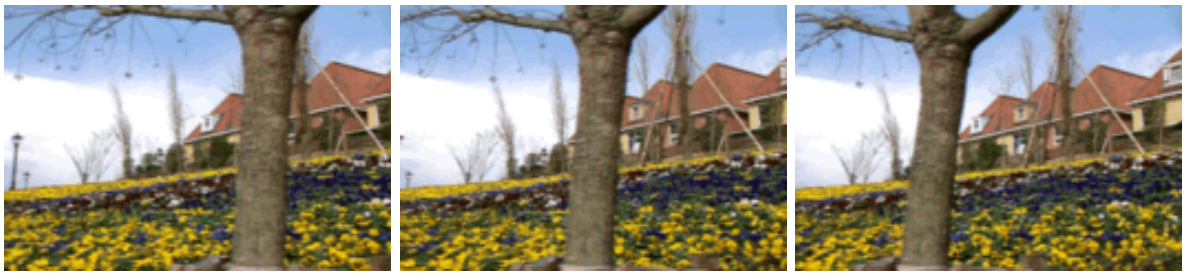
UNIVERSIDADE TÉCNICA DE LISBOA

INSTITUTO SUPERIOR TÉCNICO



3D Motion and Dense Structure Estimation:

Representations for Visual Perception and the Interpretation of Occlusions

César Augusto dos Santos Silva

(Mestre)

Dissertação para obtenção do Grau de Doutor em

Engenharia Electrotécnica e de Computadores

Orientador: Doutor José Alberto Rosado dos Santos Victor

Presidente: Reitor da Universidade Técnica de Lisboa

Vogais: Doutor Henrik I Christensen

Doutor João José dos Santos Sentieiro

Doutor Helder de Jesus Araújo

Doutor Jorge dos Santos Salvador Marques

Doutor Mário Alexandre Teles de Figueiredo

Doutor José Alberto Rosado dos Santos Victor

Maio de 2001

Agradecimentos

A concretização deste projecto contou com a contribuição preciosa de muitas pessoas, pelo que cabe aqui dirigir-lhes uma palavra de especial agradecimento. Primeiro quero agradecer a todos aqueles que contribuiram directa ou indirectamente neste projecto, quer ao nível das ideias que determinaram definitivamente o curso deste projecto, quer ao nível de críticas e trocas de opinião que moldaram a sua realização. Começo inevitavelmente com o José Alberto, que primeiro me apresentou o tema, e que depois, com toda a paciência do mundo, me orientou, me apoiou, me aturou, e, acima de tudo, me deu liberdade de pensar e de fazer. Foi e é um exemplo de organização e determinação. Agradeço a todos aqueles que emprestaram um pouco de si a este projecto: o Alexandre, o João Maciel, o Etienne, o Nuno Gracias e o Gaspar. As suas participações foram valiosas, as nossas discussões eternas, por vezes alucinadas. Foram todos eles exemplos de verdadeira curiosidade científica. Agradeço pela sua disponibilidade e amizade todo o restante pessoal do laboratório, especialmente o António e o Sjoerd. Ao ISR agradeço, pelo ar de descontracção total que nele se respira, pela vista panorâmica, pela aposta na investigação e na criatividade, coisas raras neste pedaço de terra à beira mar plantado.

Ficam para o fim mais sentimentos do que palavras, agradeço à Cecília o seu amor, agradeço aos meus irmãos a irreverência e o entusiasmo.

Dedico esta tese à minha mãe e ao meu pai.

Resumo

Nesta tese, estuda-se o problema da reconstrução tridimensional, a partir de uma câmara em movimento. Propomos estimar o movimento próprio da câmara e calcular o mapa denso de profundidades da cena observada, explorando para tal representações particulares da imagem que facilitem essa reconstrução, tendo em conta as oclusões.

A primeira parte da tese é dedicada ao estudo do problema da estimação do movimento próprio da câmara. A nossa abordagem é baseada na análise do movimento da imagem. Introduzimos um conjunto de subespaços onde são deduzidas algumas restrições relativamente ao fluxo e ao movimento da câmara. Este conjunto de subespaços pode ser descrito num único espaço topológico $\mathcal{L}$ que permite a utilização global dos vectores de fluxo. É estudado o fenómeno das oclusões, que contêm igualmente informação quanto ao movimento da câmara. Propomos um modelo formal para detectar e classificar oclusões e é deduzida uma relação entre as oclusões e o movimento da câmara.

A segunda parte da tese trata o problema da reconstrução densa. Propomos explorar representações alternativas que facilitem o problema da correspondência. Baseados na restrição de ordem, apresentamos uma representação simples e compacta — as Imagens Intrínsecas — para a estimação da correspondência estéreo, lidando com oclusões. No sentido de ultrapassar a necessidade de usar a restrição de ordem, é apresentada uma metodologia baseada em técnicas de optimização para calcular a correspondência num sistema trinocular, partindo de restrições com forte significado físico (unicidade, visibilidade e geometria das câmaras).

Foram apresentadas experiências com imagens reais para todas as metodologias propostas.

Palavras-chave: Visão por Computador, Estimação de Movimento Próprio, Reconstrução 3D, Oclusões, Correspondência Estéreo, Reconstrução com Múltiplas Vistas.

Abstract

This thesis addresses the problem of 3D reconstruction from a moving camera. We propose to estimate the motion of the camera and recover densely the depth of the scene, by exploring particular representations for the image information, where the occlusions play a central role.

The first part of the thesis is dedicated to the egomotion estimation problem. The approach is based on the analysis of the image motion. We introduce a family of subspaces where some constraints that use the global image motion information, are exploited to estimate the camera motion. This family of subspaces is described in a unique topological space — the $\mathcal{L}$-space — that allows the use of robust estimation tools. In order to deal with the occlusions, where the flow computation is difficult, we propose a formal model for detecting and classifying occlusions. We show how the occlusions provide additional information about the camera motion.

The second part of the thesis addresses the 3D reconstruction problem or dense matching. To approach the matching problem, we propose to explore alternative representations that may facilitate the correspondence process. Based on the order constraint, we present a new image representation called Intrinsic Images that can be used to solve the stereo correspondence within a simple and compact framework, dealing with occlusions. To overcome the need of using the order constraint, we propose a methodology for solving optimally the correspondence problem for a trinocular system, imposing solely physically meaningful constraints (uniqueness, visibility and geometric).

Several results with real images are presented for all methodologies proposed.

Keywords: Computer Vision, Egomotion Estimation, 3D Reconstruction, Occlusions, Stereo Matching, Multiple View Reconstruction.

Contents

Introduction

This thesis addresses the problem of retrieving the observer motion and scene structure from images, a task that is solved routinely and robustly by humans and many other animals.

The study of visual perception has a long history. During centuries, men have felt the need for an explanation of why and how things are visually perceived. Among the many questions which were discussed by physiologists, psychologists and other vision researchers till now, the oldest and more general and mysterious problem consists in how we can account for the richness of the perceived environment considering the apparent poverty of the image within our eyes.

Vision depends on a simple and almost flat retinal representation. Nevertheless the visible scene has depth, distance and consistency. Then how can vision depend on the pictures in the eyes and yet produce a scene with such complexity? The physical environment is tridimensional and projected by light on a bidimensional surface. However it is perceived in three dimensions. How can the lost third dimension be restored and reconstructed in perception? Moreover how we can perceive our own motion and tridimensional position by using vision?

Many answers for these problems can be found in different fields such as biology, psychology or engineering. But one of the most fruitful interaction of two different scientific areas addressing such issue has been observed between the biological approach to the human vision and the computational approach to the artificial vision. Thus we start in with this important relationship, which has produced some of more significant results in the history of the vision research.

Historical review

In the early days, vision researchers tried to identify and classify objects in images by the techniques of pattern recognition, which had been developed for the extraction and classification of bidimensional features [10]. They believed that the paradigm of pattern recognition could lead to systems able to understand tridimensional scenes from bidimensional image features.

However, they soon realized that a tridimensional object looks very different from viewpoint to viewpoint and 3-D meanings cannot be recovered unless some a priori knowledge or internal representation is given. Thus knowledge came to play an essential role and how to represent and organize such knowledge became a major concern.

The first organized and significant attempt to solve this problem can be found in the Gestalt psychology [22]. From the Gestalt psychology viewpoint, humans understand the 3-D environment by unconsciously matching the visual images with the vast amount of knowledge accumulated from experience.

Inspired in this principle, many researchers from the Artificial Intelligence field devised symbolic schemes in order to represent and interpret visual knowledge [11]. One of their central problems consisted in establishing such symbolic representations. Thus many combinatorial techniques were proposed, including various heuristic and complex search algorithms, relying heavily on some assumption arising from own visual experience.

Realizing that some computational complexity is inevitable as long as an internal representation is directly matched with features extracted from raw images, many researchers began to pay attention to physical laws existent on the image which could provide clues to a general tridimensional interpretation [28]. Depending on these physical laws, different computer vision areas were generated: shape from shading by using shading variations on the images, shape from texture by analyzing textures invariants, and structure from motion by computing the flow of moving objects on the image.

In contrast to the kind of knowledge devised before (which depended on specific internal representations), this new approach uses different physical constraints or assumptions about surface reflectance, illumination, perspective distortion and rigid motion.

In psychology, this view was primarly asserted by J. Gibson [22, 23] who believed that human 3-D perception occurs automatically when the visual signal triggers the right computation in our brain and such computational functionality is innate. Later, founded in Gibson's theory, David Marr [41] established a new paradigm to describe the visual perception process, which can be summarized as follows: i) primitive features are extracted (like edges or color); ii) approximate shapes or movements are computed by using some given constraints and iii)

a 3-D model is fitted to such data. Additionally, some high-level inferences can be performed based on such 3-D representation.

The major advantage of the Marr paradigm consisted in transforming the vision problem in a computer science problem, and that fact revealed very useful for computer scientists. In computer vision, one can use images captured by electric cameras and create more easily a computable model to interpret image data (by using color, motion or brightness measured by the camera). By studying computations designed to infer tridimensional information from camera images, we can learn more about the way how the visual system succeeds in interpreting images because of physical and statistical regularities present in the retinal images.

However, from a biological point of view, Marr's theory is still a set of conjectures which need scientific confirmation. Physiologically, the information within the retina is considered ambiguous and without a well-defined representation; the neural connections (from retina to the visual pathways) are ruled by strange models and their behaviors are apparently fuzzy and only partially known. Therefore, it is important to keep in mind that the approaches found in computer vision are not necessarily similar to the treatment of visual information in the brain. Camera information is very different from our own visual information. Cameras are dependent on the technology available. In contrast, our neural system suffered millions of years of evolution. So any eventual comparison between the electronic and biological systems may be at least unfair.

Nevertheless, even knowing that the computer vision problem is different from the biological vision problem in terms of "hardware", the question has been remained essentially unchangeable for researchers from both sides: how can we reconstruct and understand the tridimensional environment by using exclusively flat images? Answering this common question has been the constant challenge for either biologists or computer science researchers. Their motivations to study the problem of visual reconstruction are also very similar and arises obviously from the human visual experience. The dream of every computer vision researcher is to obtain a computer system with a performance similar to our visual system. Thus, the ultimate achievement would be to mimic the biological visual system. Such a tool would be of great interest for the better comprehension of our visual system.

However, the motivation to study computer vision can be purely technological. It is an unquestionable fact that the usual approaches for 3D reconstruction depend on the specific application domain, such as surveillance, robotics, or entertainment. To some extent, these well-known applications have ruled the recent research on computer vision, with some well-known successful results.

Next we will discuss several current "schools of thought" relatively to the nature of the information used, when addressing the problem of 3D reconstruction from images in computer vision.

When observing the projection of a set of scene points in motion, what can be determined about their 3D structure? This general problem of reconstruction can be addressed in various ways that lead to different formulations:

- Given the projections of a set of 3D points on an arbitrary set of views, how can one determine the point matches, their 3D position, and the pose and parameters of the cameras?
- When observing an image sequence, can one determine the instantaneous camera motion as well as the scene structure?

Different formulations for the general underlying problem of reconstruction make use of different data and models, since the desired final output is somewhat distinct. Hence, most methodologies for 3D reconstruction can be analyzed in terms of the type of measurements made, the models used and the representations chosen for the 3D structure and motion. Based on this analysis, a simple classification for the various existing formulations can be provided, according to:

The nature of the observations — Some methods are based on a reduced set of highly accurate image features that act as fiducials for further computations. Other methods use more globally (and directly) the information within the images for the same purpose of retrieving camera and scene structure, without requiring a previous feature selection.

The model of the camera used — A great deal of work has been done in the case of *discrete* motion, especially for stereo vision, aiming at recovering the camera positions

and orientations. Alternatively, making the assumption of *continuous* motion of camera (and image) as it moves little from one image to another, the reconstruction consists in recovering the instantaneous camera velocities, since the concepts of time and space are unified in a single framework.

The scene interpretation — The 3D scene can be recovered either sparsely or densely, depending on the photometric and geometric constraints, and the estimation algorithms used. One may provide an holistic interpretation of the scene instead of focusing on a limited set of features.

These three items motivate the principal theories for reconstruction in computer vision, depending mainly on the type of representation we adopt for each one of them, as summarized in the following table:

Observations	**Camera**	**3D Scene Interpretation**
Feature based (pre-selection required)	Discrete model (camera positions)	Sparse (reconstruct some points)
Image based (no pre-selection)	Continuous model (camera velocities)	Dense (reconstruct all points and surfaces)

The combination of all these requirements includes most of the work done in visual reconstruction. For example, if the camera is moving continuously and we want to use the image motion as observation, then our reconstruction problem consists in recovering the velocity of the camera from the optical flow. On the other hand, if we have a stereo pair with some previously selected fiducial points, then we want to estimate the matches and the pose of the cameras that generate those points. Additionally, both reconstruction algorithms can be designed to recover either the tridimensional information in some points (e.g. contours), or the complete surfaces of the objects (including in non textured zones, for example). The following paragraphs will be dedicated to discuss several approaches for visual reconstruction where different choices in terms of observations, camera and 3D scene representation are made.

Traditionally one of the mainstream approaches in computer vision consists of the feature based geometric reconstruction. This approach is connected to the uncalibrated reconstruction

of a set of cameras [25, 64, 51, 47] (usually under a **discrete** model [1]). In this field, some remarkable progress has been achieved [26], by using sophisticated mathematical tools, mainly based on projective geometry. Many intricate and elegant geometric relations have been found to describe the multiview geometry. Crucial results about the camera positions and scene structure (usually **sparse**) are clarified by using concepts perfectly described with projective geometry: the fundamental matrix, that relates a stereo pair [14, 42, 26]; the trinocular tensor, that deals simultaneously with three cameras [63, 52]; the cross-ratio of four collinear points, that is a projective invariant [12]; or the absolute conic that codifies the internal calibration parameters of the camera [15]. These and many others geometric concepts give a global consistency to the projection process and provide a coherent explanation for "what is really happening" when a set of cameras are seeing images.

A typical reconstruction algorithm starts with the extraction of **features** followed by the computation of their 3D shape. Thus, these estimates can be used to guide a **sparse** correspondence for the remaining relevant features of the image. A feature can be an interest point, an edge, an isobrightness curve, a silhouette contour or a conic. Features are useful because there is a wide choice of algorithms to estimate the 3-D structure from point, line or curve features [42]. Notice, however, that the image processing involved in the feature detection is far from being a solved problem (excluding when the required features are simply corners).

In opposition to these methods, that generally rely upon a reduced number of image features (usually pre-selected as "good" ones), other approaches use more extensively the image information (often known as **image based** approaches) for the same purpose of retrieving camera and scene structure. The most known example is the use of image motion (flow) for the 3D motion reconstruction of a moving camera — structure from motion [35, 37, 30, 5, 29, 27]. In opposition to the "geometric" approach (that structurally separates the feature selection from the reconstruction problem), the image based methods use the raw information (e.g. brightness gradients) for the purpose of retrieving information about the camera motion or scene structure, without intermediate feature extraction.

[1] although a **continuous** approach can be studied [46].

Some advantages can be found by using an image based approach. First, one can define a global minimization problem using all the image information. For example, the set of all observations can be used as input of robust estimators in order to recover a few parameters such as the translation and rotation of the camera [1, 17]. Secondly, the global structure of the image observations is still connected with the problem geometry. As an example, rigorous descriptions relating the image motion field with the camera velocity (under a **continuous** model) can be found in [67, 39]. Additionally, some information included in the image, such as the brightness variations or the normal flow computed from a moving camera, is easily computable and directly available [33, 30].

However, image based methods may suffer from various severe drawbacks: (1) Unfortunately, there are regions of the image where the observations are not particularly easy to obtain (e.g. flow in non-textured areas). Then, in many applications only **sparse** results in terms of 3D scene reconstruction can be computed. (2) Some tacit (or explicit) photometric constraints about the image have to be previously assumed — they are present either when we compute the normal flow, assuming the brightness constancy, or when we include similarity, uniqueness, order or transparency criteria for the computation of correspondences. (3) Serious difficulties can be found at occlusion boundaries (moreover the occlusion detection is not trivial).

A large amount of work has been produced in order to find other paths for reconstruction, using raw image information. As examples we can mention not only the study of motion flow (as referred before), but also the codification of the stereo images through a graph [7] (**discrete** model, **dense** 3D scene estimation), the representation of multiple view information in a dense set of voxels [36] (**discrete** model, **dense** 3D scene estimation), or the transformation of the image spatio-temporal cartesian space into another space — e.g. the frequency space [19] (**continuous** model, **dense** 3D scene estimation). All these approaches constitute relevant contributions for the image comprehension in terms of reconstruction, and truly explore the structure information coded within the images.

The next section will be dedicated to outline our approach within the general framework of the visual reconstruction.

Thesis Outline

Objectives

In this thesis, we address the central problem of computer vision of retrieving the camera motion and dense scene structure from images. Our motivation to study this problem does not arise from a purely geometric interest, typical of a feature based approach. Instead, it is focused on the study of the image dynamics, and its influence on the perception of both the observer motion and structure of the observed scene.

Throughout the thesis, we assume that the scene is static, thus excluding the study of multiple motions. Additionally we assume that some intrinsic camera parameters are previously known. This can be a reasonable assumption, provided that we choose the cameras used in the experiments.

Our principal aim is the determination of the 3D camera motion and the 3D reconstruction of the scene. These problems are always associated to the determination of a dense point-to-point correspondence (or motion), and in the presence (and making use of) occlusions. Hence we have selected three of the most challenging and representative problems in this thesis:

- The interpretation of (image) visual motion conveys information about both the 3-D scene and the observer motion. The motion of the image points forms special patterns that can reveal a global structure related to the camera motion. Thus, some important cues about the camera motion can be found by exploring the geometric properties of the image flow within an adequate image flow representation. How does the global image motion flow depend on the camera motion?

- In computer vision, occlusions are almost always seen as undesirable singularities that pose difficult challenges to image motion analysis, such as flow computation or disparity estimation. However, it is well known that occlusions are extremely powerful cues for depth and motion perception. What information can occlusions provide about the camera motion?

- Dense matching is a central issue in computer vision, namely to retrieve depth infor-

mation. Usually, the matching procedure is performed directly on the images data and is deemed to failure in regions of homogeneous brightness distribution. A challenging problem consists in finding alternative image representations that facilitate (or even trivialize) the correspondence process. How can these representations incorporate (and take advantage of) occlusion and visibility constraints ?

There are two strong ideas present along the entire thesis and behind our approach. Firstly, we believe that a major step to solve the 3D motion estimation and dense matching problems is to find adequate visual representations (or spaces) where solutions can be more easily defined or constrained. Secondly, we face occlusions as a rich source of perceptual information that should be used when estimating the observer motion or performing 3D reconstruction, dense matching or view synthesis.

Organization of the Thesis

In general, this thesis is organized in two main parts: a first part describing geometrical approaches to compute the camera motion, and a second part discussing methods to recover dense tridimensional information from the scene. The thesis is organized as follows:

First Part: A great deal of work on 3D reconstruction has been done in the case of discrete motion, especially for stereo vision, assuming that the camera intrinsic parameters are known. Making the assumption of continuous motion simplifies the problem of tracking a feature as it moves little from one image to another, unifying the concepts of time and space in a single framework. A major drawback of a continuous approach is its susceptibility to local noise (due to image noise or occlusion phenomena). However, if considered globally, the set of all observations can be used as input of robust estimators in order to recover global parameters such as the translation or the rotation of the camera. In the first part of this work, these estimators are discussed and the occlusion problem is studied. In more detail, the first part, concerned to camera motion estimation, is subdivided in two chapters:

Chapter 1 introduces the basic notion and models used in image motion analysis. We

propose to use exclusively the spatio-temporal derivatives of the image in order to recover the motion of an observer subject to arbitrary linear and angular movement. We subdivide the image plane in a set of line subspaces to estimate linearly constraints of the camera motion. Described on a different representation — the $\mathcal{L}$-space —, this approach allows for numerous observations to contribute for the complete motion estimation, improving robustness and reliability. The study proposed here gives us a global description about the connection between the general geometry of the image motion and the camera motion.

Chapter 2 discusses an important class of singularities that pose special difficulties to image motion computation: the occlusions. Rather than an undesirable artifact, we show how the occlusions provide fundamental information for camera motion estimation.

Second Part: By applying the estimators defined in the first part, we assume that the camera motion is already known. In order to estimate dense information about the 3-D structure of the scene, we could use the same image motion observations. However, these local observations can produce meaningless depth interpretations in the presence of noise. This is because depth reconstruction from 2-D is a typical inverse problem, for which solutions are known to be generally unstable with respect to noise. To cope with this inherent ill-posedness, the local correspondence measurements must be improved. This can be achieved through two complementar approaches: Firstly, finding alternative representations which support a true dense matching procedure; secondly, developing optimization techniques constraining the solution to have coherent visual properties (related to order, unicity or visibility constraints). These approaches will be discussed in the second part of this work, where results obtained from various real examples will show their potentialities. The second part, centered in the scene reconstruction, is subdivided in two chapters:

Chapter 3 introduces one of central problems in reconstruction: determining correspondences. The computation of correspondences is presented as a fundamental

issue for dense 3D reconstruction or synthesis of new views. We develop a new representation — the *Intrinsic Images* — which leads to the dense matching of a stereo pair, even with occlusions.

Chapter 4 discusses the problem of existing more than two views, introducing new constraints, namely geometric and visibility constraints. We develop an optimization approach for dense reconstruction applied to a specific trinocular example, where an optimal solution is found.

Each part is finalized by a concluding chapter where a brief discussion is provided.

Related Work

For the various problems discussed throughout the thesis, a wide range of related work takes a special relevance due to its close relation with the approaches we will propose here.

In the first part, the **egomotion problem** applied to a moving monocular observer is addressed. This is a central issue in computer vision and is deeply treated in remarkable works, such as [35, 37, 30, 5, 29, 27], just to mention a few. The first step to estimate egomotion or structure from motion is the computation of displacement between consecutive frames. Usual approaches are based either on (i) point correspondences [13]; (ii) estimation of the dense motion field, identified by the optical flow [27]; or (iii) finally in the so-called direct methods [31, 16, 18, 33].

The correspondence or optical flow estimation are, in general, ill-posed [2], and it is usually necessary to introduce very restrictive assumptions about the image or the observed scenes. The solutions obtained are often unstable and require large amounts of computation.

Direct methods are less demanding than the previous ones, and use the image brightness information directly to recover the motion parameters. The only image flow component that can be estimated based on local measurements, is that along the image gradient direction, the normal flow [30]. Thus, direct methods are usually based on normal flow or the spatio-temporal image derivatives [50]. Then how can the simple normal flow provide information about the camera motion? Fermuller and Aloimonos [18, 16, 17] introduced a method which

is based on a selection of image points that form global patterns in the image plane, using the orientation of global normal-flow vectors. This approach considers the complete camera motion estimation problem as a pattern recognition problem, introducing a set of geometric constraints that reveal a global structure hidden in normal-flow fields. In this case, the use of global data must guarantee the robustness of the algorithms.

Inspired on Fermuller and Aloimonos work, the method we follow consists in searching the image for particular geometric properties of the normal flow, tightly connected to the egomotion parameters. As proposed by Heeger and Jepson [27], one can solve separately the rotational and translational motion parameters due to bilinear nature of the image motion equations. We also subdivide the problem, but in a different way. The difference is that we split not only the solutions domain but also the search domain in various families of lines, where the normal vectors have particular geometric properties, associated to camera motion information. The space of lines can be described in a unique topological space — the $\mathcal{L}$-space —, that allows the use of nearly all available image data [57]. We present a low complexity estimator based on the simple regression problem, by applying sequentially bidimensional estimators on the $\mathcal{L}$-space.

In the second chapter of this thesis, we propose to study the importance of **occlusions** within the egomotion estimation framework. A number of algorithms have been designed to handle or eliminate occlusions in order to estimate either multiple image motions [65, 4, 60, 43] or the disparity field of a stereo pair [45, 9, 21, 34]. In contrast to these works, we do not focus on the explicit estimation of image motion (or disparity) but rather, on the role played by occlusions in motion perception. Assuming a moving monocular observer, we study the relation between the observable occlusions and the camera motion. Thus, we show how the occlusions provide, per se, fundamental information for motion estimation [58]. These conclusions partially validate some of the biological evidences already observed in the Psychology area (namely in the work due to Anderson and Nakayama [3]).

The second part is dedicated to the **dense 3D reconstruction** of the scenes. This process is intrinsically related to the correspondence problem, one of the central problems of stereo and motion analysis. Traditionally, the computation of correspondence has been associated

[24] to the following general three-step procedure:

1. Selecting image features, such as edges, interest points or brightness values.

2. Find corresponding features based on similarity and consistency criteria, where similarity considers a distance between features and consistency takes into account geometric and order constraints. This step is usually algorithmic and requires a relative weighting between similarity and consistency.

3. Compute and interpolate the disparity maps to obtain a dense field.

An additional step consists of detecting occlusion points, where the similarity and/or consistency criteria are violated.

Almost all methods for image correspondence follow the procedure described above, differing only in the nature of the image features, similarity and consistency criteria, search algorithms and photometric and geometric assumptions [45, 66, 61] (see [24] for an overview).

