SHADOW SEGMENTATION AND TRACKING IN REAL-WORLD CONDITIONS

THÈSE N^{O} 3076 (2004)

PRÉSENTÉE À LA FACULTÉ SCIENCES ET TECHNIQUES DE L'INGÉNIEUR

Institut de traitement des signaux

SECTION D'ÉLECTRICITÉ

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Elena SALVADOR

laurea in ingegneria elettronica, Università degli studi di Trieste, Italie
et de nationalité italienne

acceptée sur proposition du jury:

Prof. T. Ebrahimi, directeur de thèse
Dr H. Nicolas, rapporteur
Prof. G. Sicuranza, rapporteur
Prof. S. Süsstrunk, rapporteur

Lausanne, EPFL
2004

scielzi nol è riscjâ
al è vivi

Leonardo Zanier

Contents

Acknowledgments

This text represents what is visible of the Ph.D. work I carried out in the last four years at the Signal Processing Institute. During this time I received advice, help and support from many people. I am grateful to all of them for the contribution they gave to this work.

In particular, I wish to thank Prof. Murat Kunt who gave me the opportunity to join the Signal Processing Institute, a truly unique research environment. My greatest thanks goes to my advisor Prof. Touradj Ebrahimi, who welcomed me in his group and supported my research with constant encouragement. I appreciated his wise advice and timely suggestions.

I am grateful to all the members of my thesis committee, Dr. Henri Nicolas, Prof. Sabine Süsstrunk, and Prof. Giovanni Sicuranza, for their interest in reading and discussing my work. I owe a special thanks to Prof. Sicuranza who first introduced me to signal and image processing and gave me the opportunity to carry out my diploma project at EPFL. Without this event I would probably never have done my Ph.D. work.

I would like to thank Prof. Pierre Vandergheynst, who read a draft of this thesis, for his valuable comments and suggestions. Elisa Drelie Gelasca, Gianluca Monaci, Rosa Maria Figueras i Ventura, and Meritxell Bach Cuadra, with their careful reading and opportune revisions of drafts of my thesis chapters, helped me shape the text. I thank them all for their precious support.

Part of this thesis was developed in the framework of the European project *art.live*. I would like to acknowledge here the valuable contribution to this work given by the collaboration and the interaction with the project's partners. For their contribution, I would also like to thank the students I had the opportunity to supervise during my Ph.D.

I enjoyed working with my colleagues at the Signal Processing Institute. The international mix of this group has made this a very special experience. For this, I would like to thank all the members of the Institute. In particular, I worked together closely with Andrea Cavallaro. The numerous fruitful discussions we had, his encouragement and advice have been of primary importance both for my diploma project and my Ph.D. thesis.

My sincere gratitude goes to all my friends who shared with me these years in Lausanne, in particular to Elisa, Rosa, Oscar, Patricia, Meri, Iva, Lorenzo, Gloria and Peggy, and Mary and Francesco back home in Italy, who have been closer to me and supported me especially in bad times, when I was discouraged or frustrated of all things that did not work.

Finally, my warmest thanks goes to Gianluca and to all the members of my family. Without their love and continuous support I would have never been able to carry out this Ph.D. Thank you for encouraging and helping me in every situation.

Abstract

Visual information, in the form of images and video, comes from the interaction of light with objects. Illumination is a fundamental element of visual information. Detecting and interpreting illumination effects is part of our everyday life visual experience. Shading for instance allows us to perceive the three-dimensional nature of objects. Shadows are particularly salient cues for inferring depth information. However, we do not make any conscious or unconscious effort to avoid them as if they were an obstacle when we walk around. Moreover, when humans are asked to describe a picture, they generally omit the presence of illumination effects, such as shadows, shading, and highlights, to give a list of objects and their relative position in the scene.

Processing visual information in a way that is close to what the human visual system does, thus being aware of illumination effects, represents a challenging task for computer vision systems. Illumination phenomena interfere in fact with fundamental tasks in image analysis and interpretation applications, such as object extraction and description. On the other hand, illumination conditions are an important element to be considered when creating new and richer visual content that combines objects from different sources, both natural and synthetic. When taken into account, illumination effects can play an important role in achieving realism.

Among illumination effects, shadows are often integral part of natural scenes and one of the elements contributing to naturalness of synthetic scenes. In this thesis, the problem of extracting shadows from digital images is discussed. A new analysis method for the segmentation of cast shadows in still and moving images without the need of human supervision is proposed. The problem of separating moving cast shadows from moving objects in image sequences is particularly relevant for an always wider range of applications, ranging from video analysis to video coding, and from video manipulation to interactive environments. Therefore, particular attention has been dedicated to the segmentation of shadows in video. The validity of the proposed approach is however also demonstrated through its application to the detection of cast shadows in still color images.

Shadows are a difficult phenomenon to model. Their appearance changes with changes in the appearance of the surface they are cast upon. It is therefore important to exploit multiple constraints derived from the analysis of the spectral, geometric and temporal properties of shadows to develop effective techniques for their extraction. The proposed method combines an analysis of color information and of photometric invariant features to a spatio-temporal verification process. With regards to the use of color information for shadow analysis, a complete picture of the existing solutions is provided, which points out the fundamental assumptions, the adopted color models and the link with research problems such as computational color constancy and color invariance. The proposed spatial verification does not make any assumption about scene geometry nor about object shape. The temporal analysis is based on a novel shadow tracking technique. On the basis of the tracking results, a temporal reliability estimation of shadows is proposed which allows to discard shadows which do not present time coherence. The proposed approach is general and can be applied

to a wide class of applications and input data.

The proposed cast shadow segmentation method has been evaluated on a number of different video data representing indoor and outdoor real-world environments. The obtained results have confirmed the validity of the approach, in particular its ability to deal with different types of content and its robustness to different physically important independent variables, and have demonstrated the improvement with respect to the state of the art. Examples of application of the proposed shadow segmentation tool to the enhancement of video object segmentation, tracking and description operations, and to video composition, have demonstrated the advantages of a shadow-aware video processing.

Version abrégée

Tout au long de sa vie, l'être humain reçoit un flot continu d'informations visuelles, dues à l'interaction de la lumière et de la matière. L'analyse des phénomènes résultant de cette interaction nous apporte des informations essentielles sur notre environnement. Ainsi, l'ombre, résultat le plus évident, le plus immédiatement perceptible de cette interaction, nous permet de concevoir la notion de profondeur; dans le même ordre d'idée, le fait même que les objets soient ombrés nous permet d'appréhender leur nature tri-dimensionnelle. Bien qu'essentiel, ce type d'information n'est que rarement pris en compte de façon consciente. Ainsi, nul ne fera d'effort particulier pour éviter lesdites ombres, comme si elles constituaient un obstacle à la poursuite de notre route. Plus frappant encore, lors de la description d'une image ou d'une scène, une liste des objets sera immédiatement établie mais il ne sera que rarement fait mention des effets liés à l'illumination que l'on peut y percevoir, tels qu'ombres ou reflets.

Pour les systèmes de vision par ordinateur, traiter l'information visuelle d'une façon similaire au système visuel humain, c'est-à-dire en prenant en compte également les effets liées à l'illumination, est une gageure. Ces phénomènes sont en effet plutôt gênants dans le cadre des tâches courantes en analyse d'images, telles que segmentation ou description d'objets. En revanche, considérer ce type d'information est essentiel pour donner à une scène tout son réalisme lorsqu'il s'agit de créer de nouveaux contenus visuels par combinaison d'objets issus de différentes sources, naturelles ou artificielles.

Peu nombreuses sont les scènes naturelles dont les ombres sont absentes, et, *a contrario*, leur absence dans une scène artificielle rend celle-ci fort peu réaliste. Le travail décrit dans cette thèse s'attache à résoudre le problème de l'extraction d'ombres au sein des images numériques. Une méthode de segmentation nouvelle des ombres portées, sans supervision humaine, est proposée, tant pour des images fixes que pour des images animées. Toutefois une attention plus particulière a été portée à ce dernier cas, en raison des applications potentielles croissantes dans lesquelles une telle segmentation constituerait un apport notable, en allant de l'analyse ou de la manipulation du contenu de la vidéo à son codage, en passant par les environnements interactifs.

L'ombre est un phénomène difficile à modéliser. Son apparence varie en fonction des surfaces sur lesquelles elle est projetée. Il est donc important d'exploiter les multiples propriétés dérivant des analyses spectrale, géométrique et temporelle des ombres afin de développer des techniques efficaces conduisant à leur extraction. Pour ce faire, la méthode proposée combine une analyse de l'information couleur et de caractéristiques photométriques invariantes, à un processus de vérification spatio-temporel. Un exposé complet des solutions existantes reposant sur l'utilisation de l'information couleur est tracé, précisant les présupposés fondamentaux, les modéles de couleur adoptés, ainsi que les liens avec certains problémes de recherches tels que la constance ou l'invariance de couleur. L'analyse temporelle est pour sa part basée sur une technique de suivi d'ombre inédite grâce à laquelle une estimation de la fiabilité des ombres détectées au cours du temps permet d'écarter les

résultats ne présentant pas de cohérence temporelle. Finalement, la vérification spatiale proposée ne se fonde sur aucune hypothèse a priori quant à la géométrie de la scéne ou à la forme de l'objet à extraire. Par conséquent, la méthode est générale et peut être utilisée pour un large éventail d'applications, avec des types de données divers.

L'évaluation de l'approche s'est faite au travers d'un certain nombre de vidéos représentatives d'environnements réels, aussi bien intérieurs qu'extérieurs. Les résultats obtenus ont confirmé la validité de la méthode, notamment sa capacité à composer avec des contenus variés ainsi que sa robustesse face à différentes variables physiques. Une comparaison à l'état de l'art a permis de mettre en évidence ses apports dans le domaine. Pour finir, l'outil de segmentation d'ombre proposé a été mis en oeuvre dans différents exemples d'applications telles que la composition vidéo, ou l'aide à la segmentation, au suivi et à la description d'objets animés, pour lesquelles l'utilisation de l'information que constitue l'ombre s'est révélé avantageux.

1 Introduction

1.1 Motivations

We are nowadays witnessing to the widespread diffusion of visual information. The production of digital images and digital video, as well as the use of computer vision systems, has been made easier by the advent of digital technologies and the improved computational capability of computers, together with the diffusion of digital cameras and the advances in storage and networking. Technology progresses are at the same time favoring the creation of new, enhanced visual content that combines visual information from different sources and of different type. In this area, applications such as video post-production, realistic video conferencing, and immersive gaming are experiencing a rapid development.

Visual information, in the form of images and video, comes from the interaction of light with objects. Illumination is a fundamental element of visual information. Detecting and interpreting illumination effects is part of our everyday life visual experience. Shading for instance allows us to perceive the three-dimensional nature of objects. Shadows are particularly salient cues for inferring depth information. However, we do not make any conscious or unconscious effort to avoid them as if they were an obstacle when we walk around. Moreover, when humans are asked to describe a picture, they generally omit the presence of illumination effects, such as shadows, shading, and highlights, to give a list of objects and their relative position in the scene. The human visual system has both capabilities. It is able to analyze illumination in a scene and to discard it to reach a description of the scene's content that is more useful for action. It is also able to analyze illumination effects to get information about the scene. Millions of years of biological evolution and environmental adaptation have indeed made human vision a highly developed and complex process.

For many algorithms in computer vision, dealing with illumination effects is a challenging task. Illumination phenomena can in fact mislead fundamental tasks such as object extraction and description. For this reason, lighting conditions require careful consideration in many applications and need often to be controlled. Illumination conditions have moreover to be carefully considered when creating new visual content by combining natural and synthetic objects. When taken into account, illumination effects can play an important role in achieving realism. The challenge for computer

vision systems is then to process visual information in a way that is close to what the human visual system does, thus being aware of illumination conditions and illumination effects. Reaching this objective would enable the development of more effective computer vision systems and richer visual content.

Among illumination effects, shadows are often integral part of natural scenes and one of the elements contributing to naturalness of synthetic scenes. A growing interest has emerged over the last years within the computer vision community in the investigation of the nature of shadows in digital images. In very recent years, moreover, a number of papers published in highly regarded journals have contributed to make this a topic of great impact within different research areas, such as neurosciences, experimental psychology and vision. In parallel, also within the philosophical and history of art domains a renewed interest in shadows and their significants has emerged. The topic can therefore be considered the focal point of a converging series of multidisciplinary research works. The investigation of shadows can have a potential as a basis for a fruitful dialog between different fields of science and humanities.

1.2 Investigated approach

This thesis deals with the problem of identifying and extracting regions that correspond to shadows in images and image sequences. The goal is not to ignore illumination effects due to shadows, as illumination invariant approaches to image analysis try to do, but to separate them from the image signal. Shadows contain in fact information about the world which one does not want to loose, but it is also important to recognize that a shadow boundary is not a change in scene surface.

A shadow occurs when an object partially or totally occludes light from a source of illumination. Consequently, the most straightforward property of a shadow is that it darkens the surface on which it is cast. The difficulty for a computer vision system is then to distinguish a shadow from a naturally dark surface. Color can greatly help in this task. In this thesis, the use of color information for shadow segmentation is thoroughly investigated. The investigation has led to the study of photometric invariant color features. Their invariance in presence of shadows can be effectively exploited for shadow segmentation.

What distinguishes a shadow from a dark surface mark is not only its color characteristics. All shadows are shadows of something and are therefore related to the object that is casting them. A shadow's shape, position, and motion depend on the shape, the position, and the motion of the shadow-casting object. For an effective segmentation of shadows it is important to take also this aspect into account. In this thesis, the use of spatial and temporal properties of shadows to improve the segmentation accuracy is investigated.

On the basis of this background, a new analysis method for the segmentation of cast shadows is proposed. Two implementations, one for the segmentation of moving cast shadows in video sequences and one for the segmentation of cast shadows in still images, are developed to test the method's validity. Particular attention is dedicated to the segmentation of shadows in video. The problem of separating moving cast shadows from moving objects in image sequences is in fact of particular relevance for an always wider range of applications. A key feature of the proposed methodology for shadow extraction is its capability of working regardless of the scene's content, the camera characteristics and the illumination. The method is thus designed to be able to work in real-world environments, where the imaging conditions and the scene set-up are not under control, and without the need of human supervision.

This thesis aims at outlining the twofold importance of shadow extraction techniques. The enhancement of fundamental tasks in video analysis, such as object extraction, tracking, and de-

scription, deriving from the application of the proposed methodology, is demonstrated. Shadows need to be extracted and eliminated to improve the accuracy of object contours and the subsequent use of information about object shape and color. The extraction, tracking and description of video objects are fundamental steps for a wide range of object-based applications, ranging from video coding to video indexing, from video manipulation to video surveillance and immersive environments. All these applications can benefit from a flexible methodology that allows to distinguish objects from the shadows they cast.

The identification of shadows provides on the other hand information about and gives access to an important perceptual element of a visual scene. In applications such as object-based video editing and mixed-reality immersive environments, where new and richer visual content is created by merging objects from different sources, the ability of identifying and taking shadows into account can improve the naturalness of the merging process and have an important perceptual impact. This is demonstrated by applying the proposed method to video composition.

1.3 Main contributions

The main contributions of this work can be summarized as follows:

- Definition of a new analysis method for the segmentation of cast shadows in color images and image sequences. This is based on the analysis of shadows spectral, geometric and, in the case of video, temporal properties. It exploits color information and the properties of photometric invariant color features to provide an initial shadow hypothesis. A spatio-temporal verification stage is defined and combined to the analysis of color features to improve the accuracy of segmentation results.

- A discussion of the use of color information for shadow segmentation. It has pointed out the underlying physical models of shadows, their fundamental assumptions and the link with research problems such as computational color constancy and color invariance.

- An extensive analysis of the behavior of different photometric invariant features for shadow segmentation purposes. It has highlighted the problems related to the use of hue and saturation that are often proposed in literature.

- Definition of a novel shadow tracking strategy. The tracking method has been established on the basis of the limited amount of information available for describing shadows.

- Definition of a spatio-temporal reliability estimation of shadow segmentation results which allows to improve the overall segmentation accuracy.

- Application of the proposed method to the improvement of video object extraction, tracking and description tasks.

1.4 Organization of the thesis

This thesis is organized as follows. *Chapter 2* reviews the main elements of the image formation process and introduces the notion of color and the issue of its representation. It provides the background concepts, notions and models on which the proposed methods are based. *Chapter 3* discusses the characterization of shadows in digital images and image sequences in terms of spectral, geometric and temporal properties. Properties that can be exploited basing only on image-derived

information and with a limited number of assumptions about the scene are selected and analyzed in detail. The state of the art of shadow detection is then reviewed. *Chapter 4* is dedicated to photometric invariant features. Color invariance is introduced and the photometric invariants that are of interest for shadow segmentation purposes are discussed. *Chapter 5* presents the analysis method developed in this thesis for the segmentation of cast shadows in both moving and still color images. The spectral, spatial and temporal analysis steps are described. The performance of the proposed technique is then evaluated and compared to state of the art techniques in *Chapter 6*. The application of the proposed method for achieving a shadow-aware video processing is demonstrated in *Chapter 7*. Finally, *Chapter 8* concludes this thesis and explores directions for future work. The *Appendix* presents some of the results and ideas emerging from the multidisciplinary discussion around the nature of shadows, coming also from fields that are usually not considered in scientific investigations.

From physical scenes to digital images

2

2.1 Introduction

In this chapter, an overview of the image formation process is presented through a review of each of its elements and their interaction. The purpose of the review is to introduce the fundamental notions and models that will be of use in the follow-up of this thesis for the characterization and analysis of shadows.

Light is the first, fundamental element of vision. In its journey from sources of illumination to the eye, light collects information about the physical world around us. The collected information is captured by the eye in the form of retinal images and then transmitted to the brain which interprets it. Similarly, in a computer vision system, the information that light carries is captured by the sensors of a camera in the form of digital images which are then processed and interpreted by a computer. In this thesis, we focus on a physical phenomenon that is strictly related to light. Shadows are, in fact, discontinuities, "holes" in the flow of light through the physical world. Once light has reached a capturing device, what is then the effect of such holes on the resulting image values? In order to be able to recognize shadows in digital images, an understanding of this issue is needed. To this end, first of all, the journey of light from sources of illumination to capturing devices and image capture have to be modeled.

To model the image formation process, its three fundamental elements have to be characterized. They are: a source of light, that is a source of visible electromagnetic energy, a surface, whose properties modulate the electromagnetic energy, and the responses of a vision system to the electromagnetic energy reaching its photosensitive elements. The image values on which image analysis tools are applied are the final product of the interaction among these three elements.

In this thesis, we work on color images. Inferring physical properties of a scene from an image is made easier the more measurements are available. In a color image, each point in the scene induces three measurements. Thus, three times the amount of information with respect to a gray-level image. Its richness makes color of great importance in the analysis of shadows. Color information will indeed play a fundamental role in the shadow segmentation approach we propose in this thesis. The second aim of this chapter is then to introduce the notion of color and the models for its representation

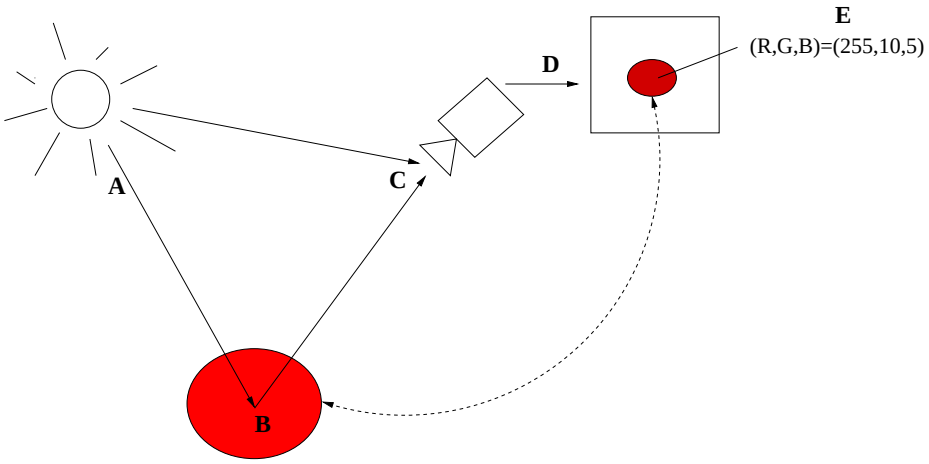

Figure 2.1: The triangle of color. The image color of a surface depends on three components: a source of visible electromagnetic energy, an object, whose properties modulate the electromagnetic energy, and a capturing device.

that have been developed to allow a meaningful processing and interpretation of color images.

Figure 2.1 provides a schematic visual representation of this chapter's contents and organization. The presentation is organized in two parts. In the first part, which comprises Section 2.2, Section 2.3, and Section 2.4, the path of light from a source of illumination to the lens of a camera is analyzed. Light sources (A) and surfaces (B) are discussed. The illumination model that will be of use in this thesis is moreover analyzed (C). In the second part of the presentation, in Section 2.5 and Section 2.6, respectively, the generation of color in the observer (D) and its representation by means of color specification systems (E) are discussed.

2.2 Light

The first element determining the appearance of a surface is given by the light, emitted by a source of electromagnetic energy, that illuminates the surface. *Light* denotes the visible part of the electromagnetic energy that encompasses wavelengths from approximately 400 nm (violet) to 700 nm (red). The visible spectrum represents only a small portion of the complete electromagnetic spectrum, which goes from gamma rays to radio waves. The electromagnetic spectrum is illustrated in Figure 2.2.

The light emitted by a source of illumination is generally composed by a mixture of energy at different wavelengths. The power emitted at each wavelength gives the *Spectral Power Distribution (SPD)* of the source. The CIE, Commission Internationale de l'Éclairage, has established a number of spectral power distributions as *CIE illuminants*. Illuminants are, therefore, standardized tables of values that represent typical SPDs of particular light sources. As an example, CIE illuminants A, D65, and F2 [24] are standardized representations of typical incandescent, daylight, and fluorescent sources, respectively. The relative spectral power distribution (normalized such that it has a value of 100 at a wavelength of 560 nm) of CIE illuminant D65 is shown in Figure 2.3 as an example.

A series of units is used to describe how energy is transferred from light sources to surface patches and what happens to the energy when it arrives at the surface. The measurement of optical radiation is a field known as *radiometry*. We briefly introduce in Section 2.2.1 some radiometric definitions and units that will be used in the following. Light sources are then discussed in Section 2.2.2.

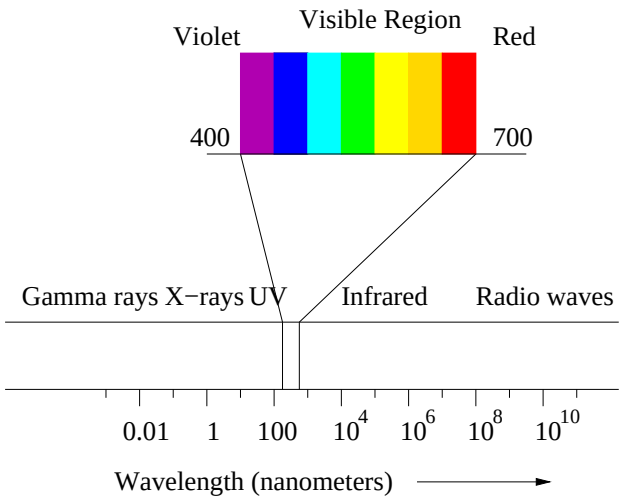

Figure 2.2: The electromagnetic spectrum. The visible spectrum represents a small portion of the complete spectrum.

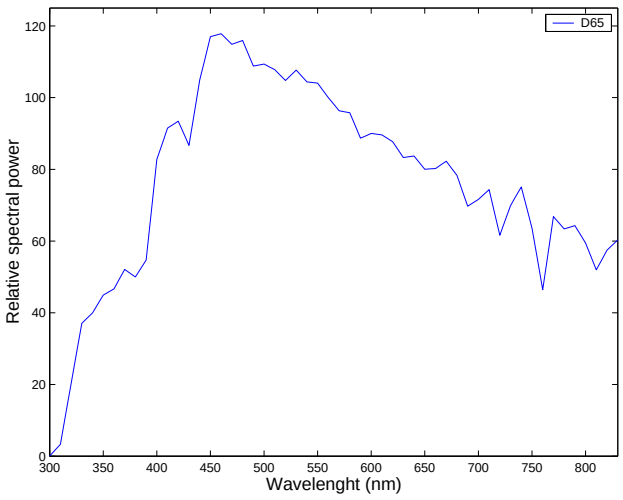

Figure 2.3: Relative spectral power distribution of CIE illuminant D65 [24]. Illuminant D65 has been statistically defined based upon a large number of measurements of real daylight.

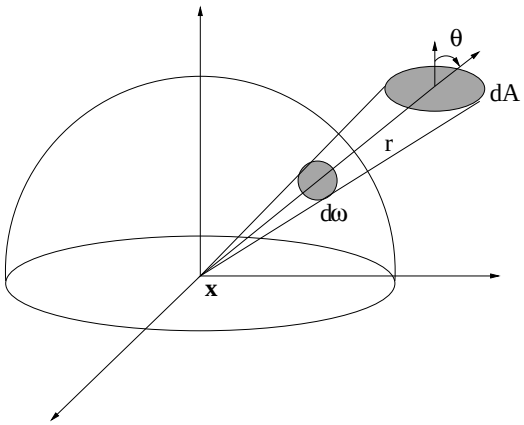

Figure 2.4: The solid angle $d\omega$ subtended by a small surface area dA from a point $\vec{x}$ in the 3D space.

2.2.1 Units

The distribution of light in space is a function of position and direction. The unit for measuring distribution of light in space is *radiance*.

Radiance: power (amount of energy per time unit) traveling at some point in a specified direction, per area unit perpendicular to the direction of travel, per solid angle unit.

The solid angle ω subtended by a surface patch at a point $\vec{x}$ in the three-dimensional (3D) space is given by the area of the patch projected onto the unit sphere whose center is at $\vec{x}$ (see Figure 2.4). If the area of the patch dA is small, then the infinitesimal solid angle $d\omega$ it subtends is easily computed in terms of the area of the patch and the distance r to it as

$$d\omega = \frac{dA \cos\theta}{r^2}, \tag{2.1}$$

where θ is the angle between the surface normal and the normal to the sphere. Solid angles are measured in steradians (sr). The units of radiance are, consequently, watts per square meter per steradian ($W/m^2 sr$). The square meters in the unit for radiance are *foreshortened*, that is perpendicular to the direction of travel. Foreshortening is needed in order to take into account the fact that a small patch viewing a source frontally collects more light than the same patch viewing the source along a nearly tangent direction.

For the majority of vision problems, it is safe to assume that light does not interact with the medium through which it travels, i.e. that it travels in vacuum. In this case, radiance has the highly desirable property that, for two points $\vec{x}_1$ and $\vec{x}_2$, which have a line of sight between them, the radiance leaving $\vec{x}_1$ in the direction of $\vec{x}_2$ is the same as the radiance arriving at $\vec{x}_2$ from the direction of $\vec{x}_1$. Radiance thus is constant along straight lines.

Radiance is used for describing both light traveling in free space and light reflected from a surface when it depends on direction. The relationship between incoming illumination and reflected light is a function of both the direction in which light arrives at a surface and the direction in which it leaves. The unit for representing incoming power is *irradiance*.

Irradiance: total incident power per surface area unit.

Irradiance has units of watts per square meter (W/m^2). Irradiance is used when describing light arriving at a surface.

The physical units introduced in this section can be extended to spectral units in order to describe energy arriving in different quantities at different wavelengths. *Spectral radiance* adds to radiance the wavelength dependency, having units of W/m^2srnm. *Spectral irradiance*, in the same way, includes the wavelength dependency and has units of W/m^2nm. These units allow to describe differences in energy with wavelength. *Spectroradiometry* is the measurement of radiometric quantities as a function of wavelength.

2.2.2 Light sources

A *light source* is a physical emitter of visible energy. Examples of light sources are incandescent light bulbs, the Sun, a clear or overcast sky, and fluorescent tubes. To characterize a light source from a radiometric point of view, a description of the radiance it emits in each direction is needed. A complete description of the radiance in each direction is, however, not always required. It is more usual to model sources as emitting a constant radiance in each direction, possibly with a family of direction zeroed, like a spotlight. The appropriate radiometric quantity in this case is the *exitance*, defined as the internally generated energy radiated per unit time and per unit area on the radiating surface.

Together with a description of the exitance, a description of the geometry of the source is required for its characterization. The geometry of the source has important effects on the spatial variation of light around the source and on the shadows cast by objects near the source. Sources are usually modeled with quite simple geometries for two main reasons. Firstly, many synthetic sources can be modeled as point sources or area sources fairly effectively. Secondly, sources with simple geometries can still yield complex effects.

A common approximation is to assume that the light source is an extremely small sphere, with no area, that is a point. Such source is known as a *point light source*. It is a natural model to use because many sources are physically small compared with the surrounding environment. A point source is referred to as being a *point source at infinity* when it can be assumed that the power at a surface due to the point source does not decrease with the distance to the source. A point source at infinity is a good model for the Sun, for example, because the solid angle that the Sun subtends is small and essentially constant wherever it appears in the field of view. Point light sources at infinity can be assumed to emit parallel light rays.

On the contrary to point light sources, *area light sources* (also referred to as *extended light sources*) have an area. They occur commonly in natural scenes (the vast majority of indoor sources are area light sources) and cast soft shadows, containing areas only partially blocked from the source. Area sources are often modeled as surface patches whose emitted radiance is independent of position and of direction. They can, in this case, be described by their exitance.

A description of the color properties of a source of illumination allows to complete its characterization. This can be done by means of illuminants. Another important quantity that can be used to characterize a source is its *correlated color temperature* [30]. The correlated color temperature of a source is the color temperature of a black-body radiator that has most nearly the same color as the source. *Black-body radiators* or *Planckian radiators* are a special type of theoretical light sources which emit energy due only to thermal excitation. Their spectral power distribution is described by Planck's equation [173] as a function of their absolute temperature (in Kelvins). The temperature of a Planckian radiator is called *color temperature* since it uniquely specifies the color of the source.

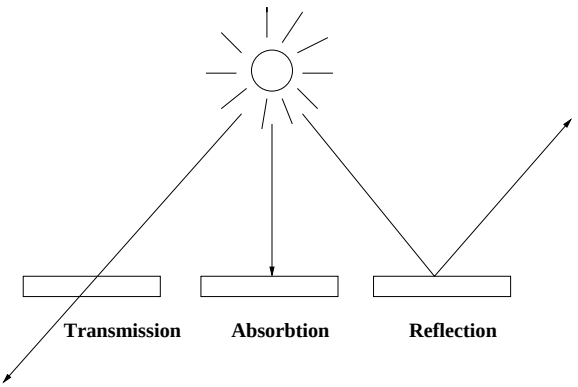

Figure 2.5: Different types of interaction between light and surfaces.

2.3 Surfaces

The second element that contributes to the appearance of a surface is given by the structural and optical properties of the surface itself, which determine the fraction of the incident illumination that is reflected by the surface patch in a certain direction. In particular, the viewing direction is of interest in image formation.

When light strikes a surface, it may be absorbed, transmitted, or reflected, as illustrated in Figure 2.5. Usually, a combination of these effects occurs. Some materials absorb light at one wavelength and then radiate light at a different wavelength. This effect is known as *fluorescence*. Furthermore, a surface that is warm enough may emit light in the visible range. The interaction of radiant energy with materials obeys the law of conservation of energy. Therefore, the amounts of absorbed, reflected and transmitted radiant power sum to the incident radiant power at each wavelength and it is typically unnecessary to measure all three. The quantities are typically measured in relative terms as percentages of the incident energy. The surface's *absorptance, transmittance* and *reflectance* are obtained.

In this thesis, we limit our analysis to opaque objects, that is we do not consider transmission. Moreover, as is commonly done in computer vision research, we discount fluorescence and emission to focus on reflection. The relationship between incoming illumination and reflected light at a given point on a surface and at each wavelength depends on the illumination and viewing geometry and on the surface's structure and material composition. A function describing this relationship provides a *reflectance model*.

The most general model of reflectance is the *Bidirectional Reflectance Distribution Function*, usually abbreviated as BRDF.

Bidirectional Reflectance Distribution Function (BRDF): ratio of the radiance in the outgoing direction to the incident irradiance at a surface point $\vec{x}$. Given two vectors $\vec{V}$ and $\vec{I}$, defining the outgoing and incoming light directions respectively (see Figure 2.6), the BRDF is denoted as $\rho_{bd}(\vec{x}, \vec{V}, \vec{I})$.

Let us consider a surface point $\vec{x}$, illuminated by radiance $L_i(\vec{x}, \vec{I})$ coming in from a differential region of solid angle $d\omega_i$ in direction $\vec{I}$ (Figure 2.6). Let us denote as i the angle of incidence between the illumination direction $\vec{I}$ and the surface normal $\vec{N}$. The irradiance at $\vec{x}$ is computed as

$$E(\vec{x}, \vec{I}) = L_i(\vec{x}, \vec{I}) \cos(i) d\omega_i. \tag{2.2}$$

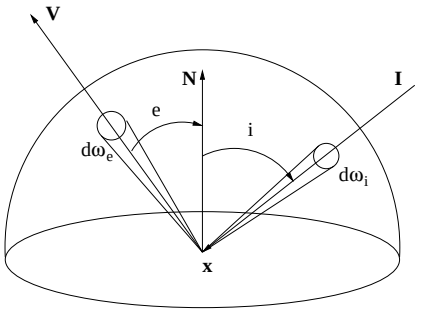

Figure 2.6: Geometry for reflected radiance and incident irradiance.

Since irradiance is expressed per area unit, whereas radiance is expressed per foreshortened area unit, as commented in Section 2.2.1, multiplying radiance by $cos(i)$ converts it to the equivalent per unforeshortened area unit. If $L_i(\vec{x}, \vec{I})$ was to emit radiance $L_r(\vec{x}, \vec{V})$ in the exit direction $\vec{V}$, its BRDF would be

$$\rho_{bd}(\vec{x}, \vec{V}, \vec{I}) = \frac{L_r(\vec{x}, \vec{V})}{E(\vec{x}, \vec{I})} = \frac{L_r(\vec{x}, \vec{V})}{L_i(\vec{x}, \vec{I}) \cos(i) d\omega_i}. \tag{2.3}$$

The BRDF has units of inverse steradiants (sr^{-1}) and can vary from 0 (no light reflected in an outgoing direction) to infinity (unit radiance in an outgoing direction resulting from arbitrarily small radiance in the incoming direction). The BRDF depends on the wavelength of the incoming light.

BRDF measurements are difficult and expensive. Therefore, simplified models are needed to describe the interaction of light with surfaces for computer vision problems. Modeling reflection may indeed be simplified for some surfaces, as discussed in the next subsections.

2.3.1 Diffuse surfaces

The light leaving many surfaces is largely independent of the exit angle. A natural measure of a surface reflection properties in this case is the *directional hemispheric reflectance* [40], denoted as $\rho_{dh}(\vec{x}, \vec{I})$. The directional hemispheric reflectance is defined as the fraction of the incident irradiance in a given direction $\vec{I}$ that is reflected by the surface, whatever the direction of reflection.

For some surfaces, the directional hemispheric reflectance does not depend on illumination direction. Examples of such surfaces include cloth, many carpets, matte paper and matte paints. In these cases, the radiance leaving the surface is independent of illumination incidence angle and the directional hemispheric reflectance, and consequently the BRDF, are constant. Such surfaces are known as *ideal diffuse (matte) surfaces* or *Lambertian surfaces**. For Lambertian surfaces, the directional hemisphere reflectance is often called the *diffuse reflectance* or *albedo*.

A Lambertian surface looks equally bright from any direction. Our perceptions of brightness, in fact, correspond roughly to measurements of radiance. The retina itself responds commensurably to the irradiance incident upon it, but in combination with the optics of the eyeball, retinal irradiance is proportional to the radiance of a surface [30]. This observation provides a rough test for the appropriateness of a Lambertian approximation.

*From Johann Heinrich Lambert (1728–77), who studied illumination phenomena in his *Photometria* (1760) [10].

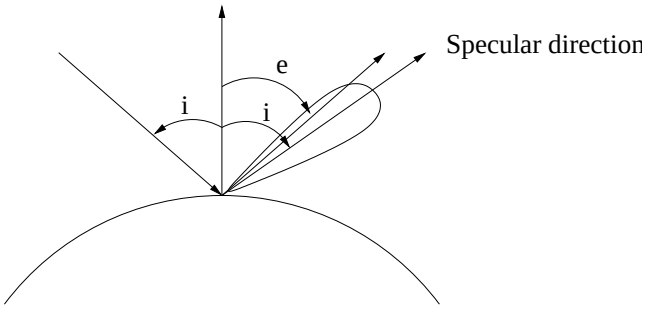

Figure 2.7: Specular surfaces commonly reflect light into a lobe of directions around the specular direction. The shape of the lobe is described in terms of the offset angle $s = (i - e)$ between exit direction and specular direction.

2.3.2 Specular surfaces

A second important class of surfaces are the glossy or mirror-like surfaces, often referred to as *specular surfaces*. Radiation arriving at a specular surface along a particular direction can leave only along the *specular direction*, obtained by reflecting the direction of incoming illumination about the surface normal. Examples of specular surfaces are mirrors and polished metal.

Only few surfaces can be approximated as ideal specular reflectors. Typically, radiation arriving in one direction leaves in a small lobe of directions around the specular direction (Figure 2.7). This results in a typical blurring effect. Larger specular lobes cause the specular image to be more heavily distorted and darker. The incoming radiance, in fact, must be shared over a large range of outgoing directions. Quite commonly, it is possible to see on specular surfaces only a specular reflection due to relatively bright objects like sources, but few other specular effects. The bright blob one sees on shiny paint or plastic surfaces is called *specularity* or *highlight*.

Relatively few surfaces are either ideal diffuse or perfectly specular. The BRDF of many surfaces can be approximated as a combination of a Lambertian component and a specular component. To model the interaction between light and surfaces we will indeed consider in this thesis a model that takes the two components into account. It is discussed in the following section.

2.4 Light and surfaces: Reflection models

Various reflection models are used in computer graphics and computer vision [26, 38, 47, 63, 69, 73, 89, 115, 124, 143] that describe the light reflected by a surface as a weighted combination of a diffuse and a specular component. They differ in the way these two components are modeled and weighted when combined.

Some models, such as the Phong model [124], do not have a physical basis, but empirically approximate some of the underlying rules of optics and thermal radiation. This can represent a limitation if the model is used to predict the color appearance of a surface. Other models, such as the Torrance-Sparrow model [38] and the Beckmann-Spizzicchino model [73], are rigorously derived but result cumbersome and impractical for computer vision applications. Approximate models, such as the Dichromatic Reflection Model [143], are still derived from physics-based reflectance models, but they are modified so as to emphasize the desired aspects of the models as well as to ignore their other unnecessary aspects. The Dichromatic Reflection Model is used for these reasons in this thesis.

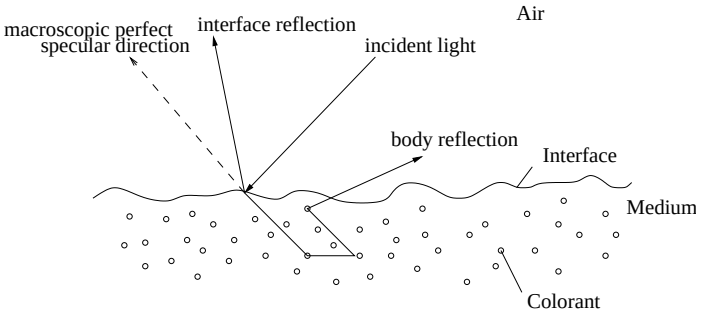

Figure 2.8: Reflection of light from a dielectric, nonuniform material.

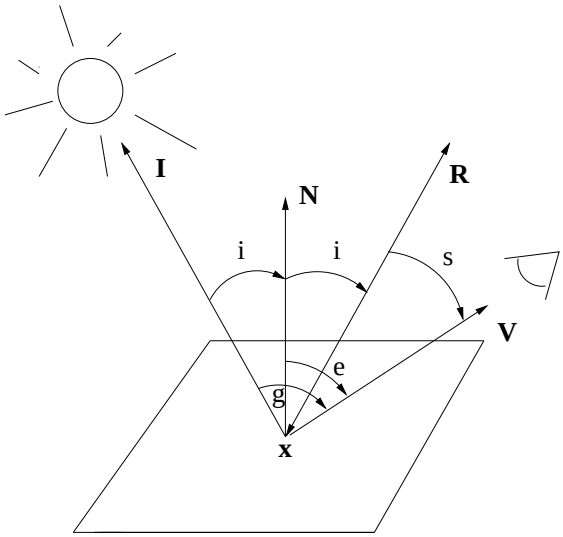

Figure 2.9: Geometry of reflection. i is the *angle of incidence* between illumination direction $\vec{I}$ and surface normal $\vec{N}$, e is the *angle of exitance* between $\vec{N}$ and viewing direction $\vec{V}$, g is the phase angle between $\vec{I}$ and $\vec{V}$, and s is the *off-specular angle* between $\vec{V}$ and $\vec{R}$, where $\vec{R}$ is the direction of perfect specular reflection.

2.4.1 The Dichromatic Reflection Model

The Dichromatic Reflection Model allows to model the physics of reflection for a wide class of dielectric, that is nonconducting, materials. It was proposed by Shafer [143] for determining the orientation of a surface in image analysis applications.

The Dichromatic Reflection Model treats *optically inhomogeneous materials*, that is materials where light interacts both with a medium that comprises the bulk of the surface matter, and with the particles of a colorant that produce scattering and coloration (see Figure 2.8). Many common materials can be described this way, including paints, varnishes, paper, ceramics, and plastics. Metals, glass, and crystals are excluded as they are optically homogeneous. Only opaque surfaces are considered in the model.

The Dichromatic Reflection Model suggests that, under all illumination and viewing geometries, the reflected light can be described as the weighted sum of two functions, an *interface reflection* function and a *body reflection* function. According to the model in fact, as illustrated in Figure 2.8, one way light is scattered from the surface is by a mirror-like reflection at the interface of the surface.

A second scattering process takes place when the rays enter the material. These rays are reflected randomly between the colorant particles. A fraction of the incident light is absorbed by the material, heating it up, and part of the light emerges. The particles in the medium absorb light selectively with respect to wavelengths. It is this property what determines an object's characteristic color.

Referring to the geometry and terminology of Figure 2.9, the Dichromatic Reflection Model states that the total radiance of the reflected light at a given point on a surface is given by

$$L(\lambda, i, e, g) = L_s(\lambda, i, e, g) + L_b(\lambda, i, e, g), \tag{2.4}$$

where λ is the wavelength. The reflected light is thus given by the sum of two independent parts:

- the radiance L_s of the light reflected at the interface between the air and the surface medium;
- the radiance L_b of the light that penetrates through the interface and that is reflected from the surface body.

Each of the two components can be then decomposed into a *composition* part and a *magnitude* part as

$$L(\lambda, i, e, g) = m_s(i, e, g)c_s(\lambda) + m_b(i, e, g)c_b(\lambda). \tag{2.5}$$

The composition term is a spectral power distribution, c_s or c_b, that depends only on wavelength but is independent of geometry. The magnitude term is a geometric scale factor, m_s or m_b, which depends only on geometry and is independent of wavelength. This independence has made the Dichromatic Reflection Model's formulation very popular. We will see how it can be exploited for deriving color invariants in Chapter 4.

The described independence property of the model is obtained at the cost of some approximations. It is important to comment them. Both interface and body reflection exhibit, in fact, an interdependence between wavelength and geometry. Interface reflection is governed by Fresnel's laws of reflection, which relate interface reflectance to the angle of incidence of the light and the index of refraction of the material. The index of refraction generally depends on wavelength and therefore interface reflection is a function of wavelength. However, since the amount of variation of the index of refraction for many materials is within a few percents across the visible spectrum, variations of c_s with wavelengths should be negligible. c_s can therefore be assumed constant. The interface reflection, in this case, has the same color as the illumination. This assumption is called the *Neutral Interface Reflection (NIR) assumption* by Lee et al. [89].

Body reflection also exhibits an interdependence between wavelength and geometry. If c_s is not constant, in fact, the color of the light passing through the interface differs somewhat from the color of the illumination. Since the total amount of light reflected at the interface varies with the angle of incidence i, the color of the light passing through the interface into the material body also varies with the angle of incidence. Thus, the color of the body reflection should vary with geometry. However, if c_s is nearly constant, this effect should be negligible as well.

It is interesting to analyze the other assumptions made by the model, which determine its scope and validity. For what concerns illumination, the model assumes that:

- there is a single light source, that can be a point source or an area source;
- the illumination has a constant SPD across the scene;
- the amount of illumination can vary across the scene.

The assumption of illumination being due to only one source of illumination is not realistic, as it will be discussed in the next subsection. For what concerns the surface properties, the model assumes that:

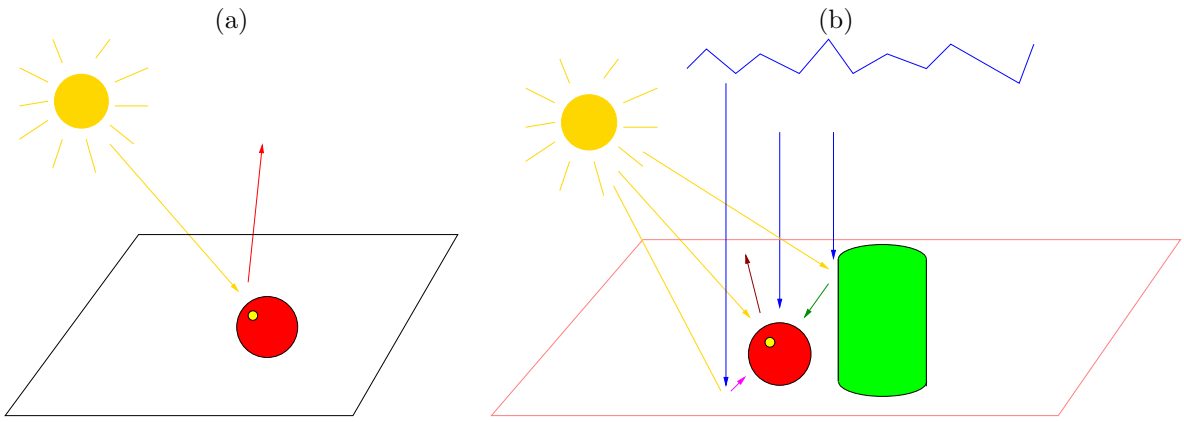

Figure 2.10: (a) Reflection at a surface due to a light source. (b) An accurate model of how the brightness and color of a surface are obtained has to take many factors into account: e.g. the light reflected by the red ball is due to two different light sources, the Sun and the sky, and to reflections from the neighboring green object and from the planar surface on which the ball lies.

- the surface is opaque;
- the surface is not optically active (no fluorescence);
- the colorant is uniformly distributed.

The assumptions about the surface are typical for reflection models and not too unrealistic. In the next subsection, the scope of the model is extended by relaxing the assumption of illumination being due only to a single light source.

2.4.2 Model extension

Eq. (2.5) allows to compute the radiance leaving a surface patch due to a source of illumination (as illustrated in Figure 2.10 (a)), but this is not enough to describe a surface's brightness and color. Radiance may arrive at a surface patch not only from a light source directly, but in other ways. It could be reflected from other surface patches, for instance. Or it could be transmitted by a transparent object. One problem that is, in particular, immediately evident with the formulation of the Dichromatic Reflection Model discussed in the previous section is that shadow regions are arbitrarily dark because they cannot see the light source. This prediction is not accurate in most of the cases. Shadows in a scene are, in fact, normally illuminated by light from other surfaces. This effect can be significant in rooms with light walls, for instance. A patch on a wall sees all the other walls and an object casting its shadows on this patch blocks only a small fraction of the visual hemisphere of the patch, until it is not close enough. Since we aim in this thesis at analyzing shadows, it is clear that the discussed model is not enough for our purposes.

