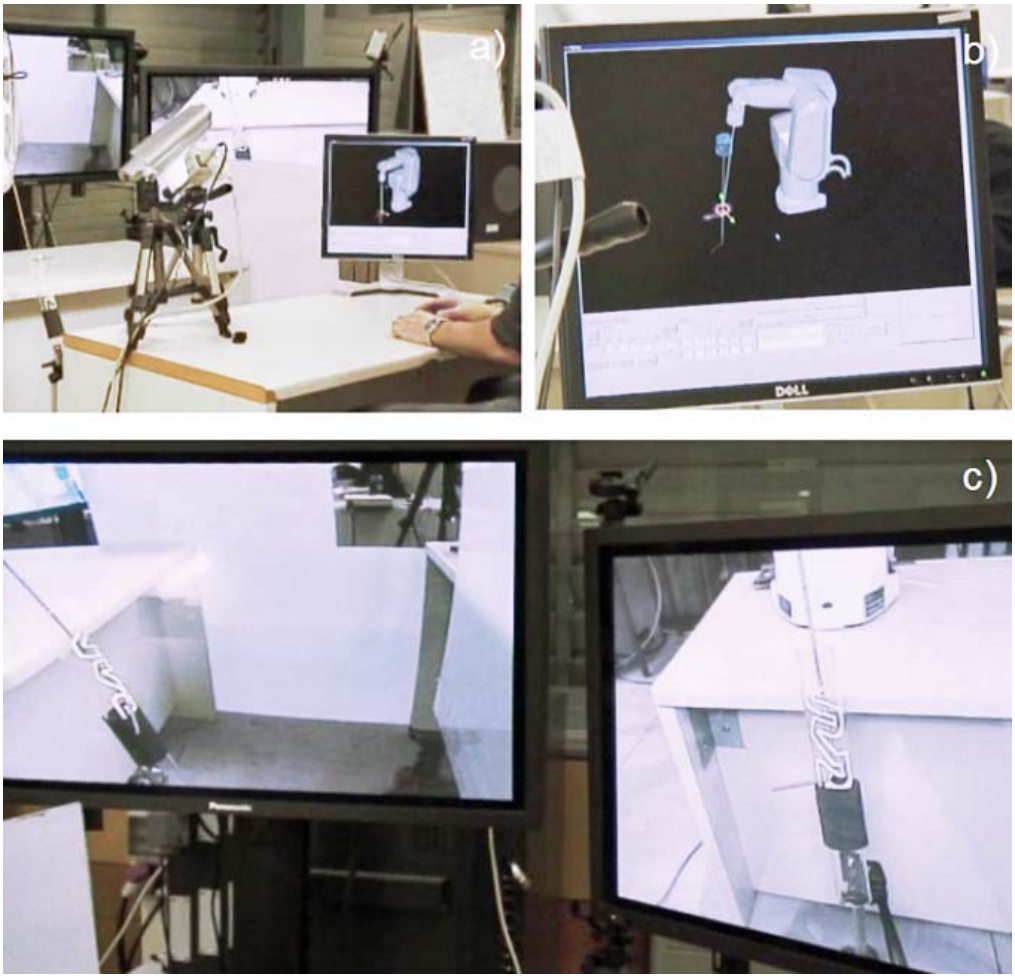

Figure 6.7. The operator's control center was equipped to manipulate the robot by a 6D mouse, obtaining visual information from a set of monitors (a). A virtual representation of the robot and its tool served as the additional aid (b), which in conjunction with the two views captured by the orthogonal cameras (c), conformed the indirect visual feedback to the operator.

The robot kinematics was restricted to move through a pivoting point. It was as though the tool was introduced into an incision. Thus, the control of motions was carried out by moving the base of the tool, which was the end-effector of the robot arm. Collisions of the tool were detected by a torque/force sensor located at its base, and the trajectory performed by the robot, as well as the selected views of the virtual scene, could be registered by the system.

Procedure. The task to be performed consisted in the alignment of the tool with respect to the cylinder. Subsequently, it should be introduced maintaining the angled tip through the curved route carved in the cylinder's wall until the end point was encountered, and afterwards, perform the route back to get out of the cylinder. There were three conditions to perform the task. The first presented the workspace directly. Thus, the operator could see the complete scene without the occlusion screen (first test). In the second, the visual feedback was provided by the monitor and depended completely on the two orthogonal views supplied by the cameras (second test). And the third presented the view of only the frontal camera and the aid provided by the virtual view of the robot (third test). These three tests were carried out by each of the participants, and each test was performed four times.

A training process was held prior to the tests. Since the features of the system were not intuitive due to robot kinematics restrictions and 6D mouse operation, it was an extensive process. Participants had thirty minutes of training to handle the system, having the opportunity to complete the task using visual feedback and were instructed about the capabilities of the aid provided by the virtual view. Only one trial of each of the three tests was performed per day, and each day participants had five minutes for training. Therefore, there were four sessions of three tests per day. According to the participant session number the order of the test was different.