In this thesis, we propose an holistic approach where global stereo image information and the similarity and consistency criteria are well defined in a common framework. This approach can be used to generate dense disparity maps (explicit reconstruction) or to synthesize new views of the same scene (without explicit reconstruction).

First, we propose a new image representation called *Intrinsic Images* that can be used to solve correspondence problems of a stereo pair [59]. This work has partially its foundations on the intrinsic curves developed by Tomasi and Manduchi [61]. They have proposed to represent a set of local descriptor vectors (brightness, derivative, second derivative, etc) through a curve in a n-dimensional space. Ideally two curves computed along two corresponding scanlines of a stereo pair can be mapped (or even coincide).

However, approaches based on curves of local descriptors vectors have obvious limitations related to rigid geometric distortion assumptions, solution ambiguity and/or high-dimensionality search algorithms. First of all, the method is only valid for constant and affine disparities (no perspective effects have been considered). Secondly, the curves have a difficult representation, specially if more than two local descriptors are considered. Finally, a curve can cross itself, clearly generating ambiguous situations.

Here we develop a simple framework that overcomes the main restrictive geometric, photometric and algorithmic constraints, mentioned before. We propose to study other kind of representations, based not only on local descriptors but also on global descriptors of the image, that we call *Intrinsic Images*, that simplify the (dense) matching process and can be used to generate new views from a stereo pair.

The main limitation with this approach is the assumption of the *order* constraint, which can be violated in the presence of occlusions (although not always), thus leading to poor results in those regions. Most of the matching methods rely explicitly on the order constraint to retrieve the 3D structure of a scene [45, 12, 7]. Many others use order-like assumptions, named in different ways such as the continuity of the disparity map [68] or the local coherency of the correspondences [49]. A different constraint is considered by Kutulakos and Seitz [36], where one explores the visibility constraint for an arbitrary set of calibrated cameras, and no approximation was made. However this approach fails when outliers are introduced in the observations since the reconstruction algorithm is greedy and is not able to deal with outlier rejection under an optimization framework.

The work by Maciel and Costeira [40] provides a good insight into an optimization approach applied to the correspondence problem. They develop a set of generic tools based on integer optimization in order to handle several constraints in a unique formulation, performing correspondence and outlier rejection (or occlusion detection) in a single step. The problem which remains to solve is to describe generic visual constraints in that optimization framework.

We propose a methodology for solving the point correspondence problem, relying exclusively on physically valid constraints (i.e. no approximations), thus avoiding the use of the order constraint. We explore and discuss generic geometric, uniqueness and visibility constraints for the stereo problem. For the example of a trinocular setup, we show how to represent these constraints in such a way that it allows us to formulate the correspondence problem as an integer optimization problem, where occlusions are naturally represented. By resorting to optimization tools, the global solution is found without imposing any additional hypotheses.

Original Contributions

Throughout the thesis several methodologies are proposed in order to estimate 3-D structure. We have selected the following contributions as the most relevant of this work:

- We introduce a search paradigm which is based on geometric properties of the normal flow field, and consists in considering a family of search subspaces to estimate the ego-motion parameters. In order to decrease the noise sensitivity of the estimation methods, we define a particular topogical space — the $\mathcal{L}$-space — to allow the use of global data for the final estimates and the use of statistical tools, based on robust regression theory.

- We present a formal model for occlusions points and develop a method suitable for occlusion detection. Through the classification and analysis of the detected occlusion points, we show how to retrieve information from occlusions about the camera translation.

- We propose two approaches to generate a dense disparity map or synthesize different views of the same scene, dealing with occlusions. First, we present a new image representation called *Intrinsic Images* that can be used to solve the stereo correspondence problem within a simple and compact framework. We extend this framework to deal with occlusions through an optimization technique based on dynamic programming.

  Secondly, we propose a methodology for solving optimally the correspondence problem for more than two views, imposing physically meaningful constraints. We have studied a paradigmatic example where the correspondence and occlusion detection is formulated as an integer optimization problem.

Part I

Camera Motion Estimation

In this part, we address the problem of egomotion estimation for a monocular moving observer. This is a real need for many robotic applications where an autonomous system must be able to estimate and/or control its motion parameters before any higher level tasks can be addressed (like a point to point moves or homing procedures). The problem of egomotion estimation consists in determining the 3D motion parameters, by observing an image sequence over time.

The method we follow consists in searching the image for particular geometric properties of the normal flow, tightly connected to the egomotion parameters. We split both the solutions domain and the search domain in various families of lines, where the normal vectors have particular geometric properties, associated to camera motion information. The problem can be formulated on a unique topological space — the $\mathcal{L}$-space — that allows for numerous normal flows vectors to contribute for the problem solution, hence improving robustness and reliability [57].

The salient aspects of the method proposed here, are the exclusive use of the spatio-temporal image derivatives, the fact that it can cope with arbitrary linear and angular observer motion and the effort that it is made on robust estimation in order to decrease the noise sensitivity.

In image motion analysis, occlusions are always undesirable singularities that pose difficult challenges to (optical or normal) flow computation. Assuming a moving monocular observer, we study the relation between the observable occlusions and the camera motion. To approach this problem, we first develop a sufficient condition for the existence of an occlusion, and thereafter a direct relation between camera translation and occlusion classification is provided.

At the end of both chapters, we report and discuss a wide variety of experiments with synthetic and real images, for various kinds of camera motion.

Chapter 1

Lines Topological Space for Egomotion Estimation

1.1 Image Motion

This section reviews the main aspects related to the image motion and its dependency both on the camera motion and scene structure. First of all, we introduce the main equations and concepts to be used in the remaining part of this chapter.

We consider a camera-centered coordinate system, as depicted in Figure 1.1. A 3-D point,

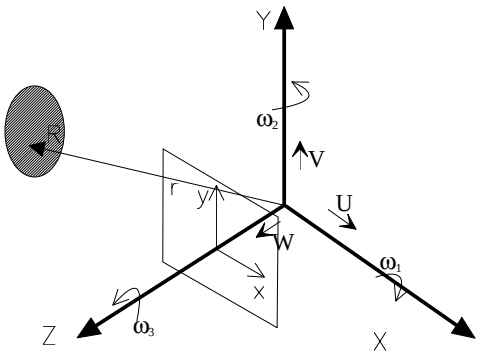


Figure 1.1: Imaging geometry and motion representation.

$\boldsymbol{R}$, is projected at the image point $\boldsymbol{r}$. The image formation is modeled as a perspective

projection, with focal length f.

Given the camera linear velocity $\boldsymbol{t} = [U\ V\ W]^T$, angular velocity $\boldsymbol{\omega} = [\omega_1\ \omega_2\ \omega_3]^T$ and the scene depth at each point, Z, the motion field induced in the image plane, can be calculated at every pixel by the following well-known vector equation [27] :

$$\boldsymbol{v(x)} = \rho(\boldsymbol{x})(\boldsymbol{x} - \boldsymbol{\Sigma}) + \boldsymbol{B(x)\omega} \tag{1.1}$$

where $\boldsymbol{x} = [x\ y]^T$ denotes an image site, $\boldsymbol{v(x)} = [u(\boldsymbol{x})\ v(\boldsymbol{x})]^T$ is the corresponding optical flow along the x and y axis, $\boldsymbol{\Sigma} = [\sigma\ \eta]^T = [fU/W, fV/W]^T$ is the the Focus of Expansion (FOE) and corresponds to the projection of the observer linear velocity, $\boldsymbol{t}$, in the image plane. The function $\rho(\boldsymbol{x})$ is the inverse of the time-to-crash, given by $\rho(x, y) = W/Z(x, y)$, and the rotation is multiplied by

$$\boldsymbol{B(x)} = \begin{bmatrix} \frac{xy}{f} & -(\frac{x^2}{f} + f) & y \\ (\frac{y^2}{f} + f) & -\frac{xy}{f} & -x \end{bmatrix} \tag{1.2}$$

1.1.1 Egomotion estimation from the optical flow field

As seen before, the motion field $\boldsymbol{v}$ is a linear function of $\boldsymbol{\omega}$ and bilinear in ρ and $\boldsymbol{\Sigma}$. With N sample points we can solve the bilinear equation (1.1), using the following system of equations:

$$\underbrace{\begin{bmatrix} \boldsymbol{v_1} \\ \vdots \\ \boldsymbol{v_N} \end{bmatrix}}_{\bar{\boldsymbol{v}}} = \underbrace{\begin{bmatrix} \boldsymbol{x_1} - \boldsymbol{\Sigma} & \dots & \boldsymbol{0} & \boldsymbol{B_1} \\ \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{0} & \dots & \boldsymbol{x_N} - \boldsymbol{\Sigma} & \boldsymbol{B_N} \end{bmatrix}}_{\boldsymbol{A(\Sigma)}} \underbrace{\begin{bmatrix} \rho_1 \\ \vdots \\ \rho_N \\ \boldsymbol{\omega} \end{bmatrix}}_{\boldsymbol{p}} \tag{1.3}$$

where $\bar{\boldsymbol{v}}$ is a $2N$ vector containing the optical flow of N points; $\boldsymbol{p}$ is a $N+3$ vector containing the 3 unknown rotational parameters and the N unknown time-to-crash values. Finally, $\boldsymbol{A(\Sigma)}$ is a $2N$-by-$(N+3)$ matrix which is completely computable if the FOE, $\boldsymbol{\Sigma}$, is known. In order to solve the system of equations (1.3) for a fixed $\boldsymbol{\Sigma}$, the following condition must be verified: $2N \geq N+3 \Leftrightarrow N \geq 3$. In general, however, $\boldsymbol{\Sigma}$ is unknown and the following cost function has

to be minimized, jointly in $\boldsymbol{\Sigma}$ and $\boldsymbol{p}$[1]:

$$\min_{\hat{\boldsymbol{\Sigma}}} \left\{ \min_{\hat{\boldsymbol{p}}} \left\{ \|\bar{\boldsymbol{v}} - \boldsymbol{A}(\boldsymbol{\Sigma})\boldsymbol{p}\|^2 \right\} \right\} \tag{1.4}$$

The main difficulty is that equation (1.1) depends on the *product* of $\rho(\boldsymbol{x})$ and $\boldsymbol{\Sigma}$, and $\rho(\boldsymbol{x})$ is an unknown 2-D function, as it depends on the depth for every pixel. Several approaches can be used to eliminate the dependency on $\rho(\boldsymbol{x})$ (or equivalently the dependency on depth). Let $\boldsymbol{d}(\boldsymbol{x}, \boldsymbol{\Sigma})$ be an unit vector, perpendicular to the direction $(\boldsymbol{x} - \boldsymbol{\Sigma})$, i.e. $\boldsymbol{d}(\boldsymbol{x}, \boldsymbol{\Sigma}) \cdot (\boldsymbol{x} - \boldsymbol{\Sigma}) = 0$ and $\|\boldsymbol{d}(\boldsymbol{x}, \boldsymbol{\Sigma})\| = 1$. Notice that $\boldsymbol{d}(\boldsymbol{x}, \boldsymbol{\Sigma})$ is well determined if $\boldsymbol{\Sigma}$ is known.

If we take the inner product of both sides of equation (1.1), the resulting equation is no longer dependent on $\rho(\boldsymbol{x})$:

$$\boldsymbol{d}(\boldsymbol{x}, \boldsymbol{\Sigma}) \cdot \boldsymbol{v}(\boldsymbol{x}) = \boldsymbol{d}(\boldsymbol{x}, \boldsymbol{\Sigma}) \cdot [\boldsymbol{B}(\boldsymbol{x})\boldsymbol{\omega}] \tag{1.5}$$

Hence, the components of the flow perpendicular to the lines going through the image pixel $\boldsymbol{x}$ and the FOE, do not depend on the camera translation. At least three points are needed to compute $\boldsymbol{\omega}$ for a fixed $\boldsymbol{\Sigma}$. However, many more points are required to search the Focus of Expansion which minimizes the variance of the rotation estimates.

This simple approach can be used for egomotion estimation [27], although the minimization process, involving the FOE search in a given domain, may be excessively time consuming. Other approaches dealing with the same problem of estimating egomotion from optical flow have appeared in the literature [32, 8]. One of the most known methods is due to Bruss and Horn [5], who presented a algorithm to the general motion which estimates translation by minimizing an appropriate residual function using iterative numerical procedures. More recently, Lourakis [38] presented an exact method to recover the FOE by solving a set of linear constraints. All these methods assume the knowledge of the complete optical flow, which is computed previously through complex algorithms imposing constraints on the structure of the scene or image. In next section we present the well known aperture problem which is the main reason why the optical flow computation is very difficult or even impossible in real applications.

[1] We use the following notation: $\min_{\hat{\theta}} \{E(\theta)\}$ represents the minimum of the function $E(\theta)$, where the index $\hat{\theta}$ is the value of θ for which $E(\theta)$ is minimal.

1.1.2 The aperture problem constraint

Assuming that the scene is composed of textured objects and that the illumination conditions vary slowly with time, one can introduce the familiar brightness-constancy hypothesis. This hypothesis implies a single linear constraint on the two components of the flow, which is the aperture problem [31]. In fact, we can only estimate the component of the optical flow, $\boldsymbol{v}$, along the vector $\boldsymbol{n}$, an unitary vector parallel to the image gradient. The projection $\boldsymbol{n} \cdot \boldsymbol{v}$ is commonly designated by the *normal-flow*, v_n. Hence, each pixel provides 1 equation on the unknown components of the flow, and the estimation problem stated in (1.3) cannot be solved, since we only have N equations instead of $2N$. Now the issue consists in estimating the camera motion uniquely from the flow information which is truly available: the normal flow.

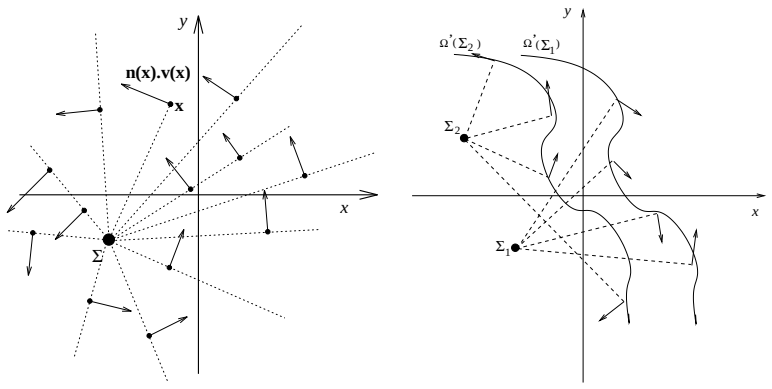
To solve this problem we can use the estimation approach based on the equation (1.5). On one hand, only 1 equation per point is needed to complete the estimation process. On the other hand, the required input data are the flow vectors *projected* along the $\boldsymbol{d(x, \Sigma)}$ directions. If, for each $\boldsymbol{\Sigma}$, instead of using the flow all over the image, we select only the image sites $\boldsymbol{\Omega(\Sigma)}$, where the normal flow and $\boldsymbol{d(x, \Sigma)}$ are collinear, then the aperture problem is no longer a limitation. The cost function can be defined as follows:

$$\boldsymbol{E(\Sigma, \omega)} = \sum_{\boldsymbol{x} \in \boldsymbol{\Omega(\Sigma)}} \left[v_n(\boldsymbol{x}) - \boldsymbol{n^t(x)(B(x)\omega)} \right]^2, \;\; \boldsymbol{\Omega(\Sigma)} = \{ \forall \boldsymbol{x} : \boldsymbol{n(x)} \cdot \boldsymbol{(x - \Sigma)} = 0 \} \tag{1.6}$$

Figure 1.2a exemplifies a set $\boldsymbol{\Omega(\Sigma)}$ for a given $\boldsymbol{\Sigma}$. It corresponds to the set of normal flow vectors that are perpendicular to a line joining the image point and the FOE. The minimization problem consists of searching the minimum of $\boldsymbol{E(\Sigma, \omega)}$ for a large array of possible $\boldsymbol{\Sigma}$ values.

1.1.3 Defining a set of search subspaces

We have seen that for a set of pixels $\boldsymbol{x} \in \boldsymbol{\Omega(\Sigma)}$, the normal flow vectors that are perpendicular to a line joining $\boldsymbol{x}$ and the FOE, are solely dependent on the camera rotation. Therefore, these vectors can be used to determine $\boldsymbol{\omega}$, according to equation (1.1). If $\boldsymbol{\Sigma}$ is the correct location of the FOE, then the residuals should be small.



(a) (b)

Figure 1.2: (a) $\Omega(\Sigma)$ for a given FOE localization. (b) Example of a search subspace.

This principle can be used to search the FOE as illustrated in Figure 1.2b: For a given location $\mathbf{\Sigma_1}$, we can define a subspace in the image (in this case, a curve), where the normal flow vectors are perpendicular to the line going through the curve points and $\mathbf{\Sigma_1}$. We define $\mathbf{\Omega'(\Sigma_1)}$ as the subset of image sites where these vectors are located. Then we can estimate the likelihood of $\mathbf{\Sigma_1}$ being the FOE and the search process can continue by using other locations $\mathbf{\Sigma_i}$ and other curves, defining the corresponding subsets $\mathbf{\Omega'(\Sigma_i)}$ (see Figure 1.2b)[2]. Hence, we can define a generic framework, described by various possible search subspaces.

In previous work [54, 55, 56, 53], we have developed an estimation approach based on these ideas, where a set of low complexity estimation algorithms are applied on several subspaces with particular geometric properties. Rather than considering the whole set of image flow data, we used only the images sites that convey relevant information about the observer motion. Then we split the search domain in several geometric figures and estimated sets of motion parameters for each one of them.

In this work, we extend this approach by formulating the estimation problem on a different space — the lines topological space — that allows for numerous line subspaces to contribute simultaneously for the estimation, improving robustness and reliability. This methodology allows the use of nearly all available normal flow vectors and permits also a better comprehension

[2]Some subspaces are not conclusive about the exact location of a given $\mathbf{\Sigma}$: this is the case of a straight line containing $\mathbf{\Sigma}$. However, finding such line can be useful to reduce the search space of the subsequent algorithms.

about the connection between the global normal flow field and the camera motion.

1.2 Definition of the Lines Topological Space

In this section, we propose to define the topological space — the $\mathcal{L}$-space — where the egomotion estimation algorithm will be designed.

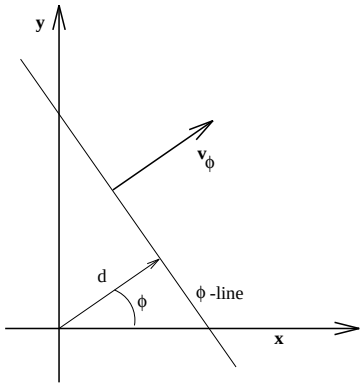


Figure 1.3: The flow vector v_ϕ lies on a line (the ϕ-line), defined by a vector with an orientation ϕ and a length d.

Consider a generic image straight line — ϕ-line — defined by a vector with orientation $\phi \in]-\pi/2, \pi/2]$ and length d (Figure 1.3). These two parameters are sufficient to represent all lines of the 2D-space. Suppose that there is a normal flow vector $\boldsymbol{v_\phi(x,\ y)}$ located on that line, being simultaneously perpendicular to it. An important property is that the normal flow vector $\boldsymbol{v_\phi}$ depends on the camera motion and ϕ-line parameters, as follows:

$$v_\phi = \underbrace{\rho(d - (\sigma\cos\phi + \eta\sin\phi))}_{\textit{translational component}} + \underbrace{\left(\frac{d}{f}(\omega_1\cos\phi + \omega_2\sin\phi) + \omega_3\right) r + \left(\frac{d^2}{f} + f\right)(\omega_1\sin\phi - \omega_2\cos\phi)}_{\textit{rotational component}}, \tag{1.7}$$

where $r = -x\sin\phi + y\cos\phi$, and $v_\phi = \boldsymbol{v(x)} \cdot (\cos\phi, \sin\phi)$[3].

Looking at the Equation (1.7), we notice that the rotational component of the normal flow vectors $\boldsymbol{v_\phi}$ is affine in r with the constant term defining a linear combination of ω_1 and ω_2.

[3]Notice that $v_\phi = \pm v_n$, where v_n defines the usual normal flow. The two values have opposite sign, whenever the image gradient angle corresponds to the suplementar angle of ϕ.

Besides the translational component vanishes when the ϕ-line goes through the FOE defined by $(\sigma,\ \eta)$.

In [53], we presented egomotion estimators based on search subspaces, minimizing the residual due to the translational component (usually non-linear). Using the same principle, we propose now a more general and global method to detect the motion parameters.

As each normal flow vector defines a unique ϕ-line, we work on the **lines topological space $\mathcal{L} \equiv \{d, \phi\}$** spanning the universe of all possible ϕ-lines. The $\mathcal{L}$-space is bounded and can be discretized according to the image grid. Similarly, each line $\{d, \phi\}$ accumulates contributions of a finite set of normal flow vectors $\boldsymbol{v_\phi}$. Additionally, each vector is linked to a coordinate, r, measured along the associated ϕ-line[4]. We propose an egomotion estimator based on the $\mathcal{L}$-space, thus allowing the use of nearly all the available image data.

1.3 Robust Estimation

We present a low complexity estimator based on the simple regression problem, by applying sequentially bidimensional estimators, defined on the $\mathcal{L}$-space.

Very often in computer vision there is the need to use estimation algorithms to recover several parameters. Most researchers apply least square (LS) techniques over the residual of some cost function. However, LS methods are optimal only when the observation error is zero-mean with normal distribution. This is rarely the case in many problems, and a few data outliers can ruin the quality of the estimates.

We will solve a simple regression problem on the normal flow observations $\boldsymbol{v_\phi}$ collected along a ϕ-line $\{d_i, \phi_j\}$. The flow on the ϕ-line depends on the desired motion parameters, according to the general equation (1.7) that can be written as follows:

$$v_{\phi_{ij}} = a_{ij} r + b_{ij} + e \tag{1.8}$$

where a_{ij} and b_{ij} are the affine parameters of the rotational component of the vectors $v_{\phi_{ij}}$, and e is an associated error (including the translational component and noise)[5].

[4] The point $r = 0$ corresponds to the nearest line point to the origin of the image plane.

[5] Notice that e may have a non-zero mean and/or a non-gaussian distribution.

The problem is to estimate both a_{ij} and b_{ij}. Various estimators have been proposed to detect and remove outliers (see [48] for an extended introduction to robust estimation). We adopted the Least Median of Squares (LMS) estimator[6]:

$$\min_{\hat{a}_{ij}\hat{b}_{ij}} \text{med}_k R_k^2 \tag{1.9}$$

where the residual $R = v_{\phi_{ij}} - (a_{ij}r + b_{ij})$ and k indexes the observations in line $\{d_i, \phi_j\}$. In this bidimensional regression problem, the LMS estimation can be solved by a non-iterative algorithm, as presented in [48].

When determining an LMS estimate of $\hat{a}_{ij}$ and $\hat{b}_{ij}$, the residuals, R_k, are compared to an estimate, ϵ_{ij}, of the error scale. The value of ϵ_{ij} must depend only on the "good" data, and can be computed robustly with the residuals R_k: $\epsilon_{ij} = \sqrt{\text{med}_k R_k^2}$. An observation is considered an outlier if the corresponding residual, R_k, is larger than ϵ_{ij}. The value of ϵ_{ij} is related to the robust standard deviation of the residuals, and is larger either when (i) the translational component is dominant and strongly non-linear or (ii) the observations are corrupted by high variance noise.

The translation influence on ϵ_{ij} vanishes when the ϕ-line intersects the FOE, because the non-linear behavior of translational component is minimized. In some cases, the translational component may depend linearly in r. Then, although ϵ_{ij} can be small, the estimates $\hat{a}_{ij}$, $\hat{b}_{ij}$ will depend on the translational motion[7]. This undesirable effect can be addressed later when the rotation parameters are estimated.

The first step of the algorithm eliminates candidate ϕ-lines with large ϵ_{ij}. We use a simple criterium to select the set S of the valid ϕ-lines:

$$S = \{\forall \{d_i, \phi_j\} : \epsilon_{ij} < \epsilon^*\}, \qquad \epsilon^* = \text{med}_{ij}\epsilon_{ij}.$$

To recover the motion parameters, we use uniquely the values $\hat{b}_{ij}$ of ϕ-lines $\in S$. According

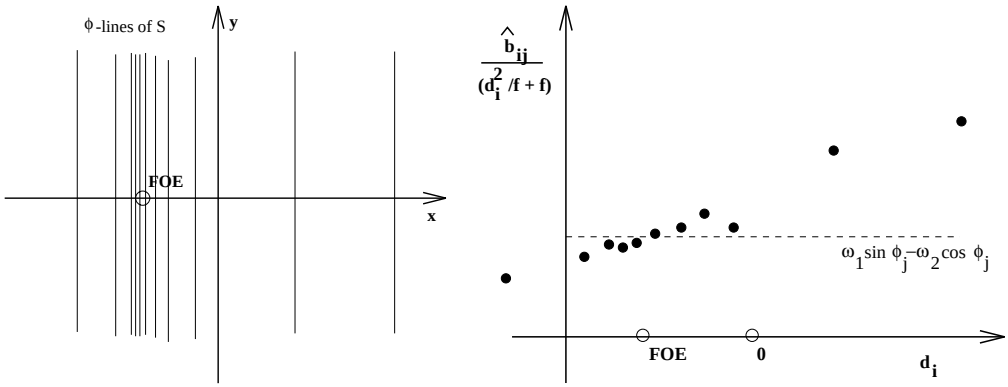
[6]Remind that the simple regression is one of central problems in statistics, and, the comparison among alternative approaches to solve it, is beyond the scope of this work.

[7]e.g the estimated value $\hat{b}_{ij}$, rather than being a pure linear combination of two rotational parameters, will be contaminated by structure and translation.

to equation (1.7), we have

$$\hat{b}_{ij} = \left(\frac{d_{ij}^2}{f} + f\right)(\omega_1 \sin\phi_{ij} - \omega_2 \cos\phi_{ij}) + \delta_{ij}, \tag{1.10}$$

where δ_{ij} is the robust mean of the error e of equation (1.8).



(a) (b)

Figure 1.4: (a) Family of ϕ-lines included in S with the same orientation. (b) The corresponding evolution of the b_{ij} estimates, scaled by $\left(\frac{d_{ij}^2}{f} + f\right)$.