An accurate description of the scene illumination is extremely difficult to obtain. Figure 2.10 (b) aims at showing how an accurate computation of the radiance at a surface can become a very complex problem when all the many factors involved are taken into account. This would be impractical for our purposes. For some environments, however, the total irradiance a patch receives from other patches is roughly constant and roughly uniformly distributed across the input hemisphere. This is true in the interior of a sphere with a constant distribution of radiance and, by accepting a model of a cube as a sphere, is roughly true for the interior of a room with white walls. In such an environment,

it is possible to model the effect of other patches on the surface patch under analysis by adding an *ambient illumination term* to each patch's radiance. The majority of reflection models that are used in computer vision applications make use of this approximation to account for all the complex ways in which light can reach an object that are not otherwise addressed by the illumination equation. More complex models are physically more accurate, but become hard to manipulate unless scenes are restricted to simple geometries. Since our goal is developing tools for shadow analysis that can be applied to real world complex scenes, we will as well make use of this approximation.

When an additional ambient diffuse light, of lower intensity, coming from all directions in equal amounts, and possibly with a different SPD than that of the light source is considered, the Dichromatic Reflection Model becomes

$$L(\lambda, i, e, g) = L_s(\lambda, i, e, g) + L_b(\lambda, i, e, g) + L_a(\lambda). \tag{2.6}$$

This extended model will be used in Chapter 3 to analyze the spectral characteristics of shadows in digital images.

The adequacy of the Dichromatic Reflection Model has been tested by Healey [63] and Tominaga and Wandell [160] and experimental results [159] show that it is valid for artificial objects like plastics and paints, and for natural object like fruits and leaves. Metals have quite different reflection properties than inhomogeneous materials. They have only the interface reflection. Light striking a metal surface can, in fact, be either absorbed or specularly reflected. This is due to the fact that electric fields cannot penetrate conductors, since the electrons inside the material move around and cancel the field. Healey [63]* therefore proposed and tested a unichromatic reflection model for metals that keeps the independence of geometry and wavelength in its formulation. Using the same notation as above, the model is formulated as

$$L(\lambda, i, e, g) = m_s(i, e, g)c_s(\lambda). \tag{2.7}$$

The Dichromatic Reflection model and Healey's unichromatic model provide a common formulation for modeling the physics of reflection for a wide variety of materials in computer vision problems. They allow to describe the information carried by light in its journey from the source of illumination to the vision system. The role of this third element contributing to the process of image formation is discussed in the next section.

2.5 Color generation

We have seen in the first part of the chapter how light is generated by sources of illumination and altered by surfaces in the scene. After multiple reflections, light finally arrives at the capturing device of the color vision system that is observing the scene. The vision system transforms the information carried by light into a color image of the physical world. The following second part of the chapter is then dedicated to the discussion of color information generation in the vision system and to its representation.

Color is the brain's reaction to a visual stimulus. It is a perceptual attribute of a visual sensation. Visual *sensation* and visual *perception* are intimately related, but they are not the same. Sensation is the process through which the senses detect visual stimuli and transmit them to the brain. Perception is the process by which sensory information is organized and interpreted by the brain. Roughly speaking, sensation furnishes the raw material of sensory experience, while perception provides the

*Among the many references that can be selected from the literature, we recommend Healey's paper for a clear analysis of the physics of reflection.

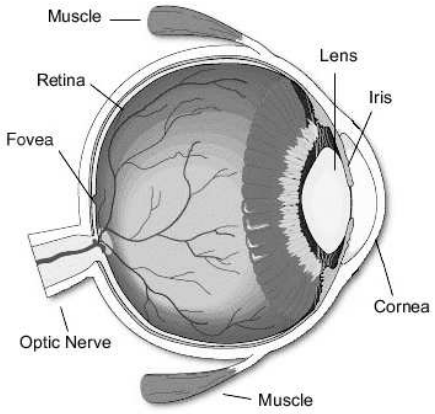

Figure 2.11: Schematic representation of the optical structure of the human eye with some key features labeled (image from [30]).

finished product. The limited knowledge we have about how the human brain gets to this finished product explains why the use of color in computer vision research remains a complex and delicate issue.

Digital instruments emulate the visual sensation by imaging the visual reality and reproducing it at high quality. In this field, the *trichromatic theory* is generally applied: the visual information is decomposed into three signals in a way very similar to what has been observed in the sensory stage of the human visual system. A wide variety of mathematical representations is then used for specifying, manipulating and communicating color. Each representation meets the requirements of a specific application. For example, color is defined in terms of the excitations of red, green, and blue phosphors for display purposes, or by its attributes of brightness, hue and saturation in user-oriented color specification. Color appearance models provide an attempt toward visual perception representation.

This second part of the chapter is organized as follows. The human eye, that is the primary sensory device, is first of all briefly described in Section 2.5.1. The acquisition of color images by color cameras is then analyzed in Section 2.5.2. Color specification systems are finally classified and reviewed in Section 2.6.

2.5.1 The human eye

The eye represents the physical interface of the human visual system. It converts electromagnetic energy into neural activity. A schematic representation of the optical structure of the human eye is shown in Figure 2.11*. The *cornea* and the *lens* act together to focus an image of the visual world on the retina, located at the back of the eye, which represents the photosensitive organ of the human visual system. A direct parallel with a camera is easily established: the cornea and the lens are equivalent to the camera's lens, the retina is equivalent to the film or other image sensor.

The *retina* is a thin layered membrane of neural cells or *photoreceptors*. The human retina has two different types of photoreceptors: the rods and the cones. They transform the optical stimuli into neuro-electrical signals that are then transmitted to the later stages of the visual system which interprets them. The *rods* are active during scotopic vision (low light levels) and do not support color perception. The *cones* are active in photopic vision (high light levels) and are responsible for

*The interested reader is referred to [56, 165] for more details.

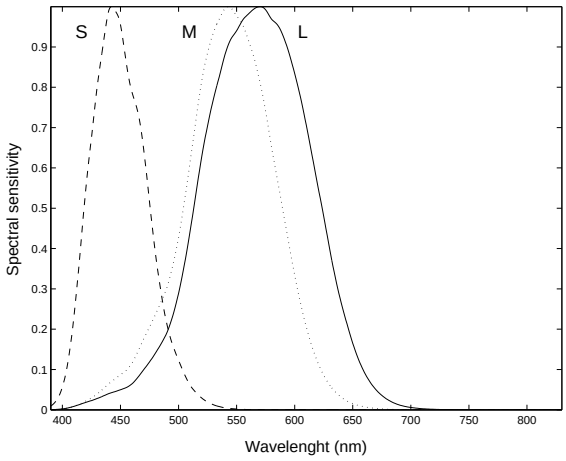

Figure 2.12: Normalized spectral sensitivity curves of the L, M, and S cones in the human eye.

color vision. In particular, three different kinds of cones have been distinguished according to their sensitivity to long, medium and short wavelengths: they are referred to as *L*, *M*, and *S cones*. In Figure 2.12, their wavelength sensitivities are shown [165].

There are about 100 million rods and 5 million cones in a human eye. Their spatial distribution varies across the retina. The highest concentration of cones occurs in the *fovea*. Conversely, there are no rods in the center of the fovea, but the rod density increases toward the periphery of the visual field. There is also a blind spot on the retina, where the neuro-electrical signal carrying the retinal image information exits the retina to reach the *optic nerve*.

The three cones are the foundations of our color vision. As any other detector of radiation, they integrate the light at all wavelengths. In this way, the entire spectrum of incident light is reduced to three signals, one for each cone, resulting in what is called *trichromacy*. The physiological basis of human color perception is thus trichromatic. This explains why it is possible to match all of the colors in the visible spectrum by appropriate mixing of three primary signals. As mentioned above, this result is exploited by digital instruments to capture and reproduce color.

The signals transmitted from the retina to the higher levels of the brain through the optic nerve are not, however, point-wise representations of the receptor signals, but the result of a complex combination between them. In this way, the input information sensed by millions photoreceptors is reduced and transmitted to about one million of optic nerve fibers without loss of visually meaningful data. This data reduction phenomenon takes advantage of differential mechanisms both in the spatial and in the spectral domain that generate signals by comparing the response of a neural cell with those of its spatial neighbors.

The mechanism of retinal coding is complex and still not well understood*. As a convenient simplification, by means of psychovisual experiments, the existence of three types of color channels, called *opponent channels*, is assumed. Referring to Figure 2.13, a *black-white or achromatic channel* is assumed to be created from the sum of the signals coming from L and M cones. The achromatic channel has the highest spatial resolution. The *red-green channel* is mainly the result of the M cones signals being subtracted from those of the L cones. Its spatial resolution is slightly lower than that of the achromatic channel. Finally, the *yellow-blue channel* results from the addition of L and M and subtraction of S cone signals. It has the lowest spatial resolution.

*The interested reader is referred to Wandell's book [165] for more details.

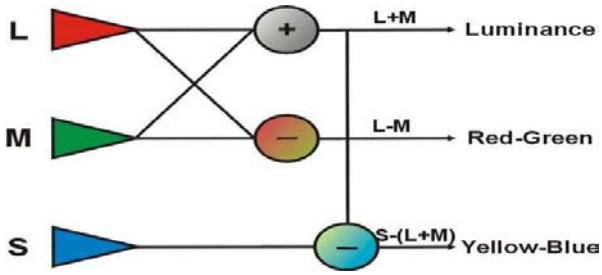

Figure 2.13: Diagram of cones interconnections in the retina leading to opponent-type signals.

2.5.2 Color image formation

Just as the human eye, a color camera contains a set of receptors which convert electromagnetic energy in electric signals. The signals are then sampled and quantized by a frame grabber that produces the final digital image. As described in Section 2.2, the signal that is captured by the camera's sensors is characterized entirely as an electromagnetic power by its Spectral Power Distribution (SPD), which is a function $E(\lambda)$ of the wavelength. Color is, in this sense, a continuous function of the wavelength of the observed signal. The process of acquisition tries to replace this continuous signal with a discrete model, thus mimicking the human eye. The spectral space, ϵ, which has infinite dimension, is replaced by a finite-dimensional color space. The trichromatic theory is typically exploited and this finite-dimensional space is a three-dimensional (3D) space.

The camera's sensors transform therefore the continuous color signal into three scalars obtained as

$$C_i = \int_{\Lambda} E(\lambda) S_i(\lambda) d\lambda, \tag{2.8}$$

where $S_i(\lambda)$ is the sensitivity of the i^{th} camera sensor, and Λ is determined by $S_i(\lambda)$, which is non-zero over a bounded interval of wavelengths λ. Typically, a red, a green, and a blue sensor are used. The measured color results in a vector of three color values, $\vec{C} = (R, G, B)$. A color camera thus establishes a *spectral integration* transformation between the space of spectral colors and the sensors response color space

$$\mathcal{R} : \mathbb{R}^{\infty} \rightarrow \mathbb{R}^3$$

defined by

$$\mathcal{R}(E(\lambda)) = (R, G, B),$$

where each component is given by Eq. (2.8). In general, the mapping from $E(\lambda)$ to image color values comprises several complex factors [58, 59], such as vignetting [6], lens fall-off, the sensitivity of the detector, and the electronics of the camera [65]. Accurately modeling the image formation process by taking these factors into account is however out of the scope of this work.

Color charge-coupled-device (CCD) cameras, which are the most widespread, use a rectangular grid of electron-collection elements laid over a thin silicon wafer to record a measure of the amount of light energy reaching each of them. To obtain color information, at each sensory element, color filters with different spectral sensitivity to the various wavelengths are interposed between the incoming illumination and the CCD element. Two types of color cameras can then be distinguished which provide a different degree of color resolution. In single-CCD cameras, color filters are layered over each pixel element of the CCD array in a mosaic pattern. Only one color channel is captured with each CCD element and the two missing color channels are estimated from the existing information in order to get a full RGB image. This process is referred to as demosaicing [5]. It introduces artifacts in the reconstructed image. 3-CCD cameras have three CCD sensors for each pixel element

which capture the three color channels and provide an higher color accuracy. To reduce cost, size and difficulties in the optics, most digital cameras are single-CCD cameras.

The color signal $E(\lambda)$ in Eq. (2.8) represents the image irradiance incident upon the camera sensors. This means that the image pixel color values on which image processing tools operate represent a measure of irradiance. What we are interested in when analyzing images is rather the radiance of surfaces in the depicted scene. It is important therefore to relate image irradiance with scene radiance. Under the assumption of thin camera lenses, it can be shown that image irradiance is proportional to scene radiance [40]. In other words, what we measure is proportional to what we are interested in. We will assume this relationship to hold in the remainder of this thesis as it is a typical assumption in computer vision problems. In this case, the signal $E(\lambda)$ that reaches the camera's sensor from a point on a surface is proportional to the surface radiance which we have modeled in Section 2.4.1 by means of Eq. (2.6).

We have reached at this point the first objective of the chapter, that is providing a formalization of the chain linking the physical world to a digital color image of it. This will allow us to analyze and to interpret image pixel color values as a function of physical phenomena, such as shadows. In order to process color information, different color representation systems have been proposed. They are reviewed in the next section.

2.6 Color representation

Since all colors can be matched by proper amounts of three primary colors, three numerical components are necessary and sufficient to define a color. It is then natural to represent colors as points in a three-dimensional vector space, called *color space* or *color model*. A color space is thus a mathematical representation of spectral colors in a finite dimensional vector space. It allows to analyze and manipulate color.

By defining different primary colors, that is basis elements of the vector space, different color models can be devised. Moreover, additional representation systems can be developed according to physical, physiological or psychological properties. A number of color specification models are in use today. Moreover, different definitions can often be found for the same model. The interested reader is referred to [163] for a detailed review, which is out of the scope of this work.

In this section, the color spaces that will be considered in this thesis are introduced. Since the reference space for defining any color specification system is provided by the CIE colorimetric standard, colorimetric spaces are also briefly introduced. The presentation follows a classification of color models in three groups:

- colorimetric models,
- device-oriented models,
- and user-oriented models.

2.6.1 Colorimetric color spaces

The branch of color science concerned with numerically specifying the color of a physically defined visual stimulus is *colorimetry*. A *colorimetry standard* was defined by the Commission Internationale de l'Éclairage, CIE, in 1931 [24] and continues to form the basis for the specification of color. The colorimetry standard allows to predict whether two color stimuli match in color for certain conditions of observation.

The CIE colorimetric system was constructed on the basis of the principles of trichromacy. Based on the hypothesis that the human retina has three kinds of color sensors and that the difference in their spectral responses contributes to the sensation of color, the CIE's trichromatic generalization states that any color stimulus can be matched in color by proper amounts of three primary stimuli.

As discussed in Section 2.5.1, more recent studies have confirmed the presence in the human retina of three types of cones and have measured [153] their spectral sensitivities (see Figure 2.12). Although it has been seen that the perception of color depends on further processing of the retinal responses, to a first order of approximation, the sensation of color, under similar conditions of adaptation, may be specified by the responses of the cones.

However, the CIE established the 1931 standard long before the accurate knowledge of the cone spectral responsitivities was available. The standard was at that time defined using the *color-matching functions* [75] determined through psychophysical color matching experiments. Human observers were asked to match the appearance of a test light by adjusting the intensities of three primary lights. The color-matching functions provide the amounts of three primaries, the so-called *tristimulus values of the spectrum*, needed to match a unit amount of power at each wavelength of the visible spectrum. Color matching functions are related to the spectral sensitivities of the three cones by linear transformations.

The CIE 1931 recommendations define a standard colorimetric observer by providing two different but equivalent sets of color-matching functions*. The two sets define two color coordinate systems, as commented in the next subsection.

CIE XYZ

The first set of CIE color-matching functions defines the *CIE RGB spectral primary system*, with red, green and blue primaries at wavelengths given by 700 nm, 546.1 nm and 435.8 nm, respectively. The RGB matching-functions have a great inconvenient: they present both positive and negative values. Since negative sources are not physically realizable, certain colors cannot therefore be matched in the matching experiment by RGB mixtures. In fact, no practical set of three primaries has been found that can reproduce all colors.

The definition of three *hypothetical primary sources*, such that all the spectral tristimulus values are positive, led to the second set of color-matching functions, which define the *CIE XYZ color coordinate system*. CIE RGB color-matching functions and CIE XYZ color-matching functions are related by a linear transformation. CIE XYZ color-matching functions are shown in Figure 2.14. The new set of primaries has the following important properties:

1. They always produce positive tristimulus values.
2. It is possible to represent any perceived color in terms of these primaries.
3. They are derived so that equal values of X, Y, and Z produce white.
4. They are arranged so that a single parameter, Y, determines the *luminance* of the color.

It is important to precisely define the concept of *luminance*, a term that is very often used but also abused in the literature. *CIE luminance* is the results of the integration of a SPD of light using the CIE XYZ color-matching curve corresponding to Y as a weighting function. The magnitude of luminance is proportional to physical power of light. The spectral composition of luminance is related to the sensitivity of human vision.

*In 1964, the CIE established a supplementary standard colorimetric observer from experiments using a visual field that subtended 10 degrees instead of the 2 degrees of the 1931 standard.

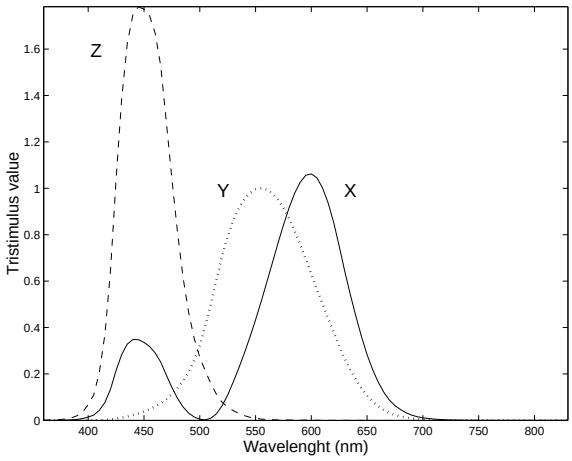

Figure 2.14: Color matching functions for the CIE 1931 XYZ standard colorimetric observer.

Color matching-functions provide the tristimulus values for spectral colors. By considering any given stimulus' spectral power as an additive mixture of various amounts of monochromatic stimuli, one can obtain the *tristimulus values for a stimulus* by multiplying the matching functions by the amount of energy in the stimulus at each wavelength and then integrating across the spectrum (*Grassman's laws of additivity and proportionality* [173]). The XYZ tristimulus values of a stimulus $E(\lambda)$ are thus computed as

$$X = \int \overline{x}(\lambda)E(\lambda), \tag{2.9}$$

$$Y = \int \overline{y}(\lambda)E(\lambda), \tag{2.10}$$

$$Z = \int \overline{z}(\lambda)E(\lambda), \tag{2.11}$$

where $\overline{x}(\lambda)$, $\overline{y}(\lambda)$, and $\overline{z}(\lambda)$ are the color matching curves.

The XYZ system can be used to represent all spectral colors, but since XYZ primaries are not physically realizable because their wavelengths have been chosen outside the visible spectrum, only a subset of the XYZ space can be physically produced. The CIE XYZ standard is the reference space for comparing and storing color information, independently from devices and applications.

CIELAB

CIE tristimulus spaces are perceptually nonuniform, that is, equal perceptual differences between colors do not correspond to equal distances in the tristimulus space. Considerable research has been directed therefore toward the development of *uniform color spaces*. The main aim in the development of uniform color spaces was to provide uniform practices for the measurements of *color differences*, something that cannot be done reliably in tristimulus spaces.

The CIE has recommended two uniform color spaces: the CIE 1976 $L^*u^*v^*$ (CIELUV) space and the CIE 1976 $L^*a^*b^*$ (CIELAB) space [24]. These spaces extend the tristimulus colorimetry to three-dimensional spaces with dimensions that approximately correlate with the perceived lightness, chroma, and hue of a stimulus. *Lightness* and *chroma* are defined as, respectively, the brightness and the colorfulness of an area judged relatively to the brightness of a similarly illuminated area that appears to be white [25]. *Brightness* is the attribute of a visual sensation according to which

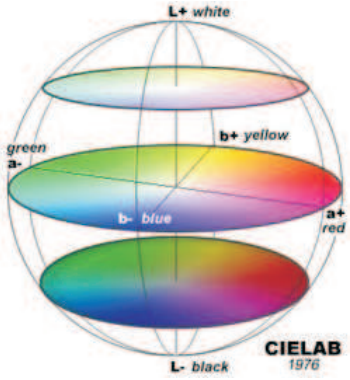

Figure 2.15: CIELAB color space.

an area appears to emit more or less light. *Hue* is the attribute of a visual sensation according to which an area appears to be similar to one of the perceived colors red, yellow, green and blue, or to a combination of two of them. Colorfulness is the attribute according to which an area appears to exhibit more or less of its hue.

All these terms are perceptual terms, that is terms that define our perceptions of colored stimuli. They are not definitions of colorimetric quantities. Unfortunately, measuring and representing visual perception is difficult. This step requires a deeper knowledge of the human visual system that is not yet available. An attempt in this direction is represented by *color appearance models*, which take into account perceptual phenomena in the specification of color. With appropriate care, the CIELAB space can be considered a simple example of color appearance model*.

The CIELAB space is defined in terms of non-linear transformations from CIE XYZ tristimuli as follows

$$L^* = 116 f(Y/Y_n) - 16 \tag{2.12}$$

$$a^* = 500[f(X/X_n) - f(Y/Y_n)] \tag{2.13}$$

$$b^* = 200[f(Y/Y_n) - f(Z/Z_n)] \tag{2.14}$$

$$f(x) = \begin{cases} x^{1/3} & \text{if } x > 0.008856 \\ 7.787x + 16/116 & \text{if } x \leq 0.008856 \end{cases} \tag{2.15}$$

$$C^*_{ab} = \sqrt{(a^{*2} + b^{*2})} \tag{2.16}$$

$$h^*_{ab} = \arctan\left(\frac{b^*}{a^*}\right). \tag{2.17}$$

Here, X, Y and Z are the tristimulus values of the considered color, while X_n, Y_n and Z_n are the tristimulus values of the reference white. The reference white allows to fix unit values of tristimulus values. L^* represents lightness, a^* approximate redness-greenness, b^* approximate yellowness-blueness, C^*_{ab} chroma, and h^*_{ab} hue. Equation (2.12) takes into account the non-linearity of human vision perceptual response to luminance.

The L^*, a^*, and b^* coordinates are used to construct a Cartesian color space (Figure 2.15). The L^*, C^*_{ab}, and h^*_{ab} coordinates are the cylindrical representation of the same space. The Euclidean distance between two points in the $L^*a^*b^*$ space was taken to be a measure of the color difference in perceptually relevant units.

*The reader is referred to [30] for a complete discussion of color appearance models.

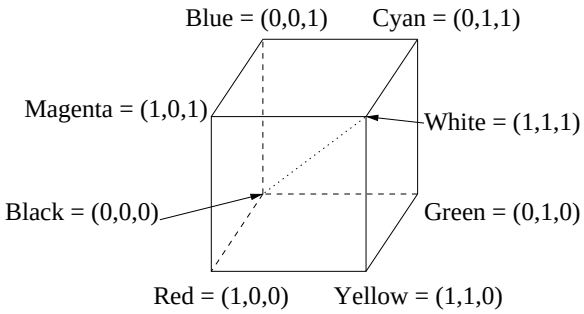

Figure 2.16: The RGB cube.

2.6.2 Device-oriented color spaces

The CIE colorimetric system represents a fundamental international standard for color measurements. Device-oriented color representation systems are associated with acquisition, reproduction and display devices. They allow to specify color in a way that is compatible with hardware tools such as television monitors, computer displays, color cameras, color scanners and color printers.

The color reproduced in a device-oriented color space depends on the equipment's characteristics and on the device set-up. It appears different if reproduced in another device space or if the device settings are changed. If the phosphors of a monitor change, for instance, the same color values produce a different color. A *calibrated* device color space is a color space whose position within the standard CIE colorimetric space is defined.

RGB

The red, green, and blue *RGB color space* is used for capture and display devices. It employs a Cartesian coordinate system, with the three axis corresponding to the red, green, and blue primaries. Since the primaries are characterized by a maximum intensity, the color solid of this system is a subset of the colors realizable by the possible primaries' mixtures. Using an appropriate scale along each primary axis, the space can be normalized, so that all colors lie in the unit cube shown in Figure 2.16. The main diagonal of the cube, with equal amounts of each primary, represents the grays: black is $(0, 0, 0)$ and white is $(1, 1, 1)$. Each color is reproduced by an additive mixture of the three primaries.

A number of RGB space variants are in use today. In the television industry, for instance, different standards have been defined by institutions in different countries. The adopted red, green, and blue primaries and the reference white are determined by the employed technology, such as the sensors in color cameras or the phosphors in cathode-ray tubes (CRTs). Recently, an international agreement has been reached on the primaries for the High Definition Television (HDTV) specification. These primaries are representative of contemporary monitors in computing, computer graphics and studio video production. The standard is known as ITU-R Recommendation BT.709 (formerly CCIR Rec. 709) [76]. It considers the CIE D65 illuminant as reference white. We will not go here into details of the different RGB primary systems specifications, but refer the interested reader to Poynton's book for a complete discussion [133].

The different RGB systems can be converted among each other using a linear transformation, assuming that the white reference values are known. Similarly, to convert from an RGB device space to the colorimetric CIE XYZ standard, a matrix transformation can be used. We report here as an example the transformation from the ITU-R BT.709 [131] standard RGB values in the range $[0, 1]$

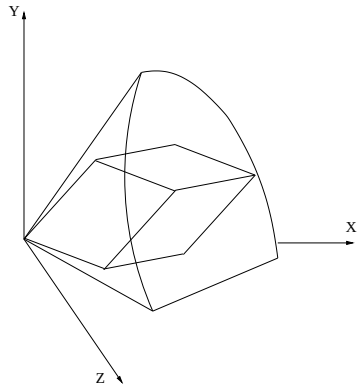

Figure 2.17: The RGB cube within the CIE XYZ space.

to CIE XYZ tristimulus values in the range $[0, 1]$:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.4125 & 0.3576 & 0.1804 \\ 0.2127 & 0.7152 & 0.0722 \\ 0.0193 & 0.1192 & 0.9502 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}. \tag{2.18}$$

The RGB values in Eq. (2.18) must not be confused with the RGB tristimulus values of the CIE RGB spectral primary system. They are based on the device primaries and are therefore device-dependent. Figure 2.17 shows a device's RGB space in the XYZ space. Only the spectral colors contained in the RGB cube can be captured or reproduced by the considered device.

The RGB representation is the most often used in image processing, computer graphics and multimedia systems. In practice, although a number of RGB space variants have been defined and are in use today, their exact specifications are usually not available to the end-user. This may cause inaccurate manipulation or reproduction of color images and pose application difficulties. An attempt to merge different, mainly device-dependent, color spaces into a single standard RGB space has been recently made by means of the *sRGB* space. The sRGB color space is based on the monitor characteristics expected in a dimly lit office. It has been standardized by the International Electrotechnical Commission (IEC)[72].

Normalized rgb

By dividing the R, G, and B coordinates by their total sum, the r, g, b quantities are obtained, which give the three components of the *normalized rgb color system.* The transformation from RGB coordinates to normalized color is thus given by

$$r = \frac{R}{R+G+B}, \tag{2.19}$$

$$g = \frac{G}{R+G+B}, \tag{2.20}$$

$$b = \frac{B}{R+G+B}. \tag{2.21}$$

This transformation projects radially a color vector in the RGB cube into a point on the unit plane shown in Figure 2.18. Two of the rgb values are sufficient to define the coordinates of the color point in this plane.

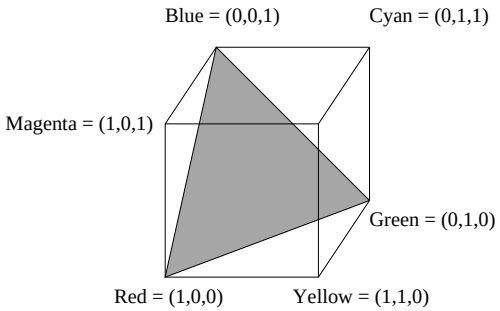

Figure 2.18: Unit plane in the RGB cube.

Since rgb is redundant because $b = 1 - r - g$, the normalized color space is also formulated as [122]

$$Y = c_1 R + c_2 G + c_3 B, \tag{2.22}$$

$$T_1 = \frac{R}{R + G + B}, \tag{2.23}$$

$$T_2 = \frac{G}{R + G + B}, \tag{2.24}$$

where c_1, c_2, and c_3 are chosen such that $c_1 + c_2 + c_3 = 1$. Y is interpreted as the luminance of the color point and *T1* and *T2* are chromatic variables.

$\mathbf{Y'UV}, \mathbf{Y'C_bC_r}$

In transmitting color images using RGB components a channel capacity that is three times that used for gray scale images is needed. To reduce these requirements, the properties of the human visual system can be exploited. There is strong evidence, as commented in Section 2.5.1, that the human visual system forms an achromatic channel and two chromatic color-difference channels in the retina. Consequently, it can be useful to convert the color signal into one component representative of luminance and two other components representative of color. The human visual system, moreover, has poor response to spatial detail in colored areas of the same luminance, compared to its response to luminance spatial detail. It is thus advantageous to transmit luminance with full detail and the two color components at lower resolution with substantially less data rate. These properties are exploited by video-oriented color representation systems.

In these systems, a weighted sum of RGB components is computed to form a signal representative of luminance. The resulting component is related to brightness but is not the CIE luminance. Many video engineers call it *luma* and denote it as Y'. However, it is important to underline that luma is very commonly called *luminance* and denoted as Y, which may cause ambiguity with the CIE notation. This issue will be discussed in more detail later in this subsection where the motivation for the adopted prime symbols to denote color components will be given. The simplest way to form the two color components is then to subtract luma from them. Since the large percentage (around 60%) of brightness is due to the green primary*, it is common to form the two color components by subtracting luma blue and red to form $(B' - Y')$ and $(R' - Y')$. These are called *chroma*. They are

*If three sources appear red, green, and blue and have the same power in the visible spectrum, the green will appear the brightest of the three because the human overall sensitivity with respect to the perceived brightness peeks in the green region of the spectrum [30].

generally sub-sampled for transmission in accordance with the weaker ability of the human visual system to discriminate spatially color information with respect to luminance spatial detail.

Various scale factors are applied to $(B' - Y')$ and $(R' - Y')$ for different applications [133]. The $Y'P_BP_R$ scale factors are optimized for component analog video. The $Y'C_BC_R$ scaling is appropriate for component digital video such as studio video, JPEG and MPEG. $Y'UV$ scaling is appropriate in the formation of composite NTSC (the American broadcast TV color system) or PAL (the European system) video signals. The YUV nomenclature is used rather loosely in the image processing community and it sometimes denotes any scaling of $(B' - Y')$ and $(R' - Y')$.

To compute $Y'C_bC_r$ from $R'G'B'$ in the range $[0, 255]$ the following matrix transformation is used [133]

$$\begin{bmatrix} Y' \\ C_b \\ C_r \end{bmatrix} = \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} + \begin{bmatrix} 0.257 & 0.504 & 0.098 \\ -0.148 & -0.291 & 0.439 \\ 0.439 & -0.368 & -0.071 \end{bmatrix} \begin{bmatrix} R' \\ G' \\ B' \end{bmatrix}. \tag{2.25}$$

The inverse of the transformation in Eq. (2.25) is used for the $Y'C_bC_r$ to $R'G'B'$ conversion

$$\begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} = \begin{bmatrix} 1.164 & 0.000 & 1.596 \\ 1.164 & -0.392 & -0.813 \\ 1.164 & 2.017 & 0.000 \end{bmatrix} \left(\begin{bmatrix} Y' \\ C_b \\ C_r \end{bmatrix} - \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} \right). \tag{2.26}$$

Y' has an excursion of 219 and an offset of +16. This coding places black at 16 and white at 235. C_B and C_R have excursions of ± 112 and offset of +128, for a range of 16 through 240 inclusive [77].

The $R'G'B'$ to $Y'UV$ mapping is defined as follows [133]

$$\begin{bmatrix} Y' \\ U \\ V \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.147 & -0.289 & 0.437 \\ 0.615 & -0.515 & -0.100 \end{bmatrix} \begin{bmatrix} R' \\ G' \\ B' \end{bmatrix}. \tag{2.27}$$

The inverse of the matrix in Eq. (2.27) is used for the $Y'UV$ to $R'G'B'$ conversion

$$\begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} = \begin{bmatrix} 1.000 & 0.000 & 1.140 \\ 1.000 & -0.394 & -0.581 \\ 1.000 & 2.028 & 0.000 \end{bmatrix} \begin{bmatrix} Y' \\ U \\ V \end{bmatrix}. \tag{2.28}$$

In video-oriented models, the transformation from RGB values to luminance-chrominance values is not applied directly on the primaries values. First, in fact, a nonlinear transfer function, the *gamma correction* [132], is applied to each of the R, G and B values, giving the nonlinear R', G', and B'. This explains the adopted prime symbols. The gamma function is applied to compensate the nonlinearity of CRTs response to the applied voltage. The CRT's phosphors response to the applied voltage follows in fact a power law, x^γ. The primary RGB signals are therefore corrected to compensate this effect by applying an inverse law, $x^{\frac{1}{\gamma}}$. For the NTSC television standard the adopted γ is equal to 2.2. For the PAL standard $\gamma = 2.8$.

To get CIE XYZ tristimulus values from $Y'C_BC_R$ or $Y'UV$, Eq. (2.26) and Eq. (2.28) have to be used to get first of all nonlinear $R'G'B'$ values and then the inverse of the gamma function has to be applied to get linear RGB values. Once RGB values, device primaries coordinates and reference white are known, CIE XYZ can be obtained by means of the appropriate matrix transformation.

Poynton [133] observes that, for transmission purposes, it is important to convey the component representative of luminance in such a way that noise introduced in transmission, processing and storage has a perceptually similar effect across the entire scale from black to white. The ideal way to do this would be to form a luminance signal as a weighted sum of RGB values and processing it

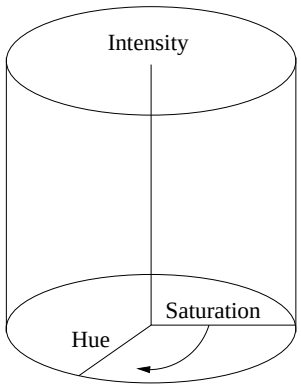

Figure 2.19: The HSI color model.

by means of a nonlinear function similar to the L^* function in the CIELAB space (see Eq. (2.12)). In video transmission, as commented above, these operations are performed in the opposite order for practical reasons, by first applying gamma correction and then by computing luma Y' as a weighted sum of nonlinear R', G' and B' values. However, Poynton observes that the nonlinear gamma function represents a good approximation of the lightness response of the human visual system. By applying gamma correction on the RGB values a representation of color that is closer to human perception is thus obtained.

2.6.3 User-oriented color spaces

None of the device-oriented color models is particularly easy to use for a user that has to numerically specify colors. These models, in fact, are not directly connected with intuitive color notions of brightness, hue, and saturation. As mentioned in Section 2.6.1, *brightness* is the attribute of a visual sensation according to which an area appears to emit more or less light. *Hue* is the attribute of a visual sensation according to which an area appears to be similar to one of the perceived colors red, yellow, green and blue, or to a combination of two of them. *Saturation* is defined as the colorfulness of an area judged in proportion to its brightness. Colorfulness is the attribute according to which an area appears to exhibit more or less of its hue. User-oriented models are then a class of models that has been developed with ease of use as a goal. These models are better suited than device-oriented spaces for human interaction. They try to build a bridge between the user and the hardware used to manipulate color.

There are many similar spaces that achieve *hue-saturation-brightness* characteristics. A comprehensive review can be found in [163]. The HSV space and the HSI space are two examples of these models which are commonly used in the literature. Their color solids are deformations of an RGB cube. The main diagonal of the cube defines the brightness axis. The color is then defined as a position on a circular plane around the axis. Hue is the angle from a reference point around the circle to the color, while saturation is the radius from the central brightness axis to the color. Approximately cylindrical coordinates are used. We describe in the following the HSI color space, which will be of interest in this thesis, as representative of this class of spaces.

HSI

Among the many similar formulations of the HSI (*hue*, *saturation*, *intensity*) space, we choose here that described in [122]. The model is defined as a cylindrical space, where the coordinates r, θ,

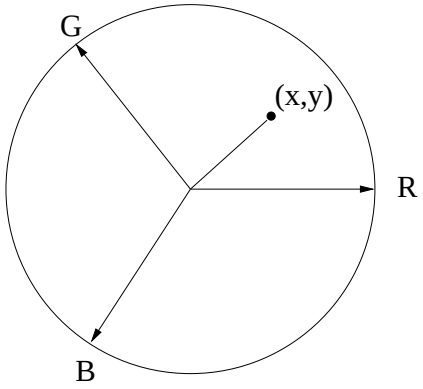

Figure 2.20: Model for hue based on a weighted average of RGB vectors [122].

and z respectively correspond to saturation, hue, and intensity. The HSI color solid is depicted in Figure 2.19.

Hue is the angle around the vertical intensity axis, with red at 0°. It gives a measure of the spectral composition of a color. It refers in fact to the wavelength of the pure color, the so-called dominant wavelength, that mixed to white produces the color under analysis. The complement of any hue is located 180° further around the cylinder. Saturation is measured radially from the vertical axis, from 0 on the axis to 1 on the surface. This component refers to the amount of white added to the dominant wavelength to produce the color under analysis. The more the amount of white, the less the saturation of the color. Intensity is 0 for black and 1 for white and is a measure of lightness.

The transformation from RGB to HSI is defined as

$$I = \frac{R+G+B}{3}, \tag{2.29}$$

$$S = 1 - \frac{min(R,G,B)}{I}, \tag{2.30}$$

$$H = \arctan\left(\frac{\sqrt{3}(G-B)}{(R-G)+(R-B)}\right). \tag{2.31}$$

In the last equation, $\arctan(y/x)$ utilizes the signs of both y and x to determine the quadrant in which the resulting angle lies. Hue is *undefined* when saturation S is zero, that is at any achromatic point along the intensity axis. The physical model used to determine the hue angle in this transformation is based on the diagram shown in Figure 2.20. If the R, G, and B radial basis vectors are equally spaced $\frac{2}{3}\pi$ apart on the unit circle, then the x and y components of an arbitrary point are given by

$$x = R - \frac{G+B}{2} = \frac{1}{2}\left[(R-G)+(R-B)\right], \tag{2.32}$$

$$y = \frac{\sqrt{3}}{2}(G-B). \tag{2.33}$$

This results in the hue angle in Eq. (2.31).

The HSI model allows users to specify color in terms of perceptual attributes and has a good compatibility with human intuition. Any color with $I = 1$ and $S = 1$ is akin to an artist's pure pigment used as the starting point in mixing colors. Adding white corresponds to decreasing S,

without changing I. Shades are created by keeping $S = 1$ and decreasing I. Tones are created by decreasing both S and I. Changing H corresponds to selecting the pure pigment with which to start. Thus, H, S and I correspond to concepts from the artist's color system, and are not exactly the same as the similar terms introduced at the beginning of the section. The formulations of H, S, and I are then flawed with respect to the properties of human color vision. Consequently, if hue and saturation have to be specified by numerical values for perceptual image computation, the polar coordinate versions of a^* and b^*, C^*_{ab} and h^*_{ab}, in the CIELAB space should be preferred.

The HSI model has some significant drawbacks, such as

- singularities in the transform, such as undefined hue for achromatic points,
- sensitivity to small deviations of RGB values near singular points,
- numerical instability when operating on hue due to the angular nature of the feature.

The main features that have made the HSI model appealing to many image processing applications are essentially

- the separability of chromatic values from achromatic values,
- the possibility of using one color feature only, hue, for segmentation purposes. Segmentation is performed on one color feature, instead of three, allowing the use of much faster algorithms.

2.7 Summary

The objective of this first chapter is to introduce the background notions, concepts and models related to the physics of image formation and the generation and representation of color in digital images. They are used in the follow-up of the thesis to derive the proposed approach to the problem of shadow segmentation.

The chapter is organized in two parts. In the first part, which comprises Section 2.2, Section 2.3, and Section 2.4, the path of light from sources of illumination to surfaces and from surfaces to capturing devices was described. Modeling the interaction of light with matter is central to this part of the chapter and to the entire image formation process. Among the variety of models proposed to this end in literature, the Dichromatic Reflection Model was described. The Dichromatic Reflection Model is derived from physics-based reflectance models and therefore is more suitable than empirical models to describe the color appearance of surfaces. Moreover, it allows to ignore the unnecessary aspects of more accurate but cumbersome, and thus impractical, physics-based models. The need to extend the model's formulation when dealing with shadows was discussed.

In the second part, in Section 2.5 and Section 2.6, the description of the image formation process was completed by introducing the role of the capturing device which converts light into color image values. This last step provided the complete formalization of the chain linking a physical scene to its digital color image. The mathematical representations of color by means of color spaces were finally discussed.

Figure 2.21: Painting was born the first time the shadow of a man was outlined on a wall (Section A.1.1).

Shadows and shadow detection

3

3.1 Introduction

The first step in the development of efficient tools for the extraction of shadows in digital images and image sequences is an understanding of how shadows appear in images and what is peculiar to them. This chapter is dedicated to the characterization of shadows and to a review of the state of the art of shadow detection.

In the previous chapter, we have pointed out the fact that shadows are due to discontinuities, "holes" in the flow of light from a source of illumination to a vision system. There, we have introduced models which allow to describe the journey of light from sources of illumination to the imaging device and to explain the resulting pixel values in digital color images as a function of physical phenomena. In this chapter, by means of the discussed models, we aim at analyzing and characterizing the effects of these "holes" on the values of the digital images we are dealing with. Moreover, since all shadows are shadows of something, additional spatial and temporal properties relating shadows to shadow-casting objects will be identified, which allow to characterize shadows for their extraction.

Due to their nature of *absence* of light and the fact that shadows do not exist in themselves but rather as shadows of something, shadows are unfortunately a difficult phenomenon to model and detect in images. The difficulty but at the same time the usefulness of analyzing shadows in different research and application domains is demonstrated by the fact that the problem of shadow detection has been increasingly addressed over the past years [134]. The state of the art of shadow detection is reviewed in the second part of the chapter.

The presentation is organized as follows. In Section 3.2, first of all, definitions concerning shadows are given and the terminology that will be used throughout this thesis is introduced. Then, a review of cues that suggest the presence of shadows in visual scenes is presented. The two components that characterize shadows in images, that is the spectral component related to the fact that shadows are due to an absence of light, and the geometric component related to the fact that shadows are generated by objects that obstruct a light source, are then more formally analyzed in Section 3.3. The state of the art methods for shadow detection proposed in the literature are finally reviewed in Section 3.4.

Figure 3.1: Light illuminating a face. Leonardo da Vinci, Codice Urbinate Latino (image from [10]).

3.2 What is a shadow?

3.2.1 Terminology and definitions

Shadows are generated by a *local* and *relative* absence of light. Shadows are, first of all, a local decrease in the amount of light that reaches a surface. Secondly, they are a local change in the amount of light reflected by a surface toward the observer.

There are three different types of absence of light:

- projected shadows, which include cast shadows;
- self shadows;
- shading.

They were clearly illustrated already in the fifteenth century by Leonardo da Vinci. We will take an example from Leonardo's work to describe them here by means of Figure 3.1. The illustrated drawing shows an illuminated human face. A is the source of light radiating toward the man's face, with incidence angles indicated by the letters from B to M.

Light reaches surfaces which obstruct its flow in two points, between I and K on the lower part of the man's nose, and between L and M on the chin. The tip of the nose prevents light from reaching the upper lip. The chin prevents light from illuminating the neck. The neck and the chin would otherwise receive some illumination. This is a first type of shadow, the **projected shadow**. In this case, the projected shadow is an **intrinsic shadow** because it is cast by an object on itself. A projected shadow which is cast by an object on a surface which belongs to a different object is called **cast shadow** or **extrinsic shadow**. Cast shadows are not illustrated by Leonardo's drawing and an example can be found in Figure 3.2. There, the sheep projects a cast shadow on the grass.

Referring again to Figure 3.1, the lower part of the nose and the lower part of the man's chin do not receive light from the source in A. In these cases, light is not occluded by an object, but rather these two parts have an orientation with respect to the source which prevents them from receiving any light from it. This is a second type of shadow, the **self shadow** or **attached shadow**. An example of self shadow is also illustrated in Figure 3.2.

Finally, a third type of absence of light is only partial. It is due to the fact that surfaces directly facing the source, but with different orientations, receive different amounts of light. The part of

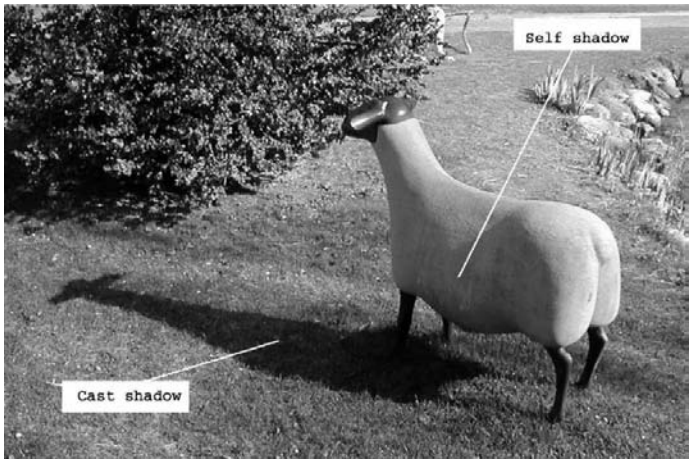

Figure 3.2: Example of self and cast shadow. Sculpture at the Fondation Pierre Gianadda, Martigny, Switzerland.

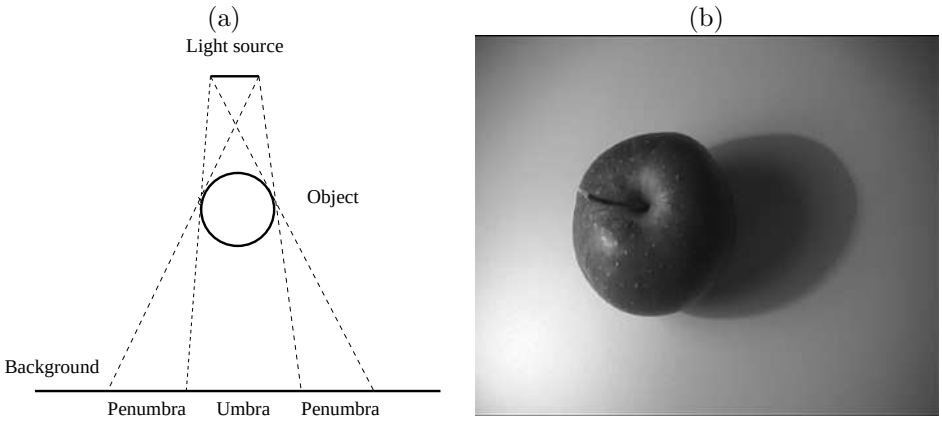

Figure 3.3: Umbra and penumbra generation. Area light sources, as the line source in (a), generate penumbra in projected shadows where the light is only partially obstructed by the shadow casting object. The umbra and penumbra structure is clearly visible in (b).

the nose between H and I appears indeed brighter when compared to the head between D and E because the orientation of light rays with respect to the surface is closer to the surface normal in those points. In this case, we do not speak of shadows but of *shading*. Shading due to the curvature of the surface is clearly visible in Figure 3.2 on the sheep's neck.

In Leonardo's illustration, the light source is what we have defined in Section 2.2.2 as a point light source. If the occluded light source has a certain extent, that is it can be modeled as an area light source, the outer portion of a projected shadow results from only the partial obstruction of light. This is the *penumbra* of the shadow. The *umbra* of a shadow is the part of the shadow where direct light is completely blocked. With direct light we denote light arriving along a direct line-of-sight from a light source. Umbra and penumbra are illustrated in Figure 3.3. A point light source only generates umbrae in projected shadows. The umbra is only illuminated by ambient light or by other light sources.

3.2.2 Shadow cues

There are a number of cues which suggest the presence of shadows in a visual scene and that could be exploited for their detection in digital images and image sequences. Funka-Lea [43] presents a complete list of them. We follow his analysis and discuss them in the following.

1. Shadows darken the surface upon which they are cast.

   The most obvious property of a surface in shadow is that it looks darker when compared to the same surface directly facing a source of illumination.

2. The change in the color of a surface due to the presence of a shadow tends to be predictable.

   This second property characterizes the relationship between shadows and lit regions on colored surfaces. Colored surfaces generally help in the task of distinguishing shadows from dark surface marks. Color information will indeed play an important role in the approach for shadow segmentation proposed in this thesis.