The completion time to correctly fulfill each task was measured in order to evaluate the performance of each test. It was complemented by a quality indicator, which was measured by the time without collision during the performance, and by the traced trajectory.

Statistical analysis. Completion time variables are expressed as mean±standard error of the mean. Differences between group means were assessed by ANOVA for multiple comparisons. A value of $P < 0.05$ was considered significant.

6.3.2. Results

The completion time differed between tests as shown in Figure 6.8a. Each of the tests presented a tendency to decrease through trials. However, an optimum performance was observed in the direct view test, as well as a better performance in the enhanced view test compared to the two cameras test. Significant differences between these tests were evaluated by the statistical analysis, which employed the data of the twenty participants. Each data point represented the average of the completion time for its corresponding trial of each test. There was a significant difference between the direct view test and the other two in all the trials ($P < 0.005$). The effect of the sensorial restrictions provided to the operator was notable. The other two tests presented a difference between them, nevertheless, statistically significant differences were only observed on the first two trials ($P = 0.025$ and $P = 0.038$, respectively). The third trial was especially where the difference was slighter and therefore, was statistically not significant ($P = 0.58$). And in the fourth trial, which appeared to be a clear difference, it was observed that it was not significant either ($P = 0.054$). It could be attributed to the order of the test performance, lack of training before the trial sessions or the knowledge acquired through trials.

A learning curve was depicted by the tendency to decrease in the completion time of all the tests. It was a performance improvement that in the direct view test changed from 163.55 ± 13.89 sec. on trial 1 to 85.21 ± 5.29 sec. on trial 4. This test presented a relative constant standard deviation from the trial 2 to trial 4. In the second test, although the performance level was not as high, there was an improvement from 303.11 ± 26.8 sec. on trial 1 to 145.58 ± 12.3 sec. on trial 4. This was a notable improvement having limited perceptual information. However, the performance level became higher with the introduction of the enhanced view. In this test, performance improved from 227.72 ± 16.21 sec. in trial 1 to 117.15 ± 7.3 sec. in trial 4. A similar order in performance level could be observed by the quality evolution shown in Figure 6.8b. While the percentage of time without collision relatively remained constant trough trials in the first test, a tendency to increase was presented in the two others. The second test, with the lowest quality level, improved from 49.1% in trial 1 to 58.04% in trial 4. And the third test improved from 59.27% in trial 1 to 72.49% in trial 4, which is a measure that approached the mean quality obtained in the first test (73.87%). Figure 6.9 shows this mean quality values, which compared to the other two tests (51.95% in the second test and 63.91% in the third test), indicated that the performance level in trials with indirect perception was higher using the enhanced view test.