In Figure 1.4a, we consider a subset of S with the same orientation $\phi = 0$. Due to the ϵ^*-selection applied before, the resulting ϕ-lines are concentrated in a neighborhood of the FOE. Figure 1.4b illustrates the evolution of $\hat{b}_{ij}$ for different d_{ij} and fixed ϕ_{ij}. When the ϕ-lines are close to the FOE, δ_{ij} approaches zero; otherwise, δ_{ij} spreads around some non-zero value. Figure 1.4b also shows that δ_{ij} has opposite signs on the left and on the right sides of the line crossing the FOE. This particular result can be easily proved algebraically by using equation (1.7) and is due to the known translational component behavior around the FOE, assuming that depth is always positive. This property can be used to identify lines going through the FOE, by a geometric search algorithm[8]. In this approach, we incorporate naturally these features into our estimation approach, assuming that the scene is sufficiently rich in terms of texture and depth variability.

[8]This geometric approach may be specially suitable when there are a few observations in the ϕ-lines that go through the FOE (or close), due to poorly textured images.

Next, using the estimated information from ϕ-lines $\in S$, we minimize two separate cost functions which depend on the translation and the rotation, as follows:

$$\min_{\hat{\sigma}\hat{\eta}} \text{med}_{\phi\in S}(d_{ij} - (\sigma\cos\phi_{ij} + \eta\sin\phi_{ij}))^2 \tag{1.11}$$

$$\min_{\hat{\omega_1}\hat{\omega_2}} \text{med}_{\phi\in S}\left(\hat{b}_{ij} - \left(\frac{d_{ij}^2}{f} + f\right)(\omega_1\sin\phi_{ij} - \omega_2\cos\phi_{ij})\right)^2 \tag{1.12}$$

The first estimator determines the FOE, $(\hat{\sigma}, \hat{\eta})$, by robustly minimizing the translation component of $\boldsymbol{v_{\phi_{ij}}}$ vectors (or, equivalently, δ_{ij} term). The second estimator recovers the rotation by using the estimates $\hat{b}_{ij}$.

When estimating the motion parameters through equations (1.11, 1.12) we also compute the corresponding residuals. By selecting the ϕ-lines with smallest residuals both for the first and the second estimators, we define a subset of S under the constraint that (i) the ϕ-lines are close to a good candidate of the FOE, and (ii) the corresponding normal flow vectors, $\boldsymbol{v_{\phi_{ij}}}$, are linearly dependent on some rotation vector.

Let R_{ij}^T and R_{ij}^ω be the residuals computed with $(\hat{\sigma}, \hat{\eta})$ and $(\hat{\omega_1}, \hat{\omega_2})$, respectively. Defining $\epsilon^T = \sqrt{\text{med}_{\phi\in S} R_{ij}^{T\,2}}$ and $\epsilon^\omega = \sqrt{\text{med}_{\phi\in S} R_{ij}^{\omega\,2}}$ we have a new subset S' given by:

$$S' = \left\{\forall\phi_{ij} \in S : \left(\frac{R_{ij}^T}{\epsilon^T}\right)^2 + \left(\frac{R_{ij}^\omega}{\epsilon^\omega}\right)^2 < 1\right\} \tag{1.13}$$

To improve the quality of $(\hat{\sigma}, \hat{\eta}, \hat{\omega_1}, \hat{\omega_2})$, we apply again the estimators (1.11) and (1.12) to a selection of vectors, $\boldsymbol{v_{\phi_{ij}}}$, lying on the ϕ-lines $\in S'$. The remaining parameter ω_3 can be easily computed if the other motion parameters are already known.

To conclude, the following special cases deserve additional attention:

Pure rotation: In this case, all normal flow vectors depend linearly on the rotation parameters. Thus, the ϕ-lines of S (after ϵ^*-selection) are widely spread (more or less uniformly). Hence, estimating the FOE using estimator (1.11) leads to a large standard deviation described by ϵ^T. To detect the pure-rotation case, one can establish two decision thresholds for ϵ^T and ϵ^ω. If ϵ^T value is above some threshold and ϵ^ω is below another one, then we decide that the camera motion is a pure rotation.

The FOE at the infinity: In the approach proposed before, the FOE location can lie outside the image domain. However, when the translation velocity is almost parallel to the image plane, the FOE goes to infinity. In this case, one can determine the direction of the vector (σ, η) rather than the exact FOE location. According to the $\mathcal{L}$-space, if the FOE is at infinity, then the ϕ-lines that intersect the FOE are parallel and the direction ϕ is orthogonal to (σ, η).

Considering the FOE estimator provided by the Equation (1.11), define the regression data matrix, $\mathcal{D}$, as the collection of the vectors $[\cos\phi_{ij} \quad \sin\phi_{ij}]$, for all ϕ-lines $\in S$. If the FOE is unbounded, then $\mathcal{D}$ is rank deficient, due to an existing dominant direction on the ϕ-lines. Therefore, before applying directly the estimator (1.11), we compute the two singular values of $\mathcal{D}$. If one of them is very small relative to the other, we consider that the FOE is approximately at the infinity and its direction given by the dominant singular vector. Remember that this result must be validated with a low variance of the rotational estimator (1.12), in order to guarantee robustness.

1.4 Experiments

We have conducted a series of experiments of different difficulty, using real image data. Five frames are used to compute the normal flow field. The $\mathcal{L}$-space is a discrete domain, bounded by the image dimensions, discretized every 0.2 radians along the ϕ-axis and every 5 pixels along the d-axis. We used the image top-left corner as the $\mathcal{L}$-space origin, as it can be defined arbitrarily on the image plane[9].

In the figures below, we represent each ϕ_{ij}-line by a single point, $(d_{ij}\cos\phi_{ij},\ d_{ij}\sin\phi_{ij})$. This simple representation provides a clear illustration of the estimation process.

Sequence 1, pure translation: In this sequence, the camera undergoes a pure translational motion in a real cluttered scene. Figure 1.5 shows the two selection steps of ϕ-lines. Since we minimize the cost function in (1.11), the corresponding dots given by $(d_{ij}\cos\phi_{ij},\ d_{ij}\sin\phi_{ij})$ tend to a circular shape having the $\mathcal{L}$-space origin and the FOE as diametrical opposite points.

[9]Notice that in the numerical computation of the rotational estimators we must take into account the actual origin (the intersection of the optical axis in the image plane).

Table 1.1 shows the rotation and translation estimates, in image coordinates. As expected,

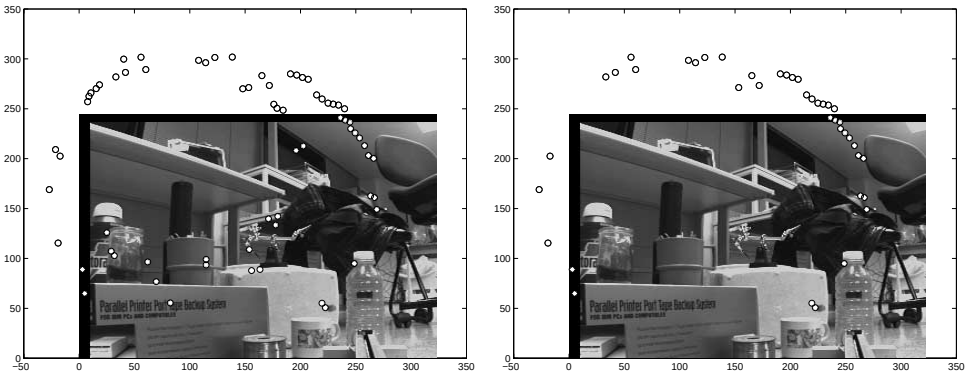


Figure 1.5: Sequence 1: Left - set S. Right - resulting set S'. The $\mathcal{L}$-space origin is located at the image top-left corner.

Motion	**Estimates**	**True values**
$(\omega_1,\ \omega_2)$	(0.00092, 0.00051) rad/fr	(0.0, 0.0) rad/fr
$(\sigma,\ \eta)$	(226.6, 80.5) pix	(245, 65) pix

Table 1.1: Estimation results associated to the first sequence.

the estimated rotation is small. The angular error between the real translation vector and its estimate is also small (< 5 degrees).

Sequence 2, pure translation: This experiment uses the "coke"-sequence (Figure 1.6). Few information is available about the camera calibration, like the focal length and the principal point. An approximated location of the FOE was obtained visually for comparison purposes. This trial shows that the algorithm can cope well with the case of pure translation. Table 1.2 shows that the rotation estimates are very small and the estimated FOE is close

Motion	**Estimates**	**Empirical values**
$(\omega_1,\ \omega_2)$	$(-0.00001,\ -0.00007)$ rad/fr	(0.0, 0.0) rad/fr
$(\sigma,\ \eta)$	(124.4, 160.0) pix	(130, 175) pix

Table 1.2: Estimation results associated to the second sequence.

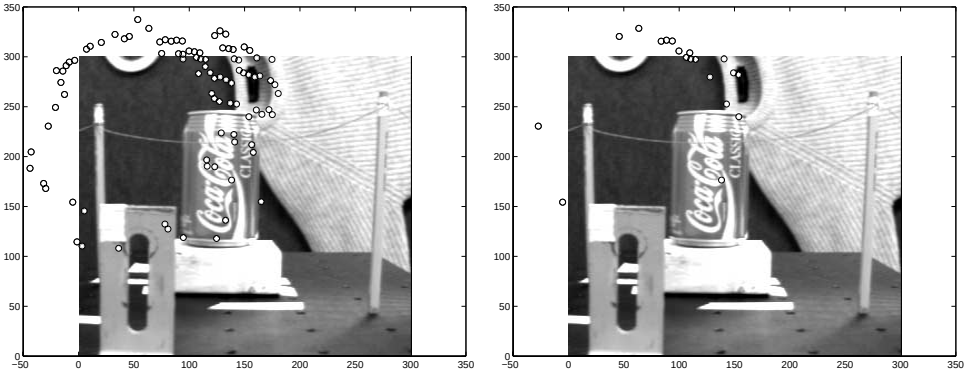


Figure 1.6: Sequence 2: Left - set S. Right - resulting set S'. The $\mathcal{L}$-space origin is located at the image top-left corner.

to the corresponding empirical value.

Sequence 3, rotation with translation: In this sequence, the camera undergoes a translational motion and rotates around the Y axis, with $\omega_2 = -0.005$ rad/frame. The rotational component of the normal flow field is dominant in relation to the translational one. Therefore, the ϕ-lines $\in S$ (Figure 1.7 on the left) are more spread than usual, and the rotation parameters estimates are very accurate (Table 1.3). After completing the estimation process,

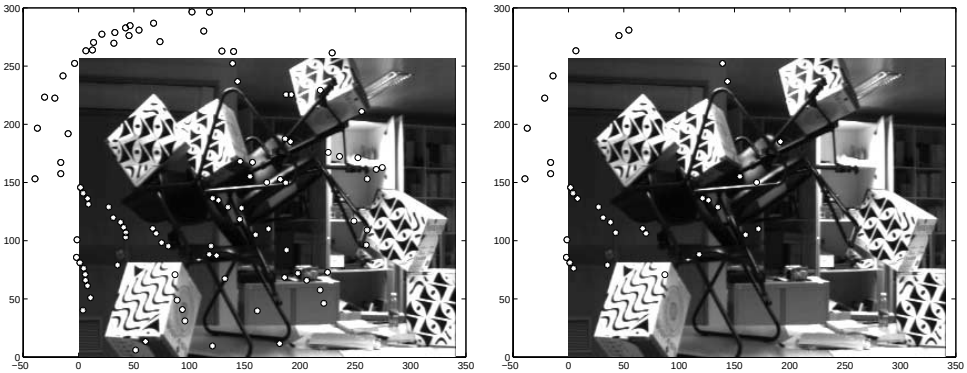


Figure 1.7: Sequence 3: Left - set S. Right - resulting set S'. The $\mathcal{L}$-space origin is located at the image top-left corner.

the FOE estimate is close to the real value.

Sequence 4, FOE at the infinity: The last experiment with the "trees"-sequence,

Motion	**Estimates**	**Real values**
$(\boldsymbol{\omega_1},\ \boldsymbol{\omega_2})$	$(0.0000,\ -0.0051)$ rad/fr	$(0.0,\ -0.005)$ rad/fr
$(\boldsymbol{\sigma},\ \boldsymbol{\eta})$	$(140,\ 120)$ pix	$(146,\ 130)$ pix

Table 1.3: Estimation results associated to the third sequence.

consists on a pure translation parallel to the image plane. For this special case, the ϕ-line dots

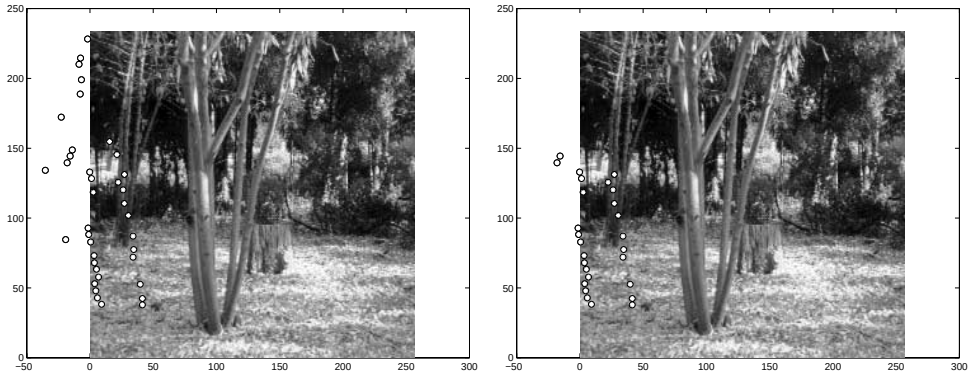


Figure 1.8: Sequence 4: Left - set S. Right - resulting set S'. The $\mathcal{L}$-space origin is located at the image top-left corner.

(Figure 1.8) form a straight line (degenerate case) perpendicular to FOE direction. Therefore, the Figure 1.8 shows that the FOE direction is close to the horizontal, as expected. The estimated rotation values were negligible ($\approx 10^{-4}$).

1.5 Conclusion

We addressed the problem of egomotion estimation for a monocular observer with arbitrary translation and rotation. Our approach is uniquely based on spatio-temporal image derivatives. The egomotion process is based on the subdivision of the problem in various bidimensional regression problems. An LMS estimator of low complexity is applied in order to estimate the motion parameters and improve the estimation robustness.

Robustness is achieved in two ways. First, we use exclusively the normal flow field together with robust statistics methods. Secondly, the problem is defined in a topological space — the

$\mathcal{L}$-space — such that the estimates are supported by a large set of image data, as opposed to some previous approaches. However, there are points where it is difficult to compute the normal flow, namely in depth discontinuities and occlusion points. Next chapter will address this issue by incorporating the occlusion information in the estimation of the FOE location.

Chapter 2

Dealing with Occlusions

2.1 A Definition for Occlusion Points

To perceive image motion, the human vision uses surface and edge representations and performs derivatives on the images captured along time. However, a non-correspondence, such as an occlusion, can play an important role in motion and depth interpretation. Anderson and Nakayama [3] have shown that occlusion is one of the most powerful cues to perceive depth and motion, and influence the earliest visual stages of stereo matching. Figure 2.1 illustrates this idea with samples of a well-known image sequence, where the occlusions (together with the image motion) give a clear perception of depth and camera motion.

Figure 2.1: Three sequential samples of the Flower Garden Sequence. The camera is going to the right. Occlusions and image motion give both a depth and camera motion perception.

In this chapter, we propose to recover camera motion information based uniquely on occlu-

sions, by observing two specially useful properties: occlusions are independent of the camera rotation, and reveal direct information about the camera translation.

We assume a monocular observer, undergoing general rotational and translational motion in a static environment. We present a formal model for occlusion points and develop a method suitable for occlusion detection. Through the classification and analysis of the detected occlusion points, we show how to retrieve information about the camera translation (FOE).

First of all, we need a formal definition of occlusion points and a methodology for detecting occlusions in image sequences.

Geometrically, an occlusion is caused by an occluding surface moving in front of an occluded surface. Additionally, if the observer is moving in a static environment, occlusions correspond to discontinuities both in the perceived motion and depth. However, unless we impose prior models to the image motion field, 3D structure or to global image features, we can only decide about the existence of a local occlusion in two consecutive frames, if the photometric properties change significantly in a local neighborhood. Thus, we can associate an occlusion point to a photometric value that perceptually "appears" or "disappears" between two consecutive frames, classified respectively as *emergent* or *submergent* occlusion point.

Hence, an occlusion has to be studied both as a geometric and photometric phenomenon. We propose to define occlusion points through a sufficient condition based on a local photometric dissimilarity over time, with precise geometric properties. This sufficient condition can be characterized rigorously for the continuous case.

First of all, we denote the spatial and temporal coordinates of an image sequence (see Figure 2.2(a)(b)) by x and t, represented by a vector $k = \binom{x}{t}$, where x is the spatial coordinate of an arbitrary scanline of the image. Additionally let us define the following auxiliary sets of space-time coordinates, representing two halves of a circle in x-t-space:

$$\begin{aligned} K^+ &= \{k : \|k\| = 1 \wedge t > 0\} \\ K^- &= \{k : \|k\| = 1 \wedge t < 0\} \end{aligned}$$

As we have already discussed, an occlusion point corresponds to photometric values that *appear* and *disappear* between frames. Let $f(k)$ denote such photometric measure of the image

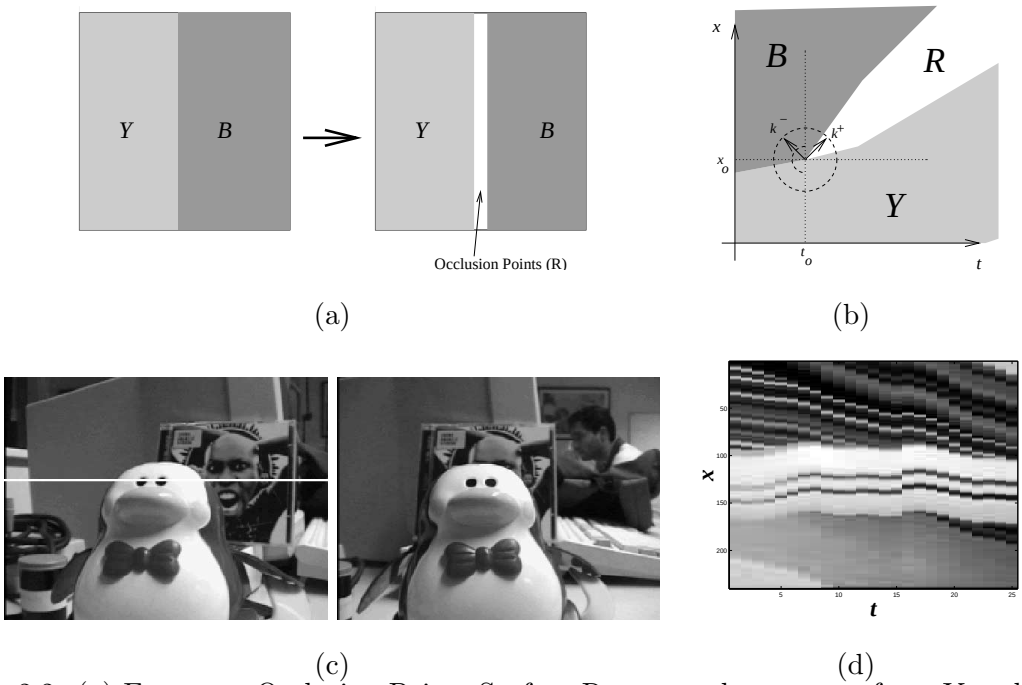


Figure 2.2: (a) Emergent Occlusion Point: Surface R appears between surfaces Y and B; (b) Image motion over time t, for a horizontal line parameterized by x. (c) Example of a real sequence with translation and rotation; (d) Image motion over time, for the horizontal scanline selected in the left pictures.

in k (for example the brightness value). Based on this notation, we present the following sufficient condition for the existence of an occlusion.

The point $k_0 = (x_0,\ t_0)$ is an emergent occlusion if

$$\exists k^+ \in K^+ : \forall k^- \in K^-, \qquad \lim_{\gamma \to 0^+} f(k_0 + \gamma k^+) \neq \lim_{\gamma \to 0^+} f(k_0 + \gamma k^-). \tag{2.1}$$

Similarly, k_0 is a submergent occlusion point if

$$\exists k^- \in K^- : \forall k^+ \in K^+, \qquad \lim_{\gamma \to 0^+} f(k_0 + \gamma k^+) \neq \lim_{\gamma \to 0^+} f(k_0 + \gamma k^-). \tag{2.2}$$

Figure 2.2(a) illustrates the meaning of the sufficient condition proposed here, with a simple example of an emergent occlusion point. The surface R appears between surfaces Y and B. Considering a given horizontal scanline, the surface R emerges at the point $(x_0,\ t_0)$, as shown in Figure 2.2(b). According to the condition defined before, this point is an emergent occlusion point, because there is a vector k^+ (with positive t) associated to a region (R), which photometric value does not exist on the half-plane $t < t_0$, in a neighborhood of the point (x_0, t_0). Figure 2.2(c) shows an example of a real sequence, where the submergent and emergent occlusion surfaces are visible on the temporal evolution of a given horizontal scanline (Figure 2.2(d)).

This sufficient condition is useful as formal model for a generic occlusion definition. However, in order to guarantee its applicability in the discrete case, we have to define more carefully the associated inequality relation. Consequently, we have developed a dissimilarity criterion inspired on a function developed by Tomasi and Manduchi [62] that was originally designed to smooth a single image, preserving the photometric discontinuities. We have changed this function in order to measure the similarity between two consecutive frames.

Suppose that pixel x_0 in frame t_0 is characterized by a photometric value $f(x_0, t_0)$. The problem to solve is to verify the existence of a similar photometric value (preferably through a perceptually meaningful way) within a given region in frame t_1. We start by defining a function $S(x_0, t_0, x_1, t_1)$ that compares the similarity of $f(x_0, t_0)$ to $f(x_1, t_1)$:

$$S(x_0, t_0, x_1, t_1) = \exp\left(-\frac{\|f(x_1, t_1) - f(x_0, t_0)\|^2}{2\sigma^2}\right)$$

where σ^2 corresponds to the variance of the associated gaussian filter. Next, we apply this similarity function to all pixels x_1 in a neighborhood $V(x_0)$ around pixel x_0 in frame t_1:

$$f_{t_1}(x_0, t_0) = \frac{\sum_{x_1 \in V(x_0)} f(x_1, t_1) \cdot S(x_0, t_0, x_1, t_1)}{\sum_{x_1 \in V(x_0)} S(x_0, t_0, x_1, t_1)} \tag{2.3}$$

When $\sigma \to 0$, this function yields (except for some pathological configurations) the photo-

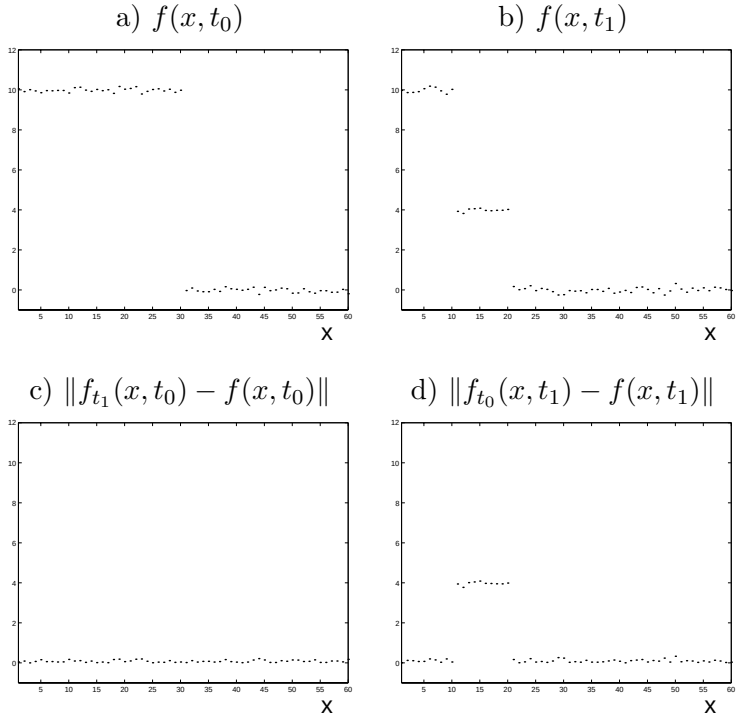


Figure 2.3: (a) Brightness function of a scanline in frame t_0; (b) Brightness function of the same scanline in frame t_1, after applying a shift in x, adding some gaussian noise in the brightness axis and introducing new brightness information (simulating an occlusion). (c-d) Computation of $\|f_{t_1}(x, t_0) - f(x, t_0)\|$ and $\|f_{t_0}(x, t_1) - f(x, t_1)\|$ for all x in frames t_0 and t_1 respectively. We assumed that $V(x)$ is the set of points within the interval $[x - 20 \;\; x + 20]$ and $\sigma = 1$.

metric value $f(x^*, t_1)$, $x^* \in V(x_0)$, that is closest to $f(x_0, t_0)$. There exists dissimilarity (and therefore (x_0, t_0) is an occlusion point), if $\|f_{t_1}(x_0, t_0) - f(x_0, t_0)\|$ is above a certain acceptable threshold T. This threshold and the variance σ^2, defined in function S, depend on perceptual criteria based on the photometric range of the images.

Figure 2.3 illustrates the performance of this dissimilarity criterion in detecting the presence of an occlusion between two simple functions (Figures 2.3a-b) which differ additionally by a translation. By subtracting directly both functions, we cannot detect easily the occlusion, because the difference is affected indistinctly by both occlusions and translation. On the contrary, by using the proposed approach, we can detect exactly the dissimilar region which corresponds to the occlusion region. Hence, $\|f_{t_1}(x, t_0) - f(x, t_0)\|$ (Figure 2.3c) is almost zero for all x, meaning that all photometric information in frame t_0 is present in frame t_1. On the other hand, the computation of $\|f_{t_0}(x, t_1) - f(x, t_1)\|$ (Figure 2.3d) shows that there is a region in frame t_1 which is dissimilar from the information present in frame t_0, revealing then an occlusion region. Furthermore, if the frame t_1 appears temporally after the frame t_0 ($t_1 > t_0$) then we can conclude that the detected occlusion is emergent, otherwise ($t_1 < t_0$) the occlusion is submergent. Notice that the occlusion detection only becomes effective if an adequate threshold is applied (in the example, $T = 1$ is an acceptable value).