3. Surface markings and texture tend to continue across a shadow boundary.

   The continuation of surface texture across a shadow boundary is another cue that can be exploited for shadow detection.

4. Shadows of extended light sources tend to have smooth boundaries.

   As commented in the previous section, shadows generated by extended light sources present a penumbra where light from the source is only partially occluded. The outer boundary of a shadow with penumbra is characterized by a decrease in intensity toward a relatively uniform darker central region, the umbra. As a consequence, the boundary looks "soft".

   The width of the penumbra depends on the geometry of the light source and on the geometry of the occluding surface. It increases with an increase in the size of the source and in the distance of the occluding surface from the surface where the shadow it cast. It decreases when the distance from the source to the occluding surface increases (Figure 3.3 (a) can help in understanding this relationship). The intensity variations in a penumbra are a complex

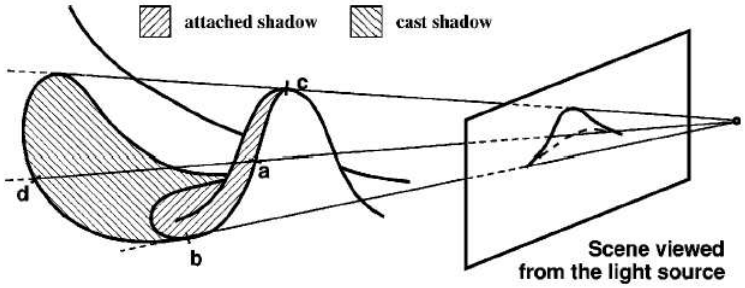

Figure 3.4: The boundary of a self or attached shadow is the outline of the shadow-casting object as seen from the light source. The boundary of the corresponding cast shadow is the projection of this contour in the direction of light rays (image from [83]).

function of the geometry of the light source and of the occluding surface, as discussed by Jiang and Ward [79]. It is therefore extremely difficult to obtain a theoretical model for an arbitrary object and an arbitrary light source.

The above-discussed cues describe shadows from a spectral point of view. For what concerns the geometry of shadows, the most obvious cue relates a shadow to its shadow-casting object.

5. A cast shadow is only possible if there is an object between the surface on which the shadow is cast and the source of illumination.

   This cue involves knowledge about the 3D position of the shadow-casting object and of the light source in the scene. This information is typically not available to image analysis algorithms and difficult to obtain automatically from images. The cue could be used in a weaker sense to rule out the possibility of a shadow if no object can be found in the image plane between the shadow and the light source. This still requires knowledge of the light source location in the image plane. In an ever weaker sense, the position of a shadow with respect to an object could be simply checked if an object can be recognized adjacent to the shadow. The possibility of a shadow could be ruled out if the shadow is inside the object and not at its boundary.

6. The shape of a shadow cast on a surface is the projection of the silhouette of the object casting it.

   As illustrated in Figure 3.4 for a simple object and a point light source, the boundary of a self shadow is the outline of the shadow-casting object as seen from the light source. The boundary of the corresponding cast shadow is the projection of this contour in the direction of light rays. The nature of the projection can be complex, especially for extended light sources. This fact is illustrated by the example of Figure 3.5. Even for a light source and an object with simple shapes as those shown in the figure, matching the shape of the resulting shadow with that of the object is not straightforward. Mamassian et al. [94] in their study on the perception of cast shadows conclude that the matching procedure would appear to be computationally prohibitive even for the simplest objects also for the human visual system.

7. Shadow boundaries tend to change direction with changes in the geometry of the surface on which they are cast.

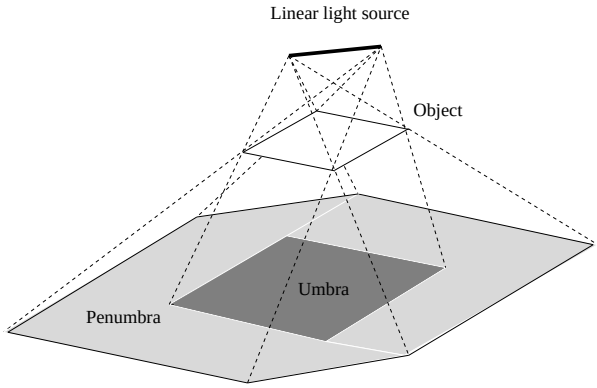

Figure 3.5: Shadow umbra and penumbra resulting from a line light source. Matching the shape of the cast shadow to that of the shadow-casting object is not straightforward even for a simple object and a simple light source.

Figure 3.6: Shadow boundaries change direction with changes in the geometry of the surface on which they are cast.

Shadows cast on surfaces inherit the shape of the surface on which they are projected. Shadow boundaries change therefore direction at surface discontinuities (see Figure 3.6). 3D information about the scene is required to exploit this cue.

When dynamic scenes are considered, additional cues can be identified.

8. Shadows cast by objects moving with respect to a fixed light source move across the scene.

   The most obvious temporal property of shadows cast by moving objects in dynamic scenes is their motion.

9. The motion of a shadow-casting object that moves relative to a fixed light source and that of its shadow are correlated.

   This last property of cast shadows has been shown to be very relevant in their perceptual interpretation. Mamassian et al. [94] observe that the fact that the relative motion of an object and its cast shadow is constrained to follow a line connecting the object to the light source represents a reliable cue to shadow labeling for human observers. The presence of such

a constrained motion in an image is a strong indicator that two moving patches are related as an object and its shadow. This percept is moreover shown to be robust to violations of shadow luminance and shape constraints. The authors suggest the possibility that the human visual system contains low-level mechanisms for detecting such correlations.

Discussion

The above-discussed visual cues provide a starting point for the development of effective techniques for shadow detection. Detecting shadows in images and image sequences is a difficult problem. In order to confidently recognize shadows, in fact, knowledge about the materials, the three-dimensional scene's layout and the lighting are needed. If the lighting, that is a characterization of the light at any point in the scene, and the material properties of a given surface are known, it can be deduced that a change in the appearance of the surface is due to a change in irradiance. With this knowledge and the determination that light from a source of illumination has been obstructed, it can be concluded that a shadow is present.

This is more knowledge that it can be expected a viewer or vision system to have when recognizing shadows in a scene. As for many other vision problems, we cannot hope to distinguish shadows from material and geometric changes with certainty. The problem is underconstrained. The more prior knowledge that can be used, the better the chances of successful analysis. Before proceeding to a formalization of visual cues leading to explicit criteria for shadow detection, therefore, the framework within which shadow detection is applied, that is the type of image content, the available a priori knowledge and whether user intervention or control on the imaging process is possible, has to be defined. The targeted framework determines the cues and the related constraints to be considered in order to solve the problem.

The methodology for shadow segmentation that we propose in this thesis is addressed to a wide range of real world scenes whose content is not a priori known. The developed tools are designed to be able to work even without knowledge of the illumination conditions, scene geometry and camera characteristics and without the need of user intervention. This feature is highly desirable for a wide range of applications, such as video production, video surveillance, and immersive gaming. Within this framework, we will make use of the first and the second spectral shadow cues (**cue 1 and 2**, as numbered above), of the first geometrical cue (**cue 5**) and of the first temporal cue (**cue 8**). In Chapter 5 we will describe how it is possible to effectively exploit such cues for shadow segmentation. In Section 3.3, first of all, a formalization of the selected spectral cues is provided together with a more detailed description of the exploited geometric characterization. The respective assumptions related to the targeted framework will be clearly stated and discussed.

In the proposed approach, we will not exploit cue 3 since our experience has shown that the description of surface texture when present in imaged shadow regions may become ineffective due to the limited amount of light reaching the surface. We will neither use cue 4, since, although the penumbra is a strong shadow cue for human observers [100], it is not always exploitable when analyzing digital images. The reasons behind our choice are the following. A penumbra region in projected shadows is typically visible in indoor scenes where light bulbs generate smooth shadow edges. In outdoor sunny scenes, on the contrary, the Sun can be reasonably approximated as a point light source at infinity, as discussed in Section 2.2.2. The penumbra the Sun generates is therefore small. Typically, if the distance between an area light source and the shadow casting-object is much bigger than the light source size and when the distance between the occluding surface and the surface where the shadow is cast is limited, the penumbra width will be small. In such cases, detecting penumbra in images as a cue to the presence of a shadow becomes difficult. In addition to the mentioned difficulties, penumbra in digital images may be confused with aliasing at the contours.

When shadows are cast on textured surfaces, moreover, it can be difficult to discriminate intensity changes due to penumbra and due to surface texture. Among the geometric cues, cue 6 requires knowledge of the 3D shape of shadow-casting objects and casting surfaces and is extremely difficult to exploit for an arbitrary scene geometry even in case this knowledge is available. We discard cue 7 as well as we deal with monocular cameras and unknown 3D scene geometry. We discard finally cue 9 since, although perceptually very relevant, the correspondence between the motion of an object and the motion of its cast shadow can be complex for an arbitrary scene and difficult to model and analyze in image sequences. In case a different framework is targeted, i.e. for instance when stereo or multiple cameras are available, the modularity of the approach we propose allows the introduction of further cues among those here discarded, such as cue 7 that can be easily exploited by means of homography [40] when shadows are cast on planar surfaces.

3.3 Modeling shadows appearance in images

Shadows are due to a relative absence of light and their spectral characteristics, that is their brightness and color, change with changes in the surface on which they are cast. The spectral characteristics of a shadow depend then on the characteristics of the light that illuminates the shadow compared to the light that would illuminate the same area if there were no obstruction. The spectral characterization of shadows involves therefore a comparison with respect to a situation where the light occlusion is not present. It is discussed in Section 3.3.1. The analysis is valid for both self and cast shadows.

As discussed in the previous section, the geometry of a shadow, that is its shape and location, is determined by the 3D shapes of the occluding surface, the light source and the surface on which the shadow is cast, and on the relative position of object, light source and viewer. Without this knowledge, it is still possible to identify some useful geometric characteristics of shadows. Such characteristics are discussed in Section 3.3.2.

3.3.1 The spectral appearance of shadows

As discussed in Chapter 2, the sensor responses of a digital camera depend both on the surfaces in a scene and on the illumination. Hence, a single surface viewed under two different illumination conditions will yield two different sets of sensor responses. If the change in the illumination conditions is due to the presence of a shadow, what can be said about the relationship between the two sets of sensor responses?

In order to formalize the above-mentioned question, we will use the instruments discussed in the previous chapter. The Dichromatic Reflection Model presented in Section 2.4.1 and described in Eq. (2.6) provides us with a description of the signal reaching the camera sensors from a surface. We will use it to compare the signals reflected by the same surface when illuminated and when shadowed. Two types of illumination, that is direct illumination and ambient illumination, are considered by the model. In this context, direct light represents light from the source of illumination that is occluded when a shadow-casting object is present in the scene. This light is absent in shadows. Ambient illumination represents all other light scattered in the neighboring environment which illuminates also shadowed points.

Let us recall here Eq. (2.6). The radiance of the light reflected at a given point on a surface in the 3D space, given some illumination and viewing geometry (see Figure 2.9), is formulated as

$$L_{lit}(\lambda, i, e, g) = L_s(\lambda, i, e, g) + L_b(\lambda, i, e, g) + L_a(\lambda), \tag{3.1}$$

where $L_a(\lambda)$, $L_s(\lambda, i, e, g)$, and $L_b(\lambda, i, e, g)$ are the ambient, interface, and body reflection terms, respectively; i is the angle of incidence between illumination direction $\vec{I}$ and surface normal $\vec{N}$ at the considered point, e is the angle of exitance between $\vec{N}$ and the viewing direction $\vec{V}$, g is the phase angle between $\vec{I}$ and $\vec{V}$, and λ is the wavelength.

If there is no direct illumination in the point under analysis because an object is obstructing the direct light, then the radiance of the reflected light becomes

$$L_{shadow}(\lambda, i, e, g) = L'_a(\lambda), \tag{3.2}$$

where $L'_a(\lambda)$ is the ambient reflection term in presence of the occluding object.

Let $S_R(\lambda)$, $S_G(\lambda)$, and $S_B(\lambda)$ be the spectral sensitivities of the red, green, and blue sensors of a color camera, respectively. According to Eq. (2.8), the color components of the reflected intensity reaching the sensors at a point (x, y) in the 2D image plane are obtained as

$$C_i(x, y) = \int_\Lambda E(\lambda, x, y) S_{C_i}(\lambda) d\lambda, \tag{3.3}$$

where $C_i \in \{R, G, B\}$ are the sensors responses, $E(\lambda, x, y)$ is the image irradiance at point (x, y), and $S_{C_i}(\lambda) \in \{S_R(\lambda), S_G(\lambda), S_B(\lambda)\}$; Λ is determined by $S_i(\lambda)$, which is non-zero over a bounded interval of wavelengths λ.

Since image irradiance is proportional to scene radiance, as commented in Section 2.5.2, for a pixel position (x, y) corresponding to the point under analysis in 3D space, the sensor measurements when the point is in direct light are

$$C_i(x, y)_{lit} = \int_\Lambda \Big[L_s(\lambda, i, e, g) + L_b(\lambda, i, e, g) + L_a(\lambda) \Big] S_{C_i}(\lambda) d\lambda \tag{3.4}$$

giving a color vector $\vec{C}_{lit}(x, y) = (R_{lit}, G_{lit}, B_{lit})$. Since we are not interested in an absolute scale factor, but we are comparing a situation where a shadow is present with one where it is absent, we can assume that the constant of proportionality between image irradiance and scene radiance in Eq. (3.4) is unity. For a point in shadow the measurements are

$$C_i(x, y)_{shadow} = \int_\Lambda L'_a(\lambda) S_{C_i}(\lambda) d\lambda \tag{3.5}$$

giving a color vector $\vec{C}_{shadow}(x, y) = (R_{shadow}, G_{shadow}, B_{shadow})$. In order to define explicit criteria for shadow segmentation, $\vec{C}_{lit}(x, y)$ and $\vec{C}_{shadow}(x, y)$ have now to be related.

To simplify the problem, let us assume that $L'_a(\lambda) = L_a(\lambda)$, that is ambient light is not influenced by the presence of the shadow-casting object. Then,

$$C_i(x, y)_{shadow} = \int_\Lambda L_a(\lambda) S_{C_i}(\lambda) d\lambda \tag{3.6}$$

and

$$C_i(x, y)_{lit} = \int_\Lambda \Big[L_s(\lambda, i, e, g) + L_b(\lambda, i, e, g) \Big] S_{C_i}(\lambda) d\lambda + C_i(x, y)_{shadow}. \tag{3.7}$$

It is straightforward that, when considering a constant ambient term to model all light coming from the environment which is not coming from the obstructed light source, each of the three RGB color components for a point on a surface, if positive and not zero, decreases when the point passes from being lit to being shadowed, that is

$$\begin{aligned} R_{shadow} &< R_{lit}, \\ G_{shadow} &< G_{lit}, \\ B_{shadow} &< B_{lit}. \end{aligned} \tag{3.8}$$

Equation (3.8) formalizes, under the considered assumptions, the first simple property in the list reported in Section 3.2.2. We will use it in Chapter 5.

In order to be able to say something more about the relationship between sensor responses for a point on a surface when in light and when in shadow, the characteristics of the direct and ambient illumination have to be considered. The spectral composition of the ambient light can be different from that of the incident light [47]. The case of outdoor sunny scenes, where shadows are illuminated by diffuse skylight which is bluer than direct light from the Sun, provides an example [31, 95, 109]. Another case is when a neighboring object is casting its color on the observed surface. This case is referred to as *inter-reflection*. Figure 3.7 (b) illustrates an example of inter-reflection in a real image. Figure 3.7 (a) shows the same scene without inter-reflection for direct comparison. The effect of the inter-reflection is visible on the upper left portion of the apple, which is brighter in (b) when compared to (a). Local effects due to inter-object reflection are extremely hard to analyze [40]*.

Figure 3.7: Example of inter-reflection in a real image. In (b) a white object is casting light on the observed object.

Relating direct and ambient illumination is a hard problem that requires a priori knowledge about the scene content. Without a priori knowledge about the scene, it is still possible to consider appropriate assumptions which allow to make the problem manageable. If the ambient illumination and the direct illumination are assumed to have the same color, formalizing the second property in the list of Section 3.2.2 becomes in fact possible. The considered condition is referred to as the *gray world condition*. The average of all the different reflectances in the scene is in fact considered to be a spectrally flat "gray". In this case, the camera response to the ambient light contribution, that is the camera response in shadowed points, is a linear combination of the responses to the body and interface reflection terms due to direct light [143].

Moreover, if regions that do not contain highlights are considered, that is if only the body reflection term is present, from the previous observation it follows that the color components for the same point in light and in shadow are related by a multiplicative constant as

$$\begin{aligned} R_{shadow} &= \alpha R_{lit}, \\ G_{shadow} &= \alpha G_{lit}, \\ B_{shadow} &= \alpha B_{lit}, \end{aligned} \tag{3.9}$$

with $\alpha < 1$.

Equation (3.9) defines a local relationship. The value of α changes from point to point on a surface according to changes in the surface orientation with respect to the illumination direction.

*A detailed analysis of mutual illumination for simple scene geometries can be found in [39, 45].

The body reflection term $L_b(\lambda, i, e, g)$ depends in fact on the incidence angle i between the normal to the viewed surface and the direction of illumination. The parameter α changes moreover in the penumbra of the shadow because of the change in the incident irradiance. If the surface upon which the shadow is cast is planar and the light source is distant from it, then the normal to the surface in its different points could be considered constant with respect to the illumination direction. In this case, a unique value of α would characterize the relationship between shadowed and lit points for the entire shadow's umbra. Equation (3.9) could be used as a global criterion for shadow segmentation in the umbra regions. Penumbra should be then separately treated.

In order to avoid the above discussed further assumption on the scene and a separate analysis of shadows penumbra, which is, as discussed in Section 3.2.2, difficult to handle in digital images, we will not make direct use of Eq. (3.9) in the methodology for shadow segmentation proposed in Chapter 5. We will utilize it indirectly through the use of color invariant features which will be discussed in detail in the next chapter. Consequently, the spectral characterization of shadows will be completed there.

Discussion

When the spectral characteristics of ambient and direct light are different, they can be modeled in the most general way as spectral power distributions as two different illuminants (see Section 2.2). Relating sensor responses for the same point in light and in shadow results therefore in relating sensor responses for the point under two different illuminants. When one of the two illuminants is unknown and the second is a known, reference illuminant, the problem is referred to as *computational color constancy problem*. It will be discussed again in more detail in Chapter 4.

Research in color constancy has shown that, under appropriate assumptions, the effects on color values of a change in illumination conditions can be modeled with a simple, but effective model, the so-called *Von Kries* or *diagonal scaling model* [30, 34]. According to this model, an illumination change from a first illuminant (1) to a second illuminant (2) can be described by an independent scaling of sensor responses in each channel as

$$\begin{bmatrix} R^{(1)} \\ G^{(1)} \\ B^{(1)} \end{bmatrix} = \begin{bmatrix} \alpha & 0 & 0 \\ 0 & \beta & 0 \\ 0 & 0 & \gamma \end{bmatrix} \begin{bmatrix} R^{(2)} \\ G^{(2)} \\ B^{(2)} \end{bmatrix}. \tag{3.10}$$

The scaling is independent of the surface reflectance but depends on camera characteristics and is affected by changes in surface orientation. Based on statistical analyses taking into account possible illuminants and surfaces, the model holds perfectly for Lambertian surfaces and cameras having sensors whose spectral sensitivities are Dirac delta functions. It holds approximately for Lambertian surfaces and real cameras having somewhat narrow-band sensors.

The diagonal scaling model is used by color constancy algorithms to pass from an image taken under an unknown illuminant to the same image under a different and known illuminant, using the coefficients α, β and γ. Since when the ambient and direct light have different spectral composition the same point when lit and when in shadow can be described as illuminated by two different illuminants, the diagonal scaling model can be used also as a general model for describing a change in illumination color at a point on a diffuse surface due to the presence of a shadow. In this case, both illuminants are unknown and it is difficult, as stated above, to determine the values of the model coefficients, which depend on the characteristics of ambient and direct light in the observed scene. Knowledge of the camera sensors [95], specific a priori information about the observed scene (e.g. the placement of calibration patches) [7, 31] or user intervention [105] are required to this end.

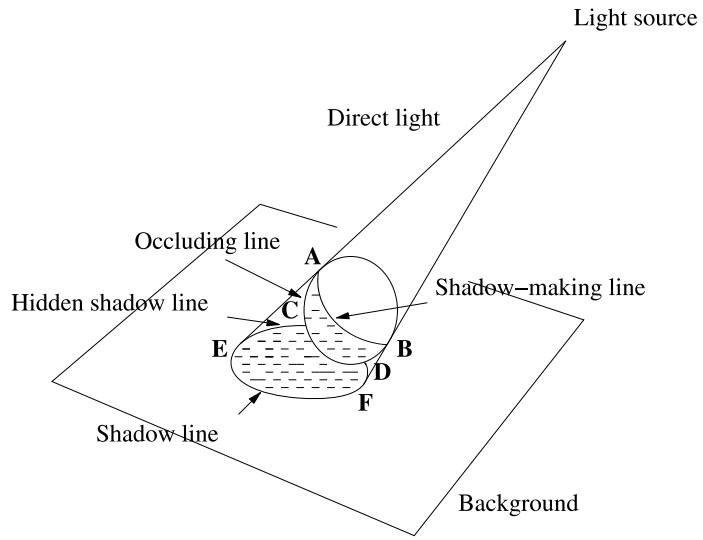

Figure 3.8: Shadow boundaries. AB is the *shadow-making line*, EF is the *shadow line*, $ACDB$ is the *occluding line*. Shadow-making lines and shadow lines are terminated at vertexes on the shadow boundary in A, B and C, D.

The assumption about ambient and direct light that we have considered for deriving Eq. (3.9) allows then to simplify the modeling of shadows and to automatically analyze shadows without knowledge of the scene content and of the camera's characteristics. As it will be demonstrated in Chapter 5, the considered assumption will be a useful approximation in the context of this work for developing efficient tools for the segmentation of shadows in a wide range of real world uncalibrated images.

3.3.2 The geometric appearance of shadows

Geometric interpretation of shadows in images and image sequences, that is recovering of information about a scene from knowledge of shadows shape and position, is an important operation in several computer vision applications, such as aerial image understanding (see Section 3.4.1) and shape reconstruction [14, 86]. The theoretical analyses of Waltz [164] and Shafer and Kanade [143, 144] were the first important works which demonstrated the advantage of introducing shadow interpretation into computer vision systems. In this section, we are interested in the inverse problem, that is exploiting the geometric characteristics of shadows in a scene as an aid in their recognition in images. Without knowledge of the 3D scene geometry and of the position of light sources and of the viewer, geometric characteristics of shadows can be derived by analyzing shadow boundaries and the position of shadows with respect to shadow-casting objects in the 2D image plane.

Following the analysis of Hambrick et al. [61], shadow boundaries can be divided into three types of segments: *shadow-making lines*, *shadow lines*, and *occluding lines*. These lines are illustrated in Figure 3.8, where the shadow of a spherical object resting on a flat surface is shown. Here, the term shadow indicates the ensemble of self and cast shadow. *Shadow-making lines*, AB, separate the illuminated surface and the non-illuminated surface of an object. They appear to be the outlines of an object if the position of the observer is aligned with the direction of the light source. The projection of the shadow-making lines along light rays defines the forward subsegment of the *shadow line*, EF. The rearward subsegment of the shadow line is projected from a hidden shadow-making line. In Figure 3.8, part of the rearward subsegment is visible between DF and CE. *Occluding*

lines, $ACDB$, block from view the rearward shadowed surfaces of the object. A subsegment of the occluding line, CD, defines the separation of the self shadow from its cast shadow when, as in the case illustrated in Figure 3.8, cast shadow and shadow-casting object are attached.

When an object lies on the surface on which its shadow is cast, the cast shadow is always attached to the object. The adjacency of an object and the position of the cast shadow at the boundary of this object is a characteristic feature of shadows that can be used to distinguish them from dark surface marks. We will exploit it in Chapter 5.

Shadow boundaries are characterized by the presence of vertexes. Shadow-making lines and shadow lines are, in fact, generally terminated at points on the shadow boundary where the derivatives are discontinuous. Vertexes on the shadow-making line are indicated by letters A and B in Figure 3.8, while vertexes on the shadow line are indicated by letters C and D. The projection of light rays in the image plane is tangent to the shadow boundary exactly at the vertexes on the shadow-making lines. Detecting such vertexes in images of arbitrary objects for which the shape of the shadow-making line may be very complex is however a challenging task.

The discussed characterization of shadows will be exploited in Chapter 5. In the following second part of the chapter, shadow detection methods in the literature are reviewed.

3.4 Shadow detection: state of the art

The problem of detecting, processing and analyzing shadows has been investigated within several research domains. From an historical point of view, the first methods for extracting shadows from images have been proposed in the field of aerial image understanding. In this context, shadows are generally detected as an aid in the recognition of objects and for the estimation of some 3D parameters of the depicted site, such as the height of buildings. With the spread of digital images in the last decade, a new interest for the detection and processing of shadows in images has recently emerged in digital photography applications, such as color correction and dynamic range compression [7], and content-based image indexing applications [136]. In the former case, identifying illumination changes due to strong shadows can help in image reproduction. In the latter case, shadows are analyzed as illumination effects that can provide useful information for indexing and retrieving images based on their content.

An especially increased interest in the extraction of moving shadows in image sequences has been motivated in recent years by the need of accurate object extraction tools for a variety of computer vision applications, including video surveillance, people tracking, traffic monitoring, and human motion capture. As discussed in Section 3.2.2, shadows cast by objects moving relative to a source of illumination have a corresponding motion. Since many techniques for the automatic extraction of objects in video make use of motion information, shadows are typically detected together with objects. A shadow cast by an object may either be in contact with the object, or disconnected from it. In the first case, the shadow distorts the object's shape and color, making the subsequent use of this information less reliable. In the second case, the shadow may be classified as a totally erroneous object in the scene. It is in this context that a lot of work has been recently proposed for explicitly detecting moving cast shadows in image sequences. Video surveillance and traffic monitoring are the applications for which the majority of contributions are proposed. Among emerging applications, interactive environments for gaming and storytelling can be cited [22, 97].

In the following, we present an overview of the different techniques proposed in the literature to solve the shadow detection problem. A first classification into two groups is proposed: model-based methods and property-based methods. *Model-based techniques* rely on models representing available a priori knowledge of the geometry of the scene, of the objects, and of the illumination direction.

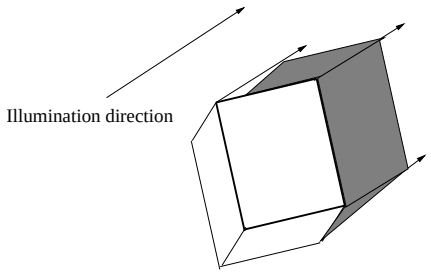

Figure 3.9: In aerial image analysis, objects of interest such as buildings can be described by means of simple rectilinear models and the known direction of illumination can be used to search for shadow evidence in dark regions as an aid in the extraction of objects and to derive height information.

They typically address frameworks which are different than that considered in this thesis. *Property-based* techniques identify shadows in a wider range of scenes by exploiting a combination of shadows spectral and geometric properties. Out of the scope of this work, and therefore not included in this review, are methods that rely on active processes such as the introduction of a second source of illumination (e.g. [175]). Moreover, since in this thesis we target image sequences taken with monocular cameras, methods that use stereo cameras (e.g. [62]) or multiple cameras (e.g. [116]) will only be briefly discussed.

3.4.1 Model-based techniques

The problem of detecting shadows in images initially arose in the domain of *aerial image understanding*. The developed methods aim at extracting cast shadows as a cue to the 3D nature of buildings with respect to planar roads or seas [111], or as an aid to the detection of buildings and changes in an observed site [11, 74, 92, 93]. The extracted shadows are typically exploited to estimate the height of objects [68] and to build a 3D site model. Cloud shadows are exploited in [147, 166] to detect clouds. In [67], vehicles are extracted by exploiting the geometric relationship with the shadows they cast.

Given the nature of the observed scenes and the available a priori information about the images (image orientation parameters, image acquisition's date and daytime), objects may be modeled by means of simple *rectilinear models* and knowledge about the *illumination direction* (e.g. the date and daytime at which the pictures have been taken and the latitude and longitude of the depicted site determine the Sun position in the scene) can be considered (see Figure 3.9). This greatly simplifies the modeling and extraction of shadows by means of geometric constraints. Mostly luminance information is used for analyzing shadows spectral characteristics, since generally the processed data are gray-level images. Shadow evidence is extracted, typically by means of thresholding, in the most dark regions of the image. The geometric relationship between potential shadows, shadow casting object models and light source direction is then analyzed. This process is mainly based on matching sets of geometric features such as edges, lines or corners to 3D object models and cast shadows predicted thanks to available information.

The use of object models and a priori information for improving, by means of cast shadow detection, the extraction of vehicles [84, 174] and other specific classes of objects, such as pedestrians [71, 149], has been exploited as well in *video surveillance* applications. In [84], 3D models of the structure of the moving objects and an illumination model which assumes parallel incoming light are used in the detection and tracking of *vehicles* in image sequences. The direction of illumination is interactively estimated off-line and allows to compute the visible contour of the 3D vehicle model's

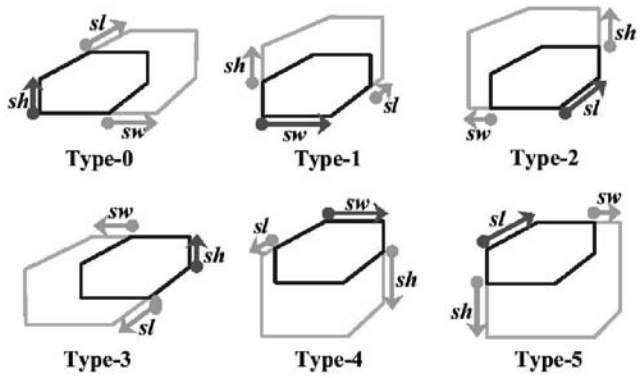

Figure 3.10: Six 2D joint-vehicle/shadow models used in [174] to separate objects from their cast shadows. The outer bounding box represents the fitted vehicle model for a vehicle that includes a cast shadow. The inner bounding box represents the fitted vehicle model for a vehicle without cast shadow. There is at least one side of the vehicle model whose location and length is not influenced by the presence of the shadow.

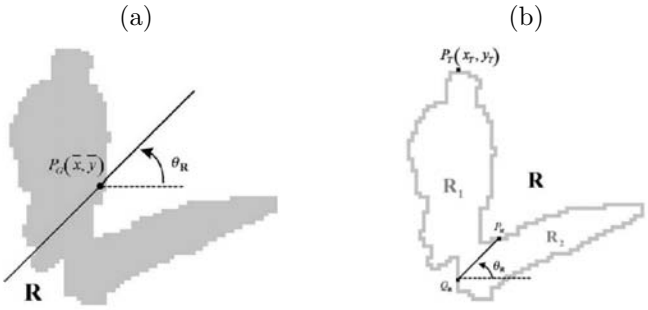

Figure 3.11: (a) Computation of the orientation of an object R representing a pedestrian on the road plane by means of its center of gravity as proposed in [71]. (b) Computation of a boundary line for cutting the shadow R_2 from the object R_1.

cast shadow projected onto the street plane. The straight line segments in which this contour is segmented are then matched with the image edge segments allowing the identification of shadows and a better object recognition and alignment. In [174], a simplified 2D joint vehicle/shadow model projected into the image plane is used to separate objects from their cast shadows without a priori knowledge of the light source position. Six different models are computed to represent possible locations of the cast shadow (see Figure 3.10).

When objects of interest are *pedestrians*, they are typically assumed to stand in an erect posture on the road plane [71, 149]. Therefore, cast shadows are attached to the persons' feet and, if the Sun is not too high in the sky, they form a characteristic angle with the vertical direction describing the objects. The shadow can thus be coarsely separated from the object by tracing an appropriate line (see Figure 3.11). The rough approximation of the extracted shadow region can then be further refined by using luminance and spatial information.

All the mentioned techniques are designed for specific applications and cannot be easily adapted to different scenarios. As stated, most of the algorithms assume that the illumination direction is

known a priori and they can handle only a specific class of objects or scenes, such as outdoor sunny scenes. Generally the detection of shadows results in a quite simplified process. In aerial images, for instance, detecting shadows is often easier than detecting buildings since shadows really appear to be the darkest areas in the image [161].

3.4.2 Property-based techniques

A combination of spectral and geometric properties of imaged shadows is exploited by property-based methods to overcome the limitations of application-specific methods and to develop approaches which can be applied to a wider class of scenes. While in the majority of the previously discussed methods shadows are detected as an aid in the extraction of objects, they are typically considered by property-based techniques as a noise component to be taken into account and removed for an accurate image and video analysis.

The human visual system is very good at describing scenes and recognizing objects disregarding illumination effects such as shadows. For instance, we do not make any conscious or unconscious effort to avoid them as they were an obstacle when we walk around. To mimic the behavior of human observers, a possible approach to the problem of removing annoying shadows is to obtain a description of images that is not influenced by the presence of shadows and of other illumination effects. An *implicit analysis of shadows* is performed in this case. Shadows are however exploited at some level in perception to retrieve information about objects shapes and locations [94]. A description of an image that is shadow-less removes therefore salient information for scene interpretation. This information could be exploited if shadows are explicitly identified, as it will be demonstrated in the following of this thesis. Techniques for the *explicit analysis of shadows* are therefore of more interest in the context of this work.

Following the characterization of shadows discussed in Section 3.2.2 and Section 3.3, we propose here to group state of the art techniques based on the type of shadow properties they exploit. Techniques that use spectral properties by means of gray-level or color features are first discussed, then techniques which make also use of geometric properties are presented. In order to provide a clear overview of specific problems and related choices, a distinction is made between approaches which deal with shadows in still images (Section 3.4.3) and approaches focusing on image sequences (Section 3.4.4). In the latter case, since the main purpose of shadow detection is the enhancement of object extraction algorithms, shadow detection techniques are typically associated with moving object extraction methods.

3.4.3 Still images

Gray-level image intensity — Gray-level image intensity is exploited for analyzing and classifying edges in [172] and for extracting shadow regions in [141]. In [172], edges are classified as belonging to a shadow or to an object by analyzing the intensity shifts across them. In [141], a shadow removal method based on a modification of the luminance histogram in images where objects occupy the upper most intensity range of the image and the image is background dominant is presented.

The reported methods represent early attempts with limited performance. Gray-level image intensity alone is, in fact, a poor source of information when trying to distinguish shadows from naturally dark surfaces in an image. More promising approaches are represented by methods that use color information.

Color — The approaches proposed in [7, 47, 52, 138, 139] aim at analyzing edges with respect to the possibility that they are due to a a material change as opposed to a shadow or other illumination effects.

The first analysis has been proposed by Rubin and Richards in [138]. A rule common to all edges arising from shadows, orientation changes, and highlights is proposed, under the assumption that the gray world condition (Section 3.3.1) holds. If the intensity at one wavelength decreases across one of these types of edges, then the intensity must also decrease at all other wavelengths across the same edge. When this condition is violated, a *spectral crosspoint* implying a material change is found. Figure 3.12 shows examples of image intensities at two different wavelengths for two points across an image edge. In (a) and (b) a spectral crosspoint indicating a material change is illustrated. In [139], the authors propose a second, independent condition. They argue that, when a pair of image regions is such that one region has greater intensity at one wavelength than at another wavelength, and the second region has the opposite property, then the two regions are likely to have arisen from different materials in the scene. They call this property the *opposite slope sign condition*. Figure 3.12 (a) and (c) show two examples of opposite slope sign conditions verified.

More formally, given two regions X and Y across an edge and intensity samples I taken at two wavelengths λ_1 and λ_2, the spectral crosspoint condition can be formulated as

$$(I_{X\lambda_1} - I_{Y\lambda_1})(I_{X\lambda_1} - I_{Y\lambda_2}) < 0 \tag{3.11}$$

and the opposite slope sign condition can be written as

$$(I_{X\lambda_1} - I_{X\lambda_2})(I_{Y\lambda_1} - I_{Y\lambda_2}) < 0. \tag{3.12}$$

According to Rubin and Richards, only the edge in Figure 3.12 (d) does not represents a material change, since the curves for the chosen points do not cross and have the same slope sign.

Figure 3.13 shows an example of opposite slope sign condition in real images which allows to distinguish a material change from a change due to a shadow. An image sequence is considered where a region of interest is selected and the behavior of the RGB components of its central point is analyzed over time. RGB components represent three different wavelengths. First, the RGB values of the point change because an object passes in front of the scene's background, then they change due to the presence of a shadow. The order of RGB color components for the selected point changes when the point passes from the background to the object, while it remains unchanged when the point passes from the illuminated background to the shadowed background.

Following and extending the work of [139], in [47] a method for distinguishing shadow boundaries from material changes in presence of ambient illumination that differs in its spectral characteristics from the incident illumination is presented. Two sets of biologically motivated operators, monochromatic opponent units and double-opponent units, are proposed to extract information about the total change in each chromatic component and about changes in their relative amounts. Given a certain knowledge of the strength of the ambient illumination, by comparing and thresholding the operators' responses, shadow boundaries are distinguished from material changes. While the gray world assumption is relaxed, an a priori estimate for the strength of the ambient illumination has to be provided to the method.

Color band ratios across region boundaries are used in [7]. An initial segmentation of the image is performed. Then, a number of tests to pairs of neighboring segments is applied for checking:

- a strict decrease in each of the three RGB color channels (Eq. (3.8));

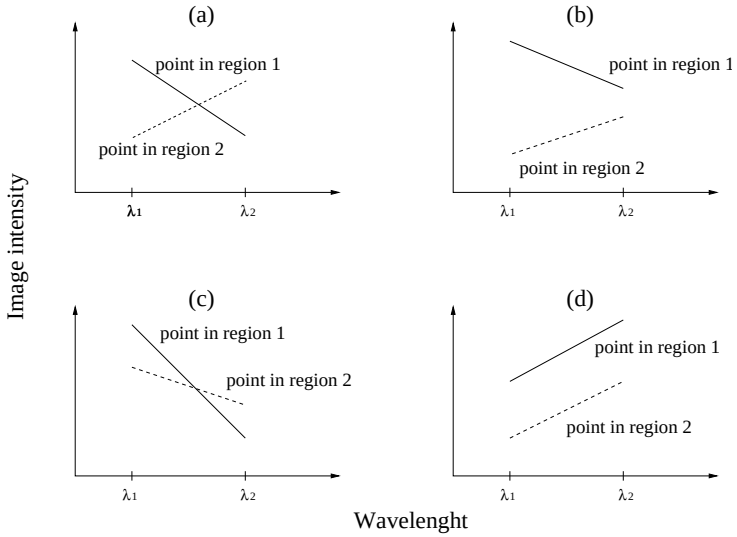

Figure 3.12: Distinguishing material changes from shadow boundaries, highlights, and surface orientation changes according to the *spectral crosspoint condition* [138] and the *opposite slope sign condition* [139] under the gray world assumption. Only the edge in (d) is not classified as material change, since any of the conditions holds.

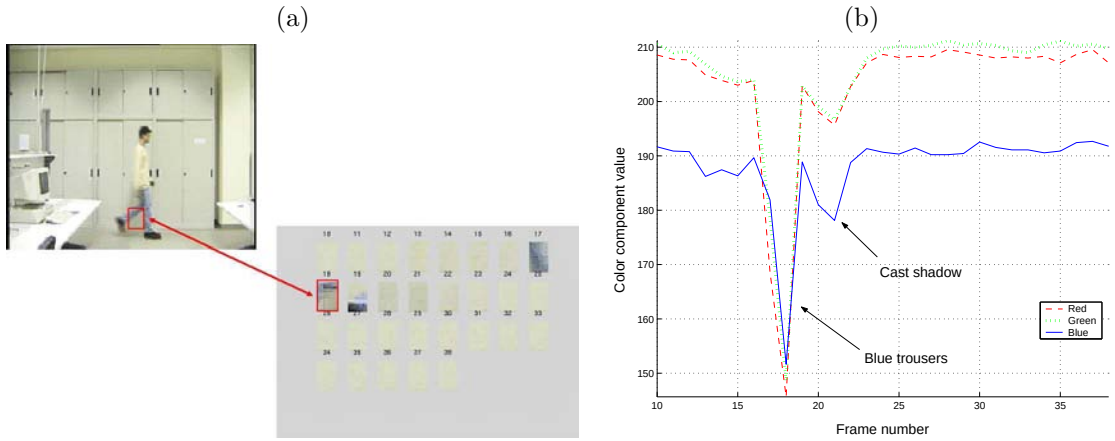

Figure 3.13: Effect of a material change and a cast shadow on the RGB color components as predicted by Rubin and Richard's opposite slope sign condition [139]. (a) Sample frame from the sequence *Laboratory* and highlighted region of interest over 28 frames of the sequence. In the selected frames, first the person's trousers and then a shadow cast upon the background cover the region. (b) The mean RGB values over a 3x3 pixels window centered in the central point of the region of interest plotted over the selected 28 frames. A change in the order of color components is visible when the window passes from the background to the object denoting a spectral crosspoint. No crosspoint appears in shadow points.

- the occurrence of multiple similar band ratios $(R_X/R_Y, G_X/G_Y, B_X/B_Y)$ in adjacent segments X, Y with different RGBs;
- a band ratio exceeding a fixed threshold.

The last exploited criterion is based on the empirical observation that material changes rarely have a ratio larger than 30 to 1, whereas the differences between bright sunlit regions and nearby shadows can exceed this value. The validity of this observation is however limited to sunny, high contrasted outdoor scenes. The authors consider a diagonal model of illumination change, as that described in Eq. (3.10), for modeling the change in sensors responses across a shadow edge. To build the diagonal model for the employed camera, a number of measurements considering a set of indoor and outdoor illuminants is required. The model is then used before any processing to first of all extract only those edges that can be due to an illumination change.

The use of photometric invariant transformations has been investigated in [52] for edge classification in video sequences. Photometric invariant transformations will be discussed in more detail in Chapter 4. An automatic color edge detection approach is first employed to extract edges in each video frame from RGB images and invariant images. Then, a rule-based edge classifier is proposed for labeling edges into shadow-geometry, highlight, or material edges according to the edge detection results on different features. Since temporal information is not exploited in the proposed method, but rather an intra-frame analysis is performed, we have included the method in this section. The following invariant features are used:

$$c_1(R,G,B) = \arctan\left(\frac{R}{B}\right) \qquad c_2(R,G,B) = \arctan\left(\frac{G}{B}\right), \tag{3.13}$$

$$o_1(R,G,B) = \frac{(R-G)}{2} \qquad o_2(R,G,B) = \frac{B}{2} - \frac{(R+G)}{4}. \tag{3.14}$$

The c_1c_2 features are shown to be invariant to shadows and shading for matte, Lambertian surfaces (this issue will be detailed in Chapter 4). o_1o_2 are taken from an opponent color space (Section 2.5.1) and are shown to be invariant to highlights for specular surfaces under the same assumptions. The rule-based classifier works as follows:

IF $E_{RGB} \neq 0$ AND $E_{c_1c_2} = 0$ THEN **shadow or geometry edge**
ELSE
IF $E_{c_1c_2} \neq 0$ AND $E_{o_1o_2} = 0$ THEN **highlight edge**
ELSE
material edge

where $E_{RGB}, E_{c_1c_2}$, and $E_{o_1o_2}$ denote the values of the RGB, c_1c_2, and o_1o_2 edge maps, respectively.

In [130, 136] the segmentation of shadows in color images of outdoor scenes is investigated.

In [136], two different and complementary techniques are presented. They are proposed as a tool for the analysis of illumination conditions for image indexing applications. The first technique is based on a Lambertian model of reflection and assumes that the gray world condition holds. It is applied in the case of overcast images. The second technique is based on the Dichromatic Reflection Model and is applied to sunny, high contrasted scenes. In this case, the gray world assumption is relaxed.

In the first technique, a color image segmentation aiming at extracting regions of the image with uniform luminance and chrominance information in the CIELAB space is performed by means of mathematical morphology tools. Then, adjacent regions characterized by different luminance and similar chrominance information are merged. This step aims at merging shadow regions having similar chrominance but lower luminance with adjacent illuminated regions on the same surface. The merging process is controlled by [139]'s opposite slope sign condition. Regions are not merged if the condition is verified. Shadows are then extracted as the darkest regions in regions of similar chrominance. The presented results show that in case of sunny scenes not all the shadows are detected.

A second technique is then proposed, which is based on the empirical observation that shadows increase saturation values. However, saturation is not directly employed, but a perceptual quantity related to it and dependent on the sensitivity of human vision, the so-called *chromatic luminance*, V_c. It is computed using the same expression for luminance applied to normalized color rgb, that is

$$V_c = 0.176r + 0.81g + 0.01b. \tag{3.15}$$

Pixels having a chromatic luminance that is larger then their luminance Y in the XYZ space, that is for which $V_c > Y$, are classified as shadow pixels. The results show that the used property is not verified for yellowish hues and that misclassifications arise in achromatic regions. If the sky is overcast, the entire image is detected as shadow.

Based on the Rayleigh scattering effect which describes the atmospheric dispersion of sunlight, the method in [130] segments shadows by means of color features in aerial images. Due to the Rayleigh effect, the atmosphere scatters much more violet/blue wavelengths of sunlight than red wavelengths. Therefore, shadows are detected in blue/violet regions of the image which have higher saturation values S than intensity values I. The HSI color space is used and the following thresholding operation is performed on each image pixel (x, y) to extract shadow pixels:

$$I(x, y) - S(x, y) \leq K. \tag{3.16}$$

The Rayleigh effect is proportional to the distance between the scene and the observer. Different threshold values $K \leq 0$ are therefore used in the detection for images taken by airborne sensors or by orbital sensors.

In [31, 95], methods to process color images for removing shadows are presented.

A camera calibration procedure [35] is used in [31] to generate a gray-scale illumination invariant shadow-free image that is used, together with the original image, to locate shadow edges. The process is based on the idea that material edges should occur in both original RGB image and invariant image, whether shadow edges should not be present in the invariant image. Thresholding out image edges that are due to shadows and re-integrating the material-edge-only map allows to obtain full color shadow-free images. For the calibration of the camera in case the sensors responsitivities are unknown, a set of images of a colored target taken at different times of the day in case of outdoors scenes, or under several different illuminants in case of indoor scenes, is required. This procedure allows to "learn" the possible variations in illumination conditions and to remove them consequently when generating the illumination invariant image. The diagonal model of illumination change of Eq. (3.10) is assumed and the assumptions of Planckian lighting (Section 2.2.2), camera narrow-band sensors and Lambertian surfaces are made. In [36], an alternative shadow removal process based on a modification to the retinex algorithm [88] is applied once shadow edges have been located, leading to similar results.

A similar illumination invariant computation has been proposed in [95] for scenes illuminated by daylight modeled by the CIE daylight standard. Knowledge of the camera sensors is needed in this case. The methods are not applicable to images from uncalibrated sources.

Geometric properties — Geometric information is generally used in a verification stage, once candidate shadow regions have been identified by means of spectral properties.

In [79], gray-level intensity and geometric constraints are used to identify and classify cast and self shadows in images of a constrained, simple environment. Initially, regions which are darker than the surrounding background are extracted. This first step is based on the assumption that most image border pixels belong to the background surface which is flat or nearly flat. Then features such as *vertexes* on the outlines of dark regions, *penumbrae*, self shadows and cast shadows as subregions of dark regions, and object regions adjacent to dark regions are searched for. Occlusion between shadows and objects is assumed to be minimal and only one area light source illuminates the scene. Finally, the consistency among *light source directions* estimated from the extracted regions is tested to confirm shadows among dark regions. To estimate the light source direction different methods are proposed. If a penumbra is present, the direction of the maximum width from an umbra to its penumbra is used. If a self shadow and a cast shadow regions are identified, then the middle point of the cast shadow boundary and the middle point of the self shadow boundary are connected to estimate the direction of light rays. If an object region is found adjacent to the dark region, then the light source direction is estimated as the direction from the middle point of the object boundary to the middle point of the cast shadow boundary. The accuracy of the estimations is generally poor due to perspective distortions and errors in the feature extraction process. This method and that proposed in [140] are the only methods in the literature dedicated to the explicit extraction of self shadows.