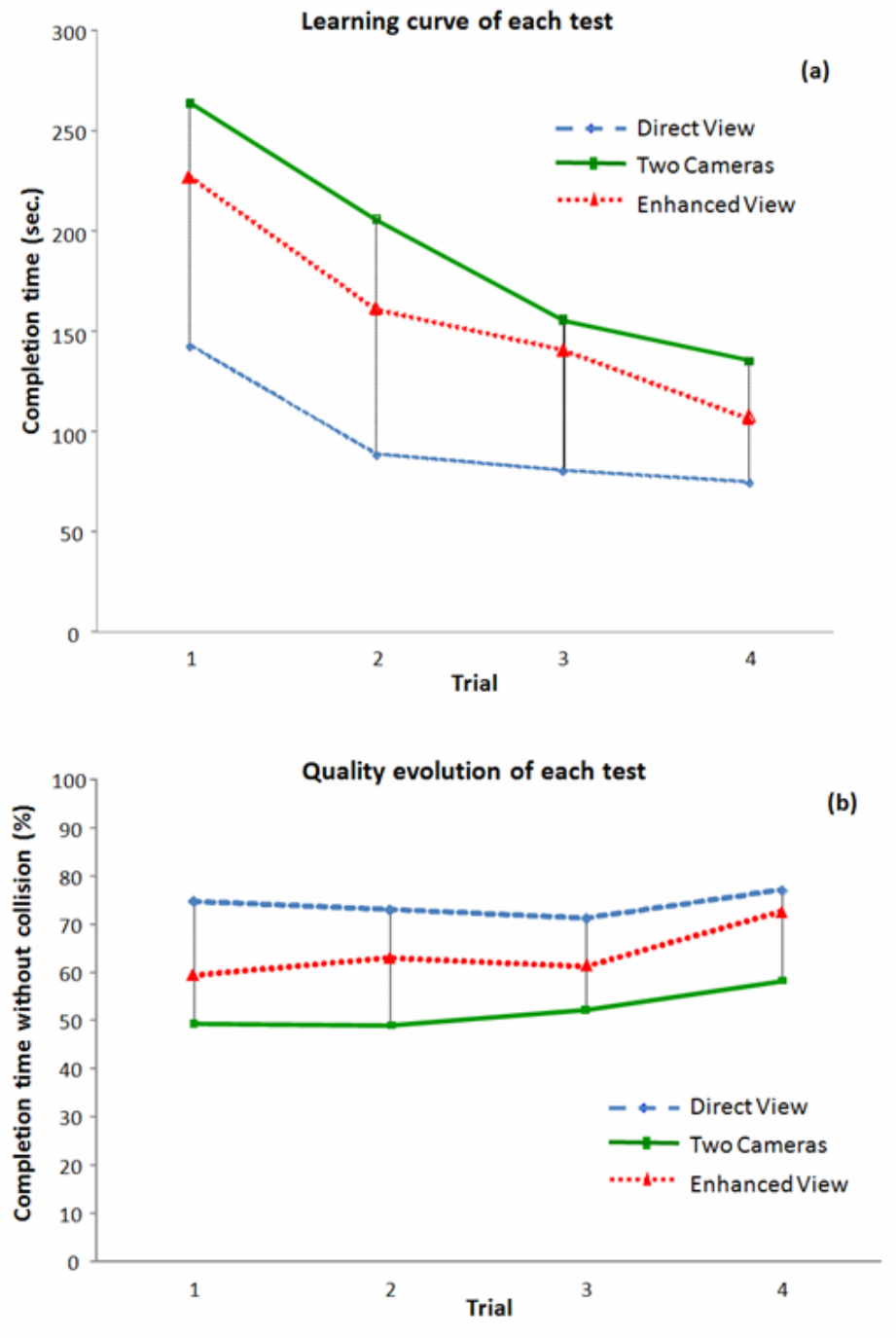

Figure 6.8. Performance level evolution of each test according to completion times and quality: a) A learning curve is expressed by the mean values obtained from the completion times; b) Quality evolution is expressed by the percentage of completion time without collision. Even though skills were improved in the three tests, a difference of performance level was observed.

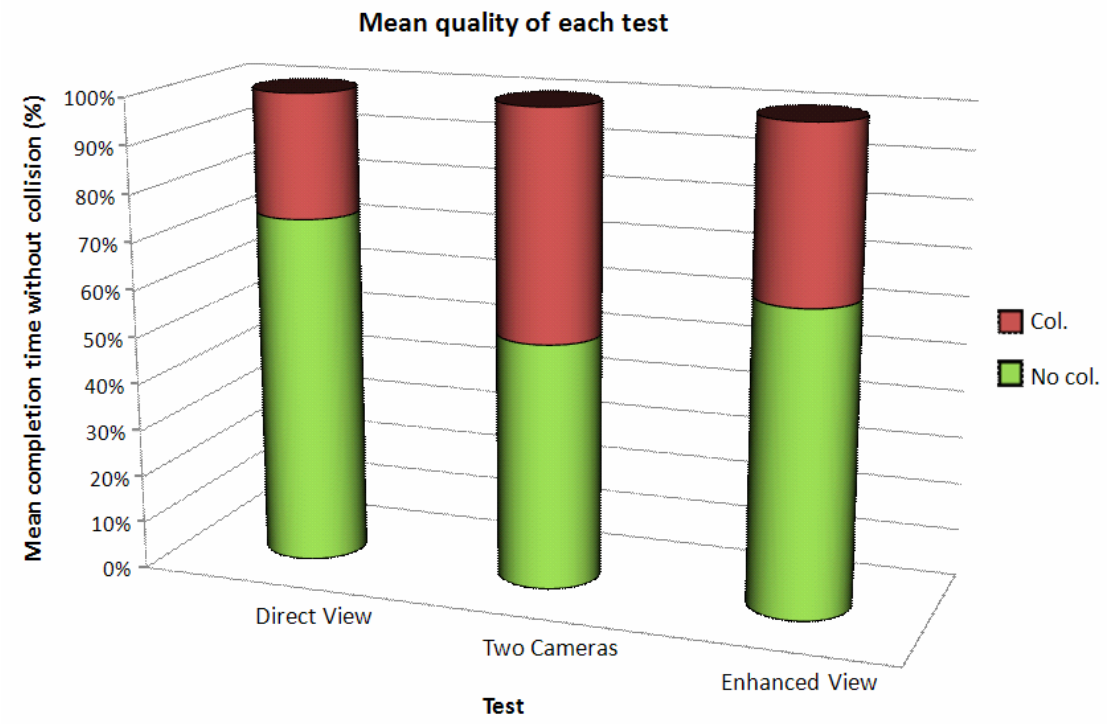

Figure 6.9. Test performance difference by the mean quality. The direct view test presented the highest percentage with a 73.87% of the completion time without collision, while the enhanced view test a 63.91% and the two cameras test, the lowest, with a 51.95%.