2.2 Egomotion Perception from Occlusions

In the previous section we have proposed a formal definition of *emergent* and *submergent* occlusion points, together with a photometric criterion for their detection.

In this section we analyze how these occlusion points can be used to retrieve information regarding the observer's 3D motion (egomotion). We consider a monocular observer under a perspective camera model, moving with arbitrary translation and rotation, in a scene with static objects.

Associated to the egomotion estimation problem, one observes one of most important properties of the occlusions:

Property 1 — *The camera rotation does not produce occlusion points. Consequently, the occlusion points are uniquely due to the camera translation.*

Property 1 states a well known fact. Only the translational part of the image motion depends on the scene depth. As occlusions are produced by depth discontinuities, only the translational component of the camera motion will give rise to occlusion effects.

Notice that one of the most important difficulties when estimating the camera motion using optic flow consists in decoupling the effects of translation from those of rotation [56]. This problem is not verified in the occlusions.

The translation of an observer is usually identified by the projection of the linear velocity on the image plane, the Focus of Expansion (FOE). In order to explore the relation between the FOE and the behavior of the occlusions, let us consider a single scanline camera to simplify the problem.

Assume that x parameterizes the scanline defined before and $v(x)$ describes the image velocity along that line. The flow $v(x)$ can be represented as a function of the camera motion parameters [30], as follows:

$$v(x) = \frac{W}{Z(x)}(x - x_{\text{FOE}}) + r(x), \tag{2.4}$$

where $r(x)$ is the motion component due to the camera rotation, x_{FOE} is the FOE projection on the scanline considered, W is the camera velocity component along the optical axis (let us assume that it is positive), and finally $Z(x)$ is the depth of the corresponding 3D point.

Assuming that x_0 is an occlusion point, one observes a discontinuity in $v(x_0)$ and a discontinuity in $Z(x_0)$ — this means $v(x_0^-) \neq v(x_0^+)$ and $Z(x_0^-) \neq Z(x_0^+)$ respectively. However these discontinuities have a different physical meaning as described by the following two properties:

Property 2 : Emergent/Submergent Occlusion

When x_0 is an emergent occlusion point, $v(x_0^-) < v(x_0^+)$; when x_0 is a submergent occlusion point, $v(x_0^-) > v(x_0^+)$.

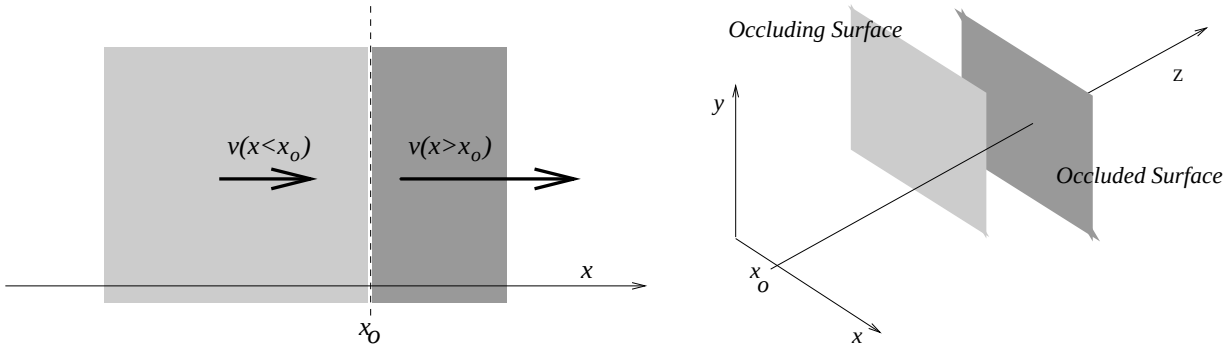
Property 3 : Left/Right Occlusion

If $Z(x_0^-) > Z(x_0^+)$ then the occluding surface is on the right[1] *of the occluded surface (because the occluding surface is naturally nearer than the occluded one). If $Z(x_0^-) < Z(x_0^+)$, then the occluding surface is on the left of the occluded surface.*

Figure 2.4 illustrates these properties with a simple example, where the occluding surface is on the left of the occluded surface ($Z(x_0^-) < Z(x_0^+)$), and x_0 is an emergent occlusion point ($v(x_0^-) < v(x_0^+)$).

[1] stipulating x_0^+ on the right of x_0^-.

In summary, when an occlusion is observed, it can be classified within four classes which consist of the combination of getting a right or left occluding surface and an emergent or submergent occlusion point.



(a) (b)

Figure 2.4: Example of an occlusion situation: (a) x_0 is an emergent occlusion point; (b) the projected occluding surface is on the left of x_0.

In the following property we show the direct relation between the camera translation (measured on the scaline by x_{FOE}) and the occlusion classification presented before.

Property 4 : Fundamental Relation between Camera Translation and Occlusions

- *An occlusion point x_0 is on the right of x_{FOE} if either (1) x_0 is emergent and the occluding surface is on the left side, or (2) x_0 is submergent and the occluding surface is on the right side.*

- *The occlusion point x_0 is on the left of x_{FOE} if either (1) x_0 is emergent and the occluding surface is on the right side, or (2) x_0 is submergent and the occluding surface is on the left side.*

Proof: To prove the property described above, let define a notation for the classification of occlusion points, based on functions $\mathcal{L}(x_0)$ and $\mathcal{E}(x_0)$ described as follows:

- $\mathcal{L}(x_0) = 1$ or $\mathcal{L}(x_0) = -1$ if the occlusion point x_0 has respectively a left or a right occluding surface;

- $\mathcal{E}(x_0) = 1$ or $\mathcal{E}(x_0) = -1$ if x_0 is respectively an emergent or a submergent occlusion point;
- $\mathcal{L}(x_0)$ and $\mathcal{E}(x_0)$ are zero if x_0 is not an occlusion point.

By using Properties (2, 3) and Equation (2.4), then we can determine whether the FOE (x_{FOE}) is located to the left or right of the occlusion point x_0:

$$\begin{aligned}
&\mathcal{L}(x_0) \cdot \mathcal{E}(x_0) = 1 \Rightarrow \\
&\Rightarrow \left(v(x_0^-) - v(x_0^+)\right)\left(Z(x_0^-)^{-1} - Z(x_0^+)^{-1}\right) < 0 \Rightarrow \\
&\Rightarrow x_0 < x_{\text{FOE}}
\end{aligned}$$

$$\begin{aligned}
&\mathcal{L}(x_0) \cdot \mathcal{E}(x_0) = -1 \Rightarrow \\
&\Rightarrow \left(v(x_0^-) - v(x_0^+)\right)\left(Z(x_0^-)^{-1} - Z(x_0^+)^{-1}\right) > 0 \Rightarrow \\
&\Rightarrow x_0 > x_{\text{FOE}}
\end{aligned}$$

QED

This result shows that classifying the occlusion point x_0 corresponds to detecting its location relatively to a projection of the FOE.

Moreover the occlusions are not affected by the camera rotation as described by Property 1. This is a huge advantage when compared to other approaches that use the optic flow to estimate the egomotion, where decoupling the rotation and translation components is a difficult problem.

However it remains the question about the algorithmic procedure to classify the occlusion point. In fact, the occlusion classification could be performed by the optic flow and depth description in a certain neighborhood. Assuming that both the optic flow and depth are unknown (and hardly computable), we propose to use exclusively the dissimilarity function developed before.

First of all, remind that the previous section describes a method to classify an occlusion as emergent or submergent. This alleviates the need to explicitly determine the local optic flow $v(x_0^-)$ and $v(x_0^+)$. Secondly, to determine whether we have a left or right occluding surface, we monitor the temporal photometric changes at the left and right side of the occlusion point.

The obvious advantage is that we no longer need to know $Z(x_0^-)$ or $Z(x_0^+)$ to reason about the nature of the occlusion.

The method we use seeks the image contour closest to the occlusion point, that preserves both photometric and geometric properties over time, thus belonging to the occluding surface. An alternative equivalent procedure consists of studying the evolution of points which do not preserve their photometric properties over time, thus belonging to the occluded surface. Notice that the complete occlusion classification can rely uniquely on the dissimilarity criterion presented before.

In order to detect automatically the location of the FOE projection along a unique scanline, we designed a function that integrates, along x (that parameterizes the scanline), the value of the dissimilarity relations from (2.3) taking into account $\mathcal{L}(x) \cdot \mathcal{E}(x)$. This function can be described for the discrete case as follows:

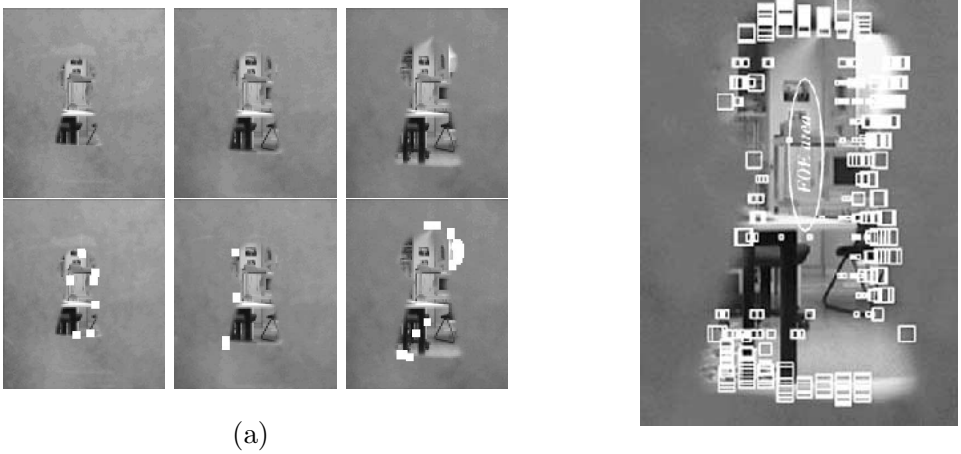
$$\begin{aligned} F(x,t) &= \sum_{\zeta=-\infty}^{x} \mathcal{L}(\zeta)\mathcal{E}(\zeta)d(\zeta,t) \\ d(\zeta,t) &= \|f_{t+1}(\zeta,t) - f(\zeta,t)\| + \|f_{t-1}(\zeta,t) - f(\zeta,t)\| \end{aligned}$$

where $t-1$ and $t+1$ correspond to the previous and the next frames.

This function decreases if x is on the right of the point where the FOE is projected on the scanline, and increases if x is on the left of the FOE projection. Thus the FOE projection is located at the absolute maximum of $F(x,t)$. By integrating the information over the image, the method becomes more robust to eventual false occlusion detections.

2.3 Experiments

In this Section, we apply the occlusion detection and classification process to four image sequences. The first sequence (the Lock Sequence, Figure 2.5(a)) shows the performance of the dissimilarity function in order to find the occlusion points. The Focus of Expansion is roughly at the center of the image and emergent occlusions are found on the boundaries of the lock hole, as expected. Considering an arbitrary line along the image, the relation between the occlusions and the FOE location can be observed: first the occlusions are emergent, second the occluded points are inside the lock hole (or the occluding surface is outside), thus completing



(a) (b)

(c)

(d)

Figure 2.5: (a) On the top, three sequential samples of the Lock Sequence and, on the bottom, associated emergent occlusion points (white squares). (b) Last frame of the Lock Sequence with previously detected emergent occlusion points superimposed on the image (more recent ones represented by larger squares). The ellipse in the figure illustrates qualitatively the expected region for the FOE location. (c) Samples of the Penguin Sequence, with associated emergent occlusion points. Each occlusion point found indicates the same FOE direction (arrow direction) given by its classification. (d) Same images with submergent occlusion points.

the occlusion classification and indicating that the FOE is somewhere in a restricted area at the center of the image (Figure 2.5(b)).

The second sequence (the Penguin Sequence, Figure 2.5(c)) was performed with static objects and a leftward moving camera with relevant rotation. In this experiment we show how the occlusion points are immune to rotational contamination of the image motion. In Figure 2.5(c) we present three samples of the sequence where the emergent occlusions appear mainly on the left side of the penguin whereas the submergent occlusions disappear on the right side (Figure 2.5(d)). This classification indicates that the FOE is on left of each occlusion point detected.

The third sequence (Figure 2.6-left) was performed with a static camera with the penguin moving to the right. In this experiment we see that occlusions can be used for the segmentation of a moving object. The function $F(x,\ t)$ was computed integrating the information included in the set of all horizontal scanlines (summing the contributions of all $F(x,\ t)$ over the vertical coordinate). Notice both the occlusion boundaries of the penguin and its velocity direction, given by the decreasing behavior of the function.

The last sequence (the Tree Sequence, Figure 2.6-right) consists in a leftward moving camera, with a large number of occlusions. Since the FOE is on the left of the image (at infinity), the function F has a decreasing behavior.

The photometric parameter used here was the brightness value, in a range of 0-255. In all experiments, we applied the same dissimilarity function with $\sigma = 5$ and the occlusion detection threshold $T = 5$. These values were chosen empirically according to the global distribution of the brightness along the sequences. In future, we plan to define automatically σ and T by using local properties of the image.

2.4 Conclusion

In this chapter we have studied the importance of the occlusions for motion detection. Based on a theoretical framework for the definition of occlusions in the continuous case, we developed a dissimilarity function for the discrete case, using local photometric and geometric properties of the image. Assuming a moving monocular camera, we show that the occlusion classification

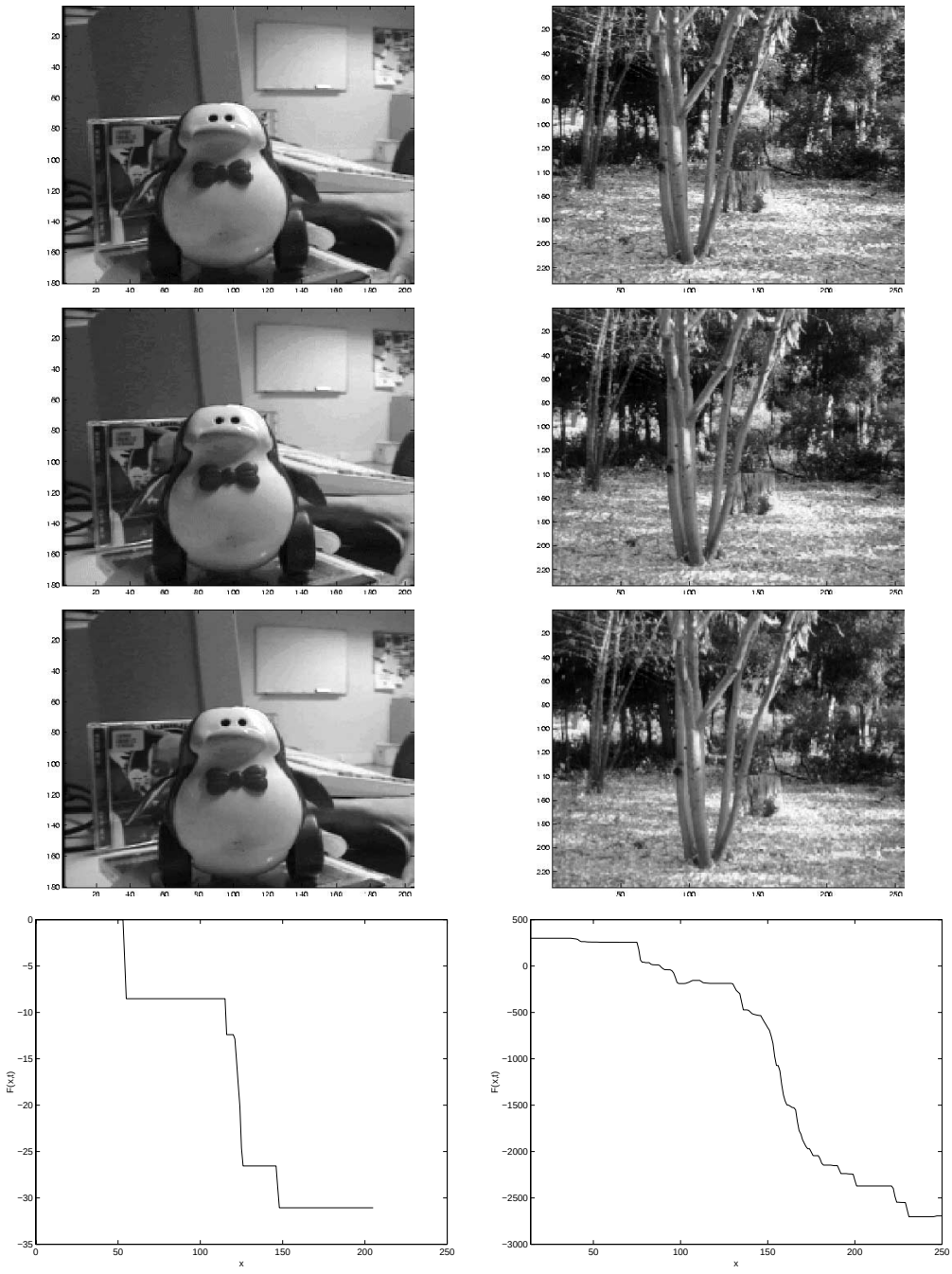


Figure 2.6: Left: Three sequential samples of the Penguin Sequence — on the bottom the function $F(x, y)$ along x (notice that the decreasing steps correspond to penguin occlusion boundaries). Right: Three sequential samples of the Tree Sequence. As expected, the function $F(x, y)$ (bottom) has a decreasing behavior whith a maximum at the beginning of the x axis.

is equivalent to the detection of a translational direction. Thus, we design a method to recover egomotion information, according to the following observations:

- Occlusions are extremely important cues for the egomotion perception.
- With a moving camera, only translation produces occlusion points. Therefore, the rotation does not influence the translational estimation.
- To detect the camera translation, no special models for motion or structure are needed.
- The camera translation can be detected even if its projection is outside the image field.
- The occlusion classification can be performed by using uniquely dissimilarity criteria (more robust than similarity criteria).

A number of experiments with real images have been performed, for various kinds of motion, that illustrate the capabilities of our approach.

Summary

We addressed the problem of egomotion estimation for a monocular observer moving under arbitrary translation and rotation, in an unknown environment. First, we propose an egomotion estimation method uniquely based on the spatio-temporal image derivatives, or the normal flow. Thus, we avoid computing the complete optical flow field, which is an ill-posed problem due to the aperture problem. We present an estimation paradigm which is based on geometric properties of the normal flow field, and consists in considering a family of search subspaces to estimate the egomotion parameters. In order to decrease the noise sensitivity, we define a particular topogical space — the $\mathcal{L}$-space — to allow the use of global data for the final estimates and the use of statistical tools, based on robust regression theory. Next we presented a formal model for occlusions points and develop a method suitable for occlusion detection. Through the classification and analysis of the detected occlusion points, we show how to retrieve information from occlusions about the camera translation.

We present and discuss a wide variety of experiments with synthetic and real images, for the various presented approaches.

Future research will address the following aspects: (a) development of more sophisticated regression algorithms to increase the egomotion estimation robustness and performance; (b) by considering an active observer, one can use further constraints in the egomotion estimation process, and the estimator proposed can be used in a closed loop navigation system; (c) study of pathological or degenerate cases [29] and the case of existing deficient variety of gradients and depth; and, finally, (d) integrate the estimation part based on occlusions with the general approach based on normal flow.

From this point on, we assume that the camera motion was estimated or, more generically,

the various camera views became weakly calibrated. Then one remains to estimate the dense 3-D information about the viewed scene. In fact, based on the flow observations it might be possible to generate the depth information, however this differential approach leads usually to meaningless local estimates because it is strongly affected by any local brightness noise. To cope with this inherent ill-posedness, the disparity measurements must be greatly improved, or, in other words, the dense correspondence problem must be explored.

Assuming calibrated cameras, the correspondence process is partially facilitated but still a very difficult problem. Next part will be completely dedicated to the dense matching for two or more cameras, where relevant questions about image representation, global optimization and the treatment of occluded points will be addressed.

Part II

Depth Reconstruction

Finding correspondences is one of central problems in stereo and motion analysis. To perceive depth, the human binocular vision performs dense matching between two or more images, captured over time. Besides, occlusions play a key role in motion and depth interpretation.

In this part of the thesis, we propose two approaches to generate dense disparity maps and new views of the same scene. The key aspect of these approaches consists in finding suitable representations to describe either the image data or visual constraints in a given optimization framework. Finding such adequate representations is a major contribution for solving difficult problems in stereo.

In the first chapter, we propose a new image representation called *Intrinsic Images* that can be used to solve correspondence problems within a natural and intuitive framework [59]. Intrinsic images combine photometric and geometric descriptors of a stereo image pair. The photometric descriptors are invariant to perspective image distortions, while the geometric descriptors can be used directly to compute disparities in a dense manner. The method is extended to cope with brightness changes and occlusions. It provides a coherent interpretation of occluded regions, based on data available from one image only. This new representation greatly simplifies the computation of dense disparity maps and the synthesis of novel views of a given scene.

The main limitation with this approach is the assumption of the *order* constraint, which can be violated in the presence of occlusions (although not always), thus leading to poor results in those regions. In those cases, only the use of more images can help generating a correct interpretation of the visual scene.

In the second chapter, we propose a methodology for solving the point correspondence problem for more than two views. We rely exclusively on physically valid constraints (i.e. no approximations), thus avoiding the use of the order constraint. First we explore and discuss geometric, uniqueness and visibility constraints for the stereo problem. For the example of a trinocular setup, we show how to represent these constraints in such a way that it allows us to formulate the correspondence problem as an integer

optimization problem, where occlusions are naturally represented. Extensive results are presented and discussed.

Chapter 3

Intrinsic Images for Stereo Matching

3.1 Stereo Matching

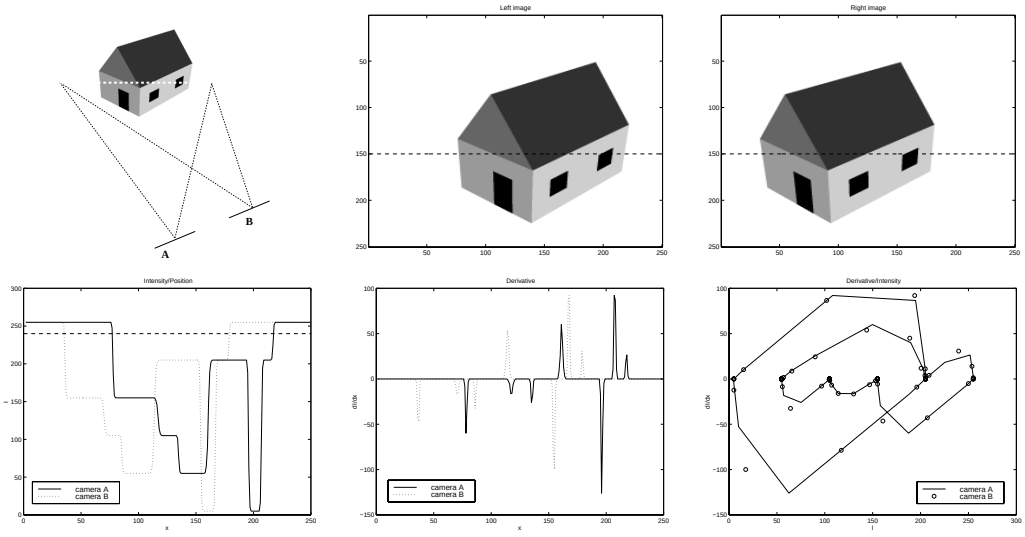
3.1.1 Motivation

One of the challenges for solving the correspondence problem is that of finding an image representation that facilitates (or even trivializes) the matching procedure. As an example, consider two corresponding epipolar lines of the stereo image pair shown in Figure 3.1. The simplest function we can analyse is the brightness function $f(y)$ and $g(x)$ defined along each (left or right) epipolar line — Figure 3.1a. This figure shows the difficulty of the matching process since the gray level functions along both epipolar lines are geometrically deformed by the 3D structure. However, we can obtain other representations for the information included along a scanline, as follows:

1. A commonly used representation, mainly in optical flow computation [56], consists in the spatial derivatives of the brightness (Figure 3.1b). As discussed in the previous part of this discussion, in some situations this matching process is not trivial, (1) first due to the aperture problem, (2) because it not computable in non-textured areas, and (3) it fails when a wide baseline stereo system is considered.

2. To search for correspondences, one could use the plot of Figure 3.1a and determine the points with equal brightness value. However, this would only work when the brightness is kept exactly constant and would lead to many ambiguities, as illustrated in the figure for the gray value 240.

3. Other possible representation consists in plotting the brightness versus its derivative [61] as shown in Figure 3.1c. In this case, the image points with the same brightness and derivatives have approximately the same coordinates, indicating a correspondence. Again, there are some ambiguous situations (shown in the figure) and the points are coincident only if the disparity is constant (no perspective effects) along these image lines.



a) b) c)

Figure 3.1: Photometric and geometric relations of the brightness values along a scanline captured by a stereo pair. a) Brightness values versus pixel positions; b) Derivatives of the brightness versus pixel positions; c) Derivatives versus brightness values.

These representations can be generalized considering other local descriptors (other than brightness, position and spatial gradient) computed along two or more corresponding scanlines. Tomasi and Manduchi [61] have proposed to represent a set of local descriptor vectors

(brightness, derivative, second derivative, etc) through a curve in a n-dimensional space, where the curve represented in Figure 3.1c is a simple example. Ideally two curves computed along two corresponding scanlines can be mapped (or even coincide).

However, approaches based on curves of local descriptors vectors have obvious limitations related to rigid geometric distortion assumptions, solution ambiguity and/or high-dimensionality search algorithms. First of all, the method is only valid for constant and affine disparities (no perspective effects have been considered). Secondly, the curves have a difficult representation, specially if more than two local descriptors are considered. Finally, a curve can cross itself, clearly generating ambiguous situations.