In [43], color information is combined with geometric information to detect cast shadows. In order to exploit shadows spectral properties, scenes are restricted to be composed by a singly colored extended light source and piece-wise constant surface reflectances. Specularities and inter-reflections are discounted. Color information is exploited in the first step of the detection process. The authors observe that Eq. (3.4), when highlights are discounted and penumbra is considered as a multiplicative parameter $\theta < 1$, that is

$$C_i(x,y)_{lit} = \int_\Lambda \Big(\theta(x,y)L_b(\lambda,i,e,g) + L_a(\lambda)\Big)S_{C_i}(\lambda)d\lambda, \tag{3.17}$$

is the parametric form of a line with parameter θ. The end point of the line at $\theta = 0$ corresponds to the shadow umbra, while the end-point close to $\theta = 1$ corresponds to the illuminated surface. Linear clusters in color space are therefore detected to segment uniformly colored candidate regions containing a shadow. The process also detects regions with self shadows and illuminated surfaces affected by shading due to changes in surface orientation. They map in fact as well to linear clusters in the considered model. Within each candidate region, the system then analyzes the spatial layout of brightness variations in order to determine a cast shadow's *umbra and penumbra* structure. As before, shading provides a similar brightness structure and cannot be distinguished from cast shadows. The method strongly relies on the presence and detectability of cast shadows penumbra. This may limit its applicability.

In order to finally insure that a candidate region corresponds to a cast shadow, geometric information is used in the second step of the method. To this end, the 3D *location of the light source* is determined by allowing the observer to cast a known shadow with a probe that can be extended

Reference	*Spectral property*	*Geometric property*	*Additional information*
[141, 172]	Surface darkening (G)		
[7]	Surface darkening (C)		Camera calibration
[138, 139]	Color appearance		
[47]	Color appearance		Ambient light strength
[136]	Color appearance		Outdoor images
[130]	Color appearance		Aerial images
[31, 36, 95]	Color invariance		Camera calibration
[52, 136]	Color invariance		
[79]	Surface darkening (G)	Vertexes in boundaries, self and cast shadow, penumbra, object adjacency, light direction estimation	
[43]	Color appearance	Penumbra, shadow-object-source relationship	Active observer

Table 3.1: A summary of the used shadow properties, the used features, and specific information or processing required by state of the art methods for still image analysis. G: gray-level image intensity, C: color information.

in the environment. Once the light source has been located, the presence of an object between the candidate shadow and the projection of the light source in the image plane is checked. The required active process limits the method's applicability.

Table 3.1 provides a summary of the exploited shadow properties, the used features, and specific information or processing required by state of the art methods. Automatically detecting shadows in still natural images from uncalibrated sources remains a very difficult problem. A wider range of real world scenes can be tackled when dealing with dynamic shadows in image sequences. In this case, the motion of shadows cast by moving objects provides an additional cue that can simplify the detection process.

3.4.4 Image sequences

Identifying moving objects from a video sequence is a fundamental and critical task in many computer vision applications. Two approaches to moving object detection can be identified with respect to the use of motion information, namely approaches based on *motion segmentation* and approaches based on *motion detection*. Motion segmentation approaches classify clusters of pixels in the image that have similar motion. Motion detection approaches identify those pixels where motion exists. For both approaches, motion is measured by analyzing the temporal changes of image intensity in the sequence [107]. Shadows cast by moving objects generate temporal changes and can mislead both motion segmentation and motion detection approaches. The detection of a moving object may then include the detection of its shadow or part of its shadow. In this case, moving shadows should be identified and removed to obtain an accurate object contour.

When the camera is fixed or its global motion has been estimated and compensated [4], the most widely adopted approaches for moving object extraction in absence of any a priori knowledge about objects of interest and environment are based on *background subtraction* and *change detection*.

Moving objects representing the scene's foreground are detected from the portion of a video frame that differs significantly from a static background model or reference image. How to deal with shadows is one of the distinguishing and challenging features of such approaches. It is in this context that several methods for the explicit detection of moving shadows have been proposed in the literature. They are reviewed here.

Gray-level image intensity — First attempts to shadow detection in image sequences limit themselves at exploiting the property that shadows darken the surface on which they are cast [41, 137] and that the ratio among intensity values between points in the shadowed region and the same illuminated region in the background is constant [137] (constant α in Eq. (3.9)). This can be considered approximately true only in the shadow's umbra.

The method proposed in [16] is dedicated to moving cast shadow detection in monochromatic video sequences and is consequently limited to the use of gray-level information. The difference in intensity between the current and the reference background frame at a pixel (x, y) due to a shadow is modeled with a linear function $sf(x, y)$ of the intensity values in the background image $g_b(x, y)$, that is

$$sf(x, y) = ag_b(x, y) + b, \tag{3.18}$$

with $-1 \leq a \leq 0$. This is a generalized expression for Eq. (3.9), which is obtained when $b = 0$. The parameters of the function are estimated from the sequence by means of a regression analysis. The estimation is done interactively at the first frame and then automatically at each frame from the shadows detected in the previous image. A planar background hypothesis is considered in order to use constant parameter values which are not influenced by changes in surface orientation. If

$$|g_c(x, y) - g_b(x, y) - sf(x, y)| < T, \tag{3.19}$$

where $g_c(x, y)$ is the value of the pixel in the current frame, the point is classified as shadow. The method's performance is limited in weak shadows and penumbra regions.

A more complete and accurate method is described in [151]. The approach aims at detecting and classifying background regions which have been covered or uncovered by a moving cast shadow from one frame of the sequence to the following. Lambertian reflection and a constant ambient illumination term which has the same color as the occluded light is considered. The gray world condition is thus considered to hold. The following assumptions on the scene are moreover made:

- a single, distant light source with non negligible size and intensity illuminates the scene;
- the background is planar and textured.

A change detection mask indicating image points that have a large frame difference between the previous frame in the sequence at time k, $I(x, y, k)$, and the current frame at time $k+1$, $I(x, y, k+1)$, is assumed to be available. In the proposed implementation, the method described in [103] is considered. Moving shadows are searched for inside the change detection mask. To this end, three criteria are combined by means of heuristic rules:

- a distinction between moving textured objects and moving cast shadows on static textured background is made by extracting and classifying luminance edges in the current and previous frame as static or moving edges according to the texture content in the frame difference (local energy in high frequencies);

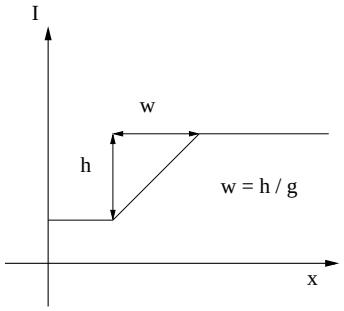

Figure 3.14: Model of an image luminance step in the direction perpendicular to a shadow contour [151]. The luminance step is defined by the step height h and the norm g of the gradient. From h and g the width w can be calculated. The shadow contour is assumed to be in the y direction.

- because of the assumed planar background and distant light source, the surface normal in shadows can be assumed constant. Therefore, shadows are looked for in regions where the luminance frame ratio $FR(x, y) = I(x, y, k+1)/I(x, y, k)$ is locally spatially constant (constant α in Eq. (3.9));
- since shadows of an extended light source have a penumbra, the width of the luminance step caused by a shadow's penumbra should be larger than that caused by object edges.

The most critical aspect of the method is the assumption of a penumbra whose width can be computed. To this end, a linear model for luminance values perpendicular to a shadow contour is assumed (see Figure 3.14). In outdoor scenes, where shadows present sharp edges due to the source of illumination that is far from the objects, this could represent a problem.

A different problem is tackled in [101, 102]. A method for removing from the static background of an image sequence shadows cast by *static objects*, such as tall buildings and trees, in order to improve the robustness of video surveillance systems is proposed. The method is composed of two parts. The first part is an off-line estimation of intrinsic images, that is time-varying reflectance and illumination images, from a sequence of images representing the scene's background with shadows at different time instants. The estimation is based on the method proposed in [170]. Using the estimated illumination images, the so-called *illumination eigenspace* is constructed, that is a database which captures the variation of lighting conditions in the illumination images. To construct the illumination eigenspace, image sequences have been stored for 120 days for 1 year. The database is then used in the second part of the method. Using the pre-constructed illumination eigenspace, an illumination image is directly estimated for each input image of the sequence to which moving objects have been removed. The input image is finally normalized in terms of illumination thanks to the computed illumination image and a shadow-free image is obtained.

Moving cast shadows are not considered in the proposed approach. The method fails when the illumination normalization is performed in presence of moving objects which cross the static shadow so that the shadow edge cast on the object largely differs from the shadow edge in the illumination image. The idea behind this method is similar to that proposed in [31]. There, thanks to the use of color information and of a camera calibration procedure, the illumination normalization was performed on a single image without the need of a sequence of images of the same scene.

Color — Outdoor environments illuminated by a far away point source (the Sun) and a diffuse source (the sky) are targeted in [109], where the Dichromatic Reflection Model with an ambient

illumination term is considered. Shadow regions are assumed to be illuminated by the diffuse sky only, inter-object reflections are neglected, and surfaces are assumed to be Lambertian. Shadows are detected inside a previously extracted moving foreground mask [108]. Initially, the method extracts as candidate shadow points

- pixels having lower values of each color component with respect to the background image (Eq. (3.8));
- by assuming that the sky is blue, pixels whose decrease in the blue channel is smaller than the decrease in the red and green channels, that is $(R_c(x,y)/R_b(x,y), G_c(x,y)/G_b(x,y)) < B_c(x,y)/B_b(x,y)$, where c, b refer to the current and the background image;
- pixels which are connected by means of a component labeling procedure that uses a spatio-temporal reflectance ratio as the connectivity criterion.

The spatio-temporal reflectance ratio P, which provides an illumination normalization, is computed by considering two neighboring pixels having intensity A_f, B_f in the current image and A_b, B_b in the reference frame as

$$Rt_A = \frac{A_f - A_b}{A_f + A_b} \qquad Rt_B = \frac{B_f - B_b}{B_f + B_b} \tag{3.20}$$

$$P(A,B) = \frac{Rt_A - Rt_B}{Rt_A + Rt_B}. \tag{3.21}$$

By means of user interaction, a training phase allows to compute an estimation of the background's body color. A singular value decomposition (SVD) is used to this end. The same SVD approach is used to extract the body color component for candidate shadow regions. The difference between the two estimated color vectors allows then to classify as shadows those regions having a small vector difference.

The approach, which is semi-automatic, is computationally expensive. As the authors state, the results are sensitive to the shadow size (as the shadows become larger the body color estimation provides more robust results), to the camera sensor characteristics, and to the background's color. For cloudy scenes and highly saturated background the test on the blue component should be by-passed.

A diagonal model of illumination change, as that described in Eq. (3.10), is used in [105] to model the appearance of a pixel when shadowed given its appearance when illuminated. Since no assumptions on the color of the light illuminating the shadow and the light illuminating the same region without occlusion is made, manual segmentation of a certain number of frames of the image sequence has to be performed in order to extract shadows and corresponding illuminated background regions and to determine empirically the model coefficients. The model coefficients are assumed to be approximately constant over flat surfaces. If the background is not flat over the entire image, different coefficients have to be computed for each flat subregion. The method's parameters require therefore a time consuming setting. They are moreover camera and scene-dependent. A probabilistic approach is then used to classify each pixel in the image into one of the three classes: background, shadow, or foreground. Gaussian distributions are assumed for illuminated and shadowed states of a pixel and a uniform distribution is assumed for objects. The method works in real-time on outdoor traffic scenes.

A probabilistic method is also proposed in [169] targeting indoor image sequences. A linear transformation is used to describe the change of intensity for a point when shadowed given its intensity in the background, as in Eq. (3.18). When color images are considered, the model is reduced

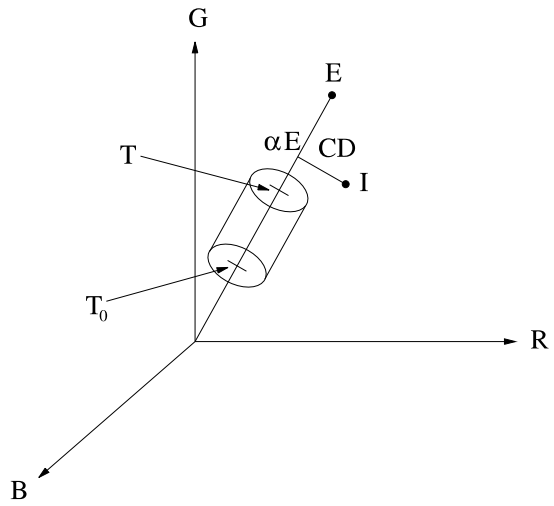

Figure 3.15: Color model proposed in [70] in the three-dimensional RGB color space. The difference between background's and current image's pixel color is decomposed into brightness (α) and chromaticity (CD) components. Pixels inside the depicted cylinder are extracted as shadow pixels.

to Eq. (3.9).

The methods proposed in [27, 28, 57, 70, 90, 127, 142] exploit the properties of invariance of some color models in presence of shadows. The used color features will be considered again in more detail in Chapter 4.

A background subtraction algorithm based on a computational color model which separates the brightness from the chromaticity component of a pixel is presented in [70]. The reference background image is statistically pixel-wise modeled by means of its mean

$$\mathbf{E}(x,y) = [\mu_R(x,y), \mu_G(x,y), \mu_B(x,y)]$$

and variance

$$\sigma(x,y) = [\sigma_R(x,y), \sigma_G(x,y), \sigma_B(x,y)]$$

in each color channel using the first N frames of the sequence. The difference between each pixel's RGB color vector $\mathbf{I}(x,y)$ in the current frame and the pixel's RGB color vector $\mathbf{E}(x,y)$ in the background model is decomposed into two components, a brightness distortion $\alpha(x,y)$, and a chromaticity distortion $CD(x,y)$, as shown in Figure 3.15.

Pixels that have similar chromaticity (small $CD(x,y)$) but lower brightness ($\alpha < T$) than those of the same pixel in the background image are classified as moving shadow pixels. A lower threshold T_0 is employed to avoid misclassification as shadows of points with low RGB values. Shadow pixels lie inside the cylinder in Figure 3.15. The method implicitly assumes Lambertian reflection and the gray world assumption to hold. Under these hypotheses, in fact, Eq. (3.9) describes the relationship between $\mathbf{I}$ and $\mathbf{E}$ and the chromaticity distortion is zero.

An approach that is based on the same assumptions and on the same principle, that is shadows lower the luminance of a pixel but do not change significantly its chrominance, is presented in [127]. Here, the $Y'C_bC_r$ space is considered. We will see in Chapter 4 how the invariance of C_b and C_r features in shadows is only approximate. Shadows are extracted by analyzing and thresholding the

luminance and chrominance ratios between current and reference frame. This is done first locally, on a pixel-by-pixel level, then globally, on a region level. The ratios in shadows are assumed to be constant (constant α in Eq. (3.9)) since the additional assumption of planar background is considered. The penumbra is discounted from the analysis and treated separately by means of morphological filtering operations. In a final temporal filtering stage, shadows detected at time t which have no intersection with shadows detected at time $t-1$ are discarded, unless they lie close to the image boundary.

Brightness, hue and saturation in the HSV color space are exploited in [27, 28]. The detection is based on the observation that shadows darken the brightness of an area without significantly modifying its hue and saturation values. This observation relies on the implicit assumption that ambient light and direct light have the same color and are white. This will become clearer in Chapter 4. Three thresholding operations are performed on previously extracted moving foreground pixels, one for each color component. A foreground pixel (x, y) is thus labeled as shadow pixel if

$$\alpha \leq \frac{V_f(x,y)}{V_b(x,y)} \leq \beta \wedge S_f(x,y) - S_b(x,y) \leq \tau_s \wedge D_H(x,y) \leq \tau_h, \tag{3.22}$$

where the subscripts f and b indicate the pixel value in the foreground and background image, $\alpha \in [0,1], \beta \in [0,1]$ and

$$D_H(x,y) = min(|H_f(x,y) - H_b(x,y)|, 360 - |H_f(x,y) - H_b(x,y)|). \tag{3.23}$$

The ratio of V values is thresholded by means of an upper and lower bound. The lower bound avoids misclassifications as shadows of pixels that have low brightness value. Thresholds are empirically determined.

The HSV space is exploited also in [90]. The authors observe that hue and saturation are not always reliable features, since they can fluctuate violently due to noise in certain regions of the image. They propose therefore to use them only in those parts of the image where their variance is below an empirically fixed threshold.

A method for real-time cast shadow detection for videoconference applications is proposed in [142]. As for [27] and [90], the method is based on the use of brightness, hue and saturation. However, saturation is used differently. It is assumed that it does not remain unchanged in shadows but it decreases. The algorithm uses approximate expressions for hue and saturation in the YUV color space in order to avoid time consuming color transformations.

A real-time system is also proposed in [57], where shadows are detected based on the following properties:

- a shadow pixel is darker than the corresponding pixel in the background reference image;
- the texture in the shadow region is correlated with texture in the background image.

Texture is analyzed by computing the normalized cross-correlation (NCC) over 7×7 pixels windows for the luminance image. To improve the method's performance, which fails to distinguish shadow and object points in regions of the object which are not textured, color is introduced. The HSL space [129] is considered which belongs to the same family of user-oriented spaces as the HSV and HSI spaces. A similarity measure between color vectors is introduced as the scalar product of the projection of RGB vectors on the chromatic plane of the HSL color space. A color normalized cross-correlation (CNCC) using the proposed similarity measure is then derived and thresholded to detect shadow pixels. As the CNCC and NCC measure the textural similarity of two regions, textured objects with multiple colors are more accurately extracted by the proposed method. Moreover, a

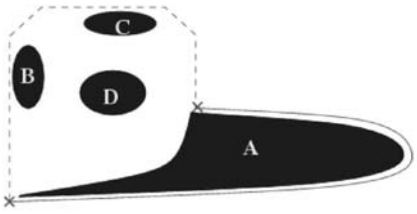

Figure 3.16: Geometric analysis of candidate shadow regions (here in blue) inside a foreground blob (here delimited by a red dashed and solid line) as proposed in [12]. Region D is discarded because it lies far from the blob's border, regions B and C are discarded because the majority of their boundary is far from the blob's boundary. Region A is confirmed as a shadow since the majority of its boundary is closed to the blob's contour.

colored textured background is better suited than an uniform background for the CNCC.

Geometric properties — As in the case of still images, a geometric analysis of shadows is typically employed in a verification stage.

The method proposed in [12] is dedicated to monochromatic image sequences of outdoor traffic scenes. It extracts in a first stage candidate moving shadows from moving foreground regions by considering gradient information in the ratio image between current frame and reference background frame. Candidate shadows are identified as uniform regions with low gradient values. In a second stage, the spatial relationship between candidate shadows and moving objects is exploited. The position of the detected shadows within the foreground regions is analyzed to this end. Candidate regions that are far from the foreground region's boundary and are small are discarded. True shadows are retained as large regions near the boundary (see Figure 3.16). An ad-hoc thresholding is used since the size of the cast shadows changes during the day.

Luminance, chrominance, and gradient density information are used in [42] as a first stage for moving cast shadow detection is outdoor traffic scenes. A combined shadow confidence score is derived for extracted foreground regions that allows to separate a cast shadow from the corresponding object. Three properties of shadows are checked:

- the luminance of the cast shadow is lower than that of the background (from Eq. (3.8));
- the chrominance of the cast shadow is similar to that of the background;
- the difference in gradient density between the cast shadow and the background is lower than the difference in gradient density between the object and the background.

As in [127], luma and chroma in the video-oriented $Y'C_bC_r$ space are used. Once candidate shadow regions having a high confidence score have been obtained, geometrical evidence is checked. Since, when object and cast shadow are attached, the cast shadow is at the boundary of the foreground region, the convex hull of the foreground edge points which are not candidate shadow points is computed. Shadow points inside the convex hull are then discounted. The use of the convex hull of candidate object regions could limit the method's performance in presence of non-convex objects, such as people. Moreover, the gradient density criterion may fail when objects are not sufficiently

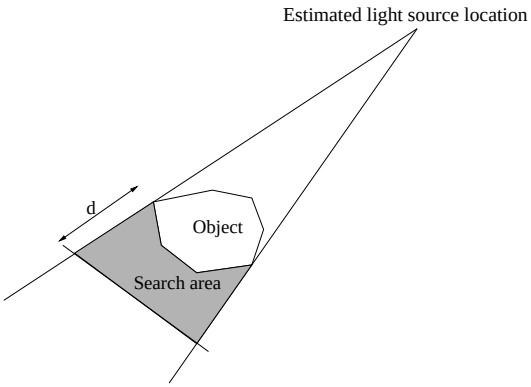

Figure 3.17: In [126], shadow detection in each frame of the image sequence is performed in a search area defined thanks to the estimated light source location in the previous frame and the known object contour.

textured.

Outdoor traffic images are considered also in [168]. An illumination assessment method is first of all used to decide whether cast shadows are present in the image based on the analysis of the brightness energy of foreground moving areas. A large brightness energy indicates the possibility that shadows exist. In this case, shadow detection is applied to improve moving foreground detection. To this end, an estimate of the illumination direction is computed by analyzing where on the foreground region's boundary the majority of dark pixels can be found. Eight possible directions are considered, equally spaced in the image plane. Having determined the direction of illumination, sample hue and saturation values of shadow points closed to the foreground region boundary portion indicated by direction of illumination are extracted. Foreground pixels having RGB values larger than the corresponding background pixels are retained as object pixels. Foreground pixels having hue and saturation values different from the average values of the samples extracted close to the shadow boundary are retained as object pixels. The Canny edge detector is then applied to both foreground and background images. The background edges are subtracted from the foreground edges to extract object edges. Finally, foreground pixels nearby object edges are retained as object pixels. An hole filling procedure produces the resulting foreground object without shadows.

In [125] and [126], the authors propose a method for the estimation of the projection of the light source direction in the 2D image plane using 2D geometric constraints among an object, its cast shadow and the light source location. To this end, the convex hull of the object is computed and directions between each couple of tangent directions to the hull's segments are computed and analyzed. Lines are searched for that touch just one point on the object and one point on the contour of the cast shadow. These lines are an approximation of the lines tangent to the shadow contour at the vertexes on the shadow-making line (see Section 3.3.2, Figure 3.8). By intersecting the regions of the image plane delimited by the lines in subsequent frames of the sequence, an estimation of the position of the light source, if the source is closed to the scene, or of the direction of parallel light rays if it is far from the scene is obtained. The estimated light direction is then used to perform a coupled light direction estimation and shadow detection. After an initialization phase in the first frame, the estimated light source direction is used for the computation of a search area for the segmentation of cast shadows in the subsequent frames. For the initialization, in [125] the shadow

Reference	*Spectral property*	*Geometric property*	*Additional information*
[41]	Surface darkening (G)		
[137]	Surface darkening (G), uniform darkening		
[16, 169]	Linear model of frame difference (G)		
[151]	Static edges on texture (G), uniform darkening	Penumbra width	
[57]	Surface darkening, texture (C)		
[27, 70, 90, 127, 142]	Color invariance		
[105]	Color appearance		Manual segmentation
[109]	Color appearance		User interaction
[42]	Color invariance, uniform region	Shadow-object relationship	
[12]	Uniform darkening (G)	Shadow-object relationship	
[168]	Surface darkening, color invariance	Illumination direction	
[126]	Uniform darkening (G)	Shadow-object-source relationship	Segmented objects

Table 3.2: A summary of the used shadow properties, the used features, and specific information or processing required by state of the art methods for image sequences analysis. G: gray-level image intensity/luminance, C: color information.

is assumed to be known at the first frame of the sequence, while in [126] this constraint is relaxed by using a reference background image and by proceeding similarly as in [127] using only luminance information. For both methods, however, knowledge of moving objects masks without shadows over the entire sequence is required.

Additional geometric properties of shadows can be exploited when stereo cameras or multiple cameras are used. We briefly introduce here approaches that have been proposed to remove shadows by means of 3D geometric analysis.

The method in [116] makes use of the characteristic property of shadows of inheriting the shape of the surface on which they are cast. Shadows on planar surfaces are planar. Therefore, in order to detect shadows of pedestrians on the road plane, the method makes use two cameras and exploits height information. The image obtained from the first camera is inversely projected to the road plane and the projected image is transformed to the view from the second camera. Shadows on the road plane occupy the same areas on the transformed image and the image acquired from the second camera, whereas object such as pedestrians with different heights from the road plane occupy different areas in these images and can therefore be identified. Dense stereo range images are used in [62] where a real-time system for tracking people is proposed. Shadows cast on the background do not cause changes in disparity images and are therefore not detected when background subtraction is applied on them. Disparity information is also used in [78], where the computation of dense stereo

range images is avoided by means of an off-line construction of disparity fields using two or more cameras. Shadows are not detected as part of the background when more than two cameras are available.

A summary of the exploited shadow properties, the used features, and specific information or processing required by state of the art methods dedicated to monocular sequences is presented in Table 3.2.

3.5 Summary

The first objective of this chapter is to discuss a characterization of shadows in digital images and image sequences as the starting point for developing efficient techniques for their analysis. To this end, a complete list of spectral, geometric and temporal visual cues that suggest the presence of a shadow in a scene was introduced and was evaluated for the purpose of their use in the context of this thesis.

Among these cues, spectral properties related to the brightness and the color of regions covered by shadows offer the greatest generality with respect to the content of the considered scenes. In a framework such that targeted in this thesis, where only image-derived information and a limited number of assumptions about the scene are considered, they provide the major amount of information for shadow detection. They were therefore selected for being used in the proposed shadow segmentation approach and investigated. To this end, the Dichromatic Reflection Model was introduced for modeling the effects of shadows on image values. The characterization of ambient light with respect to direct light is central to shadow modeling. The introduction of the gray world condition was discussed with respect to this issue.

Color information is not the only information that can be exploited for analyzing shadows. Even without knowledge of the 3D geometry of a scene, geometric shadow cues which relate the position of a shadow to that of the shadow casting object can be defined. They were discussed for the purpose of considering them as additional constraints in the proposed shadow segmentation approach. In dynamic scene, the temporal behavior of shadows was also discussed as an important constraint for their characterization.

The second objective of the chapter is to review the state of the art of shadow detection. In this context, we found useful to classify and evaluate different approaches to the problem on the basis of the properties of shadows that they exploit. The existence of two categories of approaches was outlined: model-based approaches and property-based approaches. The specificity of model-based solutions, which exploit some knowledge of the scene made available by the targeted application, limits their use to the specific applications they are designed for. Property-based approaches offer more flexible solutions. They can be applied to a larger class of scenes and adapted to new applications. A property-based approach is proposed in this thesis.

The analysis of the state of the art outlined the fact that the problem of detecting, processing and analyzing shadows has been investigated within several research domains, ranging from aerial image understanding to digital photography and video analysis. An especially increased interest in the extraction of moving shadows in video sequences has been motivated in recent years by the need of accurate video object extraction tools. Current moving object detection systems in fact typically detect shadows cast by the moving object as part of it. Although different solutions have been proposed, there exists no generally accepted methods. Due to its relevance, particular attention is dedicated in this work to the segmentation of shadows in video.

The primary role of color information that was outlined when evaluating visual cues for shadow detection is evident from the analysis of property-based methods for shadow detection, both in the

case of still images and of image sequences. In this context, different solutions have been proposed in literature with regard to both the physical models of shadows adopted and to the color features exploited. In the next chapter, this issue is investigated and the color models that are used in the proposed shadow segmentation approach are discussed.

Figure 3.18: The shadow is a metaphor of the film creation (Section A.1.2).

Photometric invariants for shadow analysis

4

4.1 Introduction

This chapter is dedicated to a discussion of photometric invariant color features. Photometric invariant color features are functions which describe the color of an imaged surface discounting changes in the imaging conditions. In the context of this thesis, we aim at investigating the role they can play in the analysis of shadows in digital color images.

As discussed in Chapter 2, the color appearance of a point on a surface depends both on the characteristics of the surface and on the imaging conditions, that is on the illumination conditions and on the geometric arrangement of surface, light source, and camera. For many image analysis and computer vision applications, the possibility of devising color models which are less sensitive than raw sensor responses to changes in the imaging parameters is highly desirable. Let us consider an example application to introduce this issue. A computer vision application that benefits from the use of color is object recognition. Departing from traditional object recognition strategies based on geometric properties, Swain and Ballard [157] were the first to introduce a simple and effective recognition scheme that identified objects entirely on the basis of color. The scheme was based on matching the color histograms of query and target images. Due to the sensitivity of color values to changes in the imaging conditions, however, when query and target objects are recorded in different pose, illumination conditions, or from a different viewpoint, the recognition accuracy degrades significantly. To overcome this limitation, functions of color values that are able to discount the effects of changes in imaging factors have been devised. Indexing on such color invariant features has shown to deliver better recognition results.

Many other computer vision problems benefit from the use of color invariant features, such as material segmentation, image retrieval, and, as we aim at discussing in this chapter, shadow segmentation. Several shadow detection methods in literature exploit indeed shadows spectral properties on the basis of the observation that, while luminance decreases in shadows, chromatic information remains approximately unchanged. For the analysis, various color spaces are used to separate the two components of the color signal, such as HSV, CIELAB, and $Y'C_bC_r$. Which criterion should one adopt when choosing a color model for the analysis? What is the underlying model of shad-

ows? What is the relationship between the features used for shadow detection and the photometric invariants proposed for object recognition? In this chapter, we aim at providing answers to these questions. The existing answers are different and a complete picture is missing in literature. The different solutions have been therefore studied and a class of invariant features has been chosen for this work. They allow to formalize the second spectral property in the list presented in the previous chapter, which states that changes in the color of a surface in shadow tend to be predictable.

This chapter is organized in the following way. In Section 4.2, the problem of color invariance is introduced and its relationship with the analysis of shadows is discussed. The selection of those color invariant features that are adopted in this thesis is discussed. Their construction is first described in Section 4.3. Their invariance with respect to shadows is then discussed in Section 4.4. Their problems and limitations are also introduced.

4.2 From color constancy to shadow analysis

The problem of color invariance emerged from research in the domain of computational color constancy and its application to color-based object recognition. Computational color constancy is therefore first of all briefly introduced in this section and the reasons that inspired the color invariant approach are explained. Different approaches to color invariance are then discussed in the light of their possible use in the analysis of shadows.

Color is one of the properties we attribute to objects, but the light from the object that reaches our eye, and thus the photoreceptors' responses, vary with illumination. Therefore, if color must describe a property of an object, the nervous system must interpret the mosaic of cone responses and estimate something about the surface reflectance function. The neural computation of color is structured so that objects retain approximately their color appearance whether we encounter them in shade or sun. Achieving the same result is extremely difficult for a computer vision system.

The property of the human visual system of approximately observing the same color under different lighting is referred to as *color constancy* [30]. In an effort to mimic this ability of the human visual system, computational color constancy algorithms aim at mapping the red, green, and blue sensor responses, RGBs, for a surface under an unknown illuminant to corresponding RGBs under a known reference light. The interested reader is referred to [8, 9] for a detailed review of computational color constancy techniques. If solved, the problem can find important applications in computer vision problems. An already cited example is object recognition. If the recognition is based on the matching of color distributions from a target object with those of a query object, a description of these features that remains the same when query and target image are recorded under two different illuminants should be adopted. A color constancy pre-processing of the images could be used to this end.

The processes through which color constancy is attained in the human visual system are unfortunately not well understood. Indeed, despite significant effort, the performance of color constancy algorithms in computer vision remains quite limited [44]. The failure of color constancy pre-processing in object recognition, due also to the complexity of color constancy algorithms which result more complex than object recognition itself, has then inspired the *color-invariant approach* [3, 32, 37, 46, 50, 110, 113, 114, 148]. While color constancy aims at computing a full three-dimensional RGB image under the reference illuminant, the goal of color invariance is less ambitious. It attempts in fact to find functions of RGB values that cancel out dependency on the imaging conditions. Two types of dependencies are considered, dependency on *geometry* and dependency on *illumination*. In this thesis, we are interested in those approaches to color invariance which can provide features that are not affected by changes in illumination due to shadows. The

specificity of shadows as illumination phenomena has therefore to be considered.

To discount the effects of illumination in an image, one possibility is to normalize each image location by a reference RGB [87] to obtain a description of the image which is independent from lighting. It is also possible to derive global statistical features for the image that do not depend on the light [33, 66, 155]. These approaches apply to images affected by a global change in illumination. While they have been used in object recognition and image retrieval for matching two images taken under different illuminants, they are not suitable for shadow analysis. Shadows are in fact a local phenomenon and the illumination compensation should apply to local illumination changes between shadowed and lit regions in the same image. This adds difficulty to the problem.

A second possibility is then to use only information at each pixel position to derive features that do not change with a change in illumination. As they are locally defined, these features can also provide invariance to geometric parameters and, most of all, are suitable for analyzing shadows. With regard to this approach, two solutions can be adopted which are based on two different models of shadow.

The first model is the model that we have discussed in Chapter 3. It considers the gray world condition to define the relationship between ambient light, illuminating the shadow, and direct light, illuminating the same point when there is no obstruction. In this case, a number of color invariants can be found that are insensitive to shadows. Among them, some features come from traditional color models such as normalized rgb, hue and saturation. Others, such as $c_1c_2c_3$ and $l_1l_2l_3$ [50], have been proposed for color-based object recognition.

The second model describes a change in illumination due to a shadow by means of the diagonal scaling model (Eq. (3.10)). The gray world assumption is relaxed and the assumptions of Lambertian surfaces and narrow band camera filters are considered. For this model, Finlayson and Hordley in [35] and Marchant and Onyango in [95] propose an invariant computation that can discount shadows. The computation is based on the assumption that the light that illuminates a shadow and the light that illuminates the same point when there is no obstruction can be described by Planckian illuminants that differ in their color temperature (Section 2.2.2). The invariant computation provides then a one-dimensional image that is invariant to light intensity and light color. By construction, the invariant coordinate remains unchanged also under the reference illuminant of the color constancy approach. With their work, the authors aim thus at bridging the gap between classical color constancy and the color-invariant approach.

Its one-dimensional nature makes this methodology applicable to the problem of shadow analysis by assuming that illumination in shadows can be modeled by a Planckian illuminant, as proposed by the authors in [31]. The invariant image computation requires knowledge of the camera sensors responsitivities or a camera calibration process by means of a colored target imaged under different illumination conditions. The proposed invariant computation cannot then be used with images from uncalibrated sources. This limits its applicability.

In the framework of this thesis we aim at segmenting shadows in a wide range of scenes, also in situations where the imaging conditions are not under control. Our attention has been therefore focused on the first category of color features. These simple transformations of RGB values are shown to provide reliable results for color image segmentation [48], color object tracking [49] and color-based object recognition [50]. In the following sections, their characteristics are discussed. In case the targeted framework allows it, i.e. when control on the camera is considered, the flexibility of the approach to shadow segmentation proposed in Chapter 5 allows the use of different invariant derivations, included that proposed in [35] and in [95].

4.3 Photometric color invariants

The starting point for the design of color invariant transformations is a model of image formation which allows to describe image pixel values as a function of imaging parameters. In Chapter 2 we have introduced such model and we have then used it in Section 3.3.1 when characterizing shadows from a spectral point of view. We reconsider it in the following sections. First of all, in the next subsection, we describe pixel values distributions in the RGB color space on the basis of such model. This allows to more intuitively explain the construction of color invariants in Section 4.3.2.

4.3.1 The Dichromatic Reflection model in color space

The Dichromatic Reflection Model's formulation for a specific point on a surface is given by

$$L(\lambda, i, e, g) = m_s(i, e, g)c_s(\lambda) + m_b(i, e, g)c_b(\lambda), \tag{4.1}$$

where $L(\lambda, i, e, g)$ is the reflected radiance at wavelength λ, m_b and m_s are the geometric scale factors of the body and interface reflection terms, and c_b and c_s are the spectral power distributions of the body and interface reflection terms. At this surface point, the geometry, that is angles i, e, and g (refer to Figure 2.9 for a definition of these angles), are determined and the magnitudes m_s and m_b of the interface and body reflection terms may be considered as scalars. Therefore, the Dichromatic Reflection Model can be rewritten at the specific point as

$$L(\lambda) = m_s c_s(\lambda) + m_b c_b(\lambda). \tag{4.2}$$

This expression defines the spectral power distribution (SPD) of the light reflected from the surface at the considered point. An image of the point taken with a linear device is composed by sensor responses that can be described by Eq. (2.8). By applying the linearity of spectral integration and by enclosing the factor of proportionality between radiance and irradiance in the scalars m_s and m_b, the measured surface color is obtained as

$$\vec{C}_L = m_s\vec{C}_s + m_b\vec{C}_b, \tag{4.3}$$

where $\vec{C}_s$ and $\vec{C}_b$ are the 3D color vectors in color space of the interface and body reflection terms, respectively.

Let us consider now the colors of a set of points on the same uniformly colored surface. Since the geometry is different at each point in the set, the scale factors m_s and m_b vary from point to point. However, the colors $\vec{C}_s$ and $\vec{C}_b$ of the interface and body reflection are the same at all points on the same surface. They are, in fact, the result of the spectral integration of $c_s(\lambda)$ and $c_b(\lambda)$ that do not vary with geometry. Without loss of generality [143], it can be considered that $0 \leq m_s, m_b \leq 1$. As a consequence, the pixel values for a set of points on a uniformly colored surface must be distributed within a parallelogram in the RGB sensor space. The parallelogram is bounded by the colors $\vec{C}_s$ and $\vec{C}_b$ of the interface and body reflection of the surface. One corner of this parallelogram is located at the origin of the color space, where $(R, G, B) = (0, 0, 0)$. The dichromatic parallelogram is illustrated in Figure 4.1.

Light reflection at matte points is primarily determined by the body reflection process [82]. Therefore, matte points form a matte line in the direction of the body reflection vector in the plane defined by the parallelogram. Highlight points exhibit both body reflection and interface reflection. However, since $m_s(i, e, g)$ is much more sensitive to a small change in the angles than is $m_b(i, e, g)$, the body reflection component is generally approximately constant in an highlight area. Highlight points thus form an highlight line in the plane in the direction of the interface reflection vector

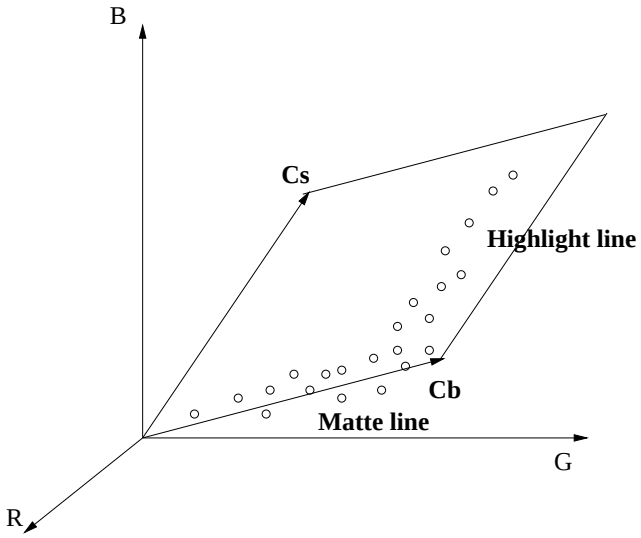

Figure 4.1: Considering dichromatic reflection, pixel values on a uniformly colored surface lie on a parallelogram in color space. The parallelogram is bounded by the colors $\vec{C}_s$ and $\vec{C}_b$ of the interface and body reflection. Matte points form a matte line in the direction of the body reflection vector. Highlight points form an highlight line in the direction of the interface reflection vector.

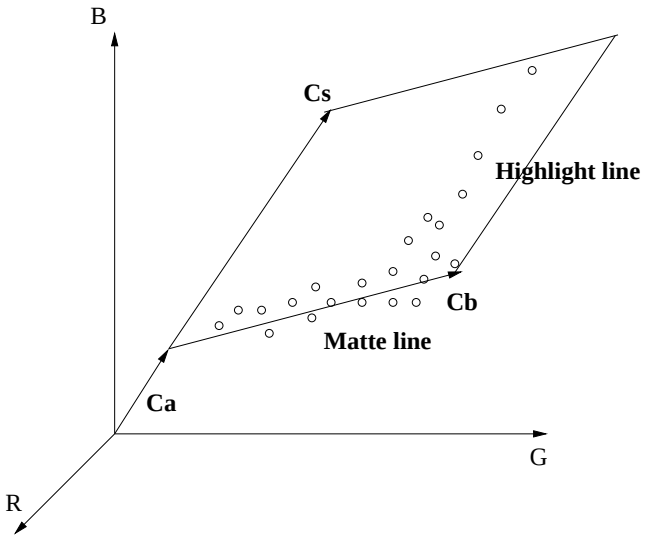

Figure 4.2: Dichromatic parallelogram considering a constant ambient illumination term.

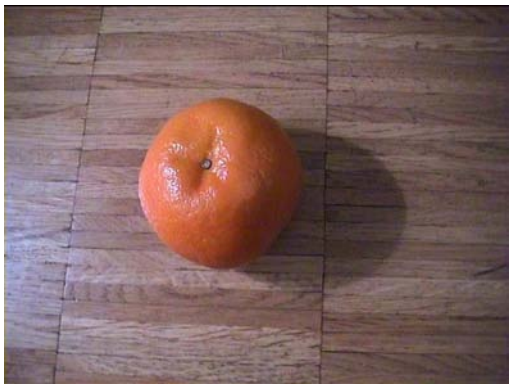

Figure 4.3: Original color image from which selected regions of interest are analyzed in Figure 4.4 and Figure 4.8.

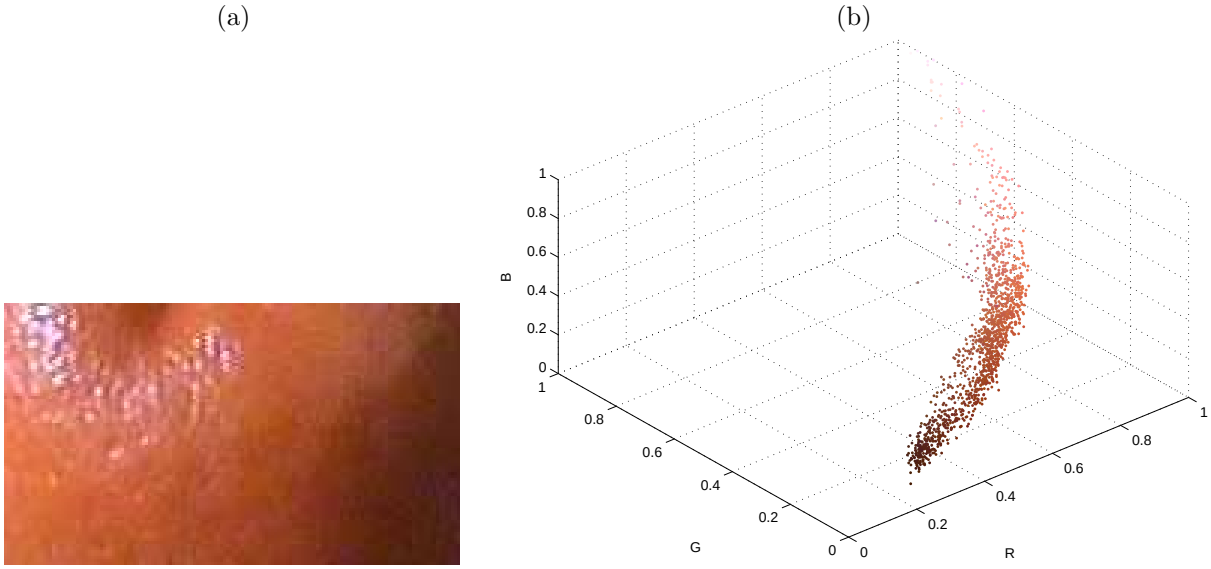

Figure 4.4: (a) A region of interest containing a uniformly colored surface affected by shading, a self shadow and highlights. (b) The corresponding pixel colors in the RGB color space.

(Figure 4.1). The extent of the line depends on the roughness of the object surface. For rough surfaces, the extent will be smaller than for very shiny surfaces.

When analyzing shadows, as commented in Section 2.4.2, the extension of the Dichromatic Reflection Model should be considered. In this case, Eq. (2.6) expresses the reflected radiance at the considered point as

$$L(\lambda, i, e, g) = m_s(i, e, g)c_s(\lambda) + m_b(i, e, g)c_b(\lambda) + L_a(\lambda), \tag{4.4}$$

where $L_a(\lambda)$ is the ambient reflection term, which is assumed independent from geometry. The model in the RGB space then becomes

$$\vec{C}_L = m_s\vec{C}_s + m_b\vec{C}_b + \vec{C}_a. \tag{4.5}$$

For a set of points on a uniformly colored surface when considering a constant ambient diffuse

illumination, the parallelogram origin is then moved to a point which represents the color of the reflected light due to ambient illumination. This situation is shown in Figure 4.2.

An example of pixel distributions in a real image (Figure 4.3) is illustrated in Figure 4.4. A region of interest containing a uniformly colored surface which exhibits highlights, shading due to surface curvature and a self shadow due to the occlusion of the direct light is shown in (a). The corresponding pixels color distribution in the RGB space are plotted in (b). The matte line and the highlight line are visible. The color cluster does not pass through the origin of the color space due to ambient light, as expected. Pixels in the self shadow region form a cloud of points at the base of the matte line. The color of the pixels in shadow, according to the considered model, is defined by $\vec{C}_a$, which does not depend on geometry and should therefore be the same for all points on the same shadowed surface. In practice, shadow pixels are spread and form the observed cloud of points.

The above-discussed model is, in fact, an idealized physical description of the world. Real images do not fully comply with it because of camera limitations on one hand and because of effects in the scene, such as inter-reflections among surfaces, that are not modeled in the considered reflection model. Real cameras have, for instance, only a limited dynamic range to sense the incoming light. In this case, if the diffuse body reflection or the interface reflection are bright, one or another color channel might saturate. The corresponding lines may collide with a face of the RGB color cube and get clipped. Typically, moreover, cameras are gamma corrected (Section 2.6.2). This means that the output of a camera is not a linear function of the input. Due to gamma correction, a curvature in the color clusters may then be introduced.

4.3.2 Construction of color invariants

Many different expressions can be derived from RGB values that are invariant to changes in the imaging conditions. In the following subsections, first, photometric invariant color features for matte, diffuse surfaces are considered, then color invariants for matte and shiny, specular surfaces.

Before proceeding, let us express here the Dichromatic Reflection Model in terms of surface spectral reflectances. We discount ambient illumination for the moment. We will consider ambient illumination in Section 4.4, as it is characteristic of illumination in shadows. Let $\rho_s(\lambda)$ and $\rho_b(\lambda)$ be the surface spectral reflectances at a point on a surface for the interface and body reflection components, respectively, and let $E(\lambda)$ be the spectral power distribution of the light incident on the surface at the considered point. Then, we have

$$L(\lambda, i, e, g) = m_s(i, e, g)\rho_s(\lambda)E(\lambda) + m_b(i, e, g)\rho_b(\lambda)E(\lambda). \tag{4.6}$$

As discussed in Section 2.4.1, the spectral energy distribution of the specular reflection component is similar to the spectral energy distribution of the incident light. Researchers usually assume that they are identical (Neutral Interface Reflection (NIR) assumption, see Section 2.4.1). As a result, the interface reflectance component $\rho_s(\lambda)$ is a constant, ρ_s, and the model formulation becomes

$$L(\lambda, i, e, g) = m_s(i, e, g)\rho_s E(\lambda) + m_b(i, e, g)\rho_b(\lambda)E(\lambda). \tag{4.7}$$

Diffuse surfaces

When diffuse surfaces are considered, reflected light is described by body reflection only. The sensor responses for a point on the surface are then given by

$$C_b^i = m_b(i, e, g) \int_\Lambda \rho_b(\lambda)E(\lambda)S_{C_i}(\lambda)d\lambda. \tag{4.8}$$

where $C_b^i \in \{R_b, G_b, B_b\}$ and $S_{C_i} \in \{S_R, S_G, S_B\}$.