Complementary information useful to differentiate performance levels of the tests was given by the registered trajectories. A determined trajectory described the traced route for each trial and was the measure of the Cartesian position of the end-effector of the robot during the completion of the task. Figure 6.10 illustrates the trajectory description of the test through two examples, one was executed with direct view and the other was occluded. As the initial position of the tool and the location of the cylinder were different and fixed in all the trials, the first process to carry out was their alignment. It consisted in the positioning of the tool parallel to the cylinder and the introduction of the angled tip of the tool through the narrow incision. Even though this trajectory information was subjective, different behaviors were distinguished. In the occluded view example, although the alignment process was not deviated, there were various fail attempts of insertion and several additional motions along the cylinder's route.

A comparison of the three tests by the performance evolution through trials could be observed by the registered trajectories. Figure 6.11 shows the trajectories of the three tests separated by trial. An improvement in trajectory performance was identified. The trajectory in the two indirect perception tests became similar to the performed by the

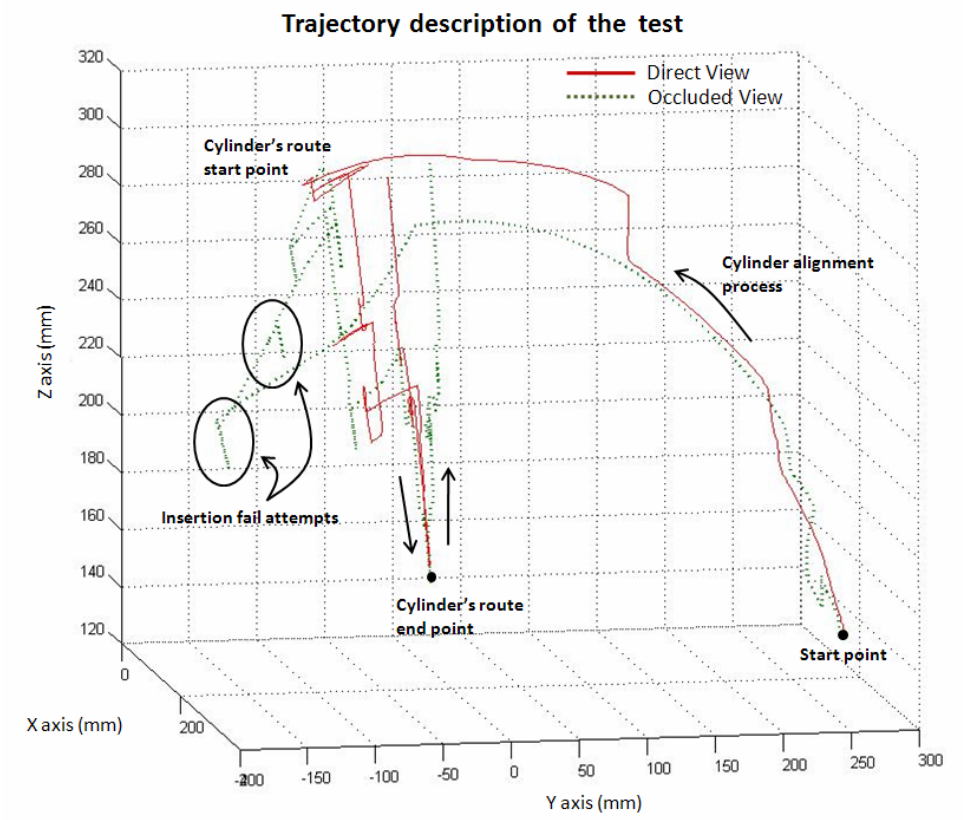

Figure 6.10. Example of two trajectories executed with two different conditions. A test carried out with an occluding view may present more difficulties than one executed with a direct view. The objective of the test can be described by these trajectories and the visual behavior of the operator can be analyzed.

direct view. Nevertheless, some notable deviations were registered. It was observed that once a participant was disoriented, it was difficult to encounter the correct position. It was a situation that was solved easier by the enhanced view. This was a fact confirmed by the trajectory performance evolution of each test between trials shown in Figure 6.12. Trajectories became relatively uniform as the learning advanced, resulting in a consistent trace for the direct view test, compared to the slightly deviated of the enhanced view test and the randomly disperse of the two cameras view.