Here we develop a simple framework that overcomes the restrictive geometric, photometric and algorithmic constraints, mentioned before. We propose to study other kind of representations, based not only on local descriptors but also on global descriptors of the image, that we call *Intrinsic Images*, that simplify the (dense) matching process and can be used to generate new views from a stereo pair.

3.1.2 Assumptions

Our main goal consists of finding a representation with the following fundamental features:

1. The new representation can be represented through a simple image.
2. It must encode both the photometric and geometric structure of the original image.
3. The original images can be recovered again from this representation.
4. The disparity can be computed easily.
5. It can handle occlusion points.

To achieve that, we propose to use both local and global descriptors of the image in a new representation, so-called an *Intrinsic Image*. By using this representation, the computation of correspondences and disparity fields can be done in a straightforward manner, and all the requirements considered above are fulfilled.

However, some assumptions have to be made. Thus, our initial effort consists of defining three general and acceptable assumptions in order to produce a correspondence framework which can be considered sufficiently reliable.

- **Calibration** The stereo system is weakly calibrated, which means that corresponding epipolar lines are known.

- **Photometric Distortion** Given two corresponding points x_0 and y_0, the respective photometric values are related by a generic non-linear function $f(y_0) = \Psi(g(x_0))$, where $f(y_0)$ and $g(x_0)$ represent a given photometric measure (e.g. the brightness value). We study the cases where the photometric distortion function, $\Psi(g)$, is the identity (constant brightness) or affine (with contrast and mean brightness differences).

- **Geometric (perspective) Distortion** Two corresponding profiles are related by a disparity mapping function defined by $x = \Phi(y)$, between the two images. We assume that $\Phi(y)$ is a monotonic strictly increasing and differentiable function.

In summary, we propose to study the intensity-based matching problem of two corresponding epipolar lines, where the matched brightness points are ruled by a generic model for geometric (perspective) distortion.

Additionally, there is a set of issues that we have to remark. First of all, the disparity mapping function $\Phi(y)$ includes important perspective and structure distortions of the scene. However the related assumptions made before are uniquely piecewise valid, considering that there are discontinuities and occlusion points on the images. Secondly, occlusion points have to be included coherently in our framework, by observing their position relatively to the corresponding points. However, unless prior knowledge exists, no depth information of an occlusion point can be recovered from two images. Finally, imposing $\Phi(y)$ to be strictly increasing, represents an order criteria, excluding order exchanges between two corresponding points. In fact an order exchange implies that an occlusion will occur between the two views. Thus, we consider those points as occlusions.

3.2 Definition of Intrinsic Images

The simplest kind of matching is that of two corresponding epipolar lines derived from two views of the same scene without occlusions. To simplify the general framework, we assume that there are no intensity changes due to viewing direction and that the disparity mapping function, $\Phi(y)$, defined as:

$$x = \Phi(y), \qquad \frac{d\Phi(y)}{dy} > 0 \tag{3.1}$$

verifies the order constraint and represents the unknown deformation at y to produce the corresponding point x. Assuming the brightness constancy hypothesis, the following nonlinear model can be expressed as

$$f(y) = g(x) = g(\Phi(y)) \tag{3.2}$$

In order to develop a simple and dense matching framework, we propose alternative representations, based not only on local descriptors but also on global image descriptors.

The simplest example of a global descriptor of a scanline is the integral of brightness. One could associate each scanline pixel, x, to the sum of all brightness values between $x = 0$ and x. However, this integral would be different for two corresponding points in the two images, due to geometric (perspective) distortion.

In the remaining of this section, we will derive a new representation, based on a different global image descriptor. Using two horizontal scanlines of the perspective projection of the synthetic scene illustrated in Figure 3.1, we show that it can deal with perspective distortion between the two images.

Let f and g be the intensity values along corresponding epipolar lines. We now assume that both functions are differentiable, obtaining the following expression from the equation (3.2):

$$\frac{df(y)}{dy} = \frac{d\Phi(y)}{dy} \left. \frac{dg(x)}{dx} \right|_{x=\Phi(y)} \tag{3.3}$$

In the absence of occlusions, we can further assume that all brightness information is preserved between two corresponding segments $]y_1\ y_2[$ and $]x_1\ x_2[$ contained respectively in left and right images. Then, by assuming that $d\Phi(y)/dy > 0$ (order constraint), one proves analytically that

the following equality holds:

$$\int_{y_1}^{y_2} \left| \frac{df(y)}{dy} \right| dy = \int_{y_1}^{y_2} \frac{d\Phi(y)}{dy} \left| \frac{dg(x)}{dx} \right|_{x=\Phi(y)} dy = \int_{x_1}^{x_2} \left| \frac{dg(x)}{dx} \right| dx \tag{3.4}$$

Now, let us define two functions α and β:

$$\alpha(y_i) = \int_{y_1}^{y_i} \left| \frac{df(y)}{dy} \right| dy \qquad \beta(x_i) = \int_{x_1}^{x_i} \left| \frac{dg(x)}{dx} \right| dx \tag{3.5}$$

where $x_i > x_1$, $y_i > y_1$ and $x_1,\ y_1$ are corresponding points. Equation (3.4) shows that when y_i and x_i are corresponding points, then $\alpha(y_i) = \beta(x_i)$. This means that, associating each scanline pixel, x, to the sum of the absolute value of the derivatives between $x = 0$ and x, we obtain the same values for corresponding points, independently of arbitrary image distortion, such as perspective effects.

In the following sections we will use these functions to build photometric and geometric image descriptors of a stereo pair. Such combined representation, called an *Intrinsic image*, will later be used for disparity estimation and synthesis of new views.

3.2.1 Photometric Descriptors

Using the stereo pair represented in Figure 3.1 as example, we can compute the functions $m_l = \alpha(y)$ and $m_r = \beta(x)$. We have shown in the previous section that $m_l = m_r$, for two

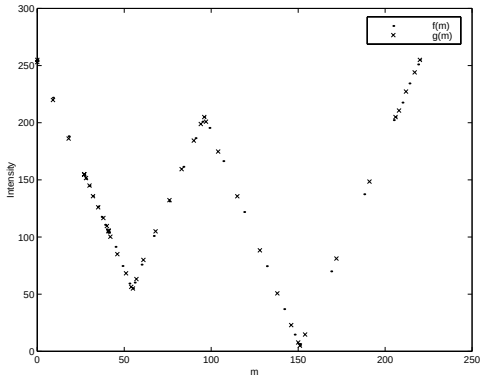


Figure 3.2: Values from functions $m = \alpha(y)$, $m = \beta(x)$ versus the image intensities.

corresponding points. Going a step further, Figure 3.2 shows the values of these functions, $m_l = \alpha(y),\ m_r = \beta(x)$, versus the image intensity values, computed for the stereo images

considered in the example. Not only the points are coincident, but also they can be represented through an image (putting together the various epipolar lines). One observes then a set of useful properties:

a) α and β are both monotonic increasing functions.

b) If $x_i,\ y_i$ are corresponding points, then $\alpha(y_i) = \beta(x_i)$, according to equations (3.4,3.5).

c) If $\alpha(y_i) = \beta(x_i)$, then $f(y_i) = g(x_i)$.

d) Let $m_l = \alpha(y)$. Every value of $m_l > 0$ corresponds to one and only one brightness value $f(y)$, meaning that the function $f(m_l)$ can be defined and represents a photometric descriptor. The same is applicable to $g(x)$.

These photometric descriptors, built from the stereo images, code photometric information available in the two images of the stereo pair, irrespective to perspective distortions. Hence in the absence of occlusions and under brightness constancy they are equal for the two images. Later, in Section 3.3, we will show how to use this property for occlusion detection.

When building a photometric descriptor for the image pair, we have lost information about the spatial domain that could lead to the computation of disparity. This aspect will be addressed in the following subsection.

3.2.2 Geometric Descriptors

We have seen a representation that codes all the photometric information of the stereo image pair, $f(m)$ and $g(m)$, and we now need to retrieve the geometrical information that is related to disparity.

Let us define the generalized functions $y'(m) = dy/dm$ and $x'(m) = dx/dm$. These functions, $x'(m)$ and $y'(m)$, are computed from images and take into account the local geometric evolution of the brightness along the scanlines.

Hence, we can form an image $x'(m)$ and $y'(m)$ by putting together various epipolar lines of an image. These descriptors convey all the necessary geometric (disparity) information available in the image pair. Each disparity, $d(x_i)$, and pixel value, x_i, can be recovered for

each value of m_i, as follows:

$$(x_i, d(x_i)) = \left(\int_0^{m_i} x'(m)dm, \int_0^{m_i} (y'(m) - x'(m))dm \right) \tag{3.6}$$

We can generalize the definitions above for images with brightness discontinuities. In order to guarantee the validity of same theory, we define $df(y)/dy$ as a generalized function, admitting Dirac deltas in some points (at the brightness discontinuities). Thus, α is also discontinuous, $f(m)$ is uniquely defined for a restricted domain, and $x'(m)$ is zero in non-imaging areas (it means, values of m not represented by a brightness value).

The geometric descriptors, together with the photometric descriptors form complete *Intrinsic Images* that contain both photometric and geometric information of the stereo pair.

3.2.3 General Properties of the Intrinsic Images

In this section we will define formally the *Intrinsic Images* obtained from the photometric and geometric descriptors presented in the previous sections, and introduce some of the interesting applications of this representation.

Let k index corresponding epipolar lines of an stereo image pair. Then, the *Intrinsic Images*, $\mathcal{X}(m,k)$ and $\mathcal{Y}(m,k)$ are defined as:

$$\mathcal{Y}(m,\ k) = (f(m,k),\ y'(m,k)) \qquad \mathcal{X}(m,\ k) = (g(m,k),\ x'(m,k)) \tag{3.7}$$

where m is computed by equation (3.5) and $(m,\ k)$ are the coordinates of the intrinsic images. It is possible to reconstruct completely the original left and right images based on $\mathcal{Y}(m,k)$, $\mathcal{X}(m,k)$, respectively. Figure 3.3 shows the photometric and geometric components of the *Intrinsic Images*, computed for the stereo image pair shown in Figure 3.1.

Given the properties described before, we state the following observation for Intrinsic Images, under the brightness constancy assumption and in the absence of occlusions:

Observation 1 — Intrinsic Images Property
If $f(y,k) = g(\Phi(y),k)$ *and* $d\Phi(y)/dy > 0$ *for all* $(y,\ k)$, *then* $f(m,k) = g(m,k)$ *and the relation between* $x'(m,k)$ *and* $y'(m,k)$ *gives the geometric deformation between corresponding points.*

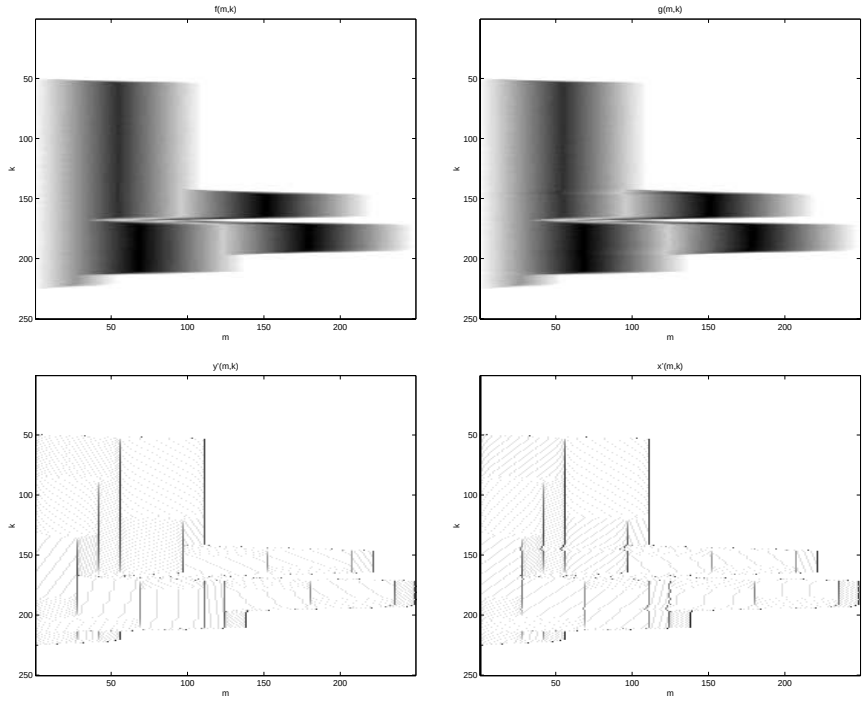


Figure 3.3: Intrinsic images. Top: photometric descriptors. Bottom: geometric descriptors.

From this observation we derive an interesting property, that will allow us to generate novel views of the observed scene, including all perspective effects.

Observation 2 — Synthesis of new Images

Assume that both cameras are parallel, related by a pure horizontal translation T, with intrinsic images $\mathcal{Y} = (f(m,k),\ y'(m,k))$ and $\mathcal{X} = (f(m,k),\ x'(m,k))$. Then, views at intermediate positions jT (where $0 \leq j \leq 1$) have the following intrinsic images, $\mathcal{I}_j$:

$$\mathcal{I}_j(m,\ k) = (f(m,k),\ jx'(m,k) + (1-j)y'(m,k)) \tag{3.8}$$

Proof: Suppose that the disparity between two corresponding points x_j and y of two generic parallel cameras is given by the well known expression $x_j = y + jT/Z$, where Z denotes the depth of the object relatively to the cameras and jT represents the translation between them. Derivating both terms of the equality in relation to m, we obtain

$$x_j^{'}(m) = y'(m) + jT\frac{dZ^{-1}}{dm} \tag{3.9}$$

Finally, by performing a weighted average between the left and the right cameras ($j = 0$ and $j = 1$ respectively), we obtain the same expression for $x_j^{'}(m)$:

$$x_j^{'}(m) = jx'(m) + (1-j)y'(m) = y'(m) + jT\frac{dZ^{-1}}{dm} \tag{3.10}$$

QED

This result provides a means of generating intermediate unobserved views simply by averaging the geometric component of the original *Intrinsic Images*. It accounts for perspective effects without an explicit computation of disparity.

We have described the essential framework for our matching approach assuming brightness constancy and in the absence of occlusions. We have defined *Intrinsic Images* and shown how to use these images to compute disparity in a direct way. Next, we will introduce occlusion information. At the end of the chapter, we will discuss some issues about the relaxation of the brightness constancy constraint.

3.3 Dealing with Occlusions

Many algorithms have been designed to handle occlusions for multiple image motion or disparity estimation [68, 45, 66, 9]. Here, we focus on introducing occlusions in *Intrinsic Images*, ensuring that there is no exceptional treatment based on heuristic criteria or adhoc thresholds.

An occlusion occurs when a surface is in front of an occluded region, which can be seen by only one camera. However, unless we impose prior models to image disparities, 3D structure or to global image features, we can detect the existence of a local occlusion based on photometric information. We will show how to include occlusion information in *Intrinsic Images*, based on the theory developed in the last section.

First of all, even in the presence of occlusion points, an intrinsic image can be defined as before. However, the *Intrinsic Images Property* stated in Observation 1, is not verified because its sufficient condition is not verified. In fact, the condition $f(y,k) = g(\Phi(y),k)$ is only piecewise valid (along corresponding profiles), but it is not valid in general, namely in occlusion points.

It is worth noticing important differences between the usual cartesian images and the associated photometric descriptor of the intrinsic images, in a stereo pair. While the intrinsic images will only differ in the presence of occlusions, cartesian images differ both by perspective effects and occlusions. This means that, in order to detect occlusions or, equivalently, photometric dissimilarities, we can rely on the photometric descriptors of the intrinsic images, where the geometric distortions have been removed.

Therefore, we propose to define global image descriptors similar to those discussed previously. Consider

$$m_l = \alpha(y_i) = \int_{y_1}^{y_i} \left| \frac{df(y)}{dy} \right| dy \qquad\qquad m_r = \beta(x_i) = \int_{x_1}^{x_i} \left| \frac{dg(x)}{dx} \right| dx \tag{3.11}$$

computed on corresponding epipolar lines of the left and right cameras, where x_1 and y_1 are the respective initial (and not necessarily corresponding) points.

In the previous section, we have shown that in the absence of occlusions, $m_l = m_r$ for corresponding points, greatly simplifying the matching procedure. In the presence of occlusions, the matching is not that trivial, but m_l and m_r can still be related by a simple function. Let m_l and m_r be parameterized by t as follows:

$$m_l = r(t) \qquad\qquad m_r = s(t) \tag{3.12}$$

The curve produced by these functions can yield uniquely three forms:

1. Horizontal inclination ($dr(t)/dt = 1$; $ds(t)/dt = 0$): there exists a region in the left camera, that is occluded in the right camera.

2. Vertical inclination ($dr(t)/dt = 0$; $ds(t)/dt = 1$): there exists a region in the right camera, that is occluded in the left camera.

3. Unitary inclination ($dr(t)/dt = 1$; $ds(t)/dt = 1$): both profiles match (no occlusions).

Figure 3.4 shows examples of two (m_l, m_r) matching scenarios with and without occlusions.

Hence, the problem to solve is that of determining the mapping curve from m_l to m_r, which can be done through several approaches. Fortunately, there is a low cost algorithm that solves it optimally, in the discrete domain.

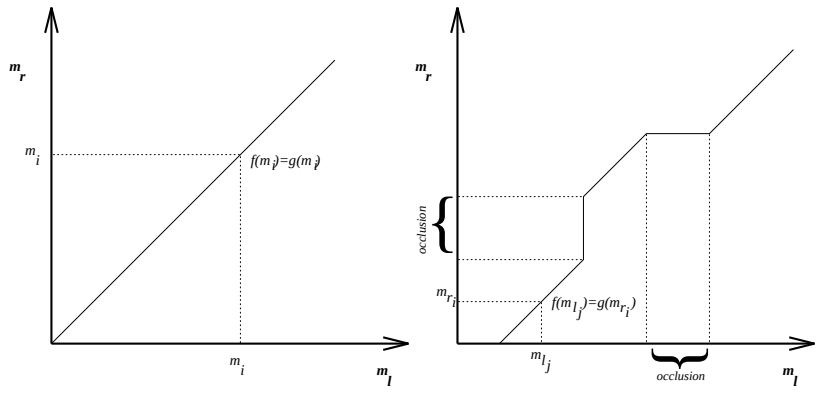


Figure 3.4: Matching scenarios without (left) and with occlusions (right).

Assume that we have two sequences $F = \{f(m_{l1}),\ f(m_{l2}), ...,\ f(m_{lp})\}$ and $G = \{g(m_{r1}),\ g(m_{r2}), ...,\ g(m_{rq})\}$ given by the photometric information along corresponding epipolar lines of the left and right intrinsic images. The corresponding points constitute a common subsequence of both F and G. Moreover, finding the set of all corresponding points corresponds to finding the maximum-length common subsequence of F and G. This problem consists in the well known longest-common-subsequence (LCS) problem [6], which can be solved efficiently using dynamic programming.

Notice that an LCS solution for our problem implies two things: (1) the corresponding points obey to the order constraint; (2) an order exchange represents existence of occlusions. These two implications can produce some ambiguity situations. However, such cases correspond mostly to perceptual ambiguities only solved with prior knowledge about the scene.

After finding an LCS solution, or, equivalently, the curve that matches m_l and m_r, we can change the definition of the *Intrinsic Images* in order to maintain the theory described in last section.

Observation 3 — Intrinsic Images Property with occlusions

Given two stereo images, if $f(y,k) = g(\Phi(y),k)$ *and* $d\Phi(y)/dy > 0$ *for almost all* $(y,\ k)$ *(except for occlusions points), then it is possible to determine a pair of intrinsic images* $\tilde{\mathcal{Y}}(t,k)$ *and* $\tilde{\mathcal{X}}(t,k)$.

$$\tilde{\mathcal{Y}}(t,k) = (\tilde{f}(t,k),\ \tilde{y}'(t,k)) \qquad \tilde{\mathcal{X}}(t,k) = (\tilde{g}(t,k),\ \tilde{x}'(t,k)) \tag{3.13}$$

where $\tilde{f}(t,k) = \tilde{g}(t,k)$ *and the relation between* $\tilde{x}'(t,k)$ *and* $\tilde{y}'(t,k)$ *gives the geometric deformation between corresponding and occlusion points of the stereo images. Knowing the functions* $r(t)$ *and* $s(t)$ *of equation (3.12), these new intrinsic images are found based on the original intrinsic images (as presented in Section 3.2), by performing the following transformation:*

$$
\begin{aligned}
(\tilde{f}(t,\ k),\ \tilde{y}'(t,k)) &= \begin{cases} (f(r(t),\ k),\ y'(r(t),k)) & if \quad dr(t)/dt = 1 \\ (g(s(t),\ k),\ 0) & if \quad dr(t)/dt = 0 \end{cases} \\
(\tilde{g}(t,\ k),\ \tilde{x}'(t,k)) &= \begin{cases} (g(s(t),\ k),\ x'(s(t),k)), & if \quad ds(t)/dt = 1 \\ (f(r(t),\ k),\ 0) & if \quad ds(t)/dt = 0 \end{cases}
\end{aligned}
\tag{3.14}
$$

This observation has two important implications. First, by computing the functions r and s as a solution of the LCS problem, we can derive a coherent framework which permits to compute disparities or generate new views as defined in last section. Secondly, the generated intermediate views exhibit consistent information in occlusion areas. However, it does not imply that this information is consistent with the real 3D structure of the occluded points (given the impossibility to recover that structure).

3.4 Experiments

Throughout this chapter, we have used a simple synthetic stereo pair in order to illustrate the various steps of our approach. By using the intrinsic images shown in Figure 3.3, we have used Equation (3.6) to compute the dense disparity map shown in Figure 3.5, without using any search algorithms. Figures 3.6a (on the left) shows a sequence of synthesized views from the same pair of images.

In order to show the behavior of occlusions in synthesized views using the proposed methodology, some known real stereo pairs are discussed. Figure 3.6a shows results obtained with the "flower garden" and "trees" sequences, where a significant amount of occlusions are present. We have applied the proposed method, based on the *Intrinsic Images* to synthesize new views from the original left and right stereo images. The occluded regions move coherently in the sequences. The perceptual quality of the reconstructed images is comparable with that of the original ones. To illustrate the ability of the method to cope with occlusions, Figure 3.6b

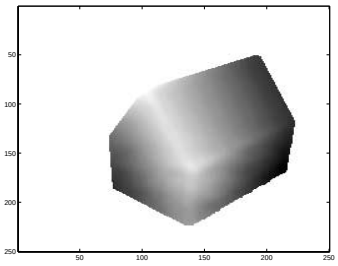


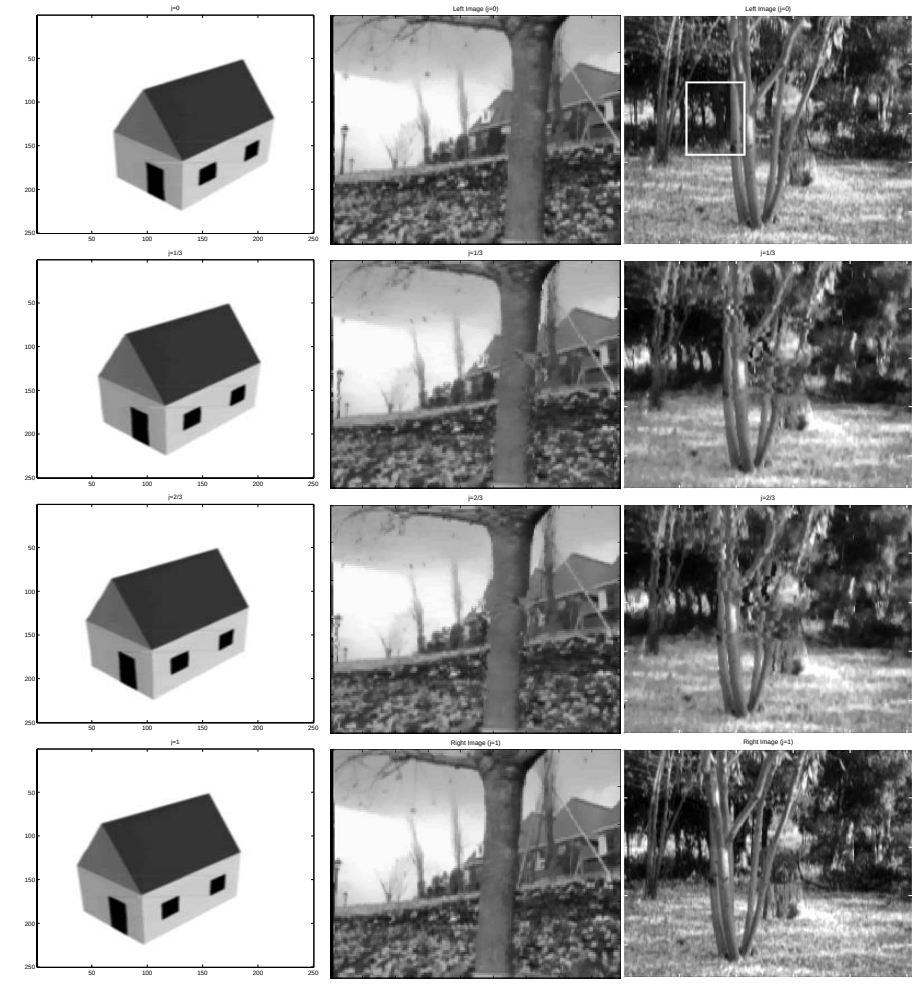
Figure 3.5: Disparity map of the synthetic pair, determined directly from the Intrinsic Images.

shows the recovered views on a detail of the images (see little square in the respective sequence), with a strong discontinuity in the depth. As expected, the occluded points disappear coherently behind the tree, without any smoothing. Notice that in the "trees" sequence there is an area, roughly on the center of the image, where some occlusions associated to order exchanges create erroneous solutions for the synthesis of that area. This can only be solved introducing more images of the sequence or with prior knowledge about the scene.