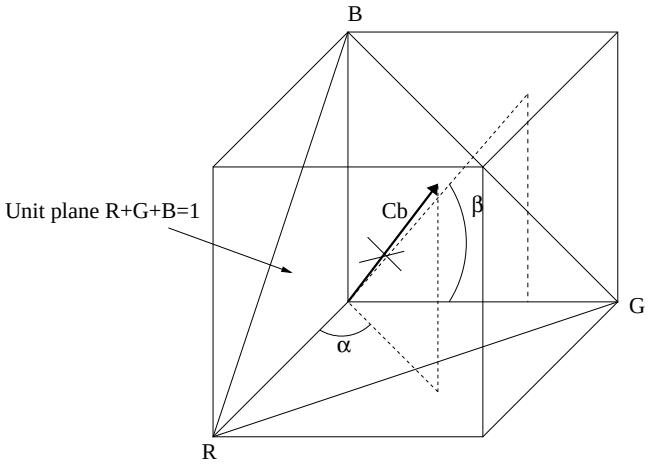

Figure 4.5: Considering dichromatic reflection, points on a uniformly colored matte surface form a line in the RGB color space in the direction of the body reflection vector $\vec{C}_b$. All points are projected onto the same point in the unit plane and have the same chromaticity coordinates. Moreover, all points can be described by the same angles α and β.

As discussed in Section 4.3.2, for points on a uniformly colored surface Eq. (4.8) describes a matte line with parameter m_b in the RGB color space (Figure 4.1). The line passes through the origin of the color coordinate system and its direction is defined by the body reflection vector $\vec{C}_b$. Changes in the imaging geometry affect the value of the parameter m_b but not the line direction. Changes in illumination affect the term $E(\lambda)$ and consequently the body reflection vector $\vec{C}_b$. However, changes in illumination intensity can be modeled by a scaling for each wavelength, that is $E'(\lambda) = \theta E(\lambda)$ and they have the same effect as a change of the geometric parameter m_b. Therefore, changes in illumination intensity only affect the position of the considered point on the matte line. The point moves toward the origin of the color space if $\theta < 1$ or in the opposite direction if $\theta > 1$.

All color vectors representing points on a matte uniformly colored surface, even when surface geometry, viewing direction, illumination direction or illumination intensity changes, share the same direction in color space. This direction is defined by $\vec{C}_b$ and is the invariant we are looking for.

The discussed physical model suggests different ways of obtaining combinations of RGB values which have the same value for all points on the matte surface. A family of color invariants [51] can be obtained by computing for each pixel position (x, y) in the image the expression

$$C^i(x, y)/C^j(x, y), \tag{4.9}$$

where $C^{i,j} \in \{R, G, B\}$ and $i \neq j$. The invariance is proved by substituting Eq. (4.8) in Eq. (4.9):

$$\frac{m_b(i, e, g) \int \rho_b(\lambda) E(\lambda) S_{C_i}(\lambda) d\lambda}{m_b(i, e, g) \int \rho_b(\lambda) E(\lambda) S_{C_j}(\lambda) d\lambda}. \tag{4.10}$$

The obtained expression depends on $S_C(\lambda)$, $\rho_b(\lambda)$ and on the spectral content of $E(\lambda)$, thus being the same for points illuminated by light of different intensity and direction, having a different surface normal and viewed from a different direction.

Any linear combination of this basic set of invariants gives a new color invariant feature, computed as

$$\frac{\sum_i a_i R_i^p G_i^q B_i^r}{\sum_j b_j R_j^s G_j^t B_j^u} \tag{4.11}$$

where $p+q+r=s+t+u$, and $p,q,r,s,t,u,a_i,b_j \in \mathbb{R}$ and $i,j \geq 1$.

A first example of instantiation of the above defined family of photometric invariant features are *normalized rgb features* (Eqs. (2.19)-(2.21)). As stated in Section 2.6.2, computing rgb from RGB values corresponds to radially projecting each color point in the 3D RGB space onto the unit plane $R+G+B=1$. When applying this color transformation on RGB values, points on the matte line are therefore projected onto the same point on the unit plane and have the same chromaticity coordinates. The body reflection vector and its projection onto the unit plane are shown in Figure 4.5.

A second family of photometric invariants can be obtained by considering the angles formed by the body reflection vector in the color space. Each point on the $\vec{C}_b$ vector can in fact be defined by two angles, α and β, and a distance from the origin of the color space, ρ (see Figure 4.5). The two angles α and β, computed from the RGB coordinates as

$$\begin{aligned} \alpha &= \arctan\left(\frac{G}{R}\right) \\ \beta &= \arctan\left(\frac{B}{G}\right), \end{aligned} \tag{4.12}$$

are a second example of photometric invariant features for matte surfaces and dichromatic reflectance.

Gevers [50] proposes the $c_1c_2c_3$ features, defined as

$$\begin{aligned} c_1 &= \arctan\left(\frac{R}{max(G,B)}\right) \\ c_2 &= \arctan\left(\frac{G}{max(R,B)}\right) \\ c_3 &= \arctan\left(\frac{B}{max(R,G)}\right), \end{aligned} \tag{4.13}$$

as an instantiation of this family of color invariants for diffuse surfaces. They are three of the six possible angles formed by $\vec{C}_b$ in the color space.

Among the well-known color features, *saturation* and *hue* in the HSI space, defined as in Eqs. (2.30-2.31), are also invariant features for diffuse surfaces. As for the other features, the demonstration of invariance is straightforward when substituting Eq. (4.8) in

$$S = 1 - 3\frac{min(R,G,B)}{R+G+B} \tag{4.14}$$

and

$$H = \arctan\left(\frac{\sqrt{3}(G-B)}{(R-G)+(R-B)}\right). \tag{4.15}$$

Saturation can be included in the first family of the discussed features, defined by Eq. (4.11), while hue belongs to the second family of angular features.

Diffuse and specular surfaces

When both matte and shiny surfaces are considered, the interface reflection vector has to be included. The RGB color components are now computed from Eq. (4.7) as

$$C_s^i = m_s(i,e,g)\rho_s \int_\Lambda E(\lambda)S_{C_i}(\lambda)d\lambda + m_b(i,e,g)\int_\Lambda \rho_b(\lambda)E(\lambda)S_{C_i}(\lambda)d\lambda. \tag{4.16}$$

As discussed in Section 4.3.2, for points on a uniformly colored surface Eq. (4.16) describes a dichromatic plane with parameters m_b and m_s in color space (Figure 4.1). The plane passes through the origin of the color coordinate system and its orientation is defined by the body and interface reflection vectors $\vec{C}_b$ and $\vec{C}_s$. Changes in the imaging geometry affect the values of parameters m_b and m_s but not the plane orientation. The same is true for changes in illumination intensity.

All color vectors representing points on a diffuse and specular surface, even when imaging geometry or illumination intensity changes, share the fact of lying on the same plane in the color space. The plane's orientation is defined by $\vec{C}_b$ and $\vec{C}_s$ and is the invariant we are looking for.

Different ways of obtaining color invariants which have the same value for all points on the matte and shiny surface can be devised. As before, we prove mathematically the intuitive notion for different color features. To this end, two assumptions have to be introduced. First of all, illumination is assumed to be *white or spectrally smooth* (i.e. approximately equal/smooth energy density for all wavelengths of the visible spectrum), that is $E(\lambda)$ has an equal value E for each wavelength. The second assumption concerns the characteristics of the camera's sensors. The *integrated white condition* is assumed to hold, that is the area under the sensors spectral functions is approximately the same:

$$\int_\Lambda S_R(\lambda)d\lambda = \int_\Lambda S_G(\lambda)d\lambda = \int_\Lambda S_B(\lambda)d\lambda = f. \tag{4.17}$$

To make the notation more compact, we define

$$K_s = E\int_\Lambda S_{C_i}(\lambda)d\lambda = Ef \tag{4.18}$$

Under the considered assumptions, the surface reflection vector lies on the diagonal of the RGB cube. If we denote with

$$K_b^i = E\int_\Lambda \rho_b(\lambda)S_{C_i}(\lambda)d\lambda, \tag{4.19}$$

Eq. (4.16) can be rewritten as

$$C^i = m_s(i,e,g)\rho_s K_s + m_b(i,e,g)K_b^i. \tag{4.20}$$

A family of color invariants for matte and shiny surfaces can be obtained by computing for each pixel position (x,y) in the image the expression

$$\frac{C^i(x,y) - C^j(x,y)}{C^k(x,y) - C^l(x,y)}, \tag{4.21}$$

where $C^k \neq C^l$. The invariance is demonstrated by substituting Eq. (4.20) in Eq. (4.21) as

$$\frac{m_s(i,e,g)\rho_s K_s + m_b(i,e,g)K_b^i - m_s(i,e,g)\rho_s K_s - m_b(i,e,g)K_b^j}{m_s(i,e,g)\rho_s K_s + m_b(i,e,g)K_b^k - m_s(i,e,g)\rho_s K_s - m_b(i,e,g)K_b^l} = \frac{K_b^i - K_b^j}{K_b^k - K_b^l}. \tag{4.22}$$

The obtained expression only depends on S_C and ρ_b, thus being the same for points on the same uniformly colored surface illuminated by light of different intensity and direction, having a different surface normal and viewed from a different direction.

Any linear combination of this basic set of invariants gives a new color invariant feature, computed as

$$\frac{\sum_i a_i(R-G)_i^p(B-R)_i^q(G-B)_i^r}{\sum_j b_j(R-G)_j^s(B-R)_j^t(G-B)_j^u} \tag{4.23}$$

where $p+q+r = s+t+u$, and $p,q,r,s,t,u,a_i,b_j \in \mathbb{R}$ and $i,j \geq 1$.

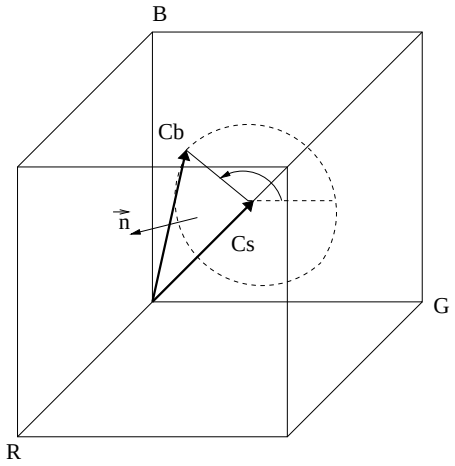

Figure 4.6: Considering dichromatic reflection and white illumination, points on a uniformly colored surface lie on a plane spanned by the body reflection vector $\vec{C}_b$ and the diagonal of the RGB cube, on which the interface reflection vector $\vec{C}_s$ lies. All points have the same hue angle. Moreover, all points can be described by the same unit normal vector $\vec{n}$.

Gevers proposes in [50] the $l_1l_2l_3$ features, defined as

$$\begin{aligned} l_1 &= \frac{(G-R)^2}{(G-B)^2+(R-B)^2+(G-R)^2} \\ l_2 &= \frac{(R-B)^2}{(G-B)^2+(R-B)^2+(G-R)^2} \\ l_3 &= \frac{(G-B)^2}{(G-B)^2+(R-B)^2+(G-R)^2}, \end{aligned} \tag{4.24}$$

as an instantiation of the discussed family of color invariants. $l_1l_2l_3$ are the squared components of the unit normal vector, $\hat{n} = (n_R, n_G, n_B)$, to the plane spanned by $\vec{C}_b$ and $\vec{C}_s$ when the Hessian normal form for the plane is considered (Figure 4.6). By substituting Eq. (4.20) in Eq. (4.24) the invariance is proved in a straightforward way.

As for diffuse surfaces, a second family of photometric invariants can be derived by considering angular features describing points on the dichromatic plane. An example of instantiation of this family of photometric invariant features is *hue*. By substituting Eq. (4.20) in Eq. (4.15) the invariance is easily demonstrated. When looking at Figure 4.6, moreover, the invariance is intuitively proved. All colors points on the plane spanned by $\vec{C}_s$ and $\vec{C}_b$ have the same hue assuming white illumination since hue is defined as a function of the angle between the main diagonal and the color point in the RGB space.

4.4 Invariance to shadows

Photometric invariant features are now analyzed with respect to the model of shadows introduced in the previous chapter. The analysis leads to a formalization of the second shadow cue in the list discussed in Section 3.2.2 and provides an analysis criterion that will be used in the proposed shadow segmentation approach.

At the end of Section 3.3.1, we have concluded that, if the gray world condition is assumed, the camera response to the ambient light contribution, that is the camera response in shadow $\vec{C}_a$, is

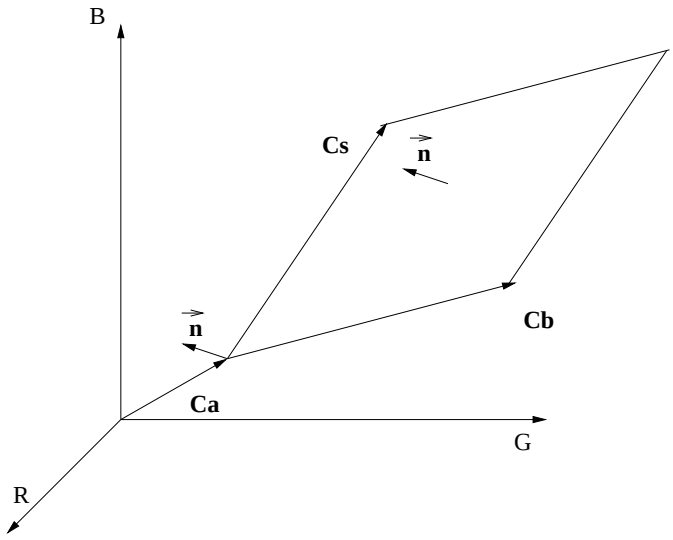

Figure 4.7: Considering dichromatic reflection and the gray world condition discussed in Section 3.3.1, the ambient reflection vector $\vec{C}_a$ is a linear combination of the body and surface reflection vectors $\vec{C}_b$ and $\vec{C}_s$. In such case, the plane containing all points on a uniformly colored surface passes through the origin of the color space. Points in shadow lie at C_a.

a linear combination of the responses to the body, $\vec{C}_b$, and interface, $\vec{C}_s$, reflection terms due to direct light. This means that the dichromatic plane containing all points on a uniformly colored surface passes through the origin of the color space, as illustrated in Figure 4.7. Shadow points illuminated by ambient light are then *coplanar* with respect to all other illuminated points on the same surface. Consequently, if the direct illumination is white and the integrated white condition holds for the used camera, the discussed set of color invariants for matte and shiny surfaces, among which hue and $l_1l_2l_3$ can be taken as instantiations, takes the same values for points directly lit and points in shadow on the same surface, whatever the surface orientation at the point is, the direct illumination's intensity and the viewing direction are, and whether the point is affected by an highlight or not.

In the case of regions that do not contain highlights, the ambient reflection term $\vec{C}_a$ is then aligned with the body reflection term $\vec{C}_b$ and, from Eq. (3.9), that is

$$\begin{aligned} R_{shadow} &= \alpha R_{lit} \\ G_{shadow} &= \alpha G_{lit} \\ B_{shadow} &= \alpha B_{lit}, \end{aligned} \tag{4.25}$$

it follows that all the discussed color invariants, among which normalized rgb, $c_1c_2c_3$, saturation, hue and $l_1l_2l_3$ are instantiations, take the same value for the same point when lit and when in shadow. This is true for all shadowed points on the same diffuse surface, whatever the surface orientation at the point is, the direct illumination's intensity and the viewing direction are.

An example of pixel distribution in the RGB space for a surface that is partially covered by a cast shadow is illustrated in Figure 4.8. A region of interest selected from the image in Figure 4.3 can be observed together with the corresponding pixel distribution in color space. Since no specularities are present in the selected region, pixels form a linear cluster for which color invariants take the same value. More examples of the behavior of color invariant features in presence of shadows in a number of real images and image sequences will be presented and discussed in the next chapter in

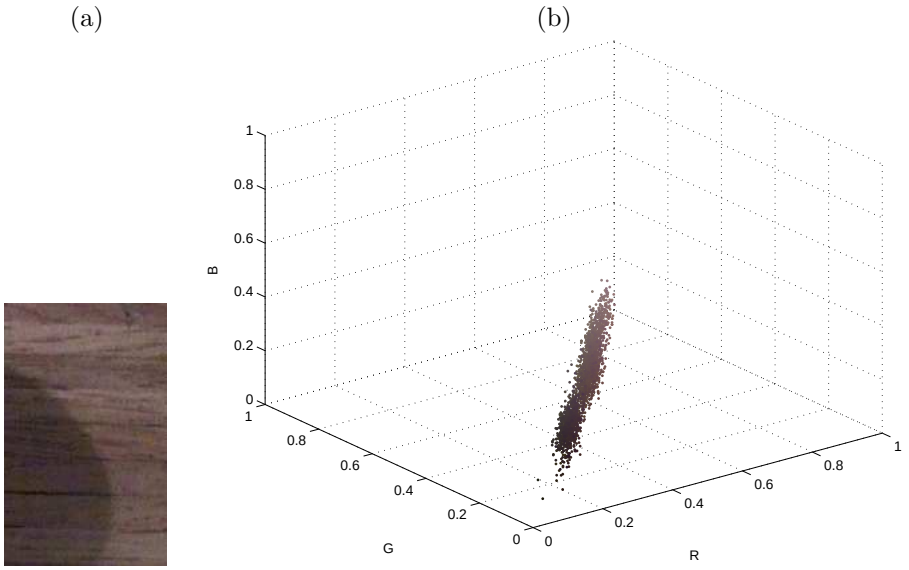

Figure 4.8: (a) A region of interest containing a surface covered by a cast shadow selected from Figure 4.3. (b) The corresponding pixel colors in the RGB color space.

Section 5.3.

The fact that the discussed features remain unchanged in presence of a shadow can be exploited for detecting shadows in digital images and image sequences. Let us define as F one of the above mentioned photometric color invariants. F_l is the value assumed by the invariant feature in a point in light, and F_s is the value in the same point in shadow. Then,

$$F_l = F_s. \tag{4.26}$$

This property is valid for all the points inside a shadow region, even when the normal to the surface changes and whether the point belongs the shadow's umbra or penumbra. It will be used, together with the property in Eq. (3.8), in Chapter 5.

4.4.1 Discussion

It is important at this point to analyze the problems and drawbacks of the discussed features related to their loss of discriminative power and the inherent instabilities caused by the non-linear transformations used in their computation.

Photometric invariant transformations become unstable in certain regions of the RGB space. It is known from Kender [80], in fact, that normalized color rgb and saturation are unstable, that is more sensitive to small perturbations such as those due to noise, for color values near the black vertex of the RGB space. Here, these features are undefined. Hue is unstable near its singularities at the entire RGB space diagonal. The analysis of noise sensitivity of $c_1c_2c_3$ and $l_1l_2l_3$ transformations has been carried out by Gevers in [51]. As normalized rgb and saturation, $c_1c_2c_3$ give rise to unstable values in presence of noise when intensity is small, while $l_1l_2l_3$ are unstable near $R = G = B$, as hue. Features that exhibit the same class of invariance are characterized by the same kind of problems in presence of noise. As the degree of invariance grows, the number of points were the color models are not defined, and exhibit a characteristic instability, grows. Table 4.1 summarizes the problems of the mentioned features, which are representative of the two groups discussed in the previous section.

Property	rgb	$c_1c_2c_3$	S	H	$l_1l_2l_3$
Undefined at (R,G,B)=(0,0,0)	Yes	Yes	Yes	Yes	Yes
Undefined at R=G=B	No	No	No	Yes	Yes
Sensitive to noise at (R,G,B)=(0,0,0)	Yes	Yes	Yes	Yes	Yes
Sensitive to noise at R=G=B	No	No	No	Yes	Yes

Table 4.1: Problems of color invariants.

The advantage of being robust to shadows and to changes in the imaging conditions is obtained for photometric invariants at the cost of a loss in their discriminative power. Invariants transform color images into a simpler feature space where some information is discarded. As the degree of invariance grows, the number of different colors that color invariants can discriminate decreases. For all invariants there are degenerate cases in which they are unable to distinguish between different material surfaces. Let us take a numerical example and consider a point having a magenta color given by $(R, G, B) = (0, 1, 1)$ and a second point having a red color given by $(R, G, B) = (1, 0, 0)$. For such points the normalized rgb values are $(0, 0.5, 0.5)$ and $(1, 0, 0)$. The $l_1l_2l_3$ values are $(0.5, 0.5, 0)$ and $(0.5, 0.5, 0)$, that is the same for the two points.

A compromise between invariance and inherent limitations of photometric invariant transformations should be searched for when using them for shadow segmentation. This issue will be discussed in detail in Chapter 5, where different features will be evaluated for selecting the color invariants to be used in the proposed approach.

To conclude this discussion, it is interesting to reconsider the color features that are used by the state of the art methods for shadow detection reviewed in Chapter 3. The reader is referred to Table 3.1 and Table 3.2 for a summary of approaches, where methods that exploit color invariance can be rapidly identified.

The techniques proposed in [42, 127] use the video-oriented $Y'C_bC_r$ space. They base their analysis on the observation that shadows modify the luma component, while the chroma components remain unchanged. Let us analyze how chroma components are influenced by a change in illumination due to shadows when considering dichromatic reflection. From Eq. (2.25), chroma components are computed from RGB values as

$$C_b = 128 - 0.148R - 0.291G + 0.439B \tag{4.27}$$

$$C_r = 128 + 0.439R - 0.368G - 0.071B \tag{4.28}$$

If we consider a matte surface for simplicity, by substituting Eq. (4.8) in Eqs. (4.27-4.28), and by appropriately developing the computations, it can be shown that C_b and C_r only assume the same value, which is zero, with a change in illumination intensity if the considered point is achromatic, that is it lies on the diagonal of the RGB cube.

The methods proposed in [27, 142, 168] make use of the invariance properties of both hue and saturation. Hue and saturation, which belong to the same color space, yet show a different class of invariance. Saturation is only invariant for matte surface, while hue is invariant also to highlights. By using both features, the more restrictive hypotheses behind saturation's invariance are implicitly considered. This issue will be reconsidered in Section 5.3.

The method in [52] uses the c_1c_2 features which are invariant for matte surfaces. Finally, when looking at Figure 3.15 and comparing it with Figure 4.7 it is clear that the method proposed in [70] also considers matte surfaces. For such surfaces, in fact, shadowed points lie on the matte line defined by the expected color E (Figure 3.15). As a consequence, their chromatic distortion CD is zero.

4.5 Summary

The objective of this chapter is to define color features that do not change their values in presence of shadows and that can provide a criterion for shadow segmentation. They allow, in fact, to formalize the second spectral property of shadows in the list of visual cues discussed in Section 3.2.2, which states that changes in the color of a surface in shadow tend to be predictable.

Invariance to illumination changes due to shadows falls within the more general problem of the invariance to imaging parameters, which comprise illumination conditions, surface orientation and viewing direction. This problem has been extensively studied for the purpose of color-based object recognition. Different approaches have been proposed in this context to obtain functions of the RGB color values that cancel out dependency on the imaging conditions. In this chapter, we have reviewed different solutions in light of their possible use for the analysis of shadows.

Shadows are a local illumination phenomenon. Consequently, only those approaches which can provide an invariant feature defined for each pixel of a color image are suitable for analyzing shadows. Our investigation has outlined two classes of such transformations, which are based on two different models of shadow. The first model is that introduced in Section 3.3.1, which considers the gray world condition to hold. The second model assumes Lambertian surfaces, narrow-band camera filters and Planckian lighting. It relaxes the gray world condition. For this latter model, the computation of the invariant parameter for each pixel position depends on the used camera. In order to design a shadow segmentation approach that can be applied also on images from uncalibrated sources, features that are invariant to shadows as described by the former model have been selected in this work.

The construction of these invariants for matte and for matte and shiny surfaces was presented. Among them, well-known features, such as hue and saturation, are used by several state of the art methods for shadow detection. Here, the common framework in which the invariance of such features, and many other color invariants showing similar properties, is derived, was discussed. The underlying physical model, the relevant assumptions, and the problems related to the use of such invariants were pointed out. In the next chapter, different invariants are evaluated on a number of test images for their use in the proposed shadow segmentation approach.

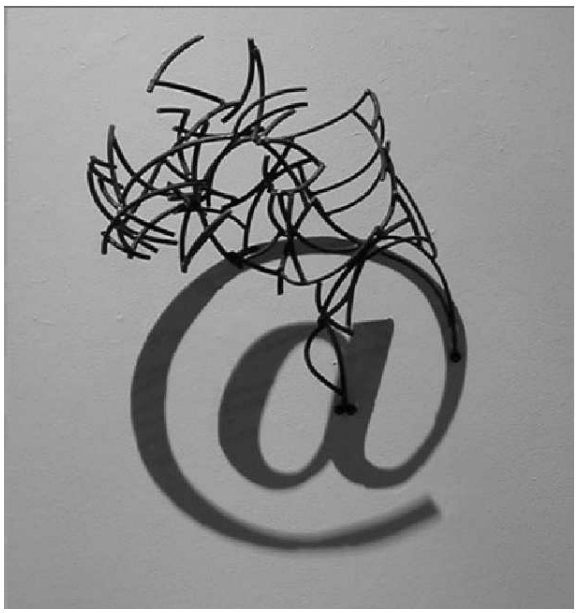

Figure 4.9: Wall sculpture in steel and shadow (Section A.1.3).

Segmentation of cast shadows

5

5.1 Introduction

In this chapter, we propose an analysis method for the segmentation of cast shadows in a wide range of natural scenes. The adopted strategy exploits spectral, spatial and temporal properties of shadows and is designed to be able to work automatically when camera, illumination and scene's characteristics are unknown.

The problem of extracting shadows from images has been investigated within several research domains. Its importance has especially come to the fore over the past years in the framework of automatic video processing and analysis methods. The accurate segmentation of moving objects in video sequences represents a key process for an always wider range of multimedia and computer vision applications. Consequently, the development of efficient methods for the identification of shadows cast by objects is of primary interest. The often unavoidable presence of shadows in natural scenes is in fact of great nuisance to automatic segmentation methods, which typically detect cast shadows as part of moving objects. In this thesis, therefore, particular attention is dedicated to the problem of segmenting *moving cast shadows* in color image sequences. The validity of the proposed approach is evaluated moreover through its implementation for the segmentation of *cast shadows in still color images*.

A specialized approach to shadow segmentation, based on object models or application domain specific knowledge, although providing successful solutions for the applications at hand, limits the generality of the proposed techniques and their extension to new applications. In particular, when shadow segmentation is performed for enhancing fundamental tasks such as object extraction and description, flexibility with respect to the nature of the considered objects and the variety of the considered scenes is expected. When the visual data content is not known a priori, such as in video coding, video editing and advanced video surveillance, this is especially desirable. The methodology proposed in this chapter is intended for such a framework and is designed to be applicable to different scenarios. To reach this objective, the criteria defined both in Section 3.3.1 and 3.3.2, and in Section 4.4 of Chapter 4 are exploited for segmenting shadows in a large class of scenes.

This chapter is dedicated to the description of the proposed methodology and of the adopted

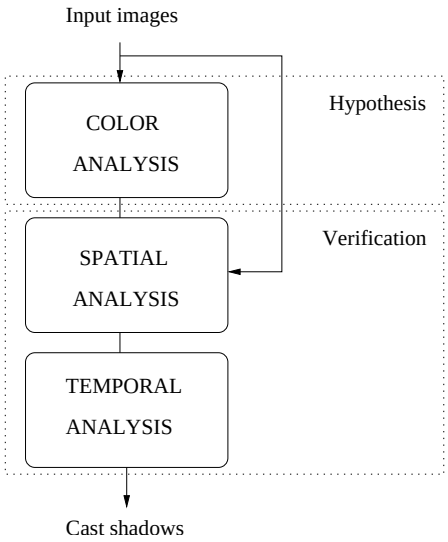

Figure 5.1: A simplified scheme of the three stages composing the proposed segmentation algorithm.

algorithmic solutions. An evaluation of the performance of the proposed system through its application to a number of test sequences and through the comparison with state of the art techniques will be provided in Chapter 6. The application of the proposed system in different contexts will be then discussed in Chapter 7.

The presentation is organized as follows. An overview of the proposed strategy is first of all presented in Section 5.2. An evaluation of color invariant features in the context of their application in the proposed methodology is provided in Section 5.3. The main phases of the analysis method are presented in Section 5.4, Section 5.5, and Section 5.6. The case of still images is analyzed in Section 5.7.

5.2 Overview of the proposed approach

The proposed segmentation approach is summarized in Figure 5.1. The analysis is organized in two main levels: an *hypothesis* level and a *verification* level. Its three main building blocks are shown. They are a *color analysis* stage, a *spatial analysis* stage, and a *temporal analysis* stage. This last stage is not present when the analysis is applied on still images. First of all, the color analysis stage generates an initial hypothesis about the presence of a shadow. Color analysis exploits shadows spectral properties on the basis of cue 1 in the list of Section 3.2.2, i.e. shadows darken the surface upon which they are cast, and by making use of the invariance properties of photometric invariant color features. The color analysis stage is discussed in Section 5.4.

Color information alone is not discriminative enough to allow for reliable shadow segmentation. After color analysis, therefore, a spatio-temporal verification is performed. As discussed in Section 3.3.2, in the spatial analysis stage we propose to exploit geometric properties of shadows related to shadow boundaries and to the adjacency of the shadow-casting object, without the need of a priori knowledge about the scene. The spatial analysis is presented in Section 5.5.

The temporal analysis stage aims at estimating the temporal reliability of the extracted potential shadows in dynamic scenes by means of a shadow tracking process. Tracking allows us to compute the life-span of each shadow. From each shadow's life-span and the relative position of objects and

shadows provided by the spatial analysis stage, a reliability estimation is derived. This reliability estimation is used to validate or to discard each shadow detected in the previous level of analysis and provides the final segmentation results. The proposed shadow tracking and temporal reliability estimation strategy are described in Section 5.6. The spatio-temporal verification process eliminates the possible ambiguities of the color analysis stage and improves the efficiency of the overall algorithm.

In the following sections, we describe in detail the adopted rules and the associated algorithmic solutions for each level of the proposed system for moving cast shadow segmentation in image sequences. The case of still images, which involves some specific assumptions and related solutions, will be analyzed in Section 5.7. Before going into the details of the method, an analysis and evaluation of invariant features in the context of their application for shadow segmentation in a wide range of real world scenes is provided in the next section. The analysis will allow to select appropriate features that will be then used in the proposed technique.

5.3 Invariant color features selection

We have seen in Chapter 4 that many different combinations of RGB color components can be defined which show invariant properties to shadows. Among them, *normalized rgb*, *hue*, and *saturation* are well known and widely used in the image processing literature. For this reason, we propose to consider them for shadow segmentation. Other transformations, such as $c_1c_2c_3$ and $l_1l_2l_3$, have been introduced for color-based object recognition. Since they have been shown to provide reliable results for image segmentation [48] we consider them as well in addition to traditional features.

By means of photometric invariant features we have introduced in Section 4.4 Eq. (4.26) which represents a theoretical model for shadow points. In practice, image pixel values only approximately comply with this model, first of all because of camera noise, but also because of other noise components. Illumination may, in fact, not always be considered white or spectrally smooth and camera sensors do not necessarily verify the integrated white condition (Eq. (4.17)). The physical model of shadows which underlies invariance is moreover a simplified description of the physical phenomenon. In particular, the gray world condition is a working hypothesis and holds only to a certain extent for real world images. The experimental analysis of the validity of such model in complex real world scenes is a challenging task that is beyond the scope of this work. Nevertheless, we propose in this section to investigate on a number of real images the behavior of different photometric invariant features with respect to Eq. (4.26). This will allow to identify possible classes of scenes which comply with the assumed model to a different extent and to define if some color invariant models result, in practice, more suitable than others for use in the segmentation of shadows when no control on the considered scenes is imposed and minimal assumptions are considered, as in the context of this thesis.

Two types of analysis have been carried out. First, the different color features values in lit and shadowed points have been computed and their behavior analyzed. Further, edge maps for the different color features have been obtained and compared. Images containing different real world illumination conditions have been considered. They have been chosen so as to contain different surface materials. Two classes of invariants have been in fact introduced in Chapter 4, invariants to shadows for diffuse surfaces, and invariants to shadows for diffuse and specular surfaces. Surfaces with different color content have also been selected since, when no control on the content of the considered scenes is imposed, we may end up at using color features also in color-deficient areas. Due to the numerical instability of some color invariants near the entire achromatic diagonal of the RGB cube, as discussed in Section 4.4.1, this factor should in fact also be taken into account when

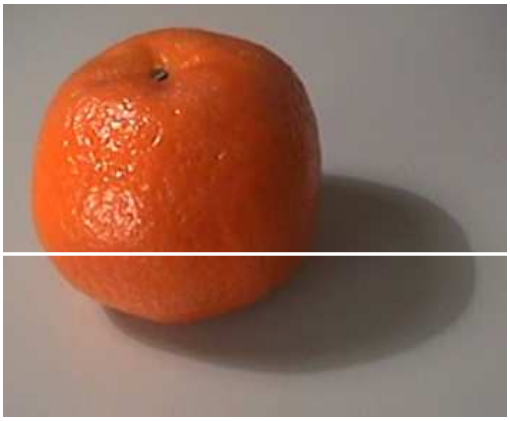

Figure 5.2: Test image *orange* and selected line for the analysis of color invariants.

using color invariants. Both still images and image sequences taken with different cameras have been analyzed so as to contain different noise levels. The results of the analyses are discussed in the following subsections by means of some representative examples.

5.3.1 Color components analysis

As first example, let us take an image of a simple close-up scene. The test image *orange* in Figure 5.2 shows an indoor scene where a shadow is cast by an object on a gray, uniform, plastic background. Shading due to the curvature of the surface and a self shadow are present on the object, which has a saturated, textured surface. Some highlights can be seen on the upper left part of the object, whose surface has a diffuse and a specular component. One direct nearby light source, an office lamp, and diffuse light from the environment which includes light from windows lighting the room, illuminate the scene.

For its analysis, we have fixed a line in the image which crosses the cast shadow and the object and observed the behavior of the different color components for each point on the line. The color features profiles for line 175 (see Figure 5.2) are shown in Figures 5.3–5.6. The intensity of the color component is plotted on the y axes while the pixel number is reported on the x axes. For comparison, the behavior of the invariant features has been displayed side by side with that of a component, such a R or I, which is sensitive to shadows. All features are in the range [0,255].

The RGB space and the invariant *normalized rgb* features are considered in Figure 5.3. The r component is shown and compared to the R component. Since the same considerations that will be done for r can be done for g and b, their profiles are not discussed here. The intensity step visible in the profile of the R channel at the cast shadow boundary is not present in the normalized r profile, showing that the invariance property of normalized color with respect to shadows holds well in this real image. Shading and the self shadow are also much less visible, as the intensity values vary much less than in the case of the R component from the self shadowed and shaded part to the lit part of the object. Normalized rgb is a photometric invariant for diffuse surfaces. However, even if the object surface contains both a diffuse and a specular component and invariance to shadows holds less, as expected, than in the diffuse background, it is nevertheless obtained to a good extent since the selected line crosses a region which does not contain highlights.

The c_2 color feature is analyzed in Figure 5.4 and compared with intensity I in the HSI space. I is sensitive to shadows and shading as R. As before, c_1 and c_3 are not shown since they have a similar behavior as c_2. As for normalized color, the invariance with respect to shadows and shading for

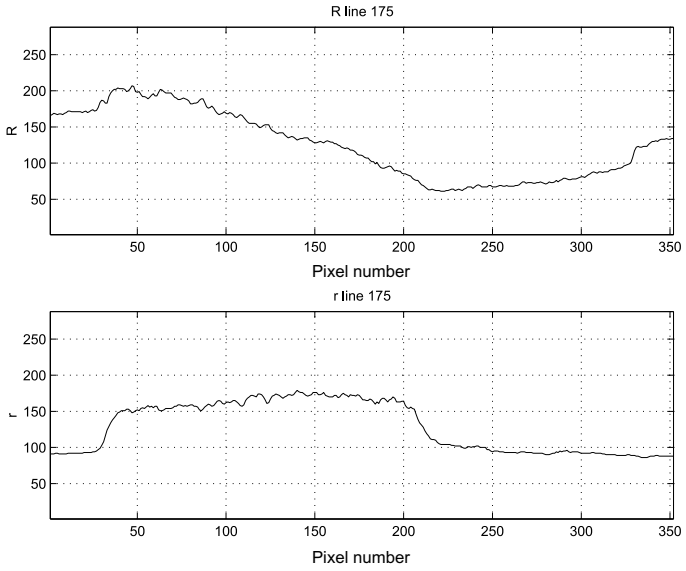

Figure 5.3: R and r components intensity profile for line 175 of image *orange*.

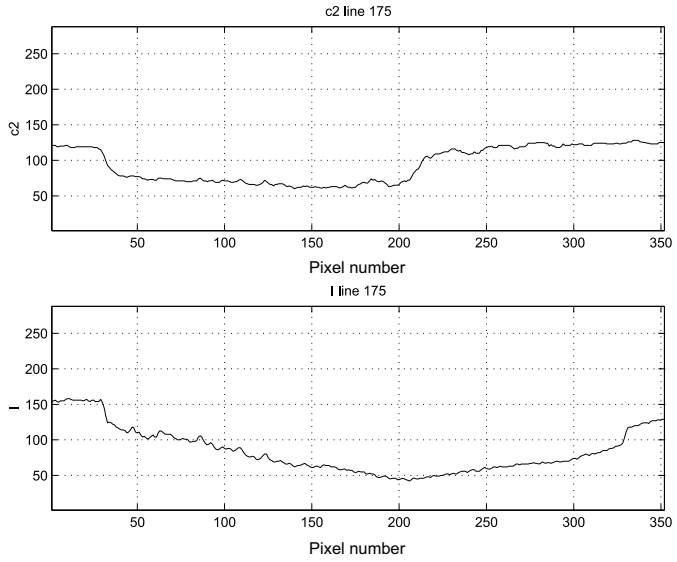

Figure 5.4: c_2 and I components intensity profile for line 175 of image *orange*.

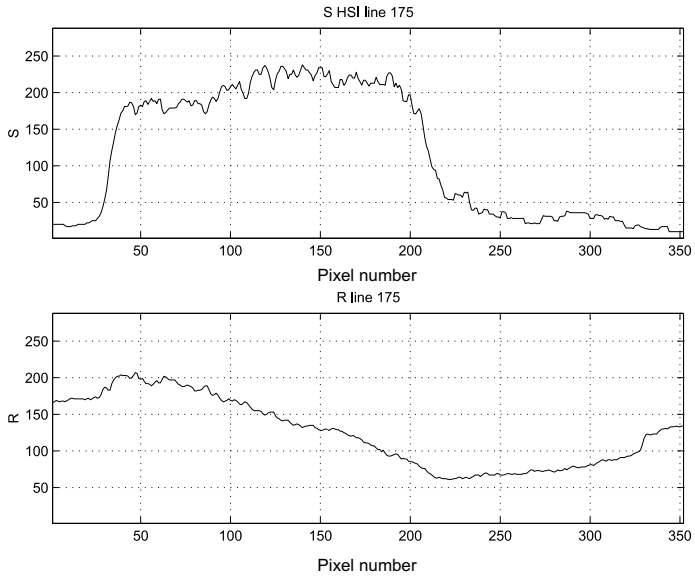

Figure 5.5: S and R components intensity profile for line 175 of image *orange*.

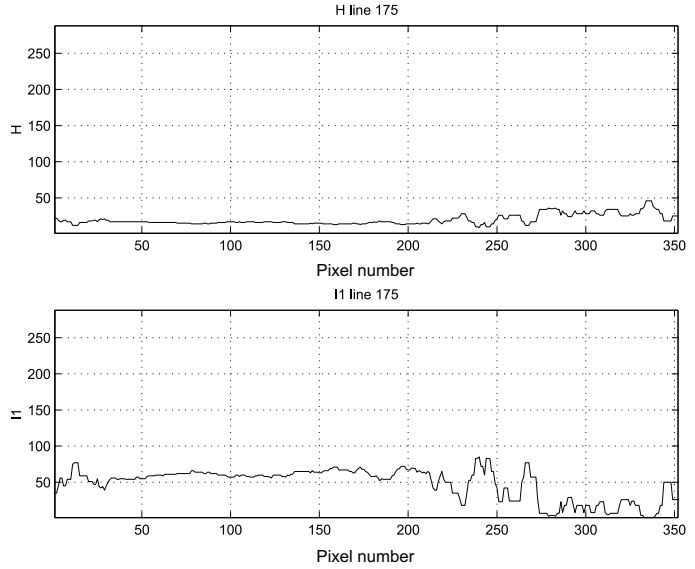

Figure 5.6: H and l_1 components intensity profile for line 175 of image *orange*.

$c_1c_2c_3$ is verified. Note the different behavior inside the object of I with respect to R in Figure 5.3. While R increases from the background to the object, I decreases. We will go back to this issue when discussing the color analysis stage in Section 5.4.2.

Saturation S in the HSI space is shown in Figure 5.5. It results as well invariant to shadows and shading. The intensity profile for the *hue* component in the HSI space is reported in Figure 5.6. The values of hue in the background region of the image show a variability that is due to the characteristic instability of H for low values of saturation (near $R = G = B$). The background is in fact gray. Inside the object, whose color is well saturated, the profile has a constant behavior, showing a high invariance with respect to the self shadow and to shading. Hue is a photometric

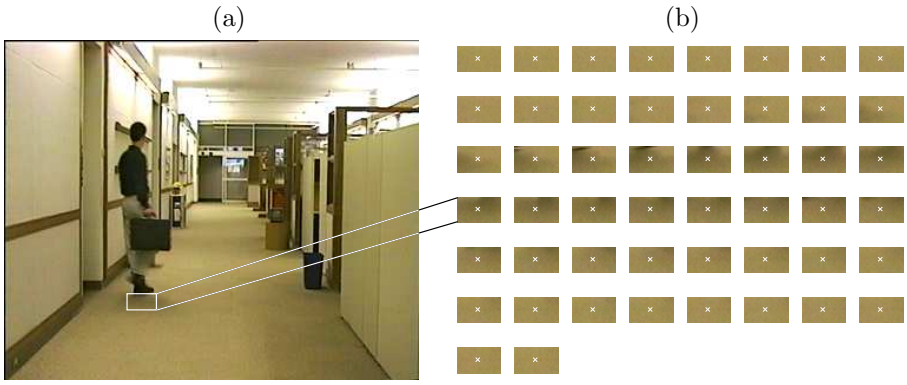

Figure 5.7: (a) Sample frame of the test sequence *Hall Monitor* and (b) selected region of interest over 50 frames. The central point is shown.

invariant for matte and shiny surfaces. The object's surface presents both a diffuse and a specular component and, as expected, hue has a higher invariance than the previously considered features. However, the higher invariance is obtained at the expenses of a decreased discrimination accuracy. By looking at the hue profile it is difficult to discriminate the object from the background, especially in its lit part.

As expected, the $l_1l_2l_3$ color features show the same problems as hue for low values of saturation in the background region of the image. The l_1 profile is illustrated in Figure 5.6. The variability of the l_1 values is higher than that of the hue values for this part of the image. The invariance to shading and to the self shadow is verified for l_1 inside the object.

To compare the case of an indoor scene illuminated by a single nearby light source, as in image *orange*, to that of an indoor scene illuminated by multiple light sources, let us consider now the test sequence *Hall Monitor* from the MPEG-4 Content Video Set. A sample frame is shown in Figure 5.7 (a). It shows a more complex scene, typical of indoor surveillance scenarios. For the analysis, we have fixed a point in a region of interest crossed by a cast shadow over time (Figure 5.7 (b)) and analyzed the behavior of the color components over a certain number of frames. The region of interest shows part of the diffuse, uniform, saturated floor's surface. While for the still image color components profiles showed how values changed in different spatial positions in the image, for image sequences we consider the same spatial position in the image and analyze the temporal variation of its values. The temporal variations of color features will be in fact analyzed for segmenting moving shadows in image sequences.

The behavior of *normalized rgb*, $c_1c_2c_3$, *saturation*, *hue* and $l_1l_2l_3$ for the central point of the selected region of interest are illustrated in Figures 5.8– 5.10. The intensity of the color component is plotted on the y axes while the frame number is reported on the x axes. The RGB components are shown for comparison in Figure 5.8. We can conclude that also in this case, the considered spectral model of shadows holds as demonstrated by the obtained invariance for all features. Saturation and $l_1l_2l_3$ seem more sensitive to fluctuations in RGB values due to camera noise than other invariants.

The test sequence *Highway* from the MPEG-7 Content Video Set is considered in Figure 5.11. In this case, the depicted scene is an outdoor traffic monitoring scene where the sky is overcast providing a very diffuse illumination (Figure 5.11 (a)). Consequently, very weak shadows are cast by vehicles on the road's asphalt, gray surface (Figure 5.11 (b)). Shadows are very diffuse but yet

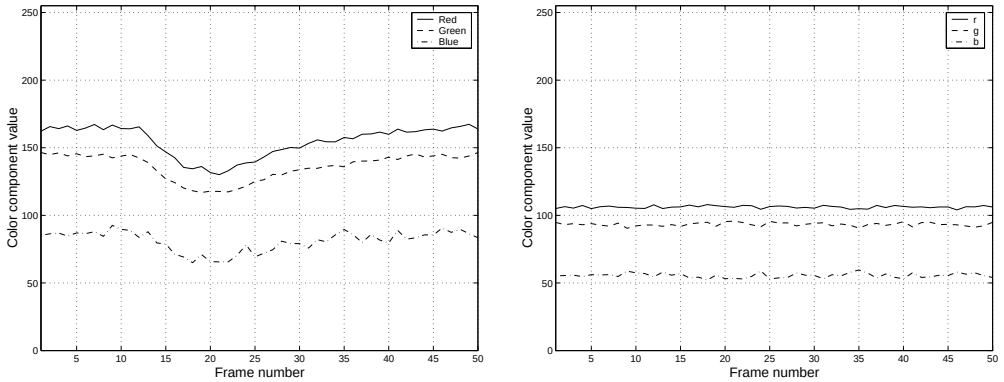

Figure 5.8: RGB and normalized rgb components intensity profile for the central point of the selected region of interest for the test sequence *Hall Monitor*.

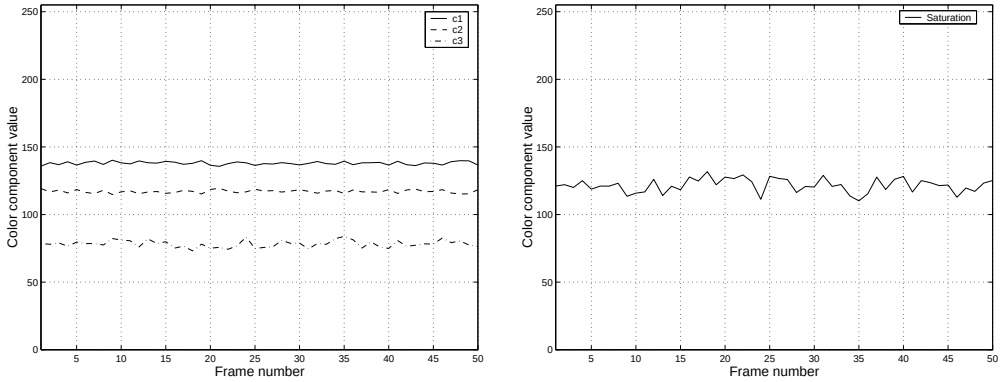

Figure 5.9: $c_1c_3c_2$ and S components intensity profile for the central point of the selected region of interest for the test sequence *Hall Monitor*.

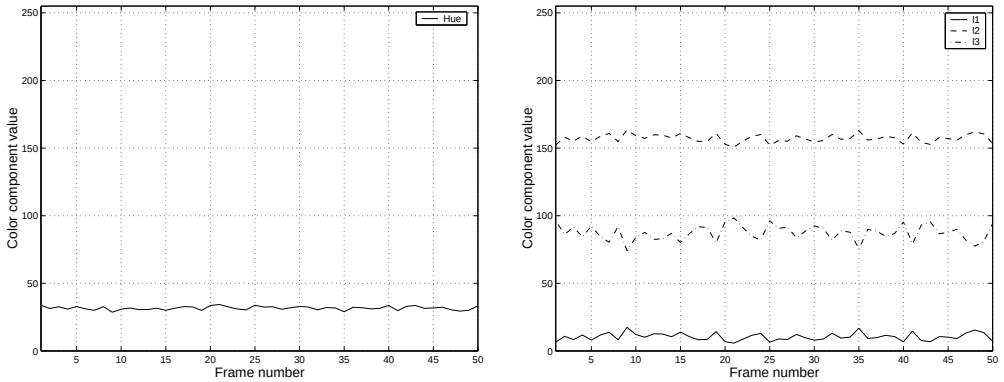

Figure 5.10: H and $l_1l_3l_2$ components intensity profile for the central point of the selected region of interest for the test sequence *HallMonitor*.