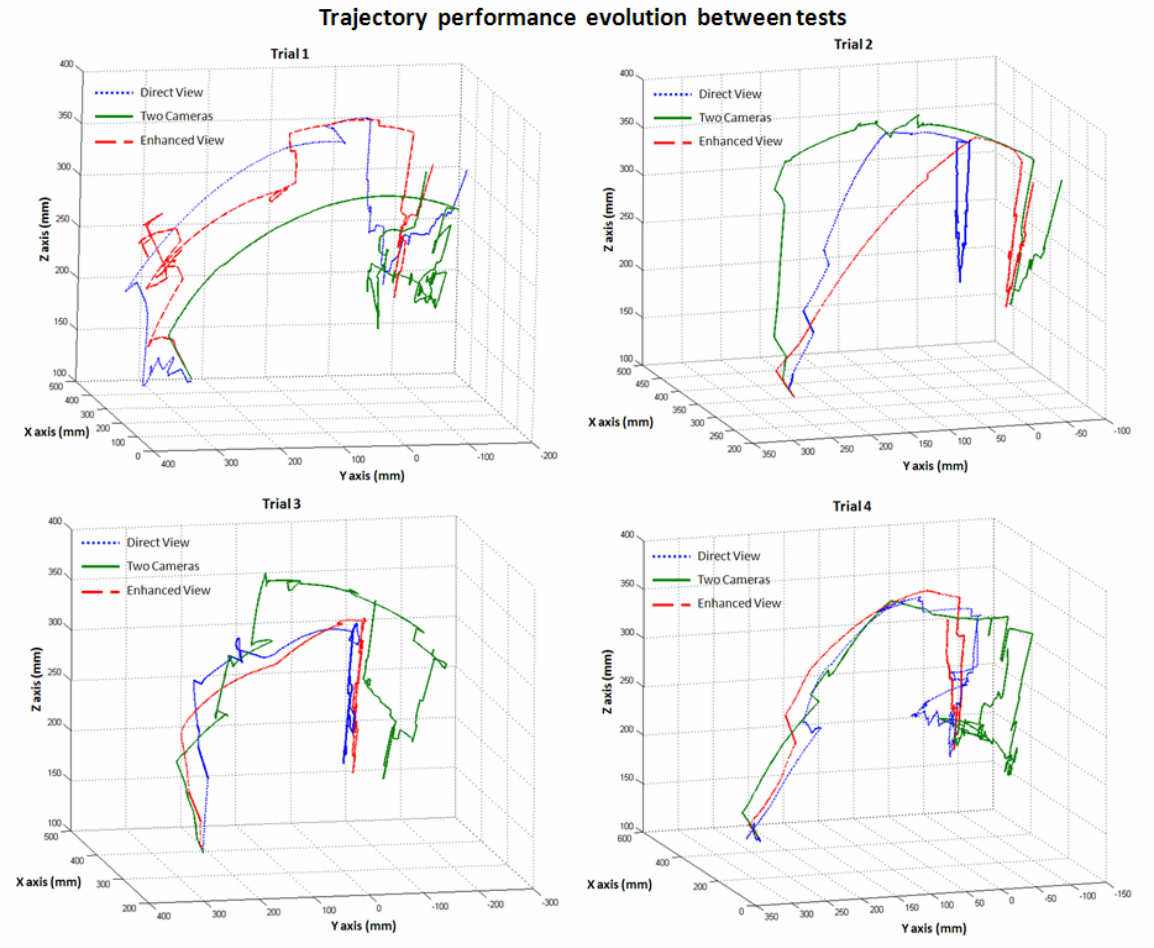

Figure 6.11. Test performance difference by the traced trajectory. Skills improved as the registered traces became more uniform through trials. However, as is shown in the trial 3 of this example, occasional disoriented behavior was observed with the two cameras test.