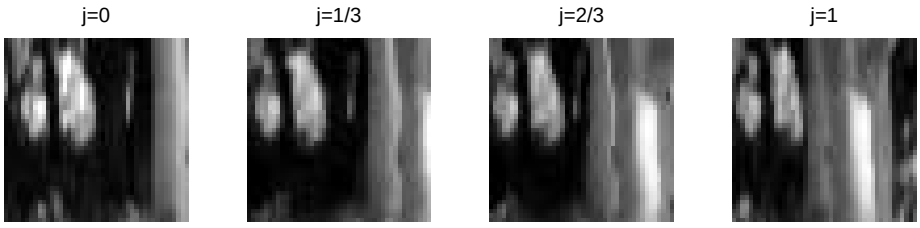
These examples illustrate how *Intrinsic Images* deal with perspective effects and occlusions, generating novel views from an original stereo pair. Next we will apply the same methodology to two different examples in order to compute a dense disparity field.

In the first example, we compute the disparity map (Figure 3.7c) from the "mask" stereo pair (Figure 3.7a,b), based on Equation (3.6). Next we apply a gaussian filter (10×10, standard deviation 5) to the computed disparity image, smoothing locally the respective disparity values, as shown in the Figure 3.7d. Figure 3.7e represents the resulting isometric plot of the scene under an arbitrary point of view. The Figures 3.7f-p show a set of different views of the scene where the image texture was superimposed. By visual inspection we can observe the realism of the estimated 3-D model, retrieved uniquely from two views with a relatively small baseline, in spite of the presence of a large amount of ambiguities (due to the homogeneous brightness and poor texture), which would represent an evident difficulty for a traditional correspondence method.

In the next example, there are multiple discontinuities in the scene, various objects in various depth levels, and a significant amount of occlusions. We have used uniquely four images from the University of Tsukuba's Multiview Image Database. Our goal consists in (1)



a)



b)

Figure 3.6: a) Synthesis of intermediate views for the synthetic images (left), the Flower Garden Sequence (center) and for the Tree Sequence (right). b) The evolution of a detail of the Tree Sequence (see little square in the respective sequence) with occlusions.

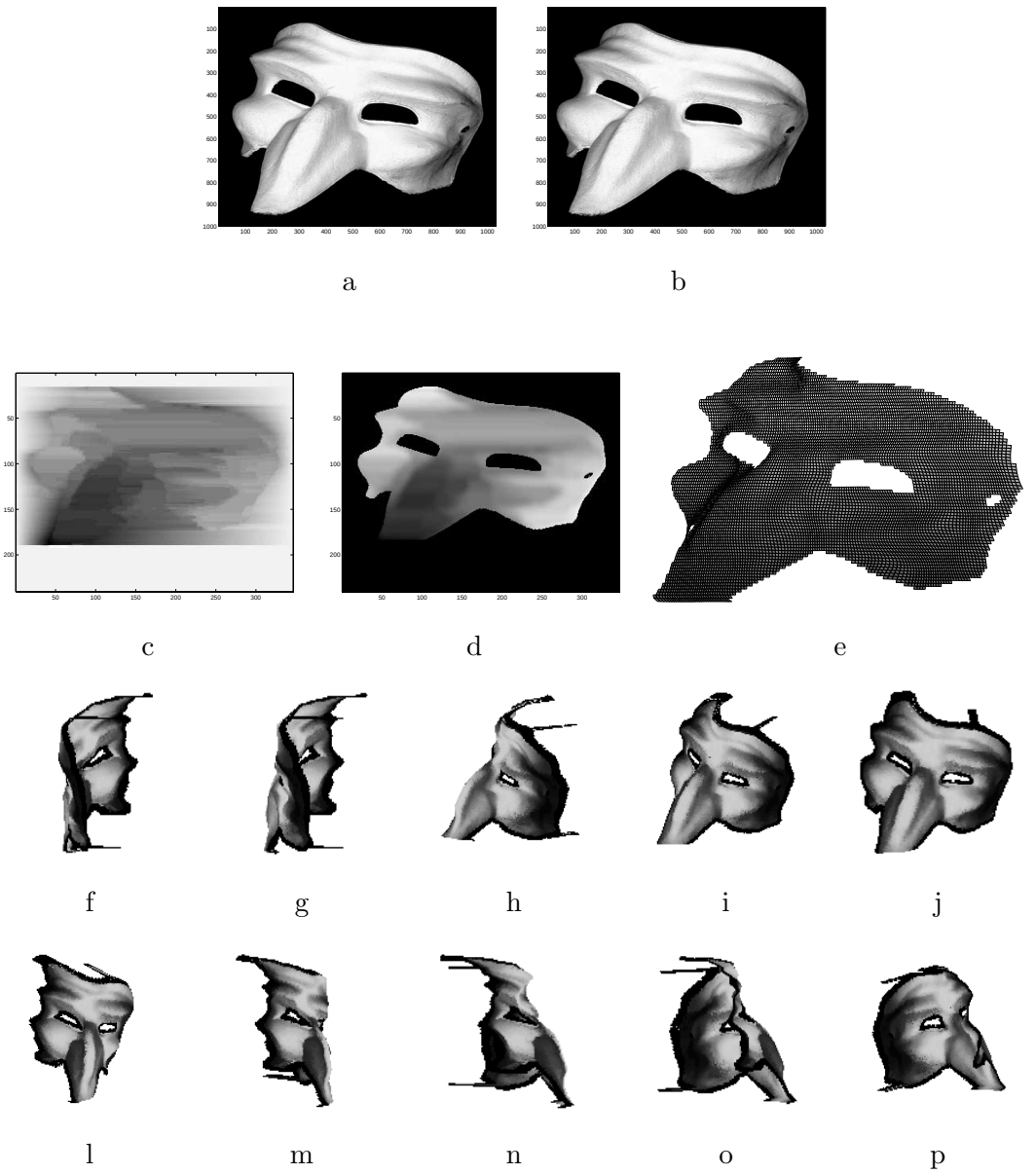


Figure 3.7: a-b) Original pair of images of a real scene. c) Estimate of the disparity map using the Intrinsic Images approach. d-e) Filtered version of the disparity map, where the background was zeroed by a threshold. f-p) Arbitrary views around the reconstructed model where the original texture was superimposed. For visualization reasons, the apparent relief of the models above are given by the disparities and not by estimates for the depths of the scene.

synthesizing intermediate views within the rectangle formed by the original four views and (2) computing the disparity map by using uniquely the upper two views of the original set.

Figure 3.8a-d shows the images used, captured by four parallel cameras located at the vertices of a rectangle. We have used the *Intrinsic Images* approach to synthesize 13 novel views between "vertical" image pairs a-c and b-d, leading to a total of 15 image pairs. This process was repeated with these 15 (original and synthesized) "horizontal" image pairs. At the end, we obtain a matrix of 15×15 camera views, starting from the 4 original images. Figure 3.8 shows an arbitrary set of 6 synthesized views that illustrate the coherent motion of vertical and horizontal discontinuity areas, specially when occlusions are present.

We have also computed a disparity map using the images shown in Figure 3.8a,b. Figure 3.9a shows the resulting disparity map, while Figure 3.9c-d shows two arbitrary views of the scene, where brighter points correspond to larger disparities. A view with superimposed texture is shown in Figure 3.9b.

These examples illustrate the potentialities of *Intrinsic Images* to generate not only novel views from a reduced set of images but also a dense disparity map of the scene, including regions of homogeneous brightness distribution.

3.5 Conclusion

In this chapter, we have considered the hypothesis of brightness constancy. In a real stereo system, however, the brightness can change due to viewing direction or different specifications of the cameras. This leads to an important question related to the following points: how the brightness constancy affects the approach developed before, and how to overcome this problem. Before concluding, we propose to discuss this issue in more detail.

A convenient model to account for photometric distortion is the following:

$$f(y) = a \cdot g(\Phi(y)) + b, \qquad a > 0 \tag{3.15}$$

This model represents an affine transformation, where a and b are the difference in contrast and brightness between the two images. Some authors prefer to estimate a priori the affine parameters [24], before applying the correspondence procedure. This can be performed by

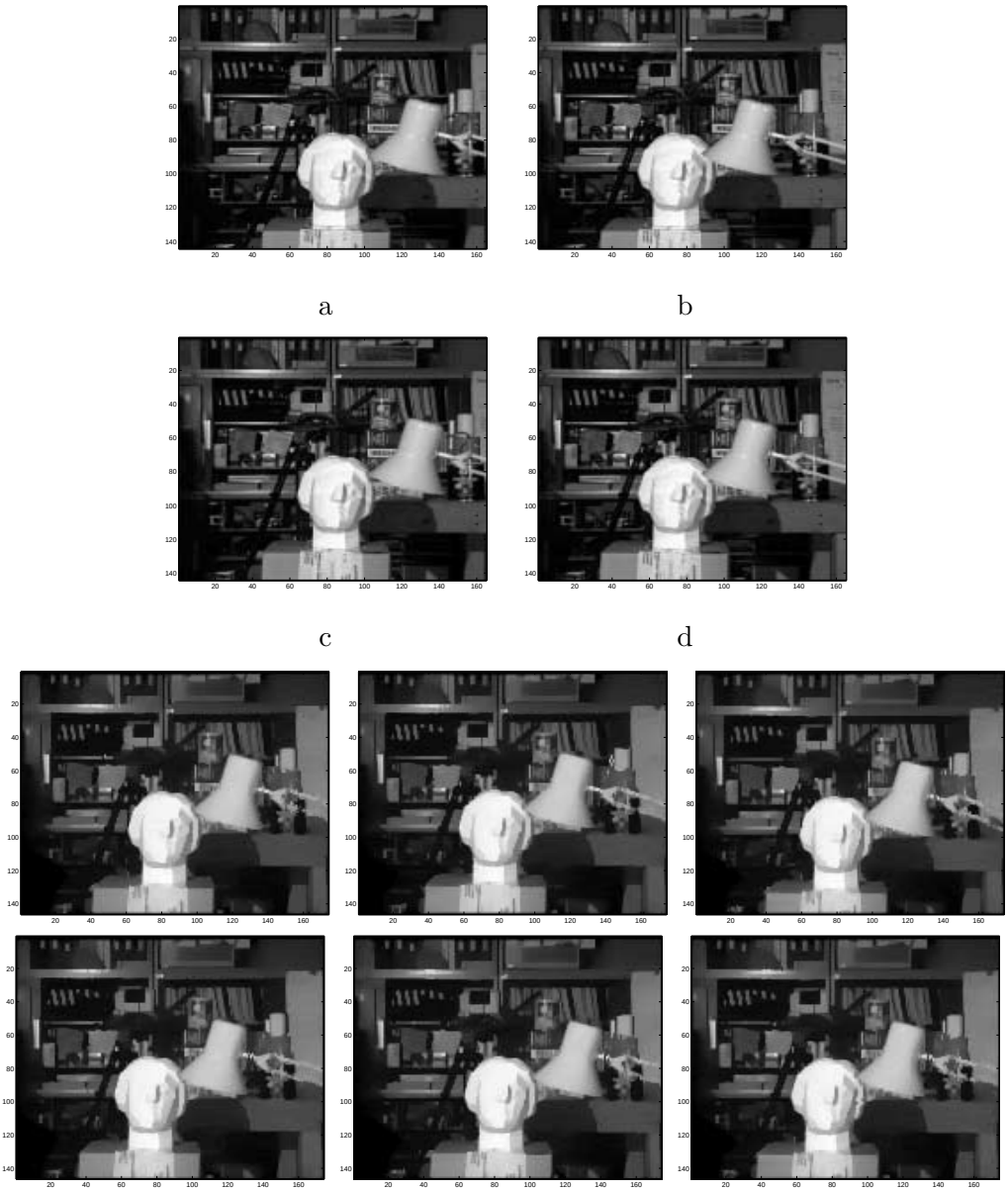


Figure 3.8: a-d) Synthesis of views by using four images provided by the University of Tsukuba. The bottom images were synthesized as if they were seen by virtual cameras located within the rectangle formed by the four original images.

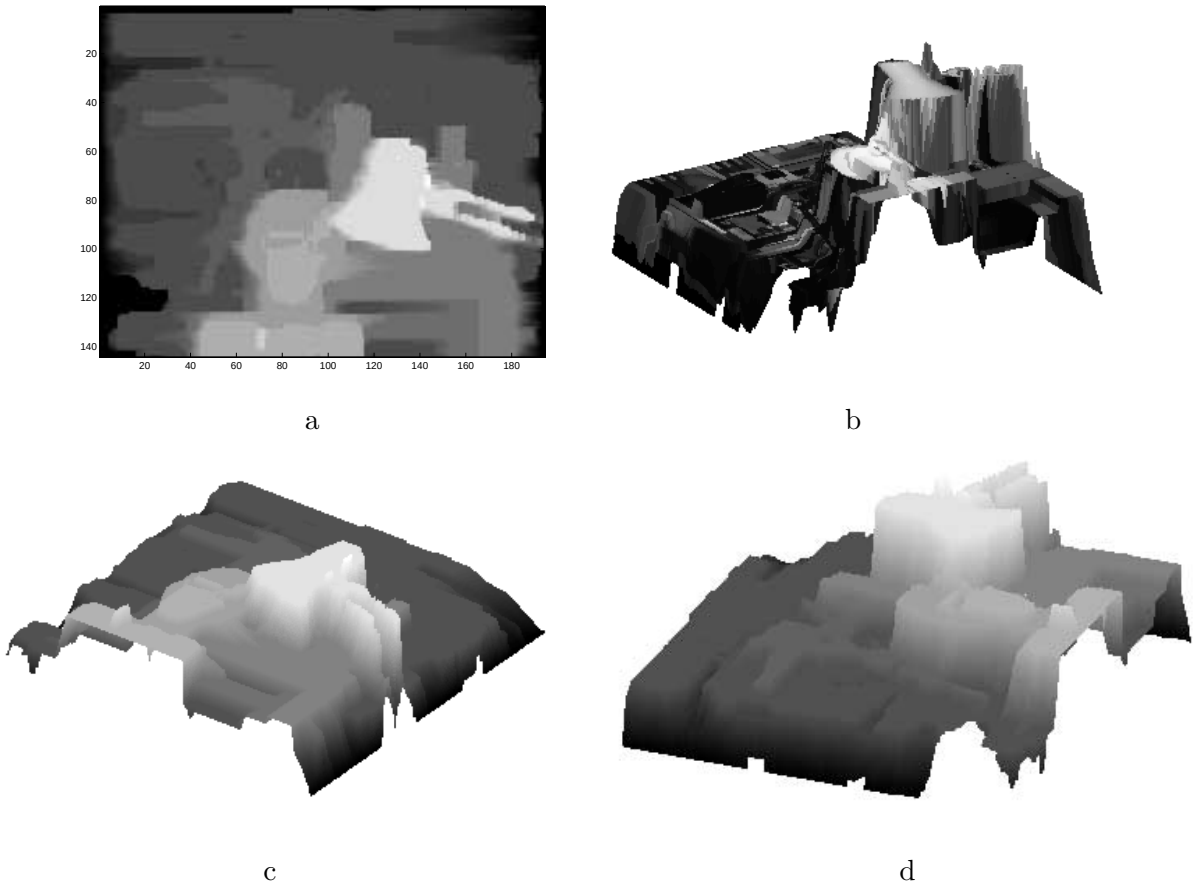


Figure 3.9: a) Disparity map computed from the image pair represented in the Figures 3.8a-b. b) Estimated 3-D model with superimposed texture. c-d) Two arbitrary views of the scene, where the brightness is proportional to the disparity of the points. Notice that the coordinates of the horizontal plane (the ground) correspond to the usual images coordinates, whereas the third one (vertical axis) represents the estimated disparity value (and not a estimate of the depth).

analyzing the global intensity function and its derivatives. However, it would be constructive to study the geometric influence of the affine distortion on the intrinsic image structure.

Considering the brightness distortion in Equation (3.15), the effect of bias, b, can be eliminated by preprocessing both signals with a zero-mean filter. Thus, we assume that the contrast term a is the dominant term.

By assuming that $f(y) = a \cdot g(\Phi(y))$ and applying equations (3.4), a simple relation between the horizontal axis of the intrinsic images is found: $m_l = a \cdot m_r$. This means that the photometric information along the scanlines of the left intrinsic image is scaled along the horizontal axis and in amplitude (the intensity values), by a, with respect to the right intrinsic image. Thus, the deformation induced by the brightness distortion in the intrinsic images, is ruled by the following equality:

$$\frac{f(m_l)}{m_l} = \frac{g(m_r)}{m_r} \tag{3.16}$$

We could apply directly on the intrinsic images a correspondence procedure by using the equation (3.16). Nevertheless this implies some search effort.

Another solution to overcome this problem consist of transforming the brightness function by a simple function which exhibits some invariant properties related to linear distortions. The logarithm function is a good candidate. In fact, when y and x are corresponding points, we have:

$$\frac{d \log |f(y)|}{dy} = \frac{d\Phi(y)}{dy} \frac{d \log |g(x)|}{dx} \tag{3.17}$$

defined wherever $f(y)$ and $g(x)$ are different of zero. Applying this relation to equation (3.4), one can conclude that it is still possible to define coherent photometric and geometric descriptors, since $m_l = m_r$ (in absence of occlusions). This means that, in general, the intrinsic image theory remains applicable.

In summary, we have proposed a new image representation — *Intrinsic Image* — which allows for solving the intensity-based matching problem under a realistic set of assumptions, such as the order constraint. The concept of intrinsic images is a useful way to approach the problem of stereo vision, and leads to a straightforward computation of disparity and new views.

Intrinsic Images of a stereo pair give exactly the complete photometric and geometric

structure of the 3D scene, independently of the observed geometric (perspective) distortion. Secondly, provided that the order constraint is not violated, a coherent interpretation of the occluded regions is made.

An Intrinsic Image is composed by a photometric and a geometric descriptor, which contain all the necessary information to reconstruct directly the original images, disparity values and other views of the same scene. We have presented some results with real stereo images. The method is very robust to the existence of occlusions and reveals high performance in multi-view synthesis.

This approach is a powerful tool to cope not only with a stereo pair of images but also with a sequence of images, both in the discrete and continuous sense.

In the future, we plan to study the potentialities of this approach applied to a larger number of images, namely in dense optical flow computation and egomotion estimation. We plan also to develop an automatic occlusion characterization based on a large number of images of same scene, in order to facilitate the creation of the associated intrinsic images with occlusion information.

Chapter 4

Reconstruction for Multiple Views

4.1 Physical Constraints

We have shown how the Intrinsic Images representation lead to a simple process of computing dense disparity maps and to synthesize novel views from a stereo pair.

The presence of occlusions brings the problem that the visual information can be available in one image only and hence disparity cannot be computed in the most straightforward way. To overcome this problem, we introduced an additional optimization process based on Dynamic Programming that can cope with the lack of image data arising from the existence of occlusions.

Both the original approach of the Intrinsic Images and the extended version to deal with occlusions are based on a number of constraints of the associated disparity mapping between an image pair. One of the main constraints used in these methods is the order constraint.

This constraint, however, is not necessarily preserved whenever occlusion points are present. Nevertheless, considering the order constraint — even if in the presence of occlusions — had two major advantages:

- It is an extremely useful assumption in the stereo problem, when there is no additional information available regarding the 3D consistency of the scene.

- The constraint can be introduced in a simple manner with dynamic programming which is an efficient and well-known optimization tool that embeds the order constraint auto-

matically.

Most of the methods rely explicitly the order constraint to retrieve the 3D structure of a scene [45, 12, 7]. Many others use order-like assumptions, named in different ways such as the continuity of the disparity map [68] or the local coherency of the correspondences [49]. A different constraint is considered by Kutulakos and Seitz [36], where one explores the visibility constraint for an arbitrary set of calibrated cameras, and no approximation was made. However this approach fails when outliers are introduced in the observations since the reconstruction algorithm is greedy and is not able to deal with outlier rejection under an optimization framework.

Suppose that no assumptions (including the order constraint) are made about the scene structure or about the projected images. What can we infer about the structure from N arbitrary positioned cameras, knowing that the scene tends to be non-smooth and exhibit significant occlusions? Can we deal with this problem by using physically meaningful constraints? Can we handle these constraints in a unique and optimal formulation? Answering these questions has especially important implications for reconstructing real objects and environments from more than two images. In contrast to the Intrinsic Images approach, where a simple and direct strategy was adopted to retrieve 3-D information from two views, in this chapter we want to develop a method that uses multiple views to perform correspondence and occlusion detection in a single step, and achieves global optimality for a set of realistic constraints.

Thus, the choice of a good set of constraints is an essential step for solving the correspondence problem:

- First, in order to deal with the intrinsic ambiguity of the correspondence problem we have to impose domain-specific constraints to restrict a final solution;
- Second, there exist real and important constraints relating the geometry of multiple views (a well-known example consists in the epipolar geometry that relates the cameras);
- Finally, the constraints must be as generic and independent of the scene geometry or topology as possible (the order assumption is not generic and depends on the scene

topology).

We will use three constraints[1] that we consider as the most important and general under a geometric point of view: the uniqueness, the visibility and the camera geometric constraints. Later we will show how they guarantee a consistent 3-D solution, avoiding the need of an approximated assumption such as the order constraint.

4.1.1 Camera Geometric Constraints

Corresponding points across multiple views are related by multiview geometric constraints (e.g the epipolar geometry) that depend on the cameras intrinsic and extrinsic parameters. Notice that these constraints are completely independent of the scene structure, as opposed to other constraints considered in this section.

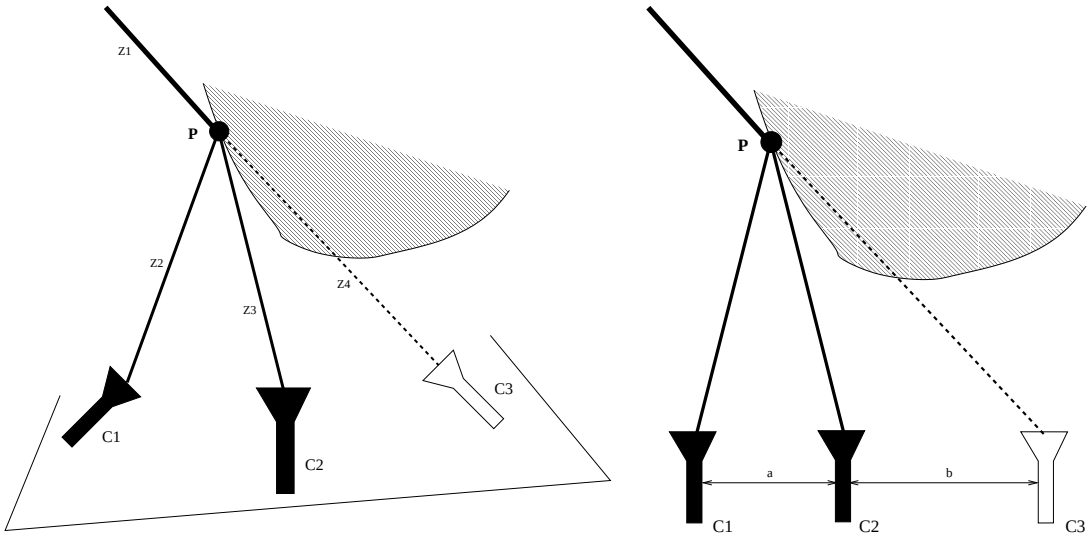


Figure 4.1: a) Generic configuration of a trinocular camera system. b) A trinocular system with collinear centers, where the optical axes are parallel and perpendicular to the baseline.

Consider for now a system of three cameras as shown in Figure 4.1a, and let us assume that all Fundamental Matrices or the correspondences among epipolar lines are known. For this generic configuration, the geometric arrangement of the cameras constrains the image

[1]Other assumptions related with photometric issues are discussed later.

projections of 3-D points and, consequently, the correspondence between image features. In particular, given a pair of matched points in two views, it may be possible to determine the position of the point in the third view. This consists in the point transfer problem [26], which can be solved by using the Fundamental Matrices, or, more generally, the Trifocal Tensor (the former is undefined for points belonging to some degenerate configurations [26]). The study of the transfer problem is out of scope of this discussion and can be found extensively treated in [26].

Here, we will focus our attention on the simple example of cameras with collinear projection centers, as illustrated in Figure 4.1b. We further assume that the epipolar lines are horizontal and that the intrinsic parameters are the same for the three cameras. Three matched points satisfy the point transfer relation derived by triangulation (or equivalently derived by using the Trifocal Tensor)[2]: $x_2 = k_a x_1 + k_b x_3$, where x_1, x_2, x_3 are pixel positions measured along the respective epipolar lines of the three cameras, and k_a and k_b are computed using the baseline distances (for the configuration shown in the Figure 4.1b: $k_a = \frac{a}{a+b}$ and $k_b = \frac{b}{a+b}$).

In summary, the geometric configuration of a trinocular system constrains the triplets of correspondences observed on the three images. Specifically, we have presented a linear relation among the correspondence positions for cameras with collinear centers.

4.1.2 Uniqueness Constraint

By uniqueness constraint we mean that a given image feature in one camera matches with at most one feature observed by another camera. In spite the fact that the uniqueness appeared already implicitly in previous approaches (e.g. in the stereo), it has special relevance when more than two views are considered. In the next section, this will become evident since the uniqueness constraint has to be explicitly imposed.

[2]Notice that the epipolar transfer (using Fundamental Matrices) fails for this example because the two epipolar lines in the third image are coincident. However this degeneracy is overcome when the trifocal tensor is used.

4.1.3 Visibility Constraint

A system with more than two camera views introduces the visibility as a new and important constraint which has to be taken into account when performing reconstruction. The visibility constraint arises from the assumption that no transparency in the scene is allowed. A correspondence observed in two cameras (which represents by triangulation a specific 3D point) restricts the visibility of a set of 3D points for the remaining cameras, thus confining the domain of possible correspondences. This idea becomes clearer if we consider the example of the Figure 4.1a.

Assume that point P is visible from cameras $C1$ and $C2$. Considering now the camera $C3$, what can we infer about its space of visibility based uniquely on the knowledge of a pair of matched points in two views?

Assuming that P is not transparent, it occludes all scene points which are beyond P and belong to the projection ray viewed by each camera. This is true even for those cameras where P is not visible. Consequently the half-line $Z1$ is not visible from the camera $C3$. This apparently trivial result is rigorously formulated in Space Carving Theory [36], for which the visibility constraint has a central role in the development of a greedy algorithm for 3D scene reconstruction.