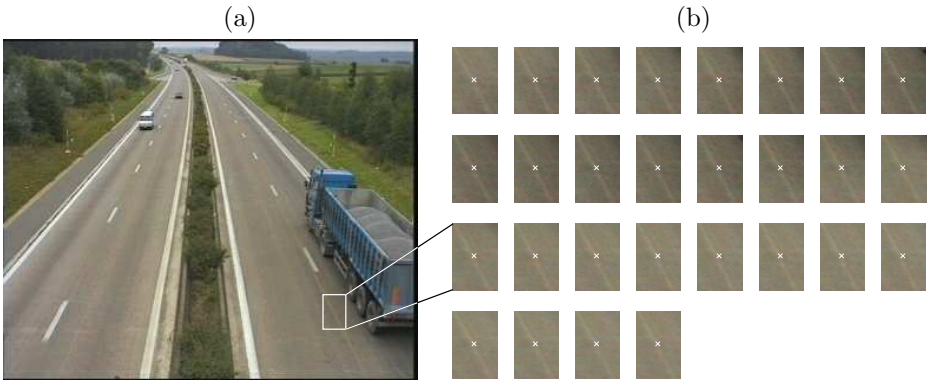

Figure 5.11: (a) Sample frame of the test sequence *Highway* and (b) selected region of interest over 28 frames. The central point is shown.

clearly visible and generating significant temporal changes in the image signal which can mislead motion analysis algorithms and consequently moving object detection results.

The observed invariance of rgb, $c_1c_2c_3$, and saturation in Figure 5.12 and Figure 5.13 shows how the case of outdoor overcast scenes can be associated from the point of view of illumination conditions to the case of indoor scenes for which the considered shadow model holds to a good extent. In this case, as for image *orange*, the selected region's color is gray and the corresponding saturation is quite low. This explains the unstable behavior of $l_1l_2l_3$ in Figure 5.14. The same problem, though less prominent, is visible in the hue profile when compared to the previous sequence where saturation values were much higher. Perez and Koch [122], who analyze hue for the purpose of color image segmentation, consider as unreliable hue values for pixels having saturation values which are below a minimum value of 20% of their total range. Here, saturation is then at the limit of their proposed threshold.

A sample frame of the test sequence *Surveillance* from the MPEG-7 Content Video Set is finally shown in Figure 5.15 (a). An outdoor sunny scene is presented where a shadow is cast by a person on the background's surface which is made of grass (Figure 5.15 (b)). The surface has an highly saturated color. The color features profiles for the central point of the selected region of interest are shown in Figures 5.16–5.18.

In this outdoor sunny scene, b among normalized rgb, c_3 among $c_1c_2c_3$ and saturation are more sensitive to the passing shadow than in the previous examples. This is especially true for saturation, which, as in all the discussed examples, also in this sequence is found to vary more than other invariants. This result seems to agree with what observed by Gevers in [50], where the different invariant color features have been evaluated and compared in color object recognition experiments. Saturation is found, in fact, to provide significantly worse recognition results than the other invariant features. The fact that color features are less invariant to shadows in this example is explained by the fact that in outdoors sunny scenes, as commented in Section 3.3.1, illumination in shadows is given by the skylight which is more saturated in the blue region of the spectrum, while illumination in lit regions is also provided by sunlight.

Hue and $l_1l_2l_3$ features are analyzed in Figure 5.18. Hue results invariant to the passing shadow and shows therefore an higher robustness with respect to other features to the working hypothesis of ambient light and direct light having the same color. An higher variability can be seen in the intensity profiles of $l_1l_2l_3$, which agrees with previously observed behaviors.

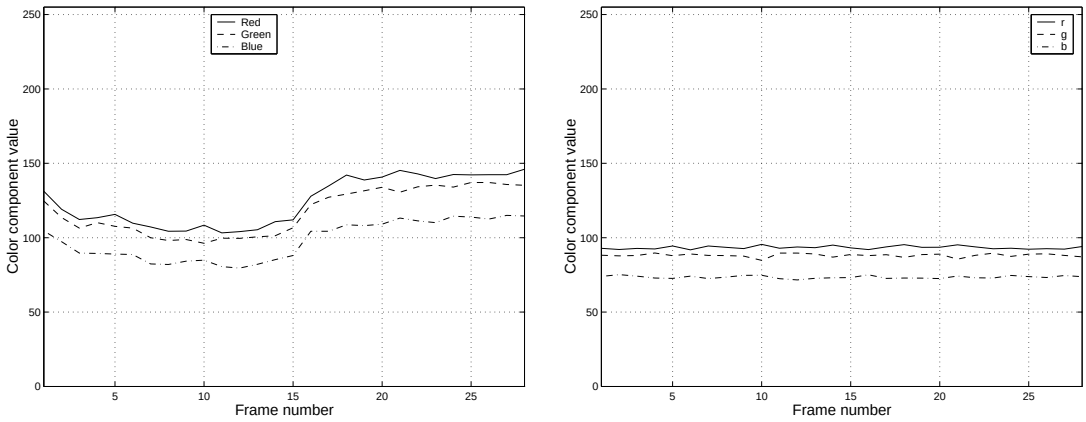

Figure 5.12: RGB and normalized rgb components intensity profile for the central point of the selected region of interest for the test sequence *Highway*.

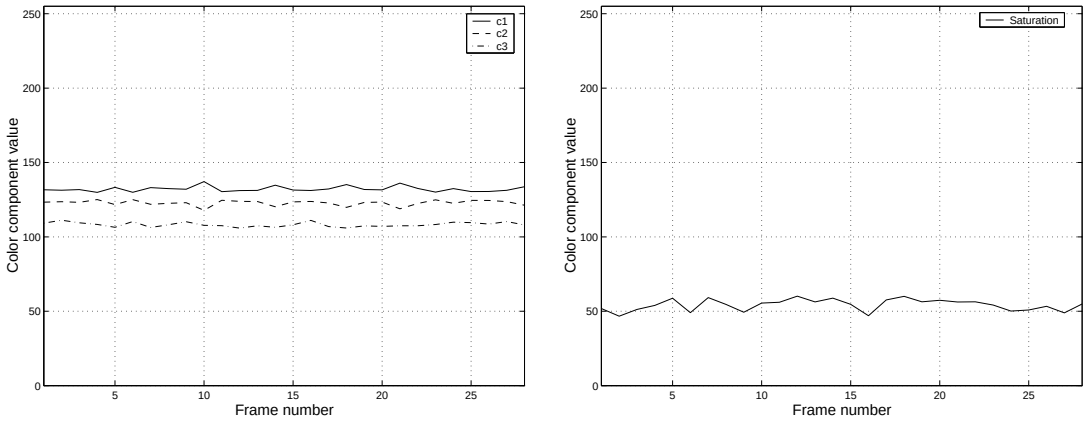

Figure 5.13: $c_1c_2c_3$ and S components intensity profile for the central point of the selected region of interest for the test sequence *Highway*.

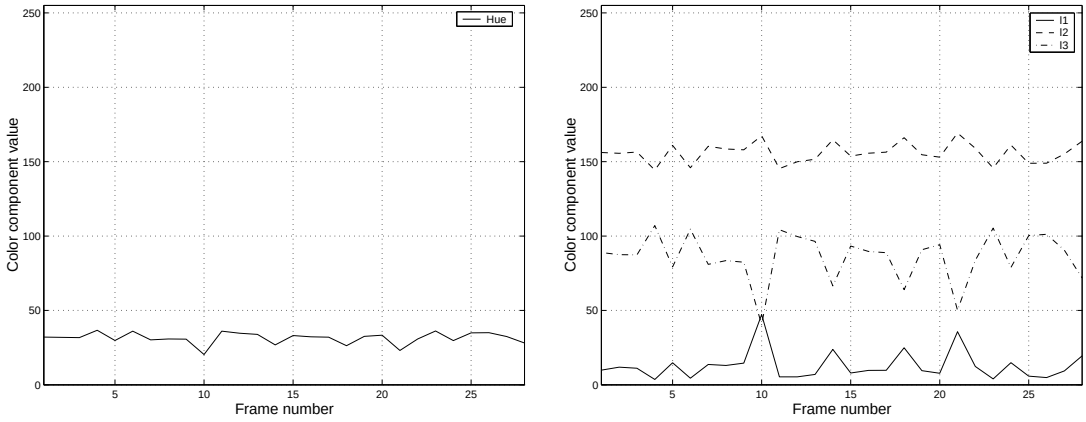

Figure 5.14: H and $l_1l_2l_3$ components intensity profile for the central point of the selected region of interest for the test sequence *Highway*.

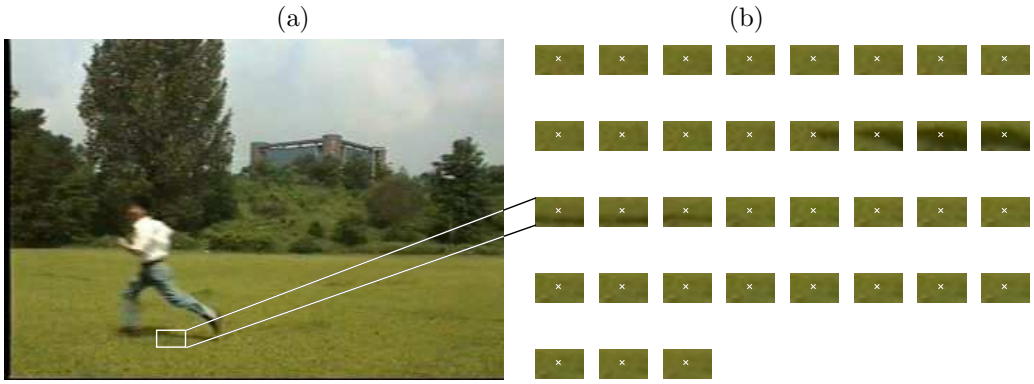

Figure 5.15: (a) Sample frame of the test sequence *Surveillance* and (b) selected region of interest over 35 frames. The central point is shown.

5.3.2 Edge maps analysis

To further investigate the problems of invariants in presence of camera noise, we have analyzed the results of an edge detection process. Edge detection will be used for the segmentation of cast shadows in still images and will be further discussed in Section 5.7. We illustrate and discuss here the results for the test image *orange*. This image has been taken with a consumer quality mono-CCD digital video recorder and is a good test bed for analyzing the behavior of color invariants in noisy conditions. The edge maps have been obtained using the Sobel edge detector (see Section 5.7.1 for the details) with thresholding parameter $\tau = 0.1$ for all components.

Since hue, unlike the other color features, is defined on a ring rather than on an interval and low values are closed to high ones, the standard difference operator used in the edge detector is not suited for computing the difference between hue values. For computing edges on hue we have therefore considered the following definition of difference, $d(x, y)$, between two angular values x and y:

$$d(x, y) = \sqrt{(\cos x - \cos y)^2 + (\sin x - \sin y)^2}, \tag{5.1}$$

yielding values in the range [0,2]. The Sobel operator is a simple operator that allows us to compare the results of the edge detection step on all the different color features. Unlike other standard operators, such as the Canny edge detector, it can, in fact, be easily modified to work with circular variables.

Computed edges for normalized color are shown in Figure 5.19. *Normalized rgb* is insensitive to shadows, but sensitive to highlights, as demonstrated in theory and in practice in the previous section. Better performance with respect to normalized color is achieved by $c_1c_2c_3$ components (Figure 5.20). Less spurious pixels due to noise inside the object are detected by the edge detection process and object edges are better defined.

The problems of the $l_1l_2l_3$ color features in regions with low saturation associated with the background region in the image can be seen in Figure 5.21. The edge maps give unsatisfactory performances in detecting object boundaries. Due to the instabilities of the $l_1l_2l_3$ transformation and its reduced discrimination accuracy, noise causes fluctuations in the values of color invariants in the background which are larger than the pixel values differences due to object boundaries. However, if we compare the edge map in Figure 5.21 with the previous ones, we can see that inside the object $l_1l_2l_3$ are much less sensitive to highlights than previously analyzed color features.

Figure 5.22 shows the computed edges in the HSI space. As it was verified for $l_1l_2l_3$, the instability

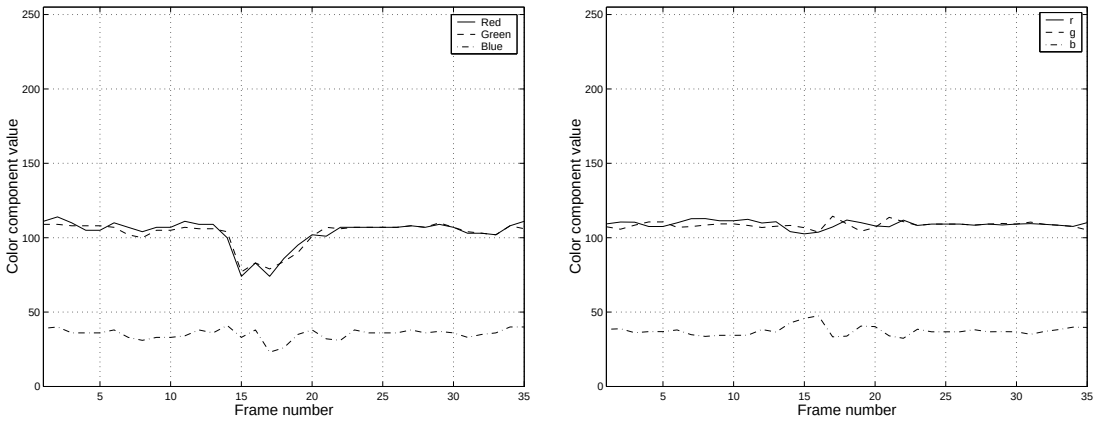

Figure 5.16: RGB and normalized rgb components intensity profile for the central point of the selected region of interest for the test sequence *Surveillance*.

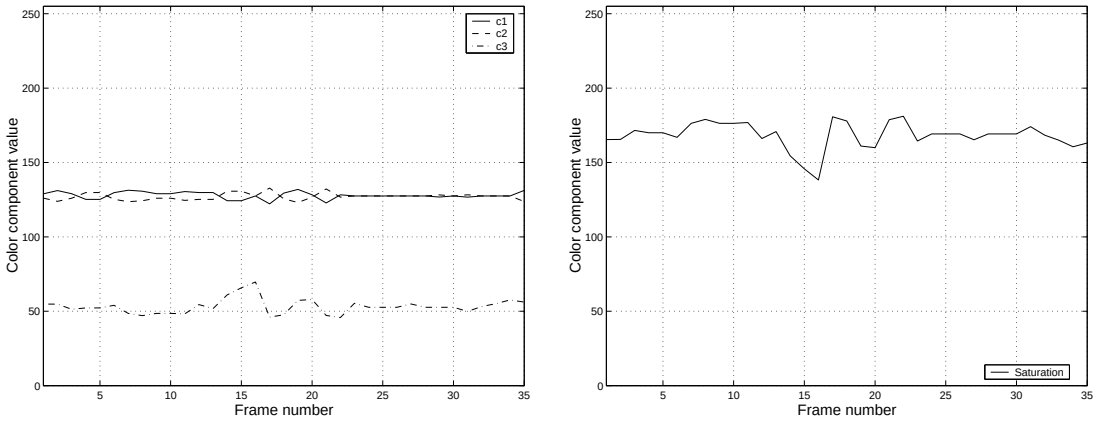

Figure 5.17: $c_1c_2c_3$ and S components intensity profile for the central point of the selected region of interest for the test sequence *Surveillance*.

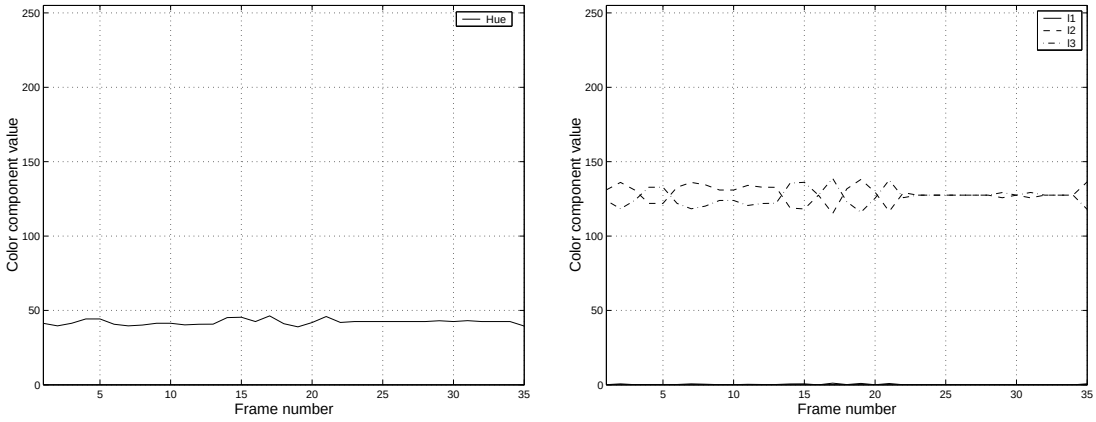

Figure 5.18: H and $l_1l_2l_3$ components intensity profile for the central point of the selected region of interest for the test sequence *Surveillance*.

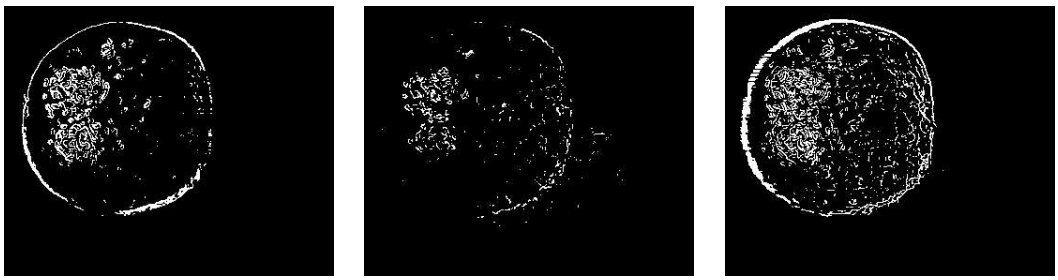

Figure 5.19: Edge maps on rgb components for image *orange*. From left to right, edges in the r, g, and b components are shown.

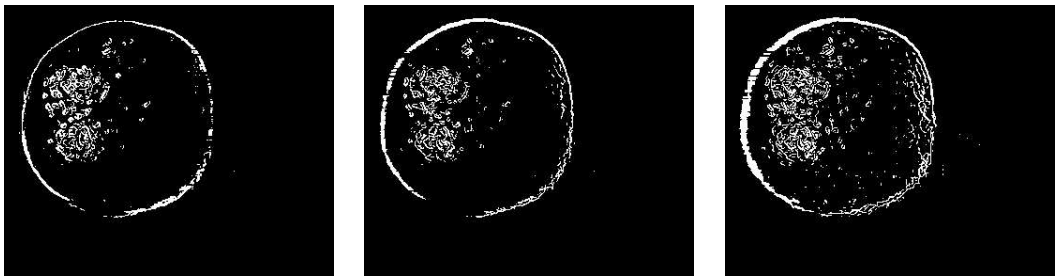

Figure 5.20: Edge maps on $c_1c_2c_3$ components for image *orange*. From left to right, edges in the c_1, c_2, and c_3 components are shown.

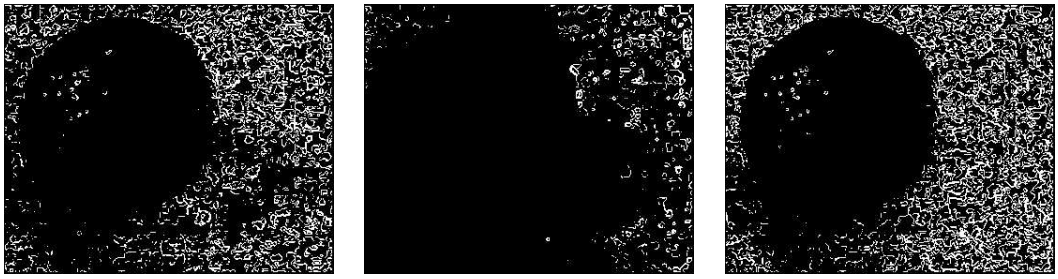

Figure 5.21: Edge maps on $l_1l_2l_3$ components for image *orange*. From left to right, edges in the l_1, l_2, and l_3 components are shown.

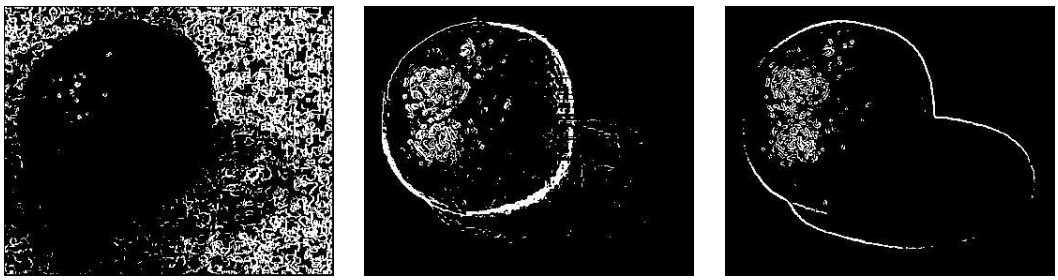

Figure 5.22: Edge maps on HSI components for image *orange*. From left to right, edges in the H, S, and I components are shown.

of H in regions of low saturation is evident from the edge maps. Its reduced discrimination accuracy is also highlighted. Saturation S does not detect shadow boundaries, yet some spurious points in the shadow region are visible. This confirms the results of the analysis presented in the previous section.

5.3.3 Discussion

To summarize the results of the proposed analysis, the following considerations can be done. *Normalized rgb* and $c_1c_2c_3$ show comparable behavior. *Saturation* behaves slightly worse with respect to the theoretical invariance. The same consideration can be done for $l_1l_2l_3$ with respect to hue. *Hue* shows, as expected, an higher invariance in presence of surfaces having a specular component. However, such advantage with respect to other features is real only in image areas which are highly saturated. In regions with neutral color and in presence of noisy conditions hue becomes unreliable. Together with its instability, another important problem of hue is its limited discrimination accuracy.

For what concerns the considered model of shadows, we have observed that the adopted physical description can represent a wide class of indoor scenes, illuminated either by a single light source or by multiple light sources, and outdoor overcast scenes, illuminated by diffuse light from the sky, for which the expected theoretical invariance holds in practice. For outdoor sunny scenes, illuminated by both the sun and the sky, the analysis confirmed the fact that the gray world assumption is less appropriate. The case of outdoor sunny images will then allow us to test the robustness of the proposed shadow segmentation method when varying the working hypotheses.

Several state of the art methods in the literature make use of saturation and hue in the detection of shadows. While hue is exploited for its invariance, the use of saturation is sometimes contradictory. Schreer et al. [142] base for instance their analysis on the empirical observation that saturation values are lowered by the presence of shadows. Risson [136], on the contrary, observes that saturation increases in outdoor sunny scenes. The discussed theoretical analysis of photometric invariants provides a more physically linked description of the different behavior of these two features which belong to the same color representation model but show a different class of invariance and consequently a different behavior. We speculate that their contradictory use is related to this fact. The above-discussed experimental analysis supports this argument by showing how using hue and saturation together can limit the effectiveness of the use of color invariance for shadow segmentation. In cases where the surface color is well saturated, in fact, the higher degree of invariance of hue could be limited by the lower invariance of saturation (this could be the case for instance in the discussed test sequence *Surveillance*). On the other hand, in cases where saturation is not affected by the presence of a shadow, the problems of hue due to noise in color deficient regions could hamper the analysis process (as for instance in the discussed image *orange*). The proposed investigation on color invariant features in real world images contributes therefore to a better use of color information for shadow segmentation.

To cope with the problematic behavior of hue and saturation, Li et al. [90] propose to reject them when detecting moving shadows in video sequences in those image regions where their values have unpredictable behavior and to limit in such regions the analysis to image brightness. Perez and Koch [122] propose a method to smooth hue values within regions of low saturation before detecting hue edges in color images. This process does not overcome however the problems related to the reduced discriminative power of hue, as shown in image *orange*. We take a different approach. Instead of limiting the use of chromatic information to highly saturated areas of the image, we propose rather to adopt features, such as *normalized rgb* and $c_1c_2c_3$, which resulted reliable in a wide range of scenes and also in color deficient regions. For these features, attention will be paid in

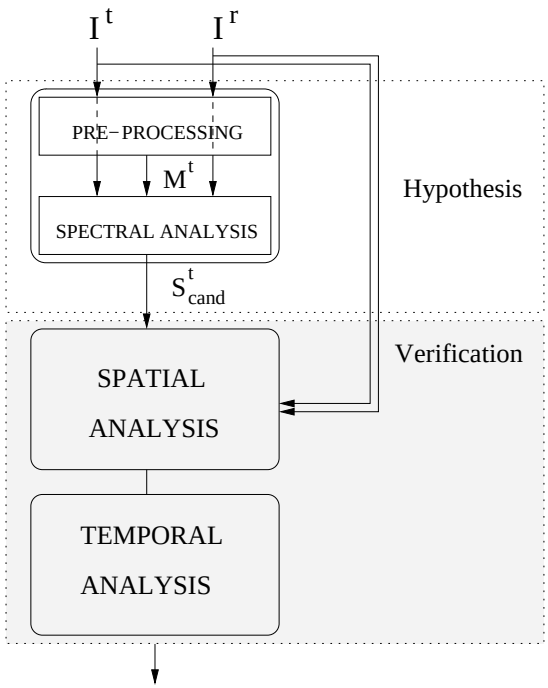

Figure 5.23: Color information allows to formulate a first shadow hypothesis for each image pixel.

regions of low intensity, where they are unstable.

It is interesting to note that also within a different research problem, that of face analysis, where color has become of standard use in face tracking to gain independence from lighting conditions, Liévin [91] arrives at similar conclusions. He observes, in fact, that in noisy conditions, given for example by the use of mono-CCD cameras, transformations such as hue provide poor results when used in skin detection. Color channel ratios, as those defined in Eq. (4.9), are reported to provide better performance. An interesting direction of investigation would be that of selecting different features for the analysis of different points in the image, according to the their color content. Hue could be used in regions of high saturation and rgb or c1c2c3 in regions of low saturation.

5.4 Color analysis

In this section, the details of the first stage of the proposed approach are presented. As commented in Section 3.3, shadows are due to a relative absence of light and their spectral analysis involves a comparison with respect to a situation where the light occlusion is not present. In an image sequence, where shadows of interest are moving shadows, by considering two different time instants it is in general possible to observe the same point in the two different illumination conditions, that is when lit and when in shadow. Two images are provided to this end as input to the system. The first image is the frame of the sequence under analysis, that is the image in which we aim at segmenting cast shadows. The second image is a reference image which represents the term of comparison.

Let us denote the current frame of the image sequence with

$$\mathbf{I}^t(x,y) = (R^t(x,y), G^t(x,y), B^t(x,y)) = (I_1^t(x,y), I_2^t(x,y), I_3^t(x,y)), \tag{5.2}$$

where R, G, B, denoted as $I_i(x,y)$, with $i = 1,2,3$, for conciseness of notation, represent the three

color channels and (x, y) indicates a generic pixel position in the 2D image plane*. When the RGB sensor responses are not readily available and the camera provides the $Y'C_bCr$ or $Y'UV$ components, then a color conversion according to Eq. (2.26) or Eq. (2.28) has to be performed. Attention has to be paid in this case to any possible sub-sampling of chroma components.

The second input to the system is given by the *reference image*. In this thesis, we consider sequences taken with a fixed camera and a static background. In this case, the reference image,

$$\mathbf{I}^r(x, y) = (I_1^r(x, y), I_2^r(x, y), I_3^r(x, y)), \tag{5.3}$$

can be either a frame in the sequence or, if such a frame is not directly available, a model resulting from a learning process [20, 28]. In the former case, it can be either the previous frame in the sequence, that is $r = t - 1$, or a background frame acquired before moving objects and moving shadows enter the field of view at a time instant $r = t_0$. If two consecutive frames are analyzed, as in [151], then regions that have been covered or uncovered by shadows from one frame to the other are detected. This means that shadows can only be completely detected if they entirely cover new background along the image sequence. Image regions that are always shadowed cannot be detected. Moreover, if shadows stop moving for a certain period of time they will be lost.

To avoid this, we adopt as reference image an image representing the static scene background which does not contain dynamic objects nor shadows due to moving objects. Static shadows, that is shadows due to static objects, such as buildings, parked cars, etc., can be present in the image. If the image is not available directly from the sequence, as commented above, it can be reconstructed by means of a learning process. A learning process allows also to cope with global illumination changes due, for instance, to changing daylight and passing clouds in outdoor scenes or artificial phenomena such as lights being switched on and off in indoor scenes. For a wide range of applications it is reasonable to assume that such reference image is available. Nevertheless, the previous frame in the sequence can be used if this is not the case. If the camera is moving, or more generally the background is moving, a global motion estimation and compensation [29] should be applied to the sequence in order to use the proposed methodology.

The first analysis stage (Figure 5.23) takes the current and the reference image as input and outputs a binary mask, $s^t_{cand}(x, y)$, containing regions of pixels which are considered potential moving cast shadow regions. These candidate regions will be then validated or discarded by the subsequent stages. The spectral properties of shadows are exploited by comparing the values of color features for each pixel in the current image with those of the corresponding pixel at the same location in the reference image. If the difference in color values is consistent with the presence of a shadow according to the assumed spectral model of shadows, the pixels are retained as candidate shadow pixels. The spectral analysis is divided into two processes. The first process exploits the property that shadows darken the surface upon which they are cast (Eq. (3.8)) and provides an initial shadow evidence. The second process considers color invariant features (Eq. (4.26)) for extracting additional evidence. The analysis is preceded by a pre-processing stage which aims at identifying those image regions where the spectral analysis may result unreliable, thus guiding the system in a reliable use of color information. The details of the implementation of the two steps are presented in the following subsections.

5.4.1 Pre-processing

The pre-processing aims at avoiding the effects of noise blow-up at unstable color invariant values. As we have amply discussed, the use of photometric invariants has a drawback related to the singularities

*Since the image is a multi-band function we represent it in bold font **I**.

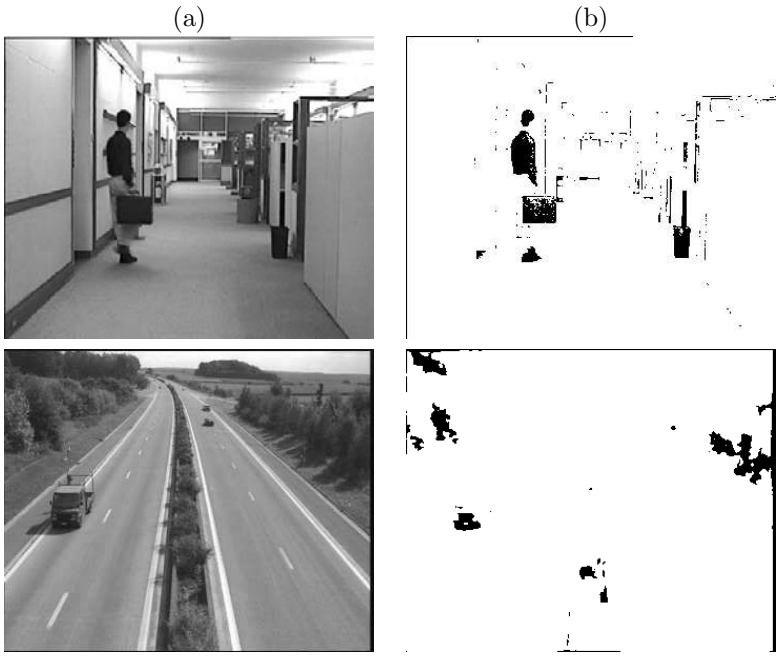

Figure 5.24: Examples of selection maps. (a) Original image and (b) corresponding selection map. The black labels identify pixels close to the black vertex of the RGB cube that are not considered in the spectral analysis stage.

in their transformations at some color values and their numerical instability in presence of noise near these singular values. The effect of noise blow-up is often ignored by shadow detection methods but it should be taken into account to avoid unreliable results.

To avoid the effects of color invariants instabilities, Otha [117] and Healey [64], which use color invariants for image segmentation, suppress unreliable values by means of thresholding. Otha, who analyzed many different color features for the purpose of color image segmentation, suggests to consider normalized rgb values only if the intensity is larger than 30 (on a range of 256 values), and rejects hue values if the saturation times $(R + G + B)$ is less than 9. The former recommendation can be then extended to $c_1c_2c_3$ features which belong to the same class of invariance as normalized rgb. The latter recommendation could be used also for $l_1l_2l_3$. Gevers [50, 51], for the purpose of color based object recognition by means of histogram matching, discards in the construction of color histograms pixels with saturation and intensity smaller than 5% of their total range. As in [117], we exclude critical pixels from the analysis by means of thresholding. Since we make use of color invariants for diffuse surfaces, as stated in Section 5.3.3, we adopt a threshold value of 50 for each of the three RGB components, that is we exclude a cubic volume close to the black vertex of the RGB cube. The threshold is based on extensive tests and is kept fixed for all our tests.

The thresholding operation is performed on both input images, $\mathbf{I}^t(x, y)$ and $\mathbf{I}^r(x, y)$, resulting in two binary masks, $m^t(x, y)$ and $m^r(x, y)$, taking a value of 1 in regions that can be further processed and a value of 0 in critical points. The two mask are then combined to form the resulting selection map $M^t(x, y)$:

$$M^t(x,y) = \begin{cases} 1 & \text{if } (m^t(x,y) = 1) \wedge (m^r(x,y) = 1) \\ 0 & \text{otherwise.} \end{cases} \tag{5.4}$$

The selection map is a binary map indicating which part of the image under analysis should be further processed by the spectral analysis stage. The points excluded by means of the selection map

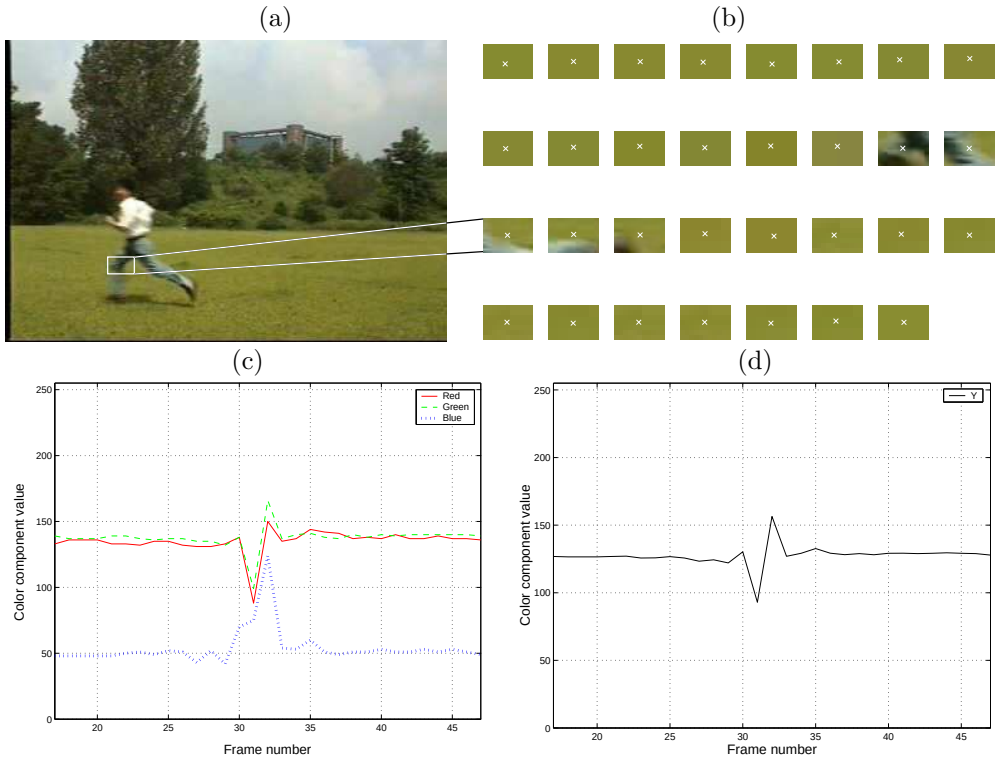

Figure 5.25: (a) Sample frame from the test sequence *Surveillance* and highlighted region of interest over 30 frames of the sequence (b). In the selected frames, the person's trousers cover the region. The mean over a 3x3 pixels square window centered in the central point of the region of interest plotted over the selected 30 frames for RGB (c) color values and Y values (d).

will be then reconsidered in the spatio-temporal verification phase. Examples of selection maps for an indoor and an outdoor image are shown in Figure 5.24. As can be seen from the images, cast shadows points do not belong to critical areas. On the contrary to what happens in the case of aerial images, we have observed in fact that moving cast shadows are generally not the darkest regions in sequences of complex real world scenes. Indoor scenes are typically illuminated by multiple light sources and shadows are therefore characterized by a strong ambient illumination contribution. In outdoor scenes, shadows still receive a good deal of light from the sky.

5.4.2 Initial evidence

The current image, the reference image and the selection map are the input data for this stage. An initial shadow evidence for pixels in the current image is obtained by analyzing RGB values.

Equation (3.8) states that each camera sensor has a lower response for a point in a shadow region with respect to the same point in light. A pixel at position (x, y) in the image under analysis $\mathbf{I}^t(x, y)$ whose values are smaller for all three color channels than those of the corresponding pixel in the reference image $\mathbf{I}^t(x, y)$ can be considered a potential moving cast shadow pixel. This results in the identification of an initial set of candidate shadow pixels

$$\mathcal{S}^t_{dark} = \{(x, y) : I^r_1(x, y) > I^t_1(x, y) \wedge I^r_2(x, y) > I^t_2(x, y) \wedge I^r_3(x, y) > I^t_3(x, y)\}. \quad (5.5)$$

As the camera output may provide only the $Y'C_bCr$ or $Y'UV$ components, the decrease of the sole luma component is typically checked in the literature [42, 127, 142] to extract potential shadow points. Alternatively, the decrease in the component corresponding to intensity in HSI-type of color spaces is analyzed [27, 136]. However, checking the property in Eq. (5.5) is not equivalent to checking the decrease of image luma or intensity. We illustrate this fact with an example. Consider a pixel whose RGB color components are $(R, G, B) = (1, 0, 1)$. The corresponding pixel's components in the reference image are $(R, G, B) = (0, 1, 0)$. The R and B components have increased their values with respect to the reference image, while the G component has decreased its value. If we compute luma by means of Eq. (2.27) for the pixel under analysis, the value we obtain is $0.299+0.114 = 0.413$, while for the pixel in the reference image it is 0.587. Even though the luma of the pixel under analysis is lower than that of the corresponding pixel in the reference image, the property in Eq. (5.5) is not satisfied. When checking the decrease of all the three channels, the increased computational complexity is compensated for by an higher accuracy.

In Figure 5.25 an example in a real image of what above stated is shown. A region of interest has been selected from the sequence *Surveillance* which is crossed by the trousers of the man running on the grass (Figure 5.25 (b)). First, a self shadowed part of the trousers is considered and then an illuminated part is shown. The behavior of RGB values for the central point of the region of interest over 30 frames is shown in Figure 5.25 (c) and compared to that of Y' in Figure 5.25 (d). While the R and G components decrease when the selected point passes from the background points to the self shadowed object region, the B component increases. Equation (3.8) is not satisfied for all the three components and the proposed analysis on RGB allows to correctly label this point. Y' decreases in the self shadowed object points and an analysis of its values only would lead to a incorrect classification of the object points as potential shadow points.

To extract candidate shadow points in $\mathcal{S}^t_{dark}$, the image difference, $\mathbf{D^t}(x, y)$, computed as

$$\mathbf{D}^t(x, y) = \mathbf{I}^r(x, y) - \mathbf{I}^t(x, y), \tag{5.6}$$

is analyzed for each point in the selection map $M^t(x, y)$, component by component. For each channel $i = 1, 2, 3$ thus the difference

$$D^t_i(x, y) = I^r_i(x, y) - I^t_i(x, y) \tag{5.7}$$

is analyzed. The three sub-analyses can be efficiently carried out in parallel. The results are then fused to obtain the final decision. In an ideal, noise-free case, the condition

$$D^t_i(x, y) = I^r_i(x, y) - I^t_i(x, y) > 0, \forall i = 1, 2, 3, \tag{5.8}$$

would suffice to state that the pixel at position (x, y) belongs to $\mathcal{S}^t_{dark}$. In real situations, the noise introduced by the acquisition process alters the above test. This noise component, the camera noise, results from the sensitivity of the sensor to temperature. The effect of camera noise is fluctuations in pixel values. These fluctuations generate values of $D^t_i(x, y)$ which are less than zero even when shadows have decreased the irradiance reaching the camera's sensors.

The effect of noise can be reduced by using a spatial support for the analysis which is larger than a single pixel. On this larger support, the average value of $D^t_i(x, y)$ is computed. The support is chosen as a window $W^{dark}_{(x,y)}$, of $q = (2N + 1)(2M + 1)$ pixels, centered at the pixel position (x, y). In $W^{dark}_{(x,y)}$, we compute the average channel difference $D^W_i(x, y)$ as

$$D^W_i(x, y) = \frac{1}{q} \sum_{i=-N}^{N} \sum_{j=-M}^{M} D^t_i(x + i, y + j). \tag{5.9}$$

The set of pixels $\mathcal{S}^t_{dark}$ is then obtained as

$$\mathcal{S}^t_{dark} = \{(x,y) : D^W_i(x,y) > b_i, \forall i = 1,2,3\}. \tag{5.10}$$

The vector $\mathbf{b} = (b_1, b_2, b_3)$ takes care of the distortions introduced by the noise. The threshold $\mathbf{b}$ can be set empirically or computed adaptively. In the former case, it will be fixed for all pixels in the image and for all the images in the sequence. In the latter case, the threshold is adapted to the image content according to some rules. The issue of adaptive threshold selection will be discussed in Section 5.5.2.

The final output of the analysis on the RGB components is a binary mask, $s^t_{dark}(x,y)$, that can be expressed as

$$s^t_{dark}(x,y) = \begin{cases} 1 & \text{if } (x,y) \in \mathcal{S}^t_{dark} \\ 0 & \text{otherwise.} \end{cases} \tag{5.11}$$

5.4.3 Additional evidence

The result of the first step of color analysis is the identification of a set of potential shadow pixels. Among potential shadow pixels also moving object pixels that are darker than the corresponding background pixels in the reference image are extracted. Moreover, erroneously detected pixels due to noise can be present in $s^t_{dark}(x,y)$. A further analysis is required to discard dark object pixels. Photometric invariant color features are then exploited to extract additional shadow evidence for image pixels.

For the analysis of invariant color features, first of all, a color transformation is used to extract from the reference image, $\mathbf{I}^r(x,y)$, and the current image, $\mathbf{I}^t(x,y)$, color invariant features. Let us denote the resulting images as $\mathbf{Inv}^t(x,y) = (Inv^t_1(x,y), Inv^t_2(x,y), Inv^t_3(x,y))$ and $\mathbf{Inv}^r(x,y) = (Inv^r_1(x,y), Inv^r_2(x,y), Inv^r_3(x,y))$.

According to Eq. (4.26), the presence of a shadow does not alter the value of the invariant color features. Invariants on the contrary change their values with changes in material properties. Photometric invariant features values for an image pixel at position (x,y) in the background image thus change their value when an object covers the background at that position in the current image. Let us then define the set of pixels $\mathcal{S}^t_{inv}$ as

$$\mathcal{S}^t_{inv} = \{(x,y) : Inv^r_1(x,y) = Inv^t_1(x,y) \wedge Inv^r_2(x,y) = Inv^t_2(x,y) \wedge Inv^r_3(x,y) = Inv^t_3(x,y)\}. \tag{5.12}$$

The set $\mathcal{S}^t_{inv}$ contains pixels for which additional shadow evidence is obtained.

The identification of pixels in $\mathcal{S}^t_{inv}$ is achieved by analyzing the absolute difference, $d^t_i(x,y)$, for each invariant feature computed as

$$d^t_i(x,y) = |Inv^r_i(x,y) - Inv^t_i(x,y)|. \tag{5.13}$$

Alternatively, the squared difference

$$d^t_i(x,y) = (Inv^r_i(x,y) - Inv^t_i(x,y))^2 \tag{5.14}$$

can be used. As for the analysis on the RGB channels, we consider a window, $W^{inv}_{(x,y)}$, centered in (x,y), and we analyze the average of differences, $d^W_i(x,y)$, given as in Eq. (5.9). Now, the set of pixels $\mathcal{S}^t_{inv}$ is obtained as

$$\mathcal{S}^t_{inv} = \{(x,y) : d^W_i(x,y) < f_i, \forall i\}, \tag{5.15}$$

where f_i takes care of the distortions introduced by noise.

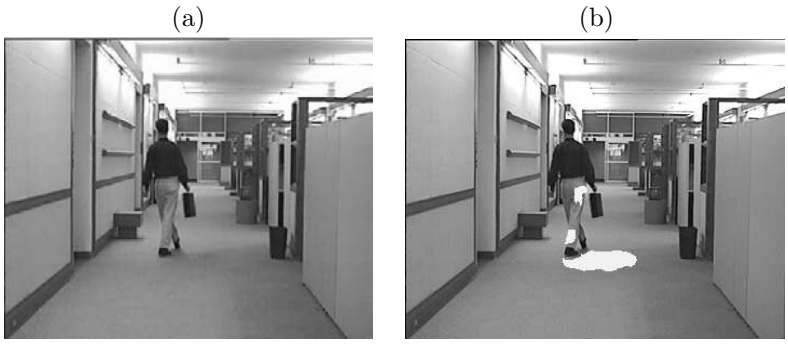

Figure 5.26: Sample result of color analysis. (a) Original image; (b) color analysis result.

On the contrary to the case of the RGB features, camera noise cannot be assumed as the only source of noise that affects the analysis of photometric invariant transformations. Camera noise is propagated through the nonlinear color conversion operations which lead to the computation of color invariants. Kender [80], in his discussion of the behavior of nonlinear color transformations, pointed out that the distribution of transformed values can show spurious modes and gaps. A deviation with respect to the theoretical invariance in real world scenes has also to be accounted for. An adaptive local thresholding is, in this case, an open problem. The setting of the threshold f_i is therefore driven by experiments on different sequences.

The final output of the analysis on the invariant features is a binary mask, $s^t_{inv}(x, y)$, that can be expressed as

$$s^t_{inv}(x,y) = \begin{cases} 1 & \text{if } (x,y) \in \mathcal{S}^t_{inv} \\ 0 & \text{otherwise.} \end{cases} \tag{5.16}$$

The shadow evidences derived by analyzing RGB color values and photometric invariant features values are finally fused. Pixels verifying the first evidence but not the second are labeled as dark object pixels. Pixels verifying both evidences are selected as candidate shadow pixels. Finally, pixels verifying the second evidence but not the first are labeled as object pixels having similar color as that of the background. From this operation a binary mask, $s^t_{cand}(x, y)$, is extracted that can be expressed as

$$s^t_{cand}(x,y) = \begin{cases} 1 & \text{if } (s^t_{dark}(x,y) = 1) \wedge (s^t_{inv}(x,y) = 1) \\ 0 & \text{otherwise,} \end{cases} \tag{5.17}$$

which contains candidate shadow regions. The candidate shadow regions are refined by eliminating small spurious blobs and by filling small holes. An example of a typical result of color analysis is shown in Figure 5.26.