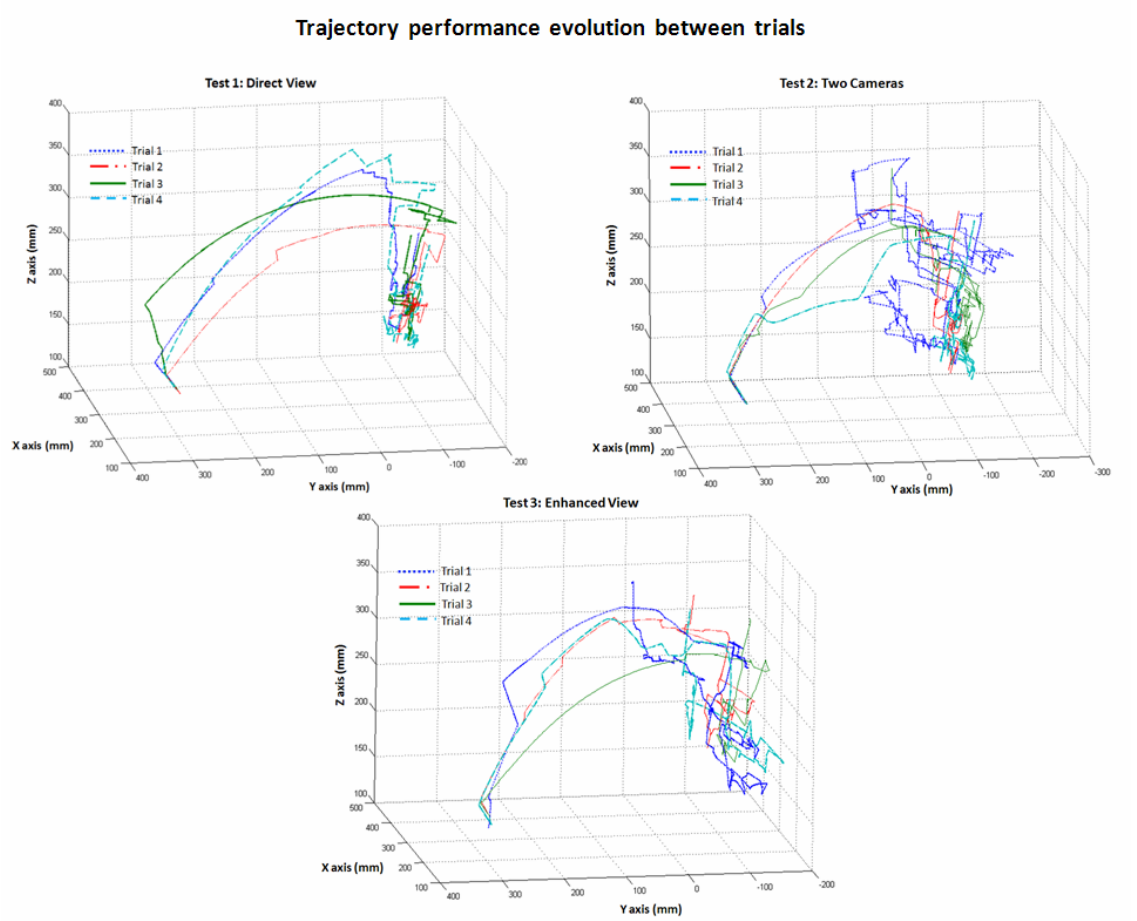

Figure 6.12. Test performance evolution by the traced trajectory. As experience was gained through trials, performance level improved. Nevertheless, the evolution was different. In this example trajectories were similar with the direct view test, while in the enhanced view they evolved from a disperse trace to a more direct, and in the two cameras test their trace was randomly dispersed through trials.

6.3.3 Discussion

The experimental setup developed to evaluate participant's performance was based on the measurement of the completion time of the task, quality of the execution and the traced trajectory. An extensive training process was carried out prior to the first trial due to the complexity of the system. The robot control was not intuitive since it was handled by a 6D mouse and its kinematics was restricted, moving the angled tip of the tool in opposition to its base. Even though participants adapted satisfactorily to the system, a shorter training session was held prior to every trial session due to the time separation between them. The action-perception task was complex and required efficient visual-motor coordination. Therefore, sessions of three trials per day, one of each test, were carried out in order to avoid fatigue. However, it was observed that on a simple session, participants learned and their skills improved by each repetition of the task and, although tests were performed in an altered order, it exerted an influence in the performance level.

The learning curve shows that participants adapted to the sensorial cues provided by the system. Completion times decreased through trials for the three tests. The direct view test evolution presented a notable difference over the two others, as was expected. Nevertheless, though a superior performance was observed in the enhanced view test with respect to the two cameras test, the difference was not as notable. This was confirmed by the statistical analysis, from which significant differences were obtained only in the first two trials. It could be attributed to the test order, in which a poor training process prior to a first test derived in a notable inferior performance compared to the third. Therefore, the memory of spatial relations between objects and actions is important [28]. A cognitive map of the route was constructed and associated actions and perception of objects in the scene.

The sensorial cues provided by the two cameras, though limited, offered a slight conception of 3D space since they were located orthogonally to the cylinder. The cognitive effect given by this perceptual information was observed in the alignment process, which presented an acceptable performance. However, this performance level decreased strongly inside the route. This effect was supported by the percentage of completion time without collision measured in all the trials. The quality level for trials carried out with the two cameras view was significantly inferior. Furthermore, it was observed that independently of the experience and learning level of the participant in this test, their capacity to solve disorientation problems due to incorrect motions was considerable lower. This could be observed through the trajectory information registered by the system, which was a useful tool that permitted visualize different traces employed to fulfill the task. These trajectories became relatively uniform as participants gained experience.

An enhancement of the perception of space was obtained through the addition of the aid. The inclusion of the virtual view exerted a strong influence in the visual guided behavior of participants. This could be observed from trajectory and quality measures. For completion times, though presented slight differences in some trials with the two cameras tests, a superior performance was also observed. Thereby, the cognitive effect provided by the aid could be considered relevant, counting on the fact that in the real case only one camera would capture the view of the workspace, in contrast to the two orthogonal. The possibility to change and select the view of the robot from the virtual representation augmented the sense of presence. Although participants presented different strategies to complete the task, the majority selected an exterior view of the robot. Thus, a conception of space and place was achieved by perceiving where they were in the scene. This agrees with Gibson's approach [29] and was achieved with only the representation of the robot and its tool. The cylinder or any other object of the scene was not represented. Therefore, the perception of one's orientation with respect to a self created reference of the world and the effect of one's actions in the environment contributed to augment the sense of presence.