Therefore determining a correspondence in two cameras has consequences to the visibility space of the remaining cameras of the system. Each correspondence constrains the visibility space of the other cameras. In our particular example, based on the determination of a single correspondence between cameras $C1$ and $C2$, the 3D points belonging to $Z1$ are excluded from the space of visibility of camera $C3$.

The constraint derived here is the so-called visibility constraint. It is quite general since no special configuration for cameras or scene was assumed. Moreover, it restricts greatly the general solution of the correspondence problem for a system with more than two cameras.

4.2 Reconstruction as an Integer Optimization Problem

In the previous section we described a set of physically meaningful constraints for a system with more than two cameras. Our final goal is to find the correspondences and occlusions through an optimal method subject to a set of realistic constraints. To achieve this, we have to define both an objective function (preferably linear) to be minimized (or maximized), and a set of constraints (preferably linear) to bound the number of solutions.

The work by Maciel and Costeira [40] provides a good insight into an optimization approach applied to a vision problem. They develop a set of generic tools based on integer optimization in order to handle several constraints in a unique formulation, performing correspondence and outlier rejection (or occlusion detection) in a single step. The problem which remains to solve is to describe some visual constraints in that optimization framework. In fact, this problem does not have a trivial solution and needs a previous effort in representing properly the image data from multiple cameras.

In order to solve our problem, we have to keep in mind three ideas:

1. The most commonly used criterion (objective function) is the image correlation, which reflects the assumption of photometric similarity. The study of more general photometric constraints is out of the scope of this work.

2. We want to solve a reconstruction problem by using all available physical constraints, without any further assumptions or approximations.

3. The method to find the correspondences must perform outlier rejection and achieve global optimality with a feasible computation effort.

To develop a method with these features, we will simplify the camera configuration. We will address the reconstruction problem for a specific trinocular setup (Figure 4.2), with three cameras having the same intrinsic parameters, with the optical centers collinear and equidistant, and the optical axes parallel and perpendicular to the baseline. A more general approach is discussed at the end of this chapter. Notice however that the collinear case constitutes a very common configuration, with simple implementation if the camera positions are controllable.

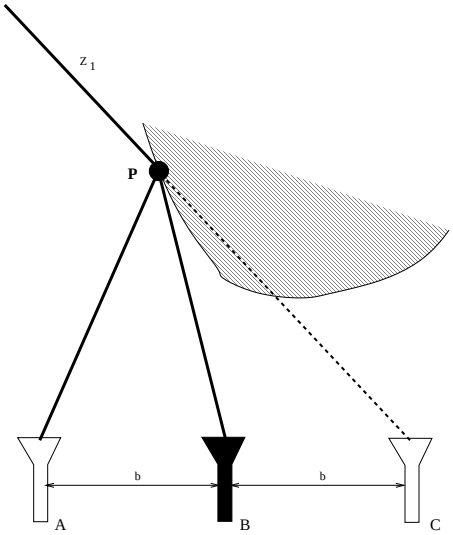


Figure 4.2: Three collinear and equidistant cameras, where the optical axes are parallel and perpendicular to the baseline.

Next, we will present a complete and optimal algorithm for the collinear trinocular setup, taking into account all the ideas discussed before, avoiding additional approximations. We first define an objective function and then present an equivalent integer problem [44] that integrates the constraints discussed in the previous section.

4.2.1 Objective Function

We use the correlation as criterion of similarity between two features. Features consist of image patches with N pixels, centered around points of interest[3]. In general, the most commonly used global objective function is the following:

$$z = \boldsymbol{p}^{\boldsymbol{T}}\boldsymbol{c} \tag{4.1}$$

where $\boldsymbol{p}$ is an indicator vector of ones and zeros, and $\boldsymbol{c}$ is the vector of all possible correlations between features of one image and those of the other image. Generically, the goal is to find the vector $\boldsymbol{p}$ that maximizes the scalar z, subject to some constraints.

[3]This does not mean necessarily that we wish to propose a feature based method, where usually a previous feature segmentation is provided. Point of interest can be viewed in a dense manner, where all pixels can be selected, independently of being "good" or "bad" features.

In the trinocular problem of Figure 4.2, we will use the three images simultaneously in a single step, to reconstruct the scene structure. How can we deal with the various correlation measurements associated to the combination of all features captured by three different cameras?

Consider cameras A, B and C of Figure 4.2. We can compute the similarity between images B and A, B and C, or A and C. However, the collection of all resulting correlations includes some redundancy. We could instead measure uniquely the similarity of B with A and C. As a result, our criterion of similarity is referred to B (because A and C are not directly compared). Referring the reconstruction to the camera B has two important advantages:

1. Each feature from image A (or equivalently from C) is compared solely to the universe of features from the image B, producing a unique set of similitude values. The most similar pair of features corresponds simply to the highest correlation value.

2. Features viewed by the middle camera B can be of two types: features also viewed by the cameras A or C, or features invisible both for A and C. The second type is very rare in real scenes (corresponds to narrow holes in the scene). This means that almost all features observed in the camera B can be reconstructed by using at least one of the two other cameras[4].

Hence we define the complete objective function, as follows:

$$z = \boldsymbol{p}^T \begin{bmatrix} \boldsymbol{c}_{AB} \\ \boldsymbol{c}_{CB} \end{bmatrix} \tag{4.2}$$

where $\boldsymbol{c}_{AB}$ is the vector of correlations between the cameras A and B, $\boldsymbol{c}_{CB}$ is the vector of correlations between the cameras C and B, and $\boldsymbol{p}$ is the associated indicator vector that has to be found in order to solve the correspondence.

The next step consists in selecting the image points we want to correlate in order to find the required correspondences. It is useless to correlate each feature of one image with all features of other image since there are camera geometric constraints which restrict greatly

[4]Notice that this is not true for the cameras A or C, where it is usual to perceive features not observable by the other cameras.

the associated admissible search domain. For the case represented in Figure 4.2, we know that the necessary search domain to find a triplet of correspondence points is bounded by the epipolar line. Secondly, the coordinates of the three points are related by $x_B = \frac{x_A + x_C}{2}$, where $x_A,\ x_B,\ x_C$ are respectively the horizontal coordinates of the cameras illustrated in the figure.

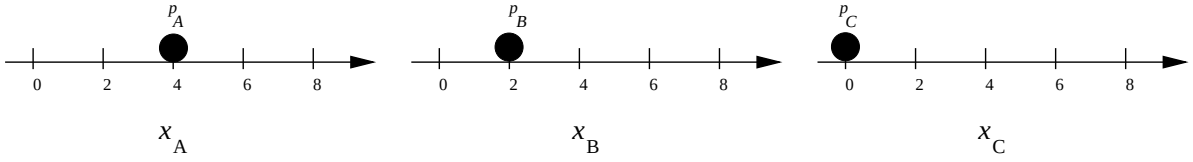
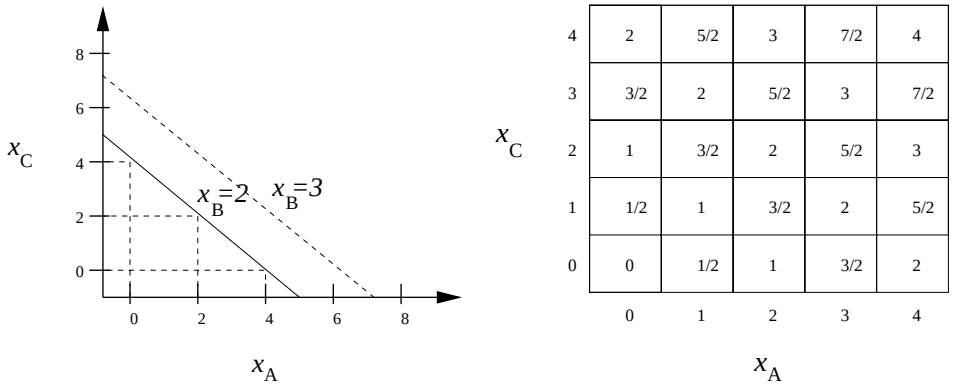


Figure 4.3: The features $p_A,\ p_B$ and p_C are projections of same 3D point.



a b

Figure 4.4: a) Valid correspondences must fulfill $x_B = (x_A + x_B)/2$. Then each value of x_B generates a line where the correspondences are valid. b) An auxiliary index matrix codifies the pixels of the three cameras, after discretizing the space of correspondences x_A versus x_C.

To generate uniquely the triplets of points for which the correlation computation is geometrically meaningful, let us concentrate our attention on a single set of three corresponding epipolar lines. Without loss of generality, for now on we will consider an image as a single horizontal line. Thus our problem consists of reconstructing the scene projected on a line of the camera.

Suppose now that we want to search the correspondences of the feature p_B of the camera B (with $x_B = 2$ — Figure 4.3) along the associated epipolar lines of cameras A and C. Valid correspondences must fulfill $x_B = (x_A + x_C)/2$, shown as a linear constraint in Figure 4.4a.

This property provides a significant reduction in the amount of correlations needed to find the correct correspondences. To incorporate this search constraint we have to represent the image data through a special form. By discretizing the space of correspondences x_C versus x_A in Figure 4.4a, an auxiliary index matrix is defined, represented in Figure 4.4b. Based on the integer values of this index matrix, one obtains the positions where the various correlations will be performed. The procedure is described as follows:

1. Discretize the corresponding epipolar lines of three cameras A, B and C, with the following indices as multiples of the same metric unit: $x_A = 0,\ 1,\ 2, ..., N$, $x_B = 0,\ 1,\ 2, ..., N$ and $x_C = 0,\ 1,\ 2, ..., N$.

2. Discretize the space of correspondences x_C versus x_A (Figure 4.4a), by constructing a matrix whose columns are indexed by x_A, and whose rows are indexed by x_C. The entries correspond to different values of x_B computed through the average between the row and the column positions (Figure 4.4b). Notice that only the integer entries are valid indices given the discretization made before. Each entry is then associated to a triplet of indices: x_A (column), x_C (row) and x_B (the average between row and column).

3. Finally, construct the correlation matrix $\boldsymbol{C_{AB}}$ by computing the correlations between images A and B, in the positions given respectively by x_A and x_B of the auxiliary index matrix built before. For the example shown in the Figure 4.4b, the following correlation matrix is obtained:

$$\boldsymbol{C_{AB}} = \begin{bmatrix} c_{02} & 0 & c_{23} & 0 & c_{44} \\ 0 & c_{12} & 0 & c_{33} & 0 \\ c_{01} & 0 & c_{22} & 0 & c_{43} \\ 0 & c_{11} & 0 & c_{32} & 0 \\ c_{00} & 0 & c_{21} & 0 & c_{42} \end{bmatrix} \tag{4.3}$$

   The index associated to each entry of $\boldsymbol{C_{AB}}$ indicates directly the coordinate positions respectively for the images A and B. The entries set to zero do no affect any further optimization (e.g. maximization) process and are maintained for matrix consistency reasons.

4. Similarly, we can build the correlation matrix $\boldsymbol{C_{CB}}$ by computing the correlations between the positions x_C and x_B. For the same example, we have:

$$\boldsymbol{C_{CB}} = \begin{bmatrix} c_{42} & 0 & c_{43} & 0 & c_{44} \\ 0 & c_{32} & 0 & c_{33} & 0 \\ c_{21} & 0 & c_{22} & 0 & c_{23} \\ 0 & c_{11} & 0 & c_{12} & 0 \\ c_{00} & 0 & c_{01} & 0 & c_{02} \end{bmatrix} \tag{4.4}$$

We have created an index structure that integrates all admissible pixel positions to compute the correlations among three corresponding epipolar lines. The resulting correlation matrices are denoted by $\boldsymbol{C_{AB}}$ and $\boldsymbol{C_{CB}}$, and the vectors $\boldsymbol{c}_{AB}$ and $\boldsymbol{c}_{CB}$ of the objective function (4.2) correspond to the respective vectorization $\boldsymbol{c}_{AB} = \mathrm{vec}(\boldsymbol{C_{AB}})$ and $\boldsymbol{c}_{CB} = \mathrm{vec}(\boldsymbol{C_{CB}})$.

To complete the definition of the objective function we have to generate the indicator vector $\boldsymbol{p}$. This is naturally provided by the correlation matrices built before. For each element of the correlation matrices we can assign an indicator element which can be 0, where no correspondence exists, or 1, where one verifies a correspondence, with this methodology, two binary matrices $\boldsymbol{P_{AB}}$ and $\boldsymbol{P_{CB}}$ are designed with the dimension of $\boldsymbol{C_{AB}}$ and $\boldsymbol{C_{CB}}$. The vector $\boldsymbol{p}$ is obtained, performing

$$\boldsymbol{p} = \begin{bmatrix} \mathrm{vec}(\boldsymbol{P_{AB}}) \\ \mathrm{vec}(\boldsymbol{P_{CB}}) \end{bmatrix}$$

In next section, we will study an optimization approach in order to maximize the function $z = \boldsymbol{p^T c}$, where $\boldsymbol{p}$ and $\boldsymbol{c}$ are built as described before. We will show then the importance of matrices $\boldsymbol{P_{AB}}$ and $\boldsymbol{P_{CB}}$ in developing a linear optimization strategy, by exploring the structure of the respective matrices.

4.2.2 An integer optimization approach

The simplest way to guarantee a solution for the correspondence problem defined before is to solve the following integer program:

$$z = \max\{\boldsymbol{c}^T\boldsymbol{p} : \ \boldsymbol{p} \in B^n\} \tag{4.5}$$

where B^n is the set of n-dimensional binary vectors. The trivial solution is found if all entries of $\boldsymbol{p}$ are 1. Thus we have to restrict the domain of the variable $\boldsymbol{p}$ by including additional constraints such as the uniqueness, the visibility and the camera geometric (transfer) constraints. If these constraints were linear with respect to the variables, the new integer program would have the following form:

$$z = \max\{\boldsymbol{c}^T\boldsymbol{p} : \ \boldsymbol{A}\boldsymbol{p} \leq \boldsymbol{b}, \ \boldsymbol{p} \in B^n\} \tag{4.6}$$

where $\boldsymbol{A}$ and $\boldsymbol{b}$ establish a generic linear constraint over $\boldsymbol{p}$. We show how to achieve a linear form for the constraints proposed in Section 4.1, based on matrices $\boldsymbol{P_{AB}}$ and $\boldsymbol{P_{CB}}$.

The first step consists in studying the necessary structure of these matrices in order to guarantee the various physical constraints, specifically the camera geometry, the uniqueness and the visibility constraints. Next we show how all these constraints have an algebraic equivalence on the matrices $\boldsymbol{P_{AB}}$ and $\boldsymbol{P_{CB}}$.

Uniqueness: The uniqueness constraint imposes that a given feature matches with at most one feature on other camera. Algebraically, this is equivalent to guarantee the following property:

Property 5 *The binary matrix* $\boldsymbol{P_{AB}}$ *has at most one logical value* 1 *per column and per diagonal*[5]*. Similarly, the binary matrix* $\boldsymbol{P_{CB}}$ *has at most one logical value* 1 *per row and per diagonal.*

Camera Geometry: The point transfer constraint can be guaranteed simply by observing the following property:

Property 6 *Three corresponding points observed in three images, produce the logical value* 1 *at the same matrix entry of both indicator matrices* $\boldsymbol{P_{AB}}$ *and* $\boldsymbol{P_{CB}}$*.*

Based on the index structure used in the construction of $\boldsymbol{P_{AB}}$ and $\boldsymbol{P_{CB}}$, the transfer relation $x_B = \frac{x_A + x_C}{2}$ of three corresponding points visible from three cameras is satisfied,

[5]The definition of diagonal of a matrix include all diagonals of the matrix (the main diagonal, the second diagonal, etc).

by guaranteeing that the values 1 are located at the same place for both indicator matrices.

Visibility: One observes a visibility constraint when one correspondence is found between a pair of images, imposing that a set of points is out of visibility in the remaining cameras. In logical terms, if a given element of $\boldsymbol{P_{AB}}$ has the value 1 (a correspondence), then some zeros are imposed in $\boldsymbol{P_{CB}}$ (non-visible points). Similarly, if a given element of $\boldsymbol{P_{CB}}$ is 1, then a set of zeros is imposed in $\boldsymbol{P_{AB}}$. The structure of this area of zeros depends on the areas which are out of visibility.

Consider again the example illustrated in Figure 4.2, where a correspondence between images A and B is detected (i.e. there is an entry of $\boldsymbol{P_{AB}}$ with value 1, say, $p = 1$). Transposing the geometric evidence shown in the figure to an algebraic description, the half-line Z_1 non visible by C corresponds to a set of non-correspondences in matrix $\boldsymbol{P_{CB}}$, with the following structure:

$$\boldsymbol{P_{CB}} = \begin{bmatrix} & \vdots & & \vdots & \cdots \\ \overbrace{}^{Z_1} & & & & \\ 0 \quad 0 & \cdots & 0 \quad 0 & q & \cdots \\ & \vdots & & \vdots & \cdots \end{bmatrix} \tag{4.7}$$

where q is the entry associated to the triplet $x_A,\ x_B,\ x_C$, coordinates of the projections of P (q will be zero if P is invisible from C and non-zero otherwise — in the example of the Figure 4.2, $q = 0$). The area pointed out in $\boldsymbol{P_{CB}}$ is forced to be zero because of a correspondence found in $\boldsymbol{P_{AB}}$. Fortunately, in matrix terms, this area is located uniquely along the row of $\boldsymbol{P_{CB}}$ associated to the entry (correspondence) considered in $\boldsymbol{P_{AB}}$. Similarly, if a correspondence is found in $\boldsymbol{P_{CB}}$, only the associated column of $\boldsymbol{P_{AB}}$ is affected.

These algebraic constraints can be stated rigorously as follows:

Property 7 *Define*

$$\boldsymbol{P_{AB}} = \{p_{ij}\}_{i=1:N,\ j=1:N}$$

$$\boldsymbol{P_{CB}} = \{q_{ij}\}_{i=1:N,\ j=1:N}$$

where ij are now the usual matrix indices. If $p_{ij} = 1$ for some ij, representing a correspondence between A and B, then $q_{il} = 0$ for all $l < j$, representing the invisible points of C where no correspondence is allowed. If $q_{ij} = 1$ for some ij, then $p_{kj} = 0$ for all $k < i$.

Now it is possible to describe the physical constraints proposed through an equivalent set of algebraic constraints on the two binary matrices: $\boldsymbol{P_{AB}} = \{p_{ij}\}_{i=1:N,\ j=1:N}$ and $\boldsymbol{P_{CB}} = \{q_{ij}\}_{i=1:N,\ j=1:N}$.

In the following steps, we derive explicitly these algebraic restrictions as linear constraints under the form of a set of inequalities[6]. This is achieved given the fact that only binary values are solution for the optimization problem ($p_{ij} \in \{0,\ 1\}$ and $q_{ij} \in \{0,\ 1\}$ for all $i = 1 : N$ and $j = 1 : N$).

1. First we introduce the uniqueness constraint with respect to the columns of $\boldsymbol{P_{AB}}$ and rows of $\boldsymbol{P_{CB}}$:

$$\sum_{i=1}^{N} p_{ij} \leq 1,\ j = 1 : N \tag{4.8}$$

$$\sum_{j=1}^{N} q_{ij} \leq 1,\ i = 1 : N \tag{4.9}$$

2. By introducing the auxiliary binary matrix $\boldsymbol{M} = \{m_{ij}\}_{i=1:N,\ j=1:N}$, where $m_{ij} \in \{0,\ 1\}$ for all $i,\ j$, the uniqueness with respect to the diagonals and Property 6 can be guaranteed simultaneously with the following inequalities:

$$p_{ij} \leq m_{ij},\ i = 1 : N,\ j = 1 : N \tag{4.10}$$

$$q_{ij} \leq m_{ij},\ i = 1 : N,\ j = 1 : N \tag{4.11}$$

$$\sum_{\forall i,j \in S_r} m_{ij} \leq 1,\ r = 1 : N \tag{4.12}$$

[6]However these linear constraints are not unique for the algebraic constraints imposed before to $\boldsymbol{P_{AB}}$ and $\boldsymbol{P_{CB}}$. The construction of an adequate set of inequalities for a given problem constitutes a fundamental issue in integer optimization field, beyond the scope of this work.

where S_r denotes the matrix diagonals: $S_r = \left\{\forall\ i,\ j \in \{1, \cdots, N\} :\ \frac{N-i+j+1}{2} = r\right\}$.

Some diagonals of $\boldsymbol{P_{AB}}$ and $\boldsymbol{P_{CB}}$ are filled with zeros due to the discretization process (associated to the zero diagonals of $\boldsymbol{C_{AB}}$ and $\boldsymbol{C_{CB}}$, as shown in example (4.3)). This can be simply imposed by[7]: $m_{ij} = 0, \forall i, j \in \{1, \cdots, N\} :\ \frac{N-i+j}{2} \in Z^+$.

3. Finally the visibility constraints are provided by:

$$\sum_{l=1}^{j-1} q_{il} + p_{ij} \quad \leq \quad 1,\ i = 1:N,\ j = 2:N \tag{4.13}$$

$$\sum_{l=1}^{i-1} p_{lj} + q_{ij} \quad \leq \quad 1,\ i = 2:N,\ j = 1:N \tag{4.14}$$

By manipulating the set of inequalities (4.8-4.14), we derive the linear constraint:

$$\boldsymbol{A}\begin{bmatrix} \boldsymbol{p} \\ \boldsymbol{m} \end{bmatrix} \leq \boldsymbol{b} \qquad \text{where } \boldsymbol{p} = \begin{bmatrix} \text{vec}(\boldsymbol{P_{AB}}) \\ \text{vec}(\boldsymbol{P_{CB}}) \end{bmatrix},\ \boldsymbol{m} = \text{vec}(\boldsymbol{M}).$$

Notice that the matrices $\boldsymbol{A}$ and $\boldsymbol{b}$ we derived admit uniquely the values -1, 0 and 1, and their structure is not unique.

In summary, we synthesized an integer program with the following form:

$$z = \max\{\boldsymbol{c}^T\boldsymbol{p} :\ \boldsymbol{A}\begin{bmatrix} \boldsymbol{p} \\ \boldsymbol{m} \end{bmatrix} \leq \boldsymbol{b},\ \boldsymbol{p} \in B^{2N^2},\ \boldsymbol{m} \in B^{N^2}\} \tag{4.15}$$

which yields the optimal solution for the collinear trinocular problem. In the following section, we discuss briefly some approaches for solving this integer program and the method adopted in this work.

4.3 Solving the integer program

To compute the optimal solution for a generic integer program, four dominant approaches can be found in the optimization literature (an important insight into this issue is found in [44]):

[7] In practice, these a priori imposed zero elements of $\boldsymbol{P_{AB}}$, $\boldsymbol{P_{CB}}$ and $\boldsymbol{M}$ must be removed before applying the optimization algorithm, reducing greatly the problem dimensionality.

1. An algorithmic approach for solving directly the integer program — The most known algorithms are the branch-and-bound and the cutting-plane algorithms. They are iterative and, in general, excessively time consuming. *The emphasis is in the architecture of the search algorithm.*

2. A linear program for solving the integer program — This approach consists in finding an equivalent linear version of the integer program, by manipulating the linear constraints of the problem. This can only be achieved in some special cases. Then, efficient algorithms can be used to solve the linear program for the optimal solution. Alternatively, one may define a linear relaxation of the integer program, which can lead to a non-optimal solution[8]. *The emphasis here is in the definition of an equivalent linear program.*

3. A non-linear program for solving the integer program — This approach consists in introducing additional non-linear constraints, producing an equivalent problem, where the search algorithms converge more efficiently. An ingenious example is presented in [20]. *The emphasis is in the definition of an equivalent non-linear program.*

4. A graph approach for solving the integer program — This approach consists in rewriting the original problem in the graph form, for which there are special methodologies for solving it. However this approach is not valid for all integer problems. *The emphasis is in the problem definition.*

All these approaches produce optimal solutions of the integer program. However, for the specific integer program presented in (4.15), no efficient approach was found to achieve the optimal solution with feasible computation. An exception can be found in the equivalent non-linear programming approach [20], where a linear function is minimized under a set of linear contraints plus a concave constraint. In the future, we plan to implement the proposed search algorithm in order to test the method. In this work, for simplicity (in terms of the problem formulation and of the algorithms available to solve it), we adopted the following

[8]Notice that, theoretically, it is always possible to find an equivalent linear program which produce the solutions of the integer program. However, finding the corresponding contraints can be per se a combinatorial problem.

linear relaxation of the integer program:

$$z = \max\{\boldsymbol{c}^T\boldsymbol{p} : \ \boldsymbol{A}\begin{bmatrix}\boldsymbol{p}\\ \boldsymbol{m}\end{bmatrix} \leq \boldsymbol{b},\ \mathbf{0} \leq \boldsymbol{p} \leq \mathbf{1},\ \mathbf{0} \leq \boldsymbol{m} \leq \mathbf{1}\} \tag{4.16}$$

In this case, $\boldsymbol{p}$ is no longer constrained to be an integer value. However, it is constructive to discuss the structure of the constraints. In fact, the constraints of the original integer program denote the integer vertices of a polyhedron. The solution of the problem is located at one of these vertices. With the linear relaxation presented in (4.16), we add a finite number of non-integer vertices to that polyhedron. The integer vertices are still vertices of the new polyhedron. This means that, when a search algorithm is applied to find the optimal solution within this polyhedron, if an integer vertex is found, then the optimal solution of (4.15) is achieved. However a non-integer vertex can eventually be found. Thus a "good" relaxation of the problem must introduce a reduced amount of non-integer vertices in order to retrieve (with high probability) an integer result when the optimization algorithm is applied.

Next we present some results by using a collinear trinocular system in several real scenes. We apply the well-known *simplex* algorithm to solve the linear relaxation of the integer program, presented in (4.16). The performance of the method will be discussed.

4.4 Experiments

In previous sections we propose a optimization approach to estimate the disparity map of a scene. To show how to retrieve disparity from the solution of the integer optimization problem, we start with an example using three corresponding epipolar lines from three consecutive images of the well-known Flower-garden sequence (Figure 4.5).