5.5 Spatial analysis

Color information alone is not discriminative enough to allow for reliable shadow segmentation. In Figure 5.26, for instance, part of the trousers of the man are detected as candidate shadows. This is due to the fact that the color of the trousers and the color of the corresponding background region are similar. In addition, the detected parts of the trousers are slightly darker than the background. These image regions have therefore the same characteristics as a shadow cast on the background and are consequently detected by the color analysis stage. To improve the accuracy of the segmentation, we therefore propose to use information about the spatial nature of shadows. Moving cast shadows

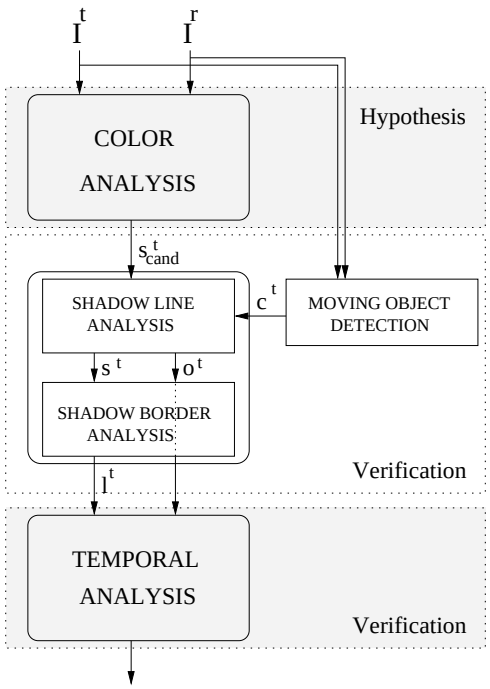

Figure 5.27: Spatial information allows to refine the initial shadow hypothesis thanks to the analysis of the relationship between shadows and shadow casting objects.

are due to the occlusion of a source of illumination by an object that is moving relative to the source. The relationship between a shadow and its shadow-casting object can be exploited for refining the results of the color analysis.

In order to analyze the spatial relationship between a cast shadow and its shadow casting object, moving objects have first of all to be identified. Many approaches have been developed for automatically detecting moving objects from image sequences. Ideally, they are expected to extract accurate object contours. Typically, as amply discussed, they extract together with object pixels also moving shadow pixels. The spatial analysis of candidate shadow regions, which include object pixels, with respect to candidate object regions, which include shadow pixels, allows then to effectively classify moving pixels in an image sequence in the two categories.

Since we consider a fixed camera and a static background, we propose to adopt to extract moving pixels a statistical model-based change detection algorithm [20] which is robust to noise and does not require fine tuning of any threshold along the sequence. This will allow us to evaluate the performance of the proposed shadow segmentation methodology independently from parameter dependencies in the object extraction algorithm.

Figure 5.27 illustrates how the change detection and the spatial analysis stage are embedded in the proposed system.

5.5.1 Moving object extraction

The goal of moving object extraction is to accurately separate moving objects from the scene's background. When the camera is fixed and the background is static, moving object extraction can be accomplished by means of change detection. Moving objects generate in fact changes in the observed image values between two different time instants. For an accurate object extraction, the

capability of detecting changes of even small entity due to objects in presence of camera noise is expected. This can be obtained by employing a statistical approach which automatically adapts to the noise level in the sequence.

The statistical-model based change detection that we consider takes as input images the same images that are the input for the shadow analysis, that is the reference image $\mathbf{I}^r(x, y)$ and the current image $\mathbf{I}^t(x, y)$. It is easily embedded in the proposed shadow segmentation algorithm. Since both objects and shadows generate temporal changes, the areas identified by means of change detection contain both moving objects pixels and shadow pixels among the ones detected as candidate shadow points. The subsequent shadow boundaries analysis will allow to distinguish shadows from moving objects and will provide as result two binary masks, one containing refined cast shadows, $s^t(x, y)$, and one containing moving objects, $o^t(x, y)$ (Figure 5.27).

In order to gain robustness with respect to noise, the change detection algorithm works according to the following principle. First of all, assumptions are made about the statistics of the noise affecting the image. Then, to evaluate the possible change at each pixel position, a distance function is calculated between the pixel in the current image and the corresponding pixel in the reference image. To improve the robustness to noise, the distance takes into account the value of other pixels in a neighborhood. The statistical properties of the distance function are then studied in order to decide, according to a statistical test, whether the pixel belongs to a changed area or to an area in the image only affected by noise. The significance level α of the test is a stable parameter and the decision threshold is automatically adapted to the noise in the image.

The analysis is performed on the RGB color components and the distance function is computed by first of all differencing the current and the reference image on each color channel separately. For simplicity of notation, we consider only one channel in the following. The results of the three analyses will be then fused to obtain the final change detection result. For each color component the distance function is then computed as

$$G_i^t(x, y) = \frac{1}{q} \sum_{i=-N}^{N} \sum_{j=-M}^{M} (I_i^t(x + i, y + j) - I_i^r(x + i, y + j))^2, \tag{5.18}$$

with $i = 1, 2, 3$. The neighborhood is chosen as a window centered in the pixel position and containing $q = (2N + 1)(2M + 1)$ pixels.

The adopted statistical model for the noise is based on the hypothesis that camera noise can be modeled as an additive random variable, $\mathbf{n}_i^t(x, y)^*$, which respects a Gaussian distribution with parameters μ_c and σ_c. It is also assumed that the noise in each color channel is spatially and temporally uncorrelated. These hypotheses are sufficiently realistic and extensively used in the literature [1, 49]. Based on these hypotheses, the observed value $I_i^t(x, y)$ of each color channel i at time instant t can be expressed as

$$I_i^t(x, y) = \hat{I}_i^t(x, y) + \mathbf{n}_i^t(x, y), \tag{5.19}$$

where $\hat{I}_i^t(x, y)$ is the true image value, that is, the value not affected by noise.

Let us suppose that there is no change in the difference image, that is $\hat{I}_i^t(x, y) = \hat{I}_i^r(x, y)$. We refer to this hypothesis as the *null hypothesis*, H_0. When H_0 is valid, the quantity

$$I_i^t(x, y) - I_i^r(x, y) = \mathbf{n}_i^t(x, y) - \mathbf{n}_i^r(x, y) = \mathbf{N}_i^t(x, y) \tag{5.20}$$

is a random variable with a Gaussian probability density function (pdf), since it is the difference of two Gaussian random variables [118]. The pdf of $\mathbf{N}_i^t(x, y)$ has mean $\mu = \mu_c - \mu_c = 0$ and variance

*We use the bold font to represent random variables.

$\sigma^2 = 2\sigma_c^2$. Given H_0, all the pixels in the considered neighborhood have changed because of noise and not because of other causes. It follows that the sum, $G_i^t(x,y)$, of the squared image difference values over the neighborhood in Eq. (5.18) becomes a random variable, $\mathbf{G}_i^t(x,y)$, described by a χ^2 distribution [171] with q degrees of freedom.

Once the distribution of the distance function $\mathbf{G}_i^t(x,y)$ has been derived, a significance test can be used to adaptively threshold the measured values of $G_i^t(x,y)$ for each pixel position and to classify them into changed and unchanged pixels. To this end, the probability of making an error when rejecting the null hypothesis if the measured distance at the pixel position is larger than a certain threshold value is computed and compared with a significance level α. The derived significance test is expressed as

$$P\{G_i^t(x,y) \geq \tau | H_0\} = \frac{\Gamma(\frac{q}{2}, \frac{\tau^2}{2\sigma^2})}{\Gamma(\frac{q}{2})} \leq \alpha. \tag{5.21}$$

Once α has been fixed, the threshold τ is automatically computed from Eq. (5.21). If the measured distance $G_i^t(x,y)$ exceeds the computed threshold, then H_0 is rejected and the pixel is labeled as changed.

The parameters of Eq. (5.21) are the number of elements q in the neighborhood, the standard deviation σ, and the significance level α. The choice of the neighborhood size q should satisfy the compromise between reliability of the statistical analysis in the neighborhood and the validity of the null hypothesis on all pixels in the neighborhood. By increasing q the statistics is more reliable and the sensitivity to noise is reduced. On the other hand, in this case, the hypothesis that all the q pixels have changed because of noise has a lower degree of confidence. This may lead to wrong detection at the border of the moving areas. A good compromise for obtaining accurate object contours is to set the value of q to 9 (N=M=1) or 25 (N=M=2).

Since it is related to the standard deviation σ_c of the camera noise, the standard deviation σ can be estimated on-line from the difference image. At the beginning of the sequence a change detection is performed on the whole image using as σ a fixed value depending on the characteristics of the acquisition system. We have adopted a value of 2. Then, the value is estimated in those areas where the hypothesis that the change is due to noise is accepted. In this way the threshold is adapted to the noise in the images.

The last parameter is the significance level, α. It represents the probability of false rejection. This makes α a stable parameter. Experimental results indicate that valid values range from 10^{-2} to 10^{-6}. The result of change detection is a binary mask, $c^t(x,y)$. An example of change detection mask is shown in Figure 5.28.

Figure 5.28: Example of change detection results. (a) Original image and (b) change detection mask. White pixels indicate changed areas.

5.5.2 Probability-based thresholding for color analysis

In Section 5.4.2, we have introduced the issue of threshold selection for binarizing the difference $D_i^W(x,y)$ in Eq. (5.10) and obtaining a binary mask of pixels got darker from the reference image to the current image. We have commented the fact that the value of the threshold $\mathbf{b}$ depends on camera noise and should be tuned for each sequence. Following the statistical approach employed by the change detection algorithm, we propose now to adopt an adaptive thresholding for the first level of spectral analysis which allows to reuse the estimation of the camera noise computed for change detection.

The reasoning is the same as the one employed for the change detection analysis. The difference lies in the fact that for color analysis we aim at determining the probability that the image difference in each color channel at a given position is *larger than zero* due to noise and not to other causes. We keep the assumptions made by the change detection and model the camera noise $\mathbf{n}_i^t(x,y)$ with a Gaussian distribution that it is spatially and temporally uncorrelated. The parameters describing the distribution are the mean μ_c and the standard deviation σ_c. The image difference, given the *null hypothesis* H_0 that only noise affects it, is again a random variable with a Gaussian probability density function which has mean $\mu = 0$ and variance $\sigma^2 = 2\sigma_c^2$. The average difference in a window $D_i^W(x,y)$ is still a Gaussian random variable with mean $\mu_D = 0$ and variance $\sigma_D^2 = \sigma^2$.

Now that we have modeled the pdf of $D_i^W(x,y)$, once the significance level α is fixed, the threshold τ_α can be determined from

$$\alpha = P\{D_i^W(x,y) \geq \tau_\alpha | H_0\} = Q(\frac{\tau_\alpha}{\sigma_D}) \tag{5.22}$$

where

$$Q(x) = \int_x^{\inf} \frac{1}{(2\pi)^{1/2}} e^{-z^2/2} dz. \tag{5.23}$$

If α is fixed to 0.05, which corresponds to an error rate of 2.5% since we are only interested in pixels who have decreased their value, τ_α is equal to $1.96 \times \sigma_D$. If α is fixed to 0.01, which corresponds to an error rate of 0.5%, τ_α is equal to $2.58 \times \sigma_D$.

Similar considerations as those done for the parameters of the change detection can be done here. The probability in Eq. (5.22) is a function of $\sigma_D^2 = \sigma^2 = 2\sigma_c^2$, that is a function of the variance of the camera noise. The estimate for σ_c^2 is readily available from the analysis of the statistical change detector. For what concerns the window size, a good compromise between robustness to noise and accuracy of detection for obtaining accurate shadow segmentation is to choose the value $q = 9$ (N=M=1) or $q = 25$ (N=M=2). Finally, a value of α in the range 0.05 to 0.01 is a valid value.

5.5.3 Shadow boundaries analysis

Once moving areas have been extracted by means of change detection, the relative position of candidate shadow regions, objects, and static background is analyzed. Moving areas may contain regions of the critical areas that were excluded by the pre-processing stage and which belong to the moving foreground. Those regions are therefore now reconsidered.

The analysis of shadow boundaries is composed by two operations. It takes as input the candidate shadows, $s_{cand}^t(x,y)$, provided by the color analysis and the change detection mask, $c^t(x,y)$. The first step checks the existence of the shadow line. This operation provides a first refinement of $s_{cand}^t(x,y)$, giving the shadow mask $s^t(x,y)$ and the object mask $o^t(x,y)$ (Figure 5.27). The object mask is refined by eliminating small spurious blobs. The second step analyzes in $s^t(x,y)$ the adjacency of candidate shadow regions boundaries with respect to the object and the background and provides

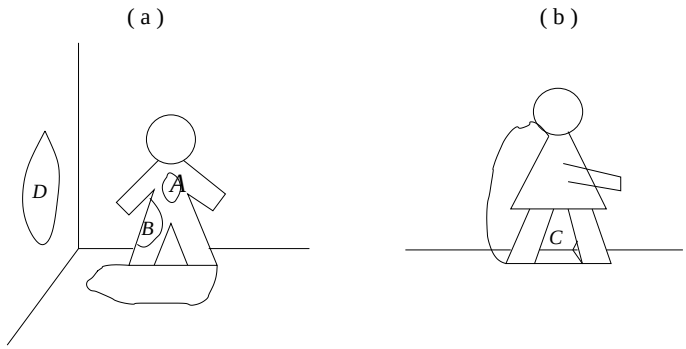

Figure 5.29: Candidate shadow regions can have different positions with respect to moving objects.

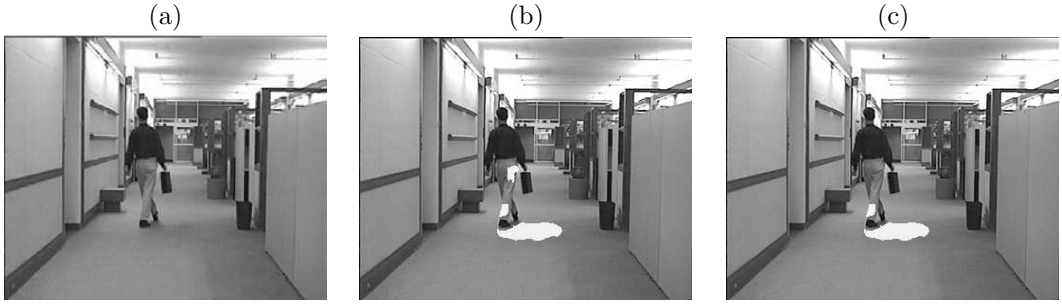

Figure 5.30: Sample result of color analysis and spatial analysis. (a) Original image; (b) color analysis only; (c) shadow line analysis result.

the mask $l^t(x, y)$ which will be processed in the temporal analysis stage. The details of the analysis are provided in the following.

Different possible cases arise with regard to the position of candidate shadow regions with respect to shadow-casting objects. They are illustrated in Figure 5.29. Cast shadows can be attached to the shadow-casting object or disconnected from it, as for shadow (D) in Figure 5.29 (a). Shadows that are attached to an object are analyzed according to the characterization of shadow boundaries discussed in Section 3.3.2. In particular, the existence of a line separating the shadow from the background, what was denoted as *shadow line* in Figure 3.8, is a necessary condition for a cast shadow. In case a candidate shadow extracted by means of the color analysis stage is fully included in an object, the shadow line is not present, and the shadow hypothesis for that region is rejected. The region is labeled as object region. This case is illustrated by shadow (A) in scene (a) of Figure 5.29. An example of the improvement obtained on the color analysis results by means of the discussed operation is illustrated in Figure 5.30. In (c) the result of the shadow line analysis is shown. The candidate shadow erroneously detected inside the object as detected by change detection has been discarded.

For the other regions that are attached to the object, the position of boundary pixels with respect to object and background pixels is analyzed. The number of boundary shadow pixels, B_b, that are adjacent to the background and the number of boundary shadow pixels, B_o, that are adjacent to the object are computed. If the majority of boundary pixels are adjacent to the object, that is if $B_o > B_b$, then a low confidence value is associated with the region under analysis. The region will be reconsidered in the temporal analysis stage.

We have found that when the above discussed condition is verified, as illustrated in Figure 5.29

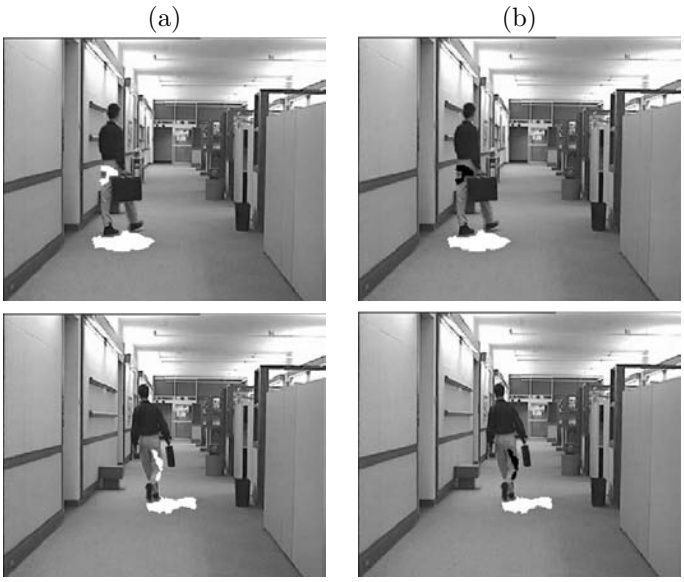

Figure 5.31: Sample results of the analysis on shadow boundaries. (a) Color analysis result; (b) shadow border analysis result. In black are displayed pixels having a low confidence level, in white pixels having a high confidence level.

(a) by means of shadow (B), the region is highly likely to belong to the object. A rejection decision for the candidate shadow under analysis cannot be however taken at this point since there are cases when shadows may share the majority of their boundary with an object boundary. This case is illustrated in Figure 5.29 (b) by means of shadow (C). Here, the shadow that is projected on the wall between the legs of the person has only a small part of its border in contact with the background. In order to deal with these cases, regions with a low confidence value are therefore reconsidered in the temporal analysis stage. The temporal reliability score will allow to take the final decision about the candidate shadow region. The combined spatio-temporal analysis allows to improve the robustness of the method.

The final result of the spatial analysis is a mask, $l^t(x, y)$, having two values: a high confidence about the considered region which is accepted as a shadow region, and a low confidence for shadows that will be accepted or rejected after temporal analysis. The mask is therefore expressed as

$$l^t(x,y) = \begin{cases} 1 & \text{if high confidence, i.e. } B_o \leq B_b \text{ for the region to which } (x,y) \text{ belongs,} \\ 0.5 & \text{if low confidence, i.e. } B_o > B_b \text{ for the region to which } (x,y) \text{ belongs,} \\ 0 & \text{otherwise.} \end{cases} \quad (5.24)$$

The object mask $o^t(x, y)$ is also provided as output.

In Figure 5.31, sample final results of spatial analysis are illustrated. In black are displayed pixels having a low confidence, in white pixels having a high confidence. These pixels are accepted as shadow pixels and do not need further verification.

5.6 Temporal analysis

The final stage of the proposed algorithm aims at providing a coherent description of the segmented shadows over time. The goal is to track shadows from frame to frame, that is to establish a

correspondence between instances of moving shadows over time. Tracking allows us to compute the life-span of each shadow. From each shadow's life-span and the relative position of objects and shadows as provided by the spatial analysis, a reliability estimation is derived. Shadows are considered reliable if they have an high confidence or if they have low confidence and significant temporal coherence. This reliability estimation is used to validate or to discard shadows detected in the previous stages.

Given the nature of shadows, shadow tracking is a difficult task. Shadows do not possess invariant shape, color, nor texture properties. These features cannot be therefore exploited for establishing a correspondence between instances of shadows over frames. The techniques proposed in the literature for tracking objects cannot therefore be directly extended to the problem of tracking shadows. A shadow tracking algorithm has to be defined based on the limited amount of information available to describe a shadow and its evolution in time.

Very little work can be found in the literature which tackles the problem of shadow tracking. Stauder [150] presents a method for detection and tracking of moving cast shadows in monocular video sequences. Temporal differences between successive frames are detected and classified into regions covered and regions uncovered by moving shadows. Entire moving shadows are tracked by temporal integration of the covered background regions while subtracting the uncovered background regions. Portions of a cast shadow that are shadowed since the beginning of the sequence are not detected by the proposed method. Scenes with only one moving object are presented and tracking is reduced to the adaptation of the 2D shape of shadows from frame to frame.

The solution we propose for tracking multiple moving cast shadows is described in Section 5.6.1. The temporal reliability estimation and the final decision on segmented shadows are then discussed in Section 5.6.2.

5.6.1 Moving cast shadows tracking

Tracking a target of interest in an image sequence means solving the correspondence problem between targets in successive frames of the sequence. In order to find the correct correspondence, it is necessary to compare the properties of the target in the current frame with those of the target in the previous frame. The comparison is not always trivial: the temporal transformation in the scene modifies, from frame to frame, the properties of the targets. Moreover, targets might cover or be covered by other targets in the scene.

In order to take into account these problems, properties that remain constant from frame to frame have to be used. For many applications in computer vision, targets of interest are video objects. Examples of such properties in the case of video object tracking are shape, color, texture, and motion. In the case of shadow tracking, shape, color, and texture properties of shadow regions change with changes in the surface upon which the shadow is cast. They cannot therefore be used for shadow tracking. For what concerns motion, the estimation and description of motion in image sequences is based on the analysis of the variations of image intensity over time [107]. In shadow regions, no reliable motion estimation can thus be performed.

Given the limited amount of information we have at our disposal for tracking shadows, we make an assumption on which we will base the proposed tracking approach. We assume that instances of the same shadow in consecutive frames overlap. This is a reasonable hypothesis for many video sequences. At each time instant, each extracted moving cast shadow is put in correspondence with previously extracted shadows. A correspondence between two shadows is established when the two corresponding regions overlap.

Based on this rule, several cases may raise as shown in the example of Figure 5.32. Let the frame at time k be composed of three extracted moving cast shadows, $S1, S2$, and $S3$. Once moving cast

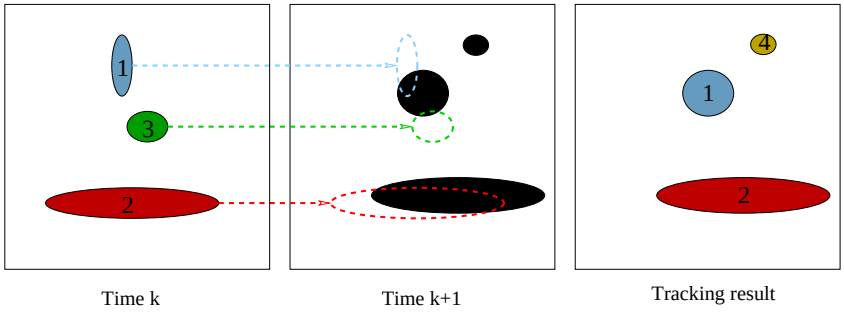

Figure 5.32: Tracking rules.

shadows are extracted at time $k+1$, their overlap with each shadow region at time k is computed. The possible configurations and the relative decisions are described below.

- If only one intersection between the regions at k and the one under analysis at $k+1$ is found (as it is for the shadow $S2$ in the example), then the corresponding track at k is continued.
- If more than one intersection is found, that is more than one track at time k has an overlap with the region at time $k+1$ under analysis (as for shadows $S1$ and $S3$) then a conflict is encountered. The conflict is solved by updating the track that has the largest intersection with the considered region (track $S1$ in the example).
- If no overlap is found, a new track is initiated in $k+1$ (as for shadow $S4$).
- If a splitting is encountered, that is when a shadow at time k is split in two disconnected regions at time $k+1$, the two regions are considered as originated by the same track if both the regions have an overlap with the shadow region at time k.
- If a merging between two shadow regions is encountered, the track corresponding to the largest intersection with the merged region is continued*.

Figure 5.33 shows a sample result of tracking for the test sequence *Hall Monitor*.

Given the mechanism described above, a *track duration* parameter is computed which defines the life-span of each shadow. The *track duration* is defined, at each frame, as the difference between the current frame number and the frame at which the track was initiated.

Shadows may be occluded by an obstacle for a certain duration. Furthermore, the algorithm for moving cast shadow extraction may fail to deliver stable results. In this second case, a previously extracted shadow is not detected in the current frame. This case is illustrated in the example reported in Figure 5.34. Here, the same shadow $S1$ is present at time $k-1$ and at time $k+1$, but not at time k. In this case, the above described mechanism would terminate a track at time k and initiate a new track at time $k+1$.

To avoid this kind of error and deal with the occlusion problem, a *track absence* parameter is defined. The *track absence* parameter is defined as the number of consecutive frames where the track

*Splitting and merging of video objects due to occlusions, collisions, and other interactions between objects, and due to errors in the object extraction stage, are the main obstacles to effective tracking of video objects. In the case of shadows, the correspondence between splitting and merging of segmented shadow regions in the image sequence and related events in the physical world is not well defined. It raises, in fact, a philosophical question: can two shadows occlude each other? Are the different shadows cast by an object a single entity or not? These issues make the definition of a shadow tracking methodology difficult.

Figure 5.33: Sample result of tracking for frames 59, 60, and 61 of the test sequence *Hall Monitor*. Different tracks are displayed with different colors.

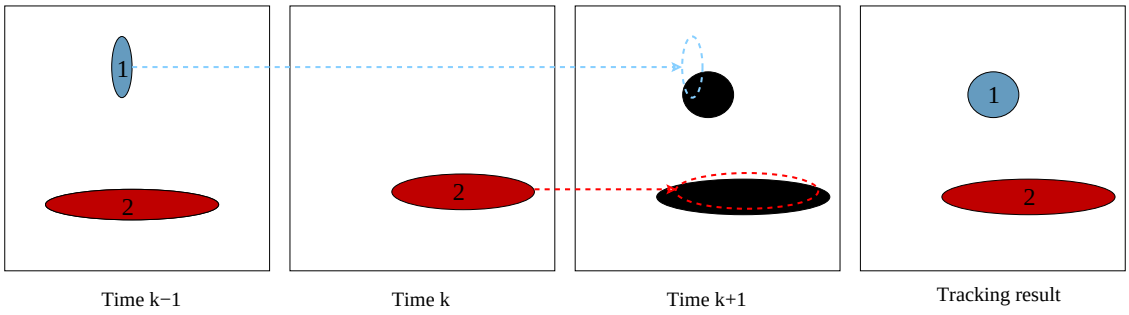

Figure 5.34: Tracking continuation example.

is not associated with any of the extracted shadow regions. Since the reason for the disappearance of the shadow is unknown, the decision to terminate the corresponding track is delayed. If the track absence exceeds a certain period of time, *absence threshold* AT, then the track is deleted.

During the absence of corresponding shadows, the missing shadow region is assumed to maintain its area and position. An alternative approach would be to predict the position of the missing shadow region by projecting its area based on its trajectory computed from previous time instants. Solving the correspondence problem between shadows defines in fact shadow trajectories. Trajectories could be used to predict future positions of shadows in the sequence. In case the description of the shadow's trajectory is not accurate, however, this solution would introduce errors in the tracking process and turn out to be of more nuisance than benefit.

Figure 5.35: Sample result of track continuation for frames 20, 21, and 22 in the test sequence *Hall Monitor*. Different tracks are displayed with different colors.

The overlap analysis is applied in the considered case as previously described. If a temporary

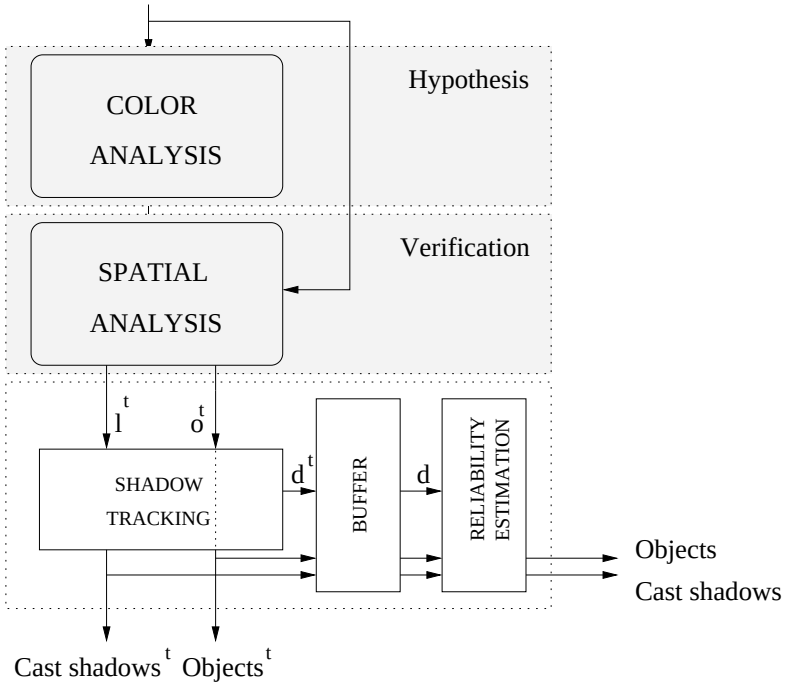

Figure 5.36: Shadow tracking establishes a correspondence between instances of shadows at different time instants. The reliability estimation of candidate shadows tracked over time allows to refine the shadow and object segmentation results over the entire sequence.

occlusion (limited in time) causes the disappearance of the shadow and the track absence is less than the absence threshold, then the reappeared shadow is associated with the correct track. An example of track continuation for the test sequence *Hall Monitor* is illustrated in Figure 5.35. If the occluded shadow region changes its dynamic behavior during the occlusion and no intersection is found when the shadow reappears, a new track is created when the shadow reappears. This error will have an influence in the computation of the *track duration* of the shadow under analysis.

5.6.2 Temporal reliability estimation

During the tracking stage, the identity j and duration d_j^t of each target shadow is stored for each frame (Figure 5.36). Once the complete sequence has been analyzed, the overall duration d_j in the sequence is computed for each shadow. These parameters are processed off-line to improve the final segmentation results for the entire sequence.

We observed that short-lived shadows are very likely to be due to failure of the shadow segmentation algorithm. Therefore, a temporal filtering of shadow segmentation results is performed to eliminate all shadows whose duration is smaller than a certain threshold, DT. The temporal filtering takes into account the results of the spatial analysis. It is in fact applied selectively on those shadow regions that had been labeled in the spatial analysis stage as having a low confidence shadow score. An example of the improvement obtained on the shadow segmentation results by means of temporal analysis is illustrated in Figure 5.37. The candidate shadow on the person's leg had a short life-span and has been eliminated by temporal filtering.

Two parameters have to be set in the temporal analysis stage. The first parameter is the absence threshold, AT. We adopted a value of AT of 2 frames for all our tests. It represents a

Figure 5.37: Sample result of color analysis and spatio-temporal verification. (a) Color analysis only; (b) spatial analysis; (c) spatio-temporal verification.

good compromise for dealing with occlusions and segmentation errors without incorrectly continuing tracks that have disappeared. The second parameter is the temporal filtering threshold, DT. The value of DT depends on the behavior of moving objects casting shadows in the image sequence. If objects are stationary for a long number of frames, the value of DT increases. If objects change their position rapidly or stay in the sequence for a short period of time, the value of DT decreases. DT has been empirically determined according to the content of the sequence.

An extensive discussion and comparison of moving cast shadow segmentation results is presented in the next chapter. In the next section, the application of the proposed methodology to still color images is described.

5.7 Cast shadow segmentation in still images

In this section and in the following one, we describe the solutions that we have developed for analyzing shadows in still color images by employing the same methodology that has been described for image sequences. The segmentation problem in this case becomes inherently more difficult.

As stated in Section 5.2, while for dynamic shadows in image sequences the observation of the same point in the two different illumination conditions, that is when lit and when in shadow, is generally possible because of the temporal dimension, this is not the case in a still color image. However, the analysis on shadows spectral properties presented in Section 3.3.1 is still valid when the comparison between color values is done between two different points on the same surface which are close enough, so that it is reasonable to assume that the imaging geometry and the ambient illumination do not change significantly from one point to the other. This is the case for points on either side of a shadow boundary. To exploit shadows spectral properties we have then applied the proposed methodology for cast shadow segmentation to the case of still images through the extraction and analysis of contours in the image. In this case, thus, the comparison for the spectral analysis of shadows is done between image pixels (x, y) and reference pixels (x_r, y_r) defined as the neighbors of the pixels under analysis. This can be expressed as $(x_r, y_r) = (x + \delta, y + \gamma)$, with δ and $\gamma \in \{0, 1, -1\}$, where δ and γ are not simultaneously equal to zero, i.e. $(x_r, y_r) \neq (x, y)$.

The analysis is organized in this case in two stages, a color analysis stage and a spatial analysis stage, which use the same principles as those used for image sequences. The two stages are described in detail in the following subsections.

5.7.1 Color analysis

As for image sequences, the analysis on color components is divided into two processes: an analysis of RGB components and an analysis of color invariants. In this case, some assumptions on the scene and the lighting are considered, namely shadows are assumed to be cast on a uniform background surface and the direct light source whose light is occluded is assumed to be strong enough or close enough to the scene so that shadow boundaries are well visible.

Initial evidence

Similarly to what was done for image sequences, the first level of the proposed strategy makes use of the property that shadows darken the surface upon which they are cast. This results in the identification of the set of pixels

$$\mathcal{S}_{dark} = \{(x,y) : I_1(x_r,y_r) > I_1(x,y), I_2(x_r,y_r) > I_2(x,y), I_3(x_r,y_r) > I_3(x,y)\}. \tag{5.25}$$

To extract the candidate shadow points in $\mathcal{S}_{dark}$, edges are first extracted from the image $\mathbf{I}(x,y) = (R(x,y), G(x,y), B(x,y)) = (I_1(x,y), I_2(x,y), I_3(x,y))$, then the property described in Eq. (5.25) is tested on the edge points.

Edges in color images can be detected in several ways [85, 129, 135, 178]. Various approaches have been proposed, including techniques extended from monochrome edge detection as well as vector space approaches, techniques based on vector order statistic operators and difference vector operators [178]. Variations to the different approaches have been introduced to improve performance in presence of noise with added algorithmic complexity.

The most simple approaches to color edge detection represent extensions from monochrome edge detection. These techniques are applied to the color channels independently and the results are fused using certain logical operators. A color edge can be considered present if an edge exists in any of the color components. Alternatively, a color edge can be considered present if the sum or the maximum of the gradients of the three color components, or the magnitude of the vector sum of the gradients of the three color components exceeds a certain threshold. Operating separately on each color channel has the advantage of reducing the complexity and of speeding up the computations by parallel processing. On the other hand, it does not take into account the correlation among color channels and, as a result, it tends to miss edges that have the same strength but in opposite direction in two of their color components. This feature does not represent a problem when edge detection is used to extract shadow edges. According to Eq. (3.8), in fact, RGB values have the same direction of change at a shadow boundary. We adopt therefore this strategy for color edge detection.

One of the representative classes of edge detectors is the Sobel operator. Having verified that it provided satisfactory results in our experiments, it has been chosen for its simplicity. The Sobel operator, a differential method based on first order derivatives, is realized by convolving the image with the following two kernel masks

$$H_h = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 1 \\ -1 & 0 & 1 \end{pmatrix}, \quad H_v = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & 1 \end{pmatrix}. \tag{5.26}$$

It produces a differential image in the horizontal and vertical direction with accentuated spatial amplitude changes. The result of the convolution of each color component image with the horizontal and vertical impulse response arrays of Eq. (5.26) are the horizontal and vertical gradients, $G_i^h(x,y)$

and $G_i^v(x,y)$ respectively, with $i = 1,2,3$. The gradient magnitude

$$G_i(x,y) = \sqrt{G_i^h(x,y)^2 + G_i^v(x,y)^2} \tag{5.27}$$

allows to define the set of edge pixels, $\mathcal{E}_i$, as

$$\mathcal{E}_i = \{(x,y) : G_i(x,y) > \tau \wedge G_i(x,y) \; local \; maximum\} \tag{5.28}$$

The threshold τ aims at eliminating noise-induced false edges. The value of τ depends on the noise level in the image and determines the sensitivity of the edge detector. The choice of the threshold in the edge detection process depends on the image characteristics and is discussed in Section 6.3. The final edge map results from a logical OR-connection operation on the three edge maps corresponding to the three color channels. The logical OR edge map tends to produce thick edges which is good for our purposes, since we aim at having as much as possible closed contours. In our technique we

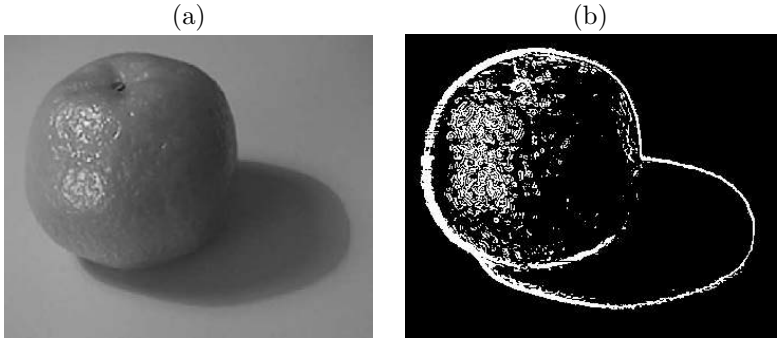

Figure 5.38: (a) Test image *orange*; (b) candidate shadow points belonging to the color edge map of the RGB image and verifying the property in Eq. (5.25).

consider that shadow contours appear in the edge maps since we assume that the direct light is strong enough to generate shadow boundaries which are not very diffuse. More sophisticated approaches to edge detection, such as Zhang's approach [176], may be more appropriate if the considered condition is not verified.

Once edges have been extracted, the property described in Eq. (5.25) is tested by analyzing the gradient image on the edges. An edge point (x,y) becomes a candidate shadow contour point, that is $(x,y) \in \mathcal{S}_{dark}$, if the gradient has the same orientation in all the three components. This is verified by analyzing the coherence of the signs of the horizontal and vertical gradients for the three color channels. A sample result of this analysis is illustrated in Figure 5.38 (b).

Additional evidence

The result of the first level of analysis is the identification of a set of candidate shadow contour pixels. This analysis leads to the detection of shadow pixels but also of object pixels. According to Eq. (4.26), contours in the color invariant images will correspond to material changes and not to shadow boundaries. This fact is exploited in the second phase of color analysis.

Color edge detection is now performed in the invariant space. A morphological dilation operation is applied on the invariant feature edge map to improve the delineation of contours. Then, isolated spurious edge pixels are removed so as to obtain the final edge map, $e^{inv}(x,y)$ (Figure 5.39 (b)). If we define $\mathcal{S}_{inv}$ as

$$\mathcal{S}_{inv} = \{(x,y) : e^{inv}(x,y)) = 0\}, \tag{5.29}$$

then the shadow hypothesis is strengthened for the set of pixels

$$\mathcal{S} = \mathcal{S}_{dark} \cap \mathcal{S}_{inv} \tag{5.30}$$

where object edges belonging to $\mathcal{S}_{dark}$ have been discarded (Figure 5.39 (c)).

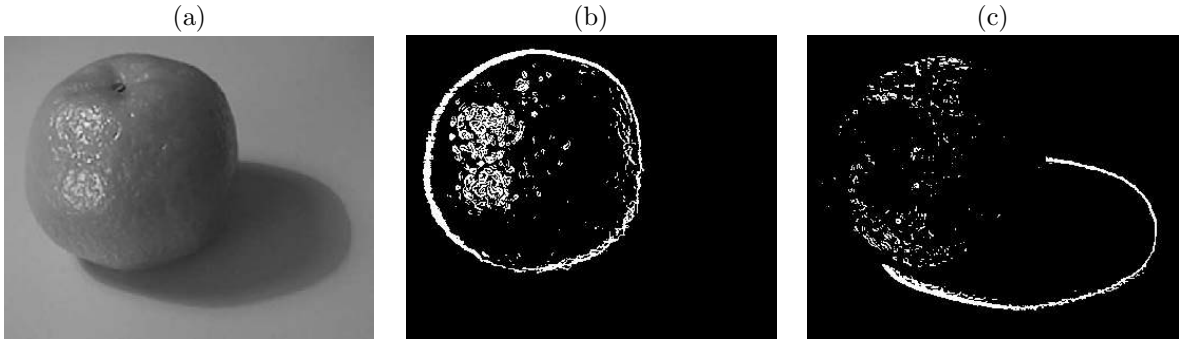

Figure 5.39: (a) Test image *orange*; (b) color edge map of the invariant features containing material boundaries for which the shadow hypothesis is weakened; (c) integration of the shadow evidence from the two color analysis steps (Eq. (5.30)).

The shadow points which form the border between shadowed background and object cannot however be found by means of this analysis. This is clear from Figure 5.39 (c). These points belong to the occluding line, CD (Figure 3.8). The occluding line does not indeed belong to $\mathcal{S}_{inv}$ since it represents a material change. In real images, moreover, $\mathcal{S}$ contains misclassified pixels due to sensor noise and approximations in the model underlying invariance. Geometrical information will be used therefore in the next analysis stage to reduce the misclassifications in $\mathcal{S}$ and to extract the missing parts of the shadow contour.

5.7.2 Spatial analysis

When dealing with image sequences, a change detection algorithm has been used to extract moving foreground regions containing objects and shadows. Then, the position of the candidate shadows with respect to objects has been analyzed. In that case, shadow casting objects have been automatically extracted by exploiting motion information. In still images, the problem of extracting shadow casting objects requires providing to the system a clear definition of what the objects of interest are. This can be obtained by means of user intervention or by exploiting specific a priori knowledge about the objects or the scene. Dealing with objects of different nature and shape, for which it is not possible to define a common model, we have considered some restrictions on the scene's layout, as done in [79], which allow us to proceed automatically. We assume that objects are imaged against a simple background where the shadows are cast and are visible closed to the shadow they cast. Under these assumptions, we exploit geometric shadow properties related to shadow boundaries and to the adjacency of each object and its cast shadow. To extend the method's applicability to scenes with complex backgrounds, color segmentation and the intervention of an user which provides the semantics, that is the meaning, of the objects of interest could be considered.

Similarly to what has been done in the analysis on image sequences, the existence of the shadow line, DE, and the hidden shadow line, CE (Figure 3.8) are checked. This is done by extracting segments in $\mathcal{S}$ and rejecting isolated and disconnected pixels, thus obtaining the subset $\mathcal{S}'$ (Figure 5.40 (b)). To this end, isolated groups of pixels are eliminated after connected component analysis. This

decision is based on a threshold whose value is set to 30% of the number of pixels of the largest connected component in $\mathcal{S}$. This value has been determined by means of extensive tests. Since it is relative to the largest component, it is adapted to the image content and does not require content-dependent setting.

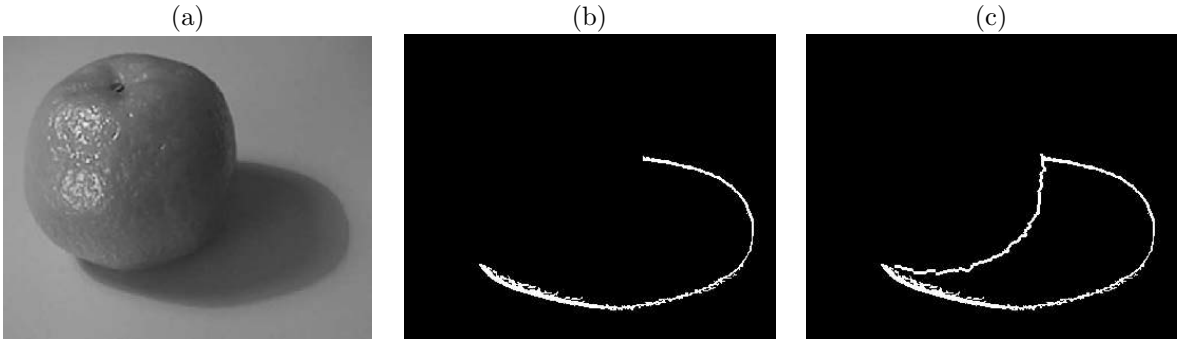

Figure 5.40: (a) Test image *orange*; (b) result of geometric analysis providing the shadow line and hidden shadow line, and (c) the complete shadow contours.

To extract the missing part of the shadow contour, the definition of occluding line, CD, is finally exploited. First, the contact points between shadow contour and object contour $e^{inv}(x, y)$ are detected. A contour dilation operation is applied to the shadow contour in order to more effectively extract contact points. Then, the position of the shadow with respect to the line that connects the two points is computed and the occluding line is extracted from $e^{inv}(x, y)$, giving the complete shadow contour (Figure 5.40 (c)). Finally, the shadow area is obtained by filling each closed shadow contour.

In the above discussion, we have considered one object only. In the case of a scene composed by multiple objects, a connected component labeling on the invariant features edge map is first of all performed and the analysis is then applied to each single object separately. If shadows do not completely lie within the image, image border pixels should be considered to close the shadow contour. Moreover, if objects and shadows occlude each other, boundary object pixels for both objects involved in the occlusion should be analyzed.

5.7.3 Cast shadow segmentation by color edge filling

A simplification of the cast shadow segmentation process has been investigated and is described in this section. A similar color analysis stage is followed by a simpler spatial processing under the assumption that objects and shadows lie within the image.

Edges are first of all extracted, as in the color analysis stage, on RGB images and photometric invariant features images. To improve edge delineation, a post-processing is applied on the edge maps. To this end, the same processing that was used as first operation in the spatial analysis stage is exploited here to eliminate isolated noise-induced edge pixels. In this case, the threshold has been fixed to 10% of the number of pixels in the largest connected component since we aim at having as much as possible close contours. A morphological dilation operation is also applied to this end (Figure 5.41).

In the second level of the analysis, the RGB edges are filled in order to obtain a binary mask, $a(x, y)$, that represents object and shadow regions in the image. The invariant features edges are filled in order to obtain a binary mask, $o(x, y)$, representing only objects in the scene.

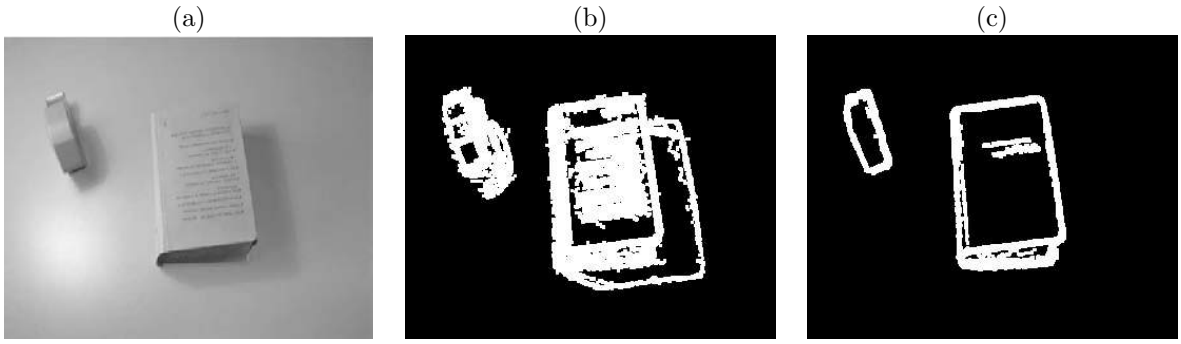

Figure 5.41: Example of edge detection followed by post-processing on RGB (b) and photometric invariants (c).

The filling is obtained by first of all performing an horizontal and vertical scanning of the edge map and setting to 1 all pixels within the most left/high edge point to the most right/low edge point. Then, additional scanning processes are performed to eliminate background regions between objects and shadows by shrinking the filled regions toward the edge map.

Shadow pixels $s(x, y)$ are then extracted as those pixels which belong to the first binary mask and do not belong to the second mask, that is

$$s(x,y) = \begin{cases} 1 & \text{if } (a(x,y) = 1) \wedge (o(x,y) = 0) \\ 0 & \text{otherwise.} \end{cases} \tag{5.31}$$

Finally, a morphological post-processing (an erosion followed by a dilation operation) is applied to the final mask to refine the results. In this case, the spatial analysis can be applied simultaneously to multiple objects in the scene and does not require each object to be selected and analyzed separately. The results of the tests of the proposed methods will be discussed in Chapter 6.

5.8 Summary

In this chapter we described an efficient methodology for segmenting cast shadows in still images and image sequences. The proposed strategy is based on a bottom-up approach composed of three successive levels: the color analysis, the spatial analysis, and the temporal analysis. This last stage is not present when the approach is applied on still images.

An initial shadow hypothesis is tested by exploiting spectral properties of shadows by means of color information. The property that shadows darken the surface upon which they are cast is first of all checked. Each camera sensor must have a lower response for a point in shadow with respect to the same point in light. Photometric color invariants are then analyzed. They must not be affected by the presence of a shadow.