6.4. REFERENCES

[1] H. R. Schiffman, *Sensation and Perception*, 5th ed. John Wiley & Sons, Inc. 2001.

[2] R. Sanghavi and R. Kelkar, "Visual-motor integration and learning disabled children," *The Indian Journal of Occupational Therapy*, vol. 37, no. 2, pp. 33-38, 2005.

[3] A. Healey, "Speculation on the neuropsychology of teleoperation: implications for presence research and minimally invasive surgery," *Presence: Teleoperators and Virtual Environment*s, MIT Press, vol. 17, no. 2, pp. 199-211, 2008.

[4] C. A. Grimberger and J. E. Jaspers, "Robotics in minimally invasive surgery," *Proc. IEEE Int. Conf. on Systems, Man and Cybernetics,* 2004.

[5] H. Mayer, I. Nagy, A. Knoll, E. U. Schirmbeck and R. Bauernschmitt, "The Endo[PA]R system for minimally invasive robotic surgery," *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems,* 2004.

[6] R. Hurteau, S. DeSantis, E. Begin and M. Gagner, "Laparoscopic surgery assisted by a robotic cameraman: concept and experimental results," *Proc. IEEE Int. Conf. on Robotics and Automation,* 1994.

[7] V. F. Muñoz, I. Garcia-Morales, J.M. Gomez-DeGabriel, J. Fernandez Lozano and A. Garcia-Cerezo, "Adaptive Cartesian motion control approach for a surgical robotic cameraman," *Proc. IEEE Int. Conf. on Robotics and Automation,* 2004.

[8] A. Casals, J. Amat, D. Prats and E. Laporte, "Vision guided robotic system for laparoscopic surgery," *IFAC Int. Cong. on Advanced Robotics,* 1995.

[9] C. Doignon, F. Nageotte, B. Maurin and A. Krupa, "Pose estimation and feature tracking for robot assisted surgery with medical imaging," in: D. Kragic and V. Kyrki (Eds.), *Unifying Perspectives in Computational and Robot Vision*, Springer-Verlag, Berlin, pp. 1-23, 2007.

[10] P. Dutkiewicz, M. Kielczewski, M. Kowalski and W. Wroblewski, "Experimental verification of visual tracking of surgical tools," *Fifth Int. Workshop on Robot Motion and Control*, 2005.

[11] J. Sun, M. Smith, L. Smith and L. P. Nolte, "Simulation of an optical-sensing technique for tracking surgical tools employed in computer assisted interventions," *IEEE Sensors Journal,* vol. 5, no. 5, 2005.

[12] S. Payandeh, Z. Xiaoli and A. Li, "Application of imaging to the laparoscopic surgery," *IEEE Int. Symp. Computer Intelligence in Robotics Automation,* pp. 432-437, 2001.

[13] P. Milgram, S. Zhai, D. Drascic and J. Grodski, "Applications of augmented reality for human-robot communication," *IEEE/RSJ Proc. on Intelligent Robots and Systems,* pp. 1467-1472, 1993.

[14] F. Devernay, F. Mourgues and E. Coste-Maniere, "Towards endoscopy augmented reality for robotically assisted minimally invasive cardiac surgery," *IEEE Proc. Int. Workshop on Medical Imaging and Augmented Reality,* pp. 16-20, 2001.

[15] A. Pandya and G. Auner, "Simultaneous augmented and virtual reality for surgical navigation," *IEEE Annual Meeting of the North American Fuzzy Information Processing Society,* 2005.

[16] R. Taylor, J. Funda, B. Eldridge, S. Gomory, K. Gruben, D. LaRose, M. Talamini, J. Kavoussi and J. Anderson, "A telerobotic assistant for laparoscopic surgery," *IEEE Engineering in Medicine and Biology Magazine,* vol. 14, no. 3, pp. 279-288, 1995.

[17] J. Funda, K. Gruben, B. Eldridge, S. Gomory and R. Taylor, "Control and evaluation of a 7-axis surgical robot for laparoscopy," *IEEE Proc. Int. Conf. on Robotics and Automation,* pp. 1477-1484, 1995.

[18] T. Ortmaier and G. Hirzinger, "Cartesian control issues for minimally invasive robotic surgery," *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems,* pp. 465-571, 2000.

[19] P. Milgram and F. Kishino, "A taxonomy of mixed reality visual displays," *IEICE Trans. Information System*, vol. 12, pp. 1321-1329, 1994.

[20] R. Schnidt, "Blue screen TV studios," http://www.heathcom.no, 1996.

[21] E.C. Urban, "The information warrior," *IEEE Spectrum*, vol. 11, pp. 66-70, 1995.

[22] J.P. Mellor, "Enhanced reality visualization in a surgical environment," AI Lab, Cambridge, MIT: 102, 1995.

[23] M. Uenohara and T. Kanade, "Vision-based object registration for real-time image overlay," in: N. Ayache (Eds.), *Computer Vision, Virtual Reality and Robotics in Medicine*, Springer-Verlag, Berlin, pp. 14-22, 1995.