In this example, the disparity between each pair of images is significant, and a relevant amount of occlusions are present. We propose to find the correspondences of the points marked in the figure (53 points per image), which correspond to the discretization of one horizontal epipolar line (associated to the vertical coordinate 60). The correlation is performed by using a 3×3 window centered at the 53 points of each image. We apply a simplex algorithm to solve the linear relaxation of the integer optimization program, as presented in (4.16). The

solutions computed for the variables contained in the vector $\boldsymbol{p}$ are integer (0 and 1), which means that they are also solution of the integer problem presented in (4.15).

Thus the optimization algorithm has produced the indicator matrices $\boldsymbol{P_{AB}}$ and $\boldsymbol{P_{CB}}$, which determine the binary solution for the correspondence problem. The solution obtained, in Figure 4.6, is optimal, given the metric, uniqueness and visibility constraints.

Figure 4.5: Three consecutive images of the Flower-Garden sequence, where we selected 53 points along corresponding epipolar lines.

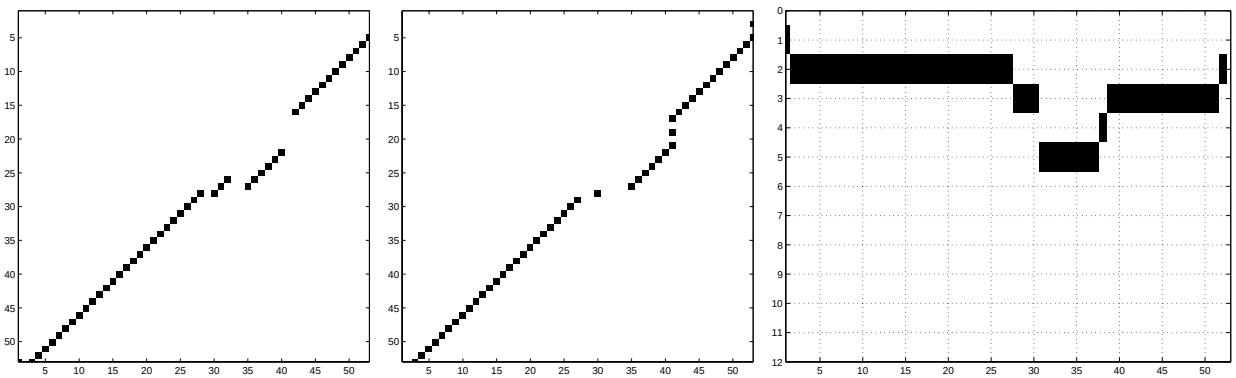


Figure 4.6: Optimal solution of the integer program. Left: $\boldsymbol{P_{AB}}$; Center: $\boldsymbol{P_{CB}}$; Right: rotation of $\boldsymbol{P_{AB}} \vee \boldsymbol{P_{CB}}$.

To get the depth information for the points selected in the central camera, we can apply the logical operator OR between both binary indicator matrices (joining the information provided by both correspondences), and perform a 45 degrees matrix rotation, in order to visualize better the disparity existent along the diagonal direction — notice that the epipolar line

coordinate of the central camera is aligned with the anti-diagonal of the matrices $\boldsymbol{P}$. The resulting matrix is shown in the Figure 4.6, where the disparity estimates codifies the depth information of the scene, for the points selected on the central image of the Figure 4.5.

This example illustrates how the algorithm retrieves disparity information for all points used in the experiment, including points without texture and points occluded by the tree for the left or right camera. This is due to the fact that the optimization process produces the most coherent solution given the available images.

Notice that even though we have used the linear relaxation of the integer optimization problem, optimal (integer) solutions were obtained. However, as discussed in the previous section, this might not happen always. This means that the relaxation of the integer program can produce some non-integer solutions, for which no correspondence configuration makes sense. To achieve the optimal solution of the integer program, a branch-and-bound approach [44] would be a possible technique. However this approach is iterative and excessively time consuming. A simple alternative consists then in considering uniquely the integer solutions of the linear program, discarding the non-integer ones. A similar strategy can be to round the non-integer to the nearest integer (0 or 1). For both cases, the resulting estimated $\boldsymbol{p}$ is sub-optimal and the correspondences found are not necessarily correct.

However, based uniquely on our experiments, we might say that such procedures produce a solution very close to the optimal one, because two typical situations were observed in our practical examples. First, in most cases the estimated solution values are all integer (which means that the solution is optimal). Secondly, when some non-integer values are found within the estimated $\boldsymbol{p}$-vector, the majority of the values (typically more than 90%) are integer and consistent with the real correspondences. Moreover, a simple post-processing step applied on the disparity image reduces usually the effect produced by eventual inconsistent solutions.

In the experiment of the Figure 4.7, we use a trinocular setup with a unique object. The most interesting feature of this example consists of the presence of a large amount of ambiguities (due to the texture of the object), which would represent an evident difficulty for an usual correspondence algorithm based on correlation and uniqueness constraints.

Due to computation time problems, we have applied a discretization procedure to the

epipolars of the images, with a period of 5 pixels. This means that the expected disparity has five times less resolution than if all points of the epipolar lines was considered. After applying the optimization approach described in the previous example, one obtains a disparity map for a set of epipolar lines. In this example, all solutions obtained with (4.16) are integer, and therefore optimal. Figure 4.8 shows the estimated 3-D model of the mask, computed for 100 equally spaced epipolar lines selected on the image plane. Each step of the graphic represents a 5-pixels variation of the disparity. In order to get more detailed relief information, the period of discretization performed along the epipolar lines must be reduced.

Figure 4.7: Images from a trinocular system, with equidistant cameras.

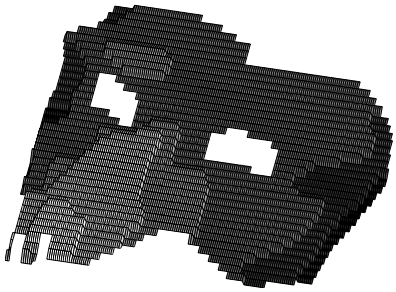


Figure 4.8: Optimal solution of the integer program, with several levels of disparity.

In the trinocular example of the Figure 4.9, there are multiple objects at various depths, and a large number of observable occlusions. We have applied the same methodology and some non-integer solutions have been observed. They were rounded to the nearest integer

(0 or 1) and a median filter was applied to the resulting disparity map in order to eliminate isolate outliers, preserving the discontinuities. Figures 4.10a-b show two arbitrary views of the estimated 3-D model of the scene, where brighter points correspond to larger disparities. A view with superimposed texture is shown in the Figure 4.10c.

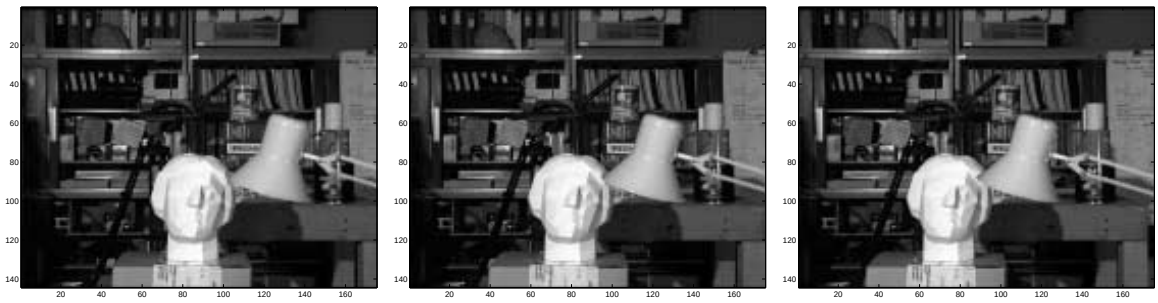
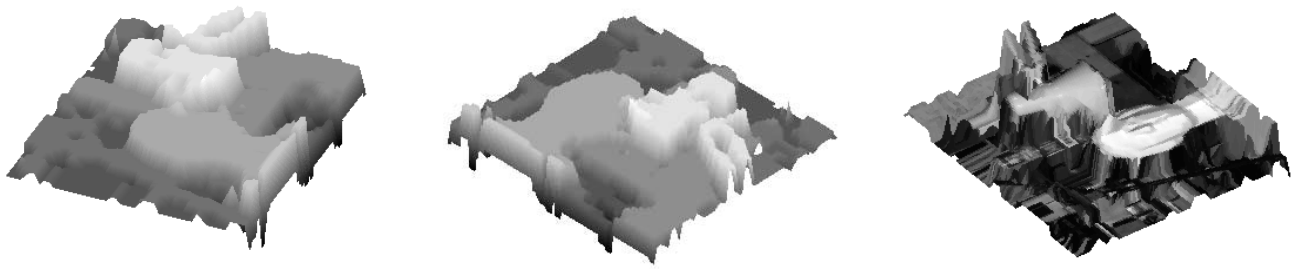


Figure 4.9: Example of a trinocular setup from University of Tsukuba's Images Database.



a b c

Figure 4.10: Arbitrary views around the reconstructed model. In c) the original texture was superimposed.

The previous examples illustrate the potentialities of an integer optimization approach to generate a dense disparity map of the scene, based on physically meaningful constraints. This approach was developed to a specific trinocular setup, for which encouraging results with real scenes were provided.

The major practical limitation found was the time of computation, which depends naturally on the number of variables (number of points selected). To overcome this problem, some future effort is needed, mainly on the design of an algorithm for solving the integer

program[9], by exploring the special structure of this problem. This can be achieved through two complementary strategies:

- Redefining the linear constraints of the problem presented in Equation (4.15) given by the matrix $\boldsymbol{A}$ and the vector $\boldsymbol{b}$ (since they are not unique for the same physical constraints).
- Designing a specific search algorithm for solving the integer program (since the expected solutions are integer and the matrices $\boldsymbol{A}$ and $\boldsymbol{b}$ admit uniquely the values -1, 0 or 1).

4.5 Conclusion

In this chapter we developed a complete optimization approach to find the correspondences of a collinear trinocular system, by using a set of physically meaningful constraints. The method we propose is based on a linear objective function and a set of linear constraints, simplifying the algorithmic tools required to solve it. The correspondence and occlusion detection are performed in a single step, achieving good and promising results with real images.

The optimization method was developed uniquely to a specific camera configuration: three equidistant and collinear cameras. One of the advantages of using the equidistant collinear trinocular case is that there are uniquely three directions in the indicator matrices we have to constrain: the rows, the columns and the diagonals. Other camera configurations lead to other algebraic constraints, applied to other directions of the matrices, raising additional discretization problems. Thus, one of the problems in defining an optimization approach for arbitrary cameras corresponds to the difficulty in representing conveniently the algebraic constraints of the indicator matrices, since the metric, uniqueness and visibility constraints remains valid.

As future work, we plan to develop an optimal scheme for solving completely the integer program, based on an equivalent linear or non-linear program. Actually, this represents more an optimization problem than a vision problem, being however very relevant because it influences directly the feasibility of the methods. The integer optimization is a mathematical field with many and exhaustive studies, using a large variety of sophisticated techniques, on both

[9] In this work we have used a standard simplex algorithm to solve the linear relaxation of the integer program.

linear and non-linear (e.g. concave) minimization. If one wants to achieve a single universal representation of the correspondence problem, handling physical meaningful assumptions in an unique and optimal formulation, then it is absolutely necessary to notice the existence of such mathematical tools. In this chapter we have tried to achieve optimal specifications for a trinocular camera configuration. This shows a promising path in direction to the complete reconstruction from multiple views, but an additional effort in image representation and optimization is needed.

Summary

In this part, we approach the correspondence problem, assuming that the cameras are completely calibrated. Finding correspondences is a central procedure in stereo and motion analysis. To perceive depth, we need to perform dense matching between two or more images, captured over time. Moreover, occlusions play a key role in motion and depth interpretation.

We propose two approaches where the global image information and the similarity and consistency criteria are well defined. Such approaches can be used to generate a dense disparity map or different views of the same scene.

In the first approach, we propose a new image representation called *Intrinsic Images* that can be used to solve correspondence problems within a simple and compact framework. The concept of intrinsic images is a useful way to approach the problem of stereo vision, and leads to a straightforward computation of disparity and new views. Intrinsic images combine photometric and geometric descriptors of a stereo image pair. The photometric descriptors are invariant to perspective image distortions, while the geometric descriptors can be used to compute directly disparities in a dense manner. We extend this framework to deal with occlusions through an optimization technique based on dynamic programming.

In the second approach, we propose a methodology for solving the point correspondence problem for more than two views, imposing physically meaningful constraints. First we explore and discuss these constraints, focusing our attention on the metric, uniqueness and visibility constraints. Next we propose to study a paradigmatic example where the correspondence and occlusion detection is formulated as an integer optimization problem. In contrast to the previous approach, where a simple and direct strategy was adopted to retrieve 3-D information from two views, here we develop a method for a special camera configuration

(three collinear cameras) which performs correspondence and occlusion detection in a single step, and achieves global optimality for a set of realistic constraints, without having to impose additional assumptions. The proposed method is based on a linear objective function and on a set of linear constraints, simplifying the algorithmic tools to solve it.

In the future, we plan (1) to study the potentialities of the Intrinsic Images applied to a larger number of images, namely in dense optical flow computation and egomotion estimation; (2) to develop an automatic occlusion characterization based on a large number of images of same scene (extension to the continuous), in order to facilitate the creation of the associated intrinsic images with occlusion information; (3) to apply an optimization approach to a trinocular system with cameras in arbitrary location; (4) to explore non-linear techniques for solving more efficiently the integer optimization problems; (5) to study a possible representation for the optimization problem based on the intrinsic image paradigm.

Conclusion

In this thesis, we have addressed the problem of 3D reconstruction from video images. We proposed several methodologies for two central problems in computer vision: camera motion estimation and dense matching or depth reconstruction.

The first part of the thesis is dedicated to the problem of egomotion estimation. The approach is based on the analysis of the image motion information and the definition of a set of subspaces, where some constraints may be exploited, that allow us to use the normal flow data globally to estimate the camera motion. Additionally, we have shown how occlusions can be treated as a powerful perceptual cue to provide additional information about camera motion, instead of being considered undesirable artifacts, as in other approaches.

The second part of the thesis is devoted to the problem of depth reconstruction or dense matching, assuming known egomotion. Dense matching is a central and difficult problem in computer vision, that has been tackled in many ways. Our approach is based on alternative representations that may facilitate *ad initio* the correspondence problem. We proposed a new image representation called *Intrinsic Images* that can be used to solve the stereo correspondence problem within a simple framework. This methodology assumes the order constraint and can be extended to cope with brightness changes and occlusions. We show how to compute dense disparity maps and synthesize novel views from an image pair directly from the *Intrinsic Image* representations. Also, the approach provides a coherent interpretation of occluded regions if the order constraint is not violated.

To overcome the need of using the order constraint, we propose another approach based on more than two views. Here, we rely solely on general constraints (geometric, uniqueness and visibility) of the stereo problem and express them in such a way that the correspondence problem can be formulated as an integer optimization problem. Optimal solutions are obtained, even in the presence of occlusions.

Throughout the thesis, we show extensive results both for camera egomotion estimation and depth reconstruction to illustrate the performance of the approaches proposed here.

There are a number of aspects present along the entire thesis that we would like to emphasize (again) for the reader:

1. Occlusions have almost always been seen as disrupting phenomena for many computer

vision problems. Instead, they provide a lot of information about camera motion and scene depth that can be exploited with important benefit. As an example we can refer the fact that occlusions are uniquely caused by camera translation hence, naturally decoupling camera rotation and translation.

Similarly, if occlusions are included in the reconstruction output, one obtains much better results, in terms of perceptual quality, as opposed to other methods that simply blur over occluding surfaces. Morevoer, important constraints are introduced when considering occlusions directly in the matching problem or for computing depth maps.

2. Representations for visual perception are also a constant across the entire thesis. We have shown how alternative representations (that we named topological subspaces) can be used for computing the camera egomotion in an efficient and robust manner. We have proposed the *Intrinsic Images* approach as another representation to code photometric and geometric image transformations that make the matching problem and view synthesis a lot simpler. Similarly, we have defined an holistic method to represent the stereo problem and associated valid constraints (i.e. no approximations), as an integer representation problem, thus making use of a large set of tools from optimization theory.

Whether or not some of this aspects may find some parallelism in our own visual system remains unraveled and may be a subject for further endeavor and research in the years to come.

Still, from an engineering point of view, one may argue that being able to correctly formulating the problem under a suitable representation, and making use of every possible piece of information, is a decisive step towards better vision systems and contribute to a better understanding of visual perception.

Bibliography

[1] Y. Aloimonos and Z. Duric. Estimating the heading direction using normal flow. *Int. Journal of Computer Vision*, 13(1):33–56, September 1994.

[2] Y. Aloimonos, I. Weiss, and A. Banddophaday. Active vision. *Int. Journal of Computer Vision*, 1(4):333–356, January 1988.

[3] B. L. Anderson and K. Nakayama. Toward a general theory of stereopsis: Binocular matching, occluding contours, and fusion. *Psych. Review*, 101:414–445, 1994.

[4] S. Beauchemin, A. Chalifour, and J. Barron. Discontinuous optical flow: Recent theoretical results. In *Vision Interface (VI97)*, Kelowna, B. C., May 1997.

[5] A. Bruss and B.K.P. Horn. Passive navigation. *Computer Vision, Graphics and Image Processing*, 21(1):3–20, January 1983.

[6] T. Cormen, C. Leiserson, and R. Rivest. *Introduction to Algorithms*. The MIT Press, 1990.

[7] I. Cox, S. Hingorani, B. Maggs, and S. Rao. A maximum likelihood stereo algorithm. *CVIU*, 63(3):542–567, 1996.

[8] K. Daniilidis. Fixation simplifies 3d motion estimation. *CVIU*, 68(2):158–169, November 1997.

[9] U. R. Dhond and J. K. Aggarwal. Stereo matching in the presence of narrow occluding objects using dynamics disparity search. *PAMI*, 17(7):719–724, July 1995.

[10] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. Wiley, NY, USA, 1973.

[11] P. Winston (ed.). *The Psychology of Computer Vision*. McGraw-Hill, NY, USA, 1973.

[12] O. Faugeras. *Three-Dimensional Computer Vision*. The MIT Press, 1993.

[13] O. Faugeras, F. Lustman, and G. Toscani. Motion and structure from motion from point and line matches. In *Proc. 1st Intern. Conf. Comput. Vision*, London, June 1987.

[14] O. Faugeras and S. Maybank. Motion from point matches: Multiplicity of solutions. *International Journal of Computer Vision*, 4:225–246, 1990.

[15] O.D. Faugeras, Q.T. Luong, and S.J. Maybank. Camera self-calibration: Theory and experiments. *ECCV*, 588:563–578, 1992.

[16] C. Fermuller. Passive navigation as a pattern recognition problem. *Int. Journal of Computer Vision*, 14(2):147–158, March 1995.

[17] C. Fermuller. Qualitative egomotion. *IJCV*, 15(1/2):7–29, June 1995.

[18] C. Fermuller and Y. Aloimonos. The role of fixation in visual motion analysis. *Int. Journal of Computer Vision*, 11(2):165–186, October 1993.

[19] D. J. Fleet, A. Jepson, and M. Jenkin. Phase-based disparity measurement. *CVGIP Image Understanding*, 53(2):198–210, 1991.

[20] C. Floudas and P. Pardalos (eds). *Recent Advances in Global Optimization*. Princeton University Press, 1992.

[21] D. Geiger, B. Landendorf, and A. Yuille. Occlusions and binocular stereo. *Int. Journal of Computer Vision*, 14(3):211–226, 1995.

[22] J.J. Gibson. *The perception of the visual world*. Houghton-Mifflin, 1950.

[23] J.J. Gibson. *The senses considered perceptual systems*. Houghton-Mifflin, 1966.

[24] R. Haralick and L. Shapiro. *Computer and Robot Vision*, volume 2. Addison-Wesley, 1993.

[25] R. Hartley, R. Gupta, and T. Chang. Stereo from uncalibrated cameras. In *IEEE CVPR*, pages 761–764, 1992.

[26] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[27] D. Heeger and A. Jepson. Subspace methods for recovering rigid motion i: Algorithm and implementation. *Int. Journal of Computer Vision*, 7(2):95–117, January 1992.

[28] B.K.P. Horn. *Robot Vision*. The MIT Press, 1986.

[29] B.K.P. Horn. Motion fields are hardly ever ambiguous. *International Journal of computer Vision*, 1(3):259–274, 1987.

[30] B.K.P. Horn and B. Shunck. Determining optical flow. *Art. Intelligence*, 17:185–203, 1981.

[31] B.K.P. Horn and E.J. Weldon. Direct methods for recovering motion. *International Journal of computer Vision*, 2(1):51–76, 1988.

[32] R. Hummel and V. Sundareswaran. Motion parameter estimation from global flow field data. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 15(5):459–476, May 1993.

[33] M. Irani, P. Anandan, and M. Cohen. Direct recovery of planar-parallax from multiple views. In *ICCV'99 Workshop: Vision Algorithms*, 1999.

[34] H. Ishikawa and D. Geiger. Occlusions, discontinuities, and epipolar lines in stereo. In *Proc. of the 5th European Conference on Computer Vision*, Germany, June 1998.

[35] J. Koenderink and A. von Doorn. Invariant properties of the motion parallax field due to the movement of rigid bodies to an observer. *Optica Acta*, (22):773–791, 1975.

[36] K. Kutulakos and S. Seitz. A theory of shape by space carving. In *Proc. of Seventh International Conference on Computer Vision*, 1999.

[37] H. Longuet-Higgins and K. Prazdny. The interpretation of a moving retinal image. *Proc. Roy. Soc. Lond. Ser. B*, (208):385–397, 1980.

[38] M. Lourakis. Egomotion estimation using quadruples of collinear image points. In *Proc. of European Conference on Computer Vision*. Springer-Verlag, 2000.

[39] Y. Ma, J. Kosecka, and S. Sastry. Linear differential algorithm for motion recovery: A geometric approach. *Int. Journal of Computer Vision*, 36(1):71–89, January 2000.

[40] J. Maciel and J. Costeira. Point correspondence by concave minimization. In *Proc. of British Machine Vision Conference*, 2000.

[41] D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman, San Francisco, USA, 1982.

[42] J. Munday and A. Zisserman (eds.). *Geometric Invariance in Computer Vision*. The MIT Press, 1992.

[43] H. Nagel and A. Gehrke. Spatiotemporally adaptive estimation and segmentation of of-fields. In *Proc. of the 5th European Conference on Computer Vision*, Germany, June 1998.

[44] G. Nemhauser and L. Wolsey. *Integer and Combinatorial Optimization*. John Wiley & Sons, Inc., 1988.

[45] Y. Ohta and T. Kanade. Stereo by intra- and inter-scanline search using dynamic programming. *PAMI*, 7(2):139–154, 1985.

[46] T. Papadopoulo. *Motion Analysis of 3D Rigid Curves from Monocular Image Sequences*. PhD Thesis, INRIA Sophia-Antipolis, France, 1995.

[47] M. Pollefeys, R. Koch, and L. Van Gool. A simple and efficient rectification method for general motion. In *ICCV*, pages 496–501, 1999.

[48] P.J. Rousseeuw and A.M. Leroy. *Robust Regression & Outlier Detection*. John Wiley & Sons, Inc, 1987.

[49] S. Roy and I. Cox. A maximum-flow formulation of the n-camera stereo correspondence problem. In *Proc. of Sixth International Conference on Computer Vision*, 1998.

[50] J. Santos-Victor, G. Sandini, F. Curotto, and S. Garibaldi. Divergent stereo in autonomous navigation : From bees to robots. *Int. Journal of Computer Vision*, 14(2):159–177, 1995.

[51] A. Shashua. Projective structure from uncalibrated images: Structure-from-motion and recognition. *PAMI*, 16(8):778–790, August 1994.

[52] A. Shashua and M. Werman. Trilinearity of three perspective views and its associated tensor. In *ICCV95*, pages 920–925, 1995.

[53] C. Silva. *Estimação do movimento próprio de um observador monocular*. Master Thesis, IST-Lisbon, Portugal, 1996.

[54] C. Silva and J. Santos-Victor. Direct egomotion estimation. In *Proc. of the 13th Int. Conference on Pattern Recognition*, Vienna,Austria, August 1996.

[55] C. Silva and J. Santos-Victor. Geometric approach for egomotion estimation using normal flow. In *Proc. of the 4th Int. Symposium on Intelligent Robotic Systems*, Lisbon,Portugal, July 1996.

[56] C. Silva and J. Santos-Victor. Robust egomotion estimation from the normal flow using search subspaces. *PAMI*, 19(9):1026–1034, September 1997.

[57] C. Silva and J. Santos-Victor. Egomotion estimation on a topological space. In *Proc. of the 14th Int. Conference on Pattern Recognition*, Brisbane,Australia, August 1998.

[58] C. Silva and J. Santos-Victor. Motion from occlusions. In *Proc. of 7th International Symposium on Intelligent Robotic Systems*, 1999.

[59] C. Silva and J. Santos-Victor. Intrinsic images for dense stereo matching with occlusions. In *Proc. of the 6th European Conference on Computer Vision*, Dublin, Ireland, 2000.

[60] W. B. Thompson. Exploiting discontinuities in optical flow. *International Journal of Computer Vision*, 30(3):163–174, 1998.

[61] C. Tomasi and R. Manduchi. Stereo without search. In *Proc. of European Conference on Computer Vision*. Springer-Verlag, 1996.

[62] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Proc. of the Sixth International Conference on Computer Vision*, Bombay, India, January 1998.

[63] P. Torr and A. Zisserman. Robust parameterization and computation of the trifocal tensor. *Image and Vision Computing*, 15:591–605, 1997.

[64] Bill Triggs. Autocalibration from planar scenes. In *EVVC (1)*, pages 89–105, 1998.

[65] J. Wang and E. Adelson. Layered representation for motion analysis. In *Proc. of the IEEE Comp. Vis. and Pat. Rec. (CVPR)*, New York, USA, June 1993.

[66] J. Weng, N. Ahuja, and T. Huang. Matching two perspective views. *PAMI*, 14(8):806–825, 1992.

[67] L. Zelnik-Manor and M. Irani. Multi-frame estimation of planar motion. *PAMI*, 22(10):1105–1116, 2000.

[68] C. Zitnick and T. Kanade. A cooperative algorithm for stereo matching and occlusion detection. *IEEE Trans. on PAMI*, 22(7), July 2000.