As a preliminary step for the spectral analysis, an evaluation of different color invariant features was performed on a number of real images. The results of the analysis outlined the fact the adopted physical description of shadows can represent a wide class of indoor scenes and outdoor overcast scenes. It is less appropriate, as expected, for outdoor sunny scenes. This case allows to test the robustness of the proposed method when varying the working hypotheses. The analysis showed moreover the different behavior of color invariants with changes in the content of the considered scenes. In particular, hue, which is widely used in literature, resulted unreliable in color deficient

scenes in presence of noisy conditions. Saturation behaved slightly worse with respect to normalized rgb and $c_1c_2c_3$, which were therefore selected for the analysis. The evaluation completes the analysis of color invariant features for shadow segmentation that was initiated in Chapter 4.

The initial shadow hypothesis provided by the color analysis stage is verified by exploiting spatial and, in the case of image sequences, temporal properties of shadows. The position of a candidate shadow region with respect to the shadow casting object is considered for the spatial analysis, which does not require any knowledge of the structure of the object or of the scene. Then, the temporal behavior of shadows is analyzed.

The temporal verification exploits a tracking strategy that has been defined on the basis of the limited amount of information available to describe a shadow and its evolution over time. Tracking allows to compute the life-span of each shadow. From each shadow's life span and the relative position of objects and shadows provided by the spatial analysis stage, a reliability estimation is derived. This reliability estimation is used to validate or to discard each shadow detected in the previous levels of analysis and provides the final segmentation results.

In the case of still images, the proposed methodology for cast shadow segmentation was applied through the extraction and analysis of contours in the image. The analysis is organized in this case in two stages, a color analysis stage and a spatial analysis stage, which use the same principles as those used for image sequences. Some assumptions on the scene and the lighting are considered, as the segmentation problem in this case becomes inherently more difficult. The performance of the proposed technique is evaluated in the next chapter.

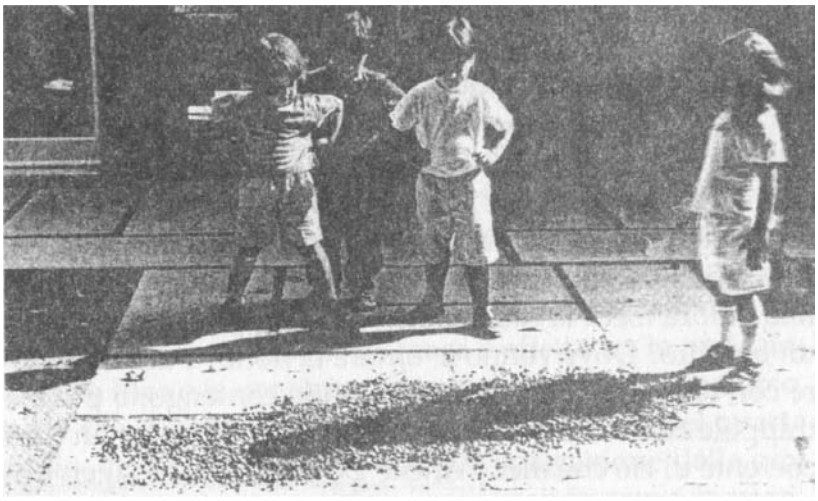

Figure 5.42: Shadows are created by the night (Section A.2.1).

6 Performance evaluation

6.1 Introduction

In the previous chapter, a methodology for the segmentation of cast shadows in video sequences and still images has been proposed. In this chapter, the performance of the proposed method is evaluated.

To this end, the results of the tests of the proposed method are presented and analyzed. Subjective assessment of the segmentation accuracy, objective evaluation with respect to a ground-truth segmentation and comparison with state of the art techniques are introduced.

Tests on image sequences are first discussed in Section 6.2. Still images are then considered in Section 6.3.

6.2 Results on image sequences

The results of cast shadow segmentation and tracking in image sequences are organized as follows. The performance of the first two blocks of the system, that is the color analysis and the spatial analysis (see Figure 5.1), is first discussed in Section 6.2.1 and Section 6.2.2. These two stages provide an on-line segmentation. The results are evaluated subjectively, objectively, and compared to state of the art methods.

Then, in Section 6.2.3, the improvements introduced by the off-line temporal verification stage are discussed by means of subjective and objective comparison with the results of the first part of the segmentation algorithm.

In the tests, sequences from the MPEG-4 and MPEG-7 Content Video Set are used, as well as test sequences from the test set of the ATON project* and the European project *art.live*†. The sequences are in CIF format (288×352 pixels), unless otherwise stated in the remainder of the section.

*http://cvrr.ucsd.edu/aton/

†European project IST 10942 *art.live* (Architecture and authoring Tools for prototype for Living Images and new Video Experiments), http://www.tele.ucl.ac.be/PROJECTS/art.live/

6.2.1 Segmentation results

Figures 6.1 (b)– 6.4 (b) show the shadow masks obtained by color analysis (Section 5.4) followed by spatial verification (Section 5.5) for two indoor sequences, one outdoor sequence where the sky is overcast, and two outdoor sunny scenes. The detected shadows are superimposed over the original image and color-coded in white.

The reference image is the first frame of the sequence, acquired before the objects enter in the field of view, except for the test sequence *Surveillance*. In this case, frame 210 is chosen which does not contain moving objects nor shadows due to moving objects. The parameters to choose for this first part of the algorithm are the size of the observation windows for the initial evidence, W^{dark} (Section 5.4.2 and Section 5.5.2), and for the additional evidence, W^{inv} (Section 5.4.3), and the value of the threshold f_i for the photometric invariant color features test (Section 5.4.3). The $c_1c_2c_3$ features have been used in our tests. The values of the above-mentioned parameters are the same for the indoor and outdoor overcast scenes and they are the result of an extensive analysis: W^{dark} is 5×5 pixels, W^{inv} is 7×7 pixels, and $f_i = 7$ for all components. For the outdoor sunny scenes, as predicted by the analysis on color invariants presented in Section 5.3, an higher value of f_i is required to cope with the fact that the assumed gray world condition is not verified. A value of f_i of 18 is therefore adopted.

The segmentation results for four sample frames of the test sequence *Hall Monitor* are shown in Figure 6.1 (b). This sequence represents a typical indoor surveillance sequence. The method correctly identifies the shadows which moving objects cast on the floor and on walls. In the second image it is possible to notice an error due to the fact that the color of the trousers of the man and the color of the corresponding background region are similar. In addition, the trousers are slightly darker than the background. The spatial analysis stage does not succeed in eliminating the candidate shadow, because of the existence of the shadow line. The temporal verification stage will overcome this problem.

A different scenario is depicted in Figure 6.2. People walking in a room cast several shadows which are caused by their interaction with multiple light sources. This is a scene representing a typical environment for interaction purposes, where users act in front of a display. For good visibility of the display the interaction area has to be rather dark or spot lights have to be used, which cause significant shadows. Objects are large and close to the camera. In this scene, a model-based method for shadow segmentation would fail due to the complexity of the scene. The proposed method is based on shadow properties and therefore it can be applied to complex scenes, when shadows and objects occlude each other.

An outdoor scene is depicted in Figure 6.3. Vehicles of different dimensions are running on a highway. Lighting conditions are different when compared to the previous indoor sequences. Shadows are very weak due to the diffuse illumination coming from the overcast sky. Despite the difficulty in recognizing them when looking at the images of Figure 6.3 (a), shadows generate changes in image values that can mislead motion detection algorithm and consequently moving object detection results. Also in this case shadows are correctly extracted. Misclassified pixels can be noted on the truck's and vehicles' wheels which have a gray color that is very similar to that of the road's asphalt. A post-processing based on edge detection could be applied to refine the segmentation. The results demonstrate that the proposed method can be applied on a large class of scenes, without changing the values of the parameters. In the second image, it can be noted that the shadow that the truck casts on the road in front of the white car has been correctly detected. In case the shadow segmentation algorithm is applied as a post-processing stage to improve the results of a video object extraction algorithm, this detection is an important result. The two vehicles are, in fact, very likely to be extracted as an unique object by video object segmentation techniques.

Figure 6.1: Shadow segmentation results for the test sequence *Hall Monitor*. (a) Original image; (b) shadow mask (white pixels) superimposed on the original image.

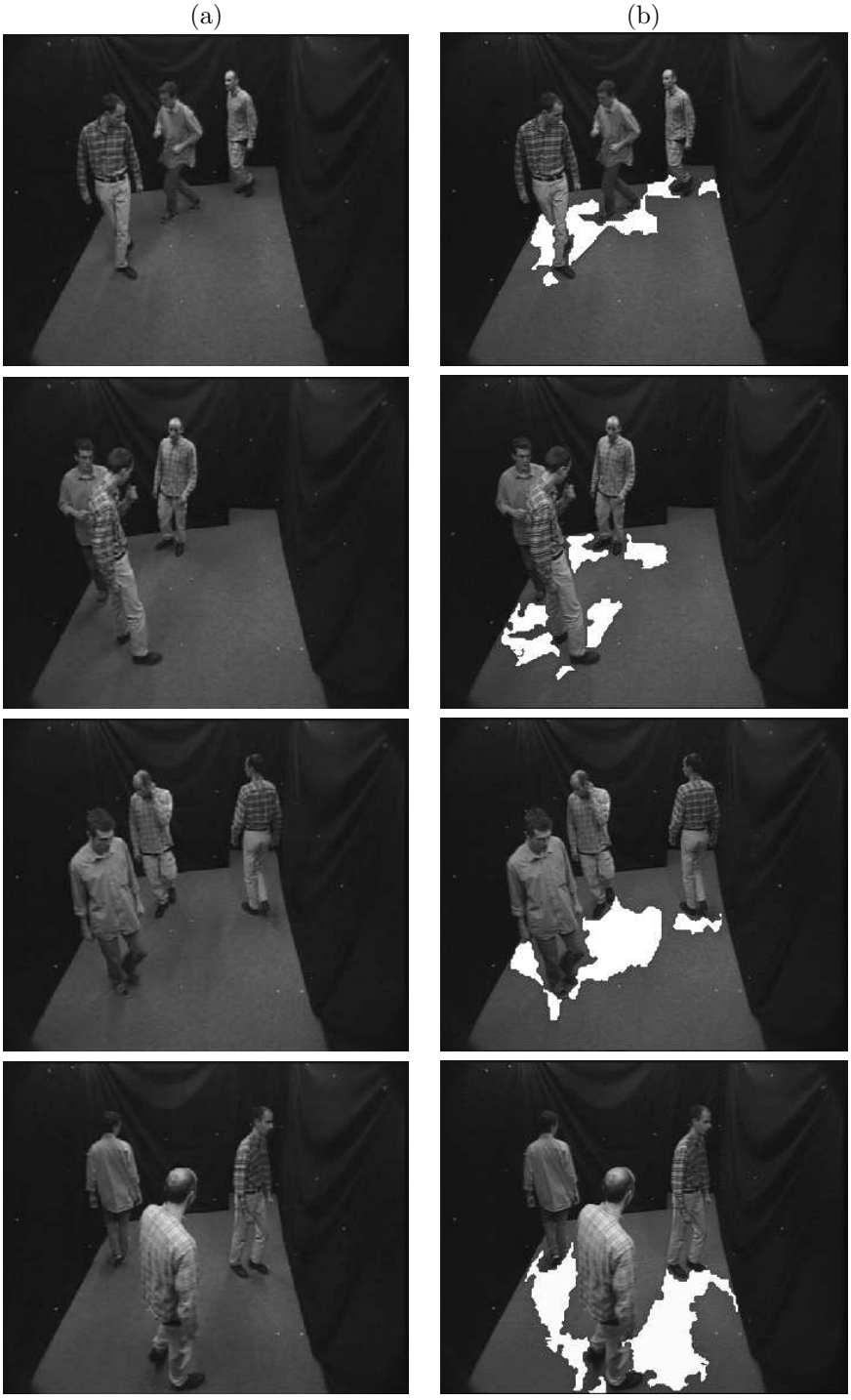

Figure 6.2: Shadow segmentation results for the test sequence *Group*. (a) Original image; (b) shadow mask (white pixels) superimposed on the original image.

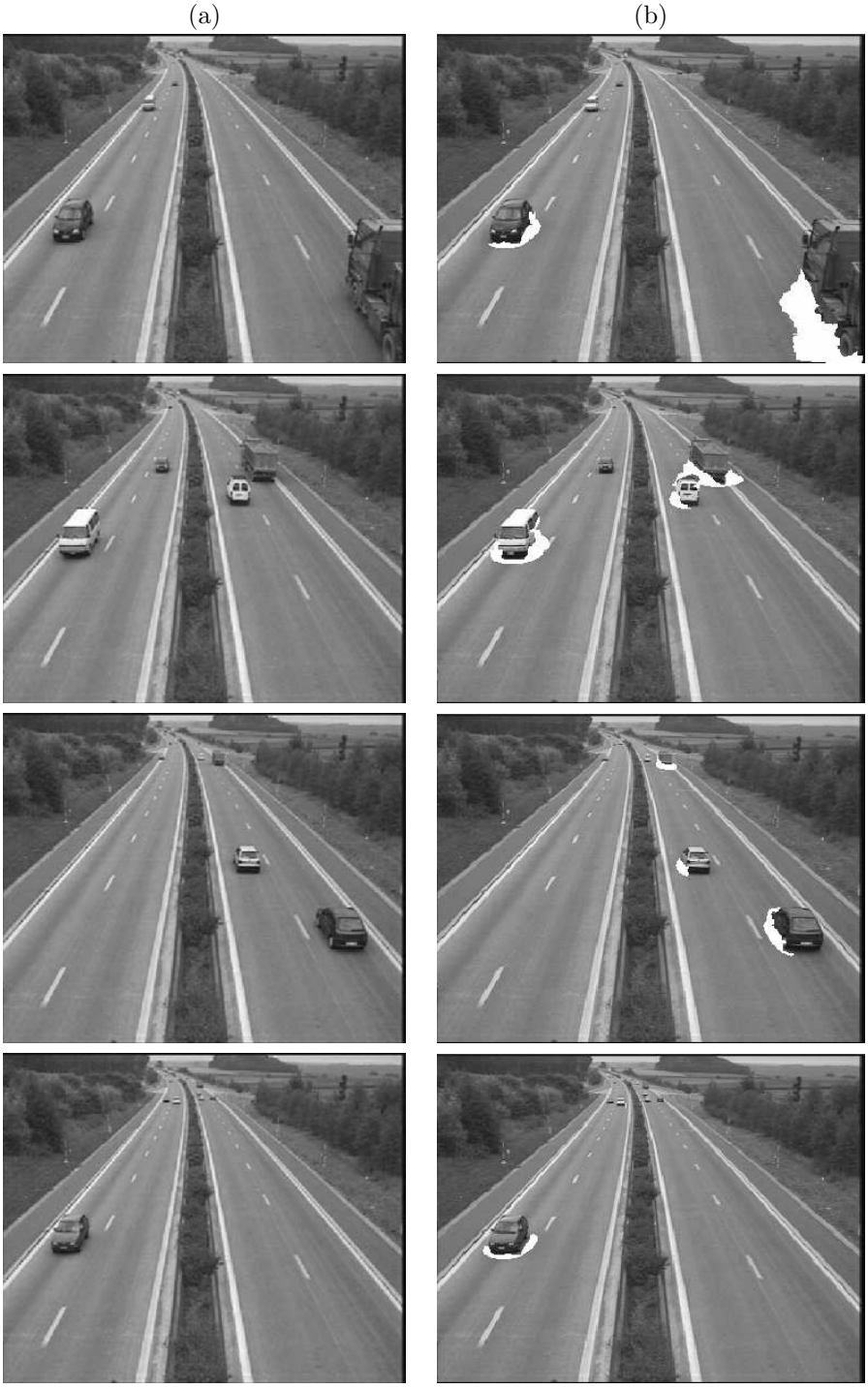

Figure 6.3: Shadow segmentation results for the test sequence *Highway*. (a) Original image; (b) shadow mask (white pixels) superimposed on the original image.

Figure 6.4: Shadow segmentation results for the test sequence *Surveillance*. (a) Original image; (b) shadow mask (white pixels) superimposed on the original image.

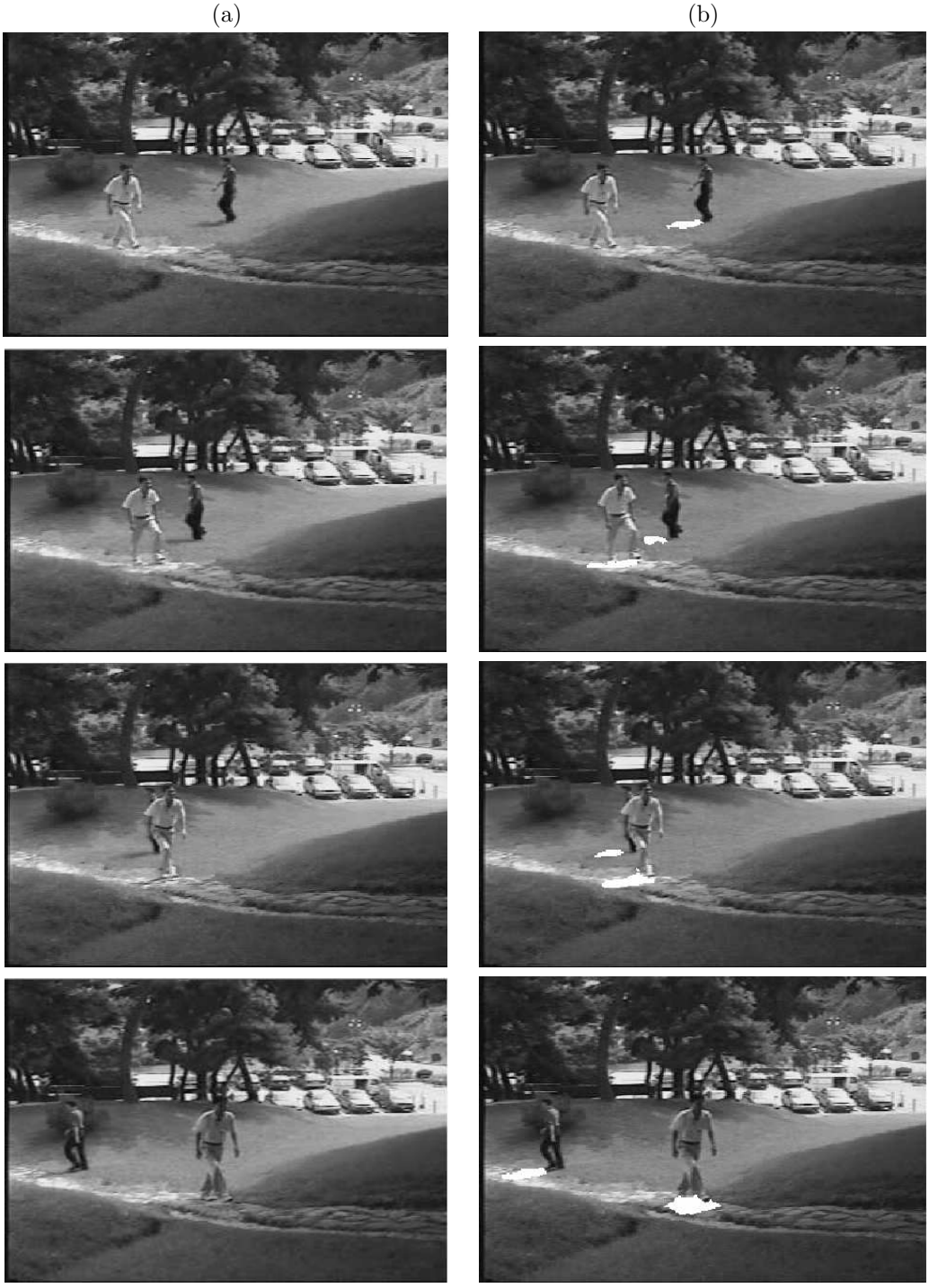

Figure 6.5: Shadow segmentation results for the test sequence *Surveillance2*. (a) Original image; (b) shadow mask (white pixels) superimposed on the original image.

The false adjacency would make the subsequent tracking, counting or classification operations of individual objects difficult. The refinement by means of shadow segmentation allows then to avoid this kind of error.

To demonstrate the performance of the proposed method in outdoor sunny scenes, the results of the tests on the sequences *Surveillance* and *Surveillance2* are shown in Figure 6.4 and Figure 6.5. The image format is, in this case, 352×240. Due to the high value of the threshold, some object pixels have been misdetected on the head of the person in the third image of Figure 6.4. Also in this case, the temporal verification stage will overcome the problem.

The illumination conditions in Figure 6.5 are more complex with respect to the previous scene. Here, big static shadows are cast on part of the image and light illuminating the person on the foreground changes considerably from the first two frames to the third and fourth image. The method's efficacy in complex real world conditions is demonstrated by the reported results.

6.2.2 Objective performance evaluation and comparison

In the previous section, segmentation results have been evaluated on the basis of subjective assessment only. To quantitatively analyze the performance of the method with different parameter sets and to objectively compare its results with those of other state of the art methods, an objective evaluation criterion has to be defined.

Objectively and quantitatively assessing the accuracy of the results of a shadow segmentation algorithm is not a simple task. Ideally, an exact, correct segmentation would be used as ground-truth information against which to judge the actual segmentation results. The disparity between the ground-truth segmentation and the actual segmentation would be computed to evaluate the accuracy of the results. The generation of a ground-truth for shadow regions in real world scenes is, however, a very difficult task. In many cases, in fact, the outer boundary of a shadow occurs at points of infinitesimal decrease in the amount of illumination (see Figure 6.2 or Figure 6.3 for instance). As a result, the exact boundary of a shadow cannot be manually determined in a reliable way.

As a solution to this problem, we propose that a more significant performance analysis can be obtained by combining the shadow detection method with an object extraction method and by evaluating the object segmentation accuracy. Obtaining a ground-truth segmentation of video objects is, in fact, a more reliable operation than defining an ideal shadow segmentation. The combination of the shadow segmentation algorithm with an object segmentation method allows also to demonstrate an important application of moving cast shadow segmentation.

Metric — The evaluation of the object segmentation accuracy is based on computing the pixel-wise deviation of the segmentation result from the corresponding ground-truth segmentation. The deviation is computed by taking into account two types of errors, namely false positives and false negatives. *False positives*, ϵ_p, are pixels incorrectly detected as belonging to an object. *False negatives*, ϵ_n, are pixels belonging to an object that have not been detected.

Let $card(C^t)$ represent the number of pixels detected as object pixels, and $card(C_g^t)$ the number of pixels belonging to the ground-truth segmentation, at each time instant t. The ensemble of false positive pixels, ϵ_p^t, at t can be expressed as

$$\epsilon_p^t = card(C^t \cap \overline{C_g^t}), \tag{6.1}$$

where $\overline{C_g^t}$ is the complement of C_g^t. The ensemble of false negative pixels, ϵ_n^t, at t can be then expressed as

$$\epsilon_n^t = card(\overline{C^t} \cap C_g^t). \tag{6.2}$$

The deviation from the reference segmentation at each time instant t can be computed as

$$\epsilon^t = \begin{cases} 0 & \text{if } card(C^t) = card(C_g^t) = 0 \\ \frac{\epsilon_n^t + \epsilon_p^t}{card(C^t) + card(C_g^t)} & \text{otherwise.} \end{cases} \tag{6.3}$$

where $\epsilon^t \in [0, 1]$.

The value of ϵ^t is proportional to the amount of segmentation errors with respect to the ground-truth segmentation. The quality of the results is inversely proportional to the deviation between actual and ground-truth segmentation. The *accuracy* of the segmentation is then quantified by

$$\nu^t = 1 - \epsilon^t, \tag{6.4}$$

with $\nu^t \in [0, 1]$. The larger ν^t, the higher the accuracy. When $\nu^t = 1$, then there is a perfect match between segmentation results and ground-truth segmentation.

Evaluation — By means of the above introduced accuracy metric, we aim now at evaluating the results of the combination of the proposed cast shadow segmentation method with the statistical model-based change detector embedded in the method and discussed in Section 5.5.1. First of all, the influence of parameter values on the method's performance is evaluated and then a comparison with state of the art methods is discussed.

The ground-truth segmentation for the test sequence *Hall Monitor* has been obtained manually and has been made available by the European project COST 211*. Since the test sequence *Hall Monitor* is a challenging sequence for moving cast shadow segmentation due to the color content of objects and background, which are quite similar, we consider this sequence in our analysis. As commented in the above discussion of shadow segmentation results, the misclassification of object points as shadow points is a major problem when segmenting shadows. The sequence provides therefore a significant test case.

In Table 6.1, the mean values for the 300 frames of the test sequence *Hall Monitor* of false positives, false negatives, and object segmentation accuracy for different sets of parameters are reported. False positives and false negatives are reported as percentage of the segmented area in the ground-truth. The obtained results show that the method's performance remains stable for different parameter configurations. From the results we can observe that the parameter which has the major influence on the performance is the threshold f_i.

W^{dark}	W^{inv}	f_i	$\%\epsilon_p$	$\%\epsilon_n$	ν
3×3	7×7	7	22.22	6.94	0.865
5×5	7×7	7	23.84	5.71	0.865
5×5	5×5	7	21.63	7.59	0.863
5×5	7×7	4	28.55	4.97	0.850
5×5	7×7	5	25.35	5.10	0.862
5×5	7×7	6	23.34	5.81	0.866

Table 6.1: System performance with different parameter sets.

In order to further evaluate the performance of the proposed algorithm, the video object extraction results have been compared to those of four state of the art object extraction methods which include a moving cast shadow detection technique. The four approaches are the method in [27]

*http://www.iva.cs.tut.fi/COST211/

(DNM1), which exploits the invariance properties of hue and saturation, the technique in [151] (DNM2), which makes use of luminance information, texture information, and the penumbra of shadows, and the statistic approaches in [105] (SP), which makes use of a diagonal model to characterize illumination changes in shadows, and in [70] (SNP), which exploits the invariance properties of a computational color model which separates chromaticity from brightness. The methods have been analyzed and compared in [134]. They are described in more detail in Section 3.4.4. The adopted acronyms correspond to those used to denote the different techniques in [134].

To perform the comparison, the object segmentation accuracy ν^t (Eq. 6.3) is computed over the sequence. The results for the test sequence *Hall Monitor* are presented in Figure 6.6. Object and shadows masks obtained with the different methods for frame 55 are also shown to help evaluation. The large error in the first frames of the sequence is due to the fact that these frames correspond to the entrance of the man in the scene. The first part of the man entering the scene is his shoe. The shoe has a color that is very similar to that of the background. For this reason, the detection algorithms may be misled and do not detect the shoe that is instead present in the ground-truth segmentation.

The mean values of accuracy over the entire sequence corresponding to the plots of Figure 6.6 are reported in Table 6.2. The combination of the proposed shadow segmentation method with the adopted change detector results in a more accurate object detection over time when compared to state of the art shadow-invariant object detection algorithms.

	DNM1	DNM2	SP	SNP	Proposed
ν	0.78	0.60	0.59	0.63	0.86

Table 6.2: Mean values of object segmentation accuracy for test sequence *Hall Monitor* for the proposed method, the method in [27] (DNM1), the method in [151] (DNM2), the method in [105] (SP), and the method in [70] (SNP).

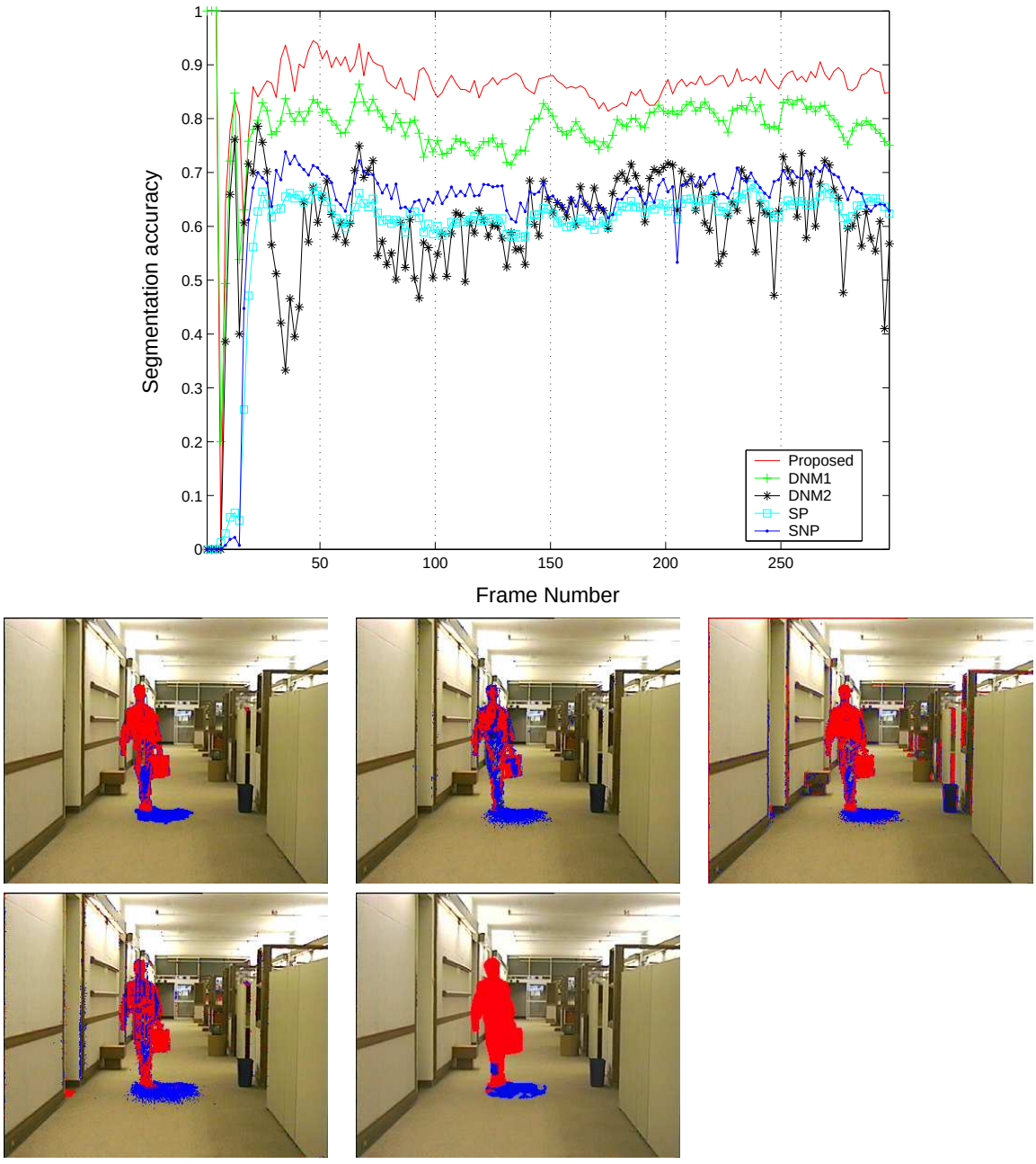

Figure 6.6: Comparison of video object segmentation accuracy ν for test sequence *Hall Monitor*. Top: objective spatial accuracy comparison among the proposed approach, the method in [27] (DNM1), the method in [151] (DNM2), the method in [105] (SP), and the method in [70] (SNP). Bottom: subjective object and shadow segmentation comparison among the five methods for frame 55. Object pixels are displayed in red, shadow pixels in blue.

6.2.3 Segmentation and tracking results

Figures 6.7– 6.10 illustrate the improvements over the cast shadow segmentation results that are obtained by means of the temporal analysis stage, once candidate shadows have been tracked over time. For each test sequence, the first column shows the results of the first stage of the shadow segmentation process and the second column shows the final results after the off-line processing performed thanks to the tracking stage.

One parameter has to be set for the temporal verification stage, that is the temporal filtering threshold DT (Section 5.6.2). As discussed in Section 5.6.2, the threshold DT has been empirically determined as a function of the level of activity in the sequence which is reflected by the distribution of shadow durations in the sequence. The values used to obtain the results in this section are 38, 20, 4 and 5 for *Hall Monitor*, *Highway*, *Group* and *Laboratory*, respectively. In the former two sequence, objects remain for a long time in the scene. In the *Group* sequence the main big shadow formed by the fusion of people shadows has a long duration, while segmentation errors change rapidly of position with the movements of the persons. In the last sequence, people enter and exit the scene rapidly. The image format for test sequence *Laboratory* of the ATON project is 320×240.

In the results of the first columns of the reported figures errors of the segmentation algorithm are shown. These figures indicate difficult cases for the proposed approach. Parts of moving objects are misclassified as moving shadows where the color of the object is similar to that of the background. This often happens in portions of the object that are self shadowed. Examples of this type of error can be seen in Figure 6.7, Figure 6.8, in Figure 6.9, second row, and in Figure 6.10, first row. The results illustrated in the second columns of the figures show the improvements achieved thanks to tracking. The described failures of the moving cast shadow extraction stage have been eliminated by the analysis of the temporal behavior of shadows.

The results on the test sequence *Laboratory* in Figure 6.10, second row, allow to make two observations about the proposed approach. The first observation concerns the use of the border analysis in the spatial analysis stage discussed in Section 5.5.3 to postpone the rejection of a candidate shadow after the temporal analysis has been performed. The efficacy of the analysis is illustrated in these results. The candidate shadow between the person's legs is, in fact, an example of critical case when a true shadow shares the majority of its boundary with an object boundary. This case was illustrated in Figure 5.29 (b) by means of shadow (C). Since the majority of the border's pixels is in contact with the object, then a low confidence value is associated with the region under analysis. The region is therefore reconsidered in the temporal analysis stage. The temporal reliability score then succeeds in taking a correct final decision about the candidate shadow region.

The second observation concerns the limits of the temporal filtering stage. A failure of the tracking stage is, in fact, shown in Figure 6.10, second row. The moving cast shadow extraction algorithm has correctly detected the shadows cast by the person on the background close to his left hand and right foot. However, since the majority of edge pixels of these shadow regions are connected to the object and their temporal duration is very short, they have been removed by the off-line processing. This kind of error occurs typically to small shadow regions and does not significantly affect the overall method's performance.

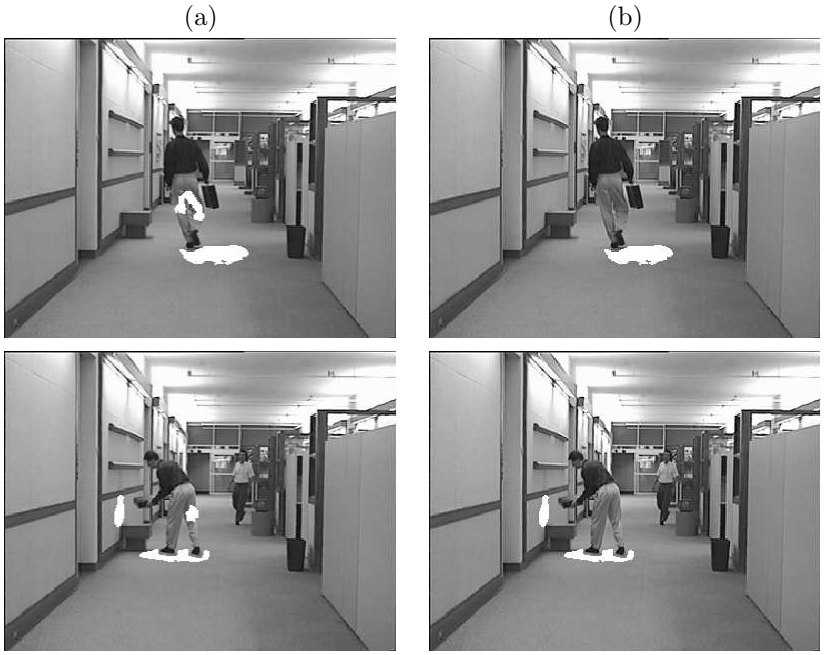

Figure 6.7: Shadow segmentation and tracking results for test sequence *Hall Monitor*. (a) Moving cast shadow extraction. (b) Final result after tracking.

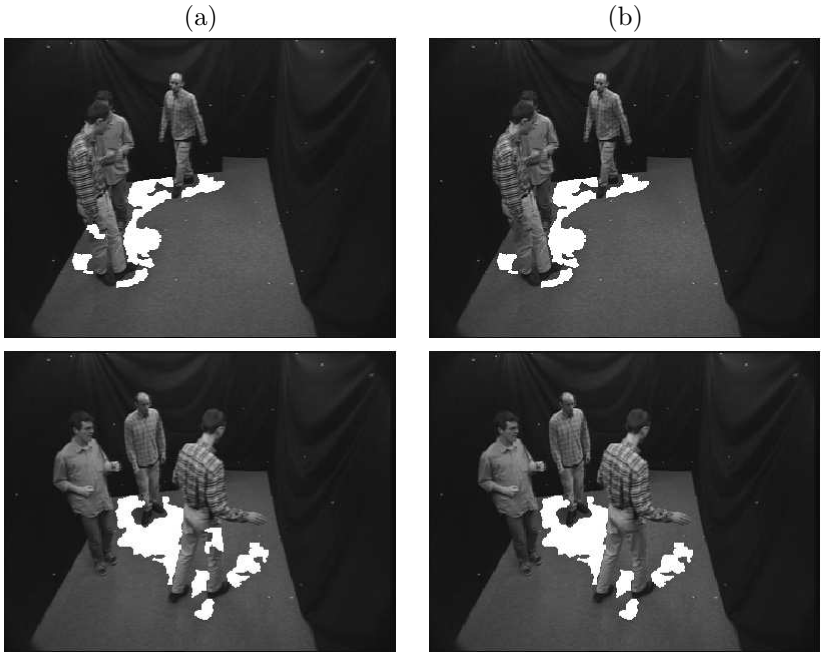

Figure 6.8: Shadow segmentation and tracking results for test sequence *Group*. (a) Moving cast shadow extraction. (b) Final results after tracking.

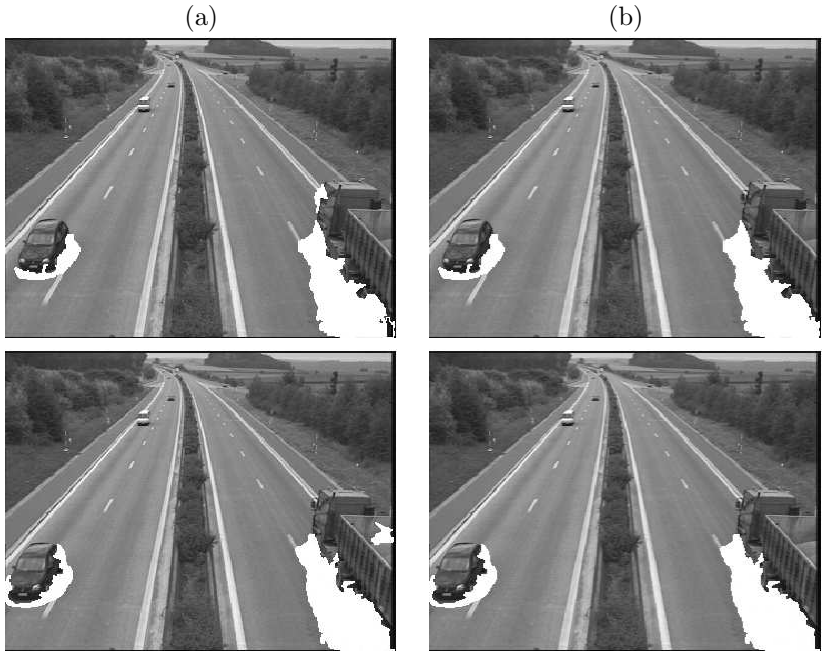

Figure 6.9: Shadow segmentation and tracking results for test sequence *Highway*. (a) Moving cast shadow extraction. (b) Final results after tracking.

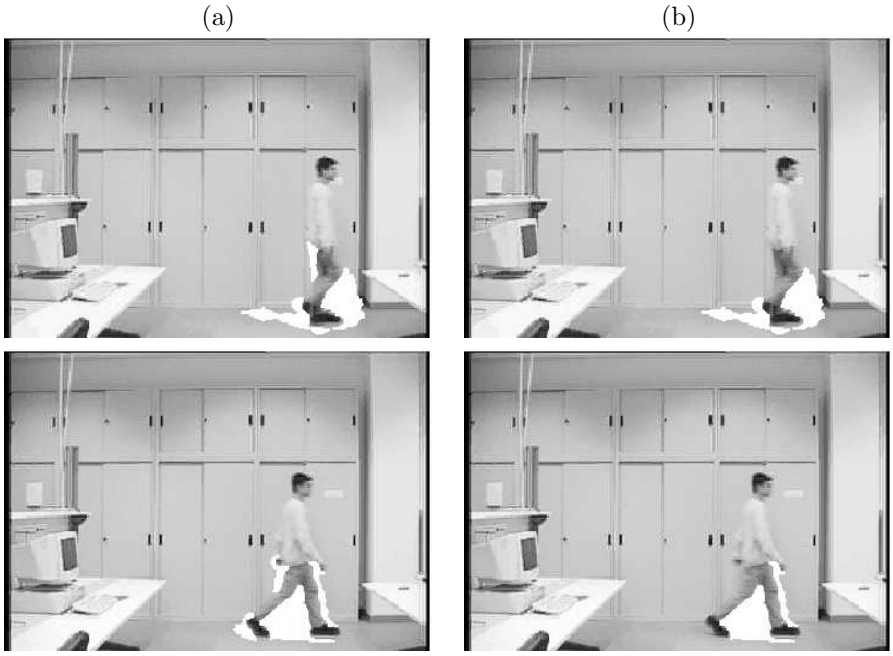

Figure 6.10: Shadow segmentation results for test sequence *Laboratory*. (a) Moving cast shadow extraction. (b) Final results after tracking.

Objective evaluation — To quantitatively evaluate the improvements obtained by means of tracking, the object segmentation accuracy ν^t (Eq. 6.4) has been computed. The results for the test sequence *Hall Monitor* are shown in Figure 6.11. Here, we compare the following results. First, the spatial accuracy of results for the change detection. Second, the spatial accuracy of results for the change detection combined to the moving cast shadow extraction method of the first stage of the proposed algorithm. Third, the spatial accuracy obtained by using tracking. In the second row of the figure, object masks obtained with the three different methods for frame 125 are shown for subjective comparison.

From the plots it can be noted that the shadow segmentation process brings improvements to the object segmentation quality in particular in the initial and final part of the sequence. In these parts, in fact, shadows detected by the change detector as part of the moving objects are quite large and cause significant detection errors. In the central part of the sequence, one person is leaving the room while the other is entering the room and both are far from the camera, thus casting small shadows. Between frames 80 and 130, the misclassifications of object points on the trousers of the man as shadow points affect the quality of the object extraction results, which are slightly worse than those of the change detector alone. It is in this case that the tracking stage brings its major improvement. In the second half of the sequence, where errors due to misclassifications of object points as shadow points do not occur, the results with and without tracking remain the same.

In the very first part of the sequence, for a couple of frames, it can be noted that the temporal analysis stage worsens the performance of the shadow segmentation algorithm. These frames correspond to the entrance of the man in the room. While the man enters the scene, shadows are cast on the wall which have a very short duration and are removed by the temporal analysis. This error could be corrected by adding a control in the method which detects objects entering or leaving the scene and suspends the temporal filtering for some frames during these events.

The mean values of accuracy over the entire sequence corresponding to the plots of Figure 6.11 are the following: change detection 0.82, shadow segmentation 0.86, shadow segmentation and tracking 0.87. The tests confirm that the results of a change detection algorithm can be progressively improved by first extracting moving shadows and by then tracking them.

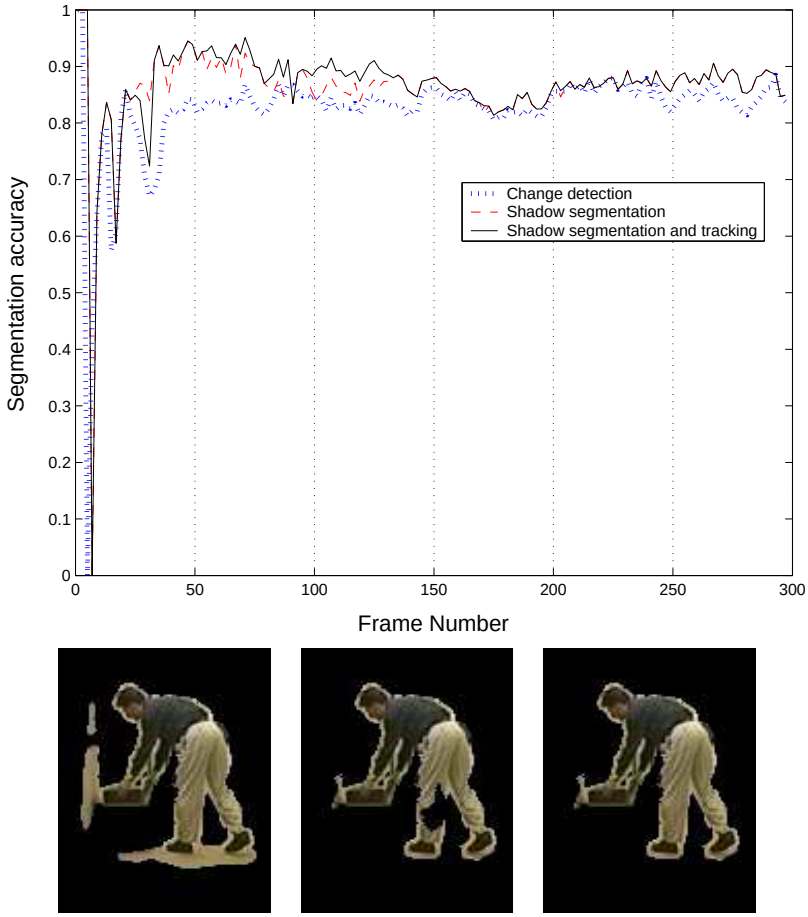

Figure 6.11: Comparison of video object segmentation accuracy ν for test sequence *Hall Monitor*. Top: objective spatial accuracy comparison among change detection results, change detection results combined to the first stage of the proposed moving cast shadow segmentation method, and to the complete moving cast shadow segmentation and tracking method. Bottom: subjective object extraction comparison among the three methods for frame 125.

6.3 Results on still images

The results of the application of the proposed methodology on still images are discussed in this section. The images used for the tests have been obtained using a SONY DCR-PC7E digital video recorder, a commercial digital camera. The images are in CIF format (288×352 pixel). Objects are made of different materials and present different colors.

The parameters of the proposed algorithm for still images are the edge detection thresholds used to binarize the edge gradient obtained with the Sobel operator (Section 5.7.1). The output of the edge detector is characterized by two types of errors. A false positive occurs when an edge is declared, but no edge is present. A false negative occurs when an edge is present, but no edge is declared. A low value for the threshold leads to a high false positive rate and a low false negative rate and viceversa. The values of the thresholds have been determined empirically based, therefore, on the following reasoning. The threshold value for the invariant features analysis must be large enough to minimize the occurrence of false positives detected due to noise far outside the object contours. The threshold for the RGB color space analysis should be small enough to minimize the occurrence of false negatives and to obtain closed contours. The values of thresholds for the different test images reported in this section are shown in Table 6.3. As can be noted when looking at the corresponding images, the threshold for RGB images is related to the strength of the cast shadow. The weaker the shadow, the lower the threshold. The threshold on photometric invariants is less straightforward.

	RGB	$c_1c_2c_3$
Image1	0.06	0.12
Image2	0.05	0.12
Image3	0.02	0.14
Image4	0.07	0.07
Image5	0.03	0.07
Image6	0.03	0.09

Table 6.3: Value of the thresholds for color edge detection on the different test images.

Figure 6.12 shows the results of the proposed algorithm for a selection of test images. The original image (Figure 6.12 (a)) and the superimposition of shadow masks on the original image (Figure 6.12 (b)) are displayed. The obtained results show that cast shadows are correctly detected by the proposed algorithm.

Smeared edge markings can be observed in the extracted shadows, particularly for Figure 6.12, bottom. This type of error is caused by the use of a small threshold for edge detection in the RGB space. Shadows are, in fact, quite weak in this image. To overcome this problem, a morphological post-processing depending on the application at hand may be used to improve the final segmentation results.

Segmentation by color edge filling

The results of the extraction of cast shadows by means of edge filling (Section 5.7.3) are shown in Figure 