[24] R.B. Welch, *Perceptual Modification: Adapting to Altered Sensory Environments*, Academic Press, New York, 1978.

[25] W.A. IJsselsteijn, M. Lombard and J. Freeman, "Toward a core bibliography of presence," *CyberPsychology & Behavior*, vol. 4, pp. 317-321, 2001.

[26] A.A Navarro and J. Aranda, "Assisting minimally invasive surgery through exterior orientation to enhance perception," *IEEE Int. Conf. Robotics and Automation*. pp. 2642-2647, 2007.

[27] A. Hernansanz, X. Giralt, A. Rodríguez and J. Amat, "RPQ: robotic proximity queries – development and applications," *Int. Conf. Informatics in Control, Automation and Robotics*. pp. 59-66, 2007.

[28] C.G.L. Cao and P. Milgram, "Disorientation in minimal access surgery: a case study," *Proc. IEA/HFES*, 2000.

[29] J.J. Gibson, *The Ecological Approach to Visual Perception*, Houghton-Mifflin, Boston, 1979.

Chapter 7

Conclusion

The 3D visual effect supplied by the appearance of objects under perspective projection is an essential sensory component to perceive space. Particularly, the effect created by the projection of angles provides a convincing 3D conception that has led to direct this research considering this simple geometric feature as a monocular cue. The response of the human visual system to the visual stimulus created by the angular variation can be appreciated by simply observing the vertices of a cube or the corners of a table. This is a visual process dedicated to determine spatial information and therefore, an ideal model from which the computer vision approach developed in this thesis was inspired.

The link established between depth cues represented by image features and biological processes performed by the human visual system have led to qualify the developed 2D-3D metric relation as a computational approach to direct perception. It has been an appropriate principle from which this approach has been based on and has represented its ultimate aim, which was the artificial emulation of a specific function of sight. The perspective distortion model resultant from the 2D-3D geometric relation developed and evaluated through the angular variation analysis, has demonstrated to be a reliable technique focused to extract 3D information from bidimensional images. Since it is a method based on the analysis of the proper content of images and their invariant entities, the concept of direct perception appeared and showed to be suitable to describe the angular variation. Although the modeling of this geometric property has confirmed the fact that the optic array is a direct provider of spatial information, it only describes a visual behavior under a determined stimulus. Thus, it is not a conclusive proof to the direct perception theory. However, it is a convincing support.

A set of methods was developed in order to extract spatial information from images using angular variation. Each method has determined visual properties from the angular projection. Parting from conventional strategies, relating image and object features, to invariant properties of geometric configurations, these methods have demonstrated that the perspective distortion of angles is an explicit and reliable source of spatial information. This information specifically describes spatial orientation of 3D features, which in conjunction with external 3D data provides position and scale. The set of developed conventional strategies has been based on the rotational motion analysis and its derived cone-shape line motion structure. Both of these methods are practical and useful in determined applications. However, the analysis of the perspective distortion offers a biologically inspired method to calculate spatial information directly from the projective nature of angles. Therefore, the development of a perspective distortion model, in addition to be a computational approach to direct perception, has shown to be a simple, fast and robust technique. Since it is focused in changes produced in the proper content of the image, it is independent of explicit external data. Thus, from the set of developed methods, this technique can be qualified as an improvement or evolution of the other methods, where 2D-3D relations were applied indirectly through a set of geometric constraints. As a result, the projective nature of angles was modeled and from its characteristics was concluded that a single angle possess quantifiable information only in restrictive conditions. Therefore, it requires additional information provided by the inclusion of more angles or produced by rotational motion.

Implications of the significant quality provided by the content of images to represent sensorial information were studied. The interpretation of monocular cues, as it is performed by the human visual system, has been resembled by a computer vision algorithm and thus, 3D information of the scene has been calculated. In action-perception applications this is a useful tool that supplied in a suitable form improves the visual-motor coordination and hence, increases performance skills to complete determined tasks. Experimental results have shown a positive effect on the cognitive mediation between action and perception under tasks aided with a complementary view of the scene. As the only needed information was the orientation between the self-moved tool and the camera, a key role is given to the hand-eye relation. Therefore, the visualization of one's own actions in a self-referenced environment provides relevant cognitive information necessary to augment presence and improve the visual-motor coordination in mediated environments.

The potential of this computational approach, as has been demonstrated in the previous chapter, is prominent in various applications; parting from simple calibration tasks to presence enhancement in immersive teleoperations. The studied link between projected visual cues and their interpretation to determine spatial information has resulted in the

development of a computational approach, which instigates to extend future research in the analysis of different geometric entities as additional descriptors of visual perception processes. This modeling of geometric variations captures the visual behavior in response to stimuli generated by the interaction with the environment and permits to identify and understand certain processes employed by the human visual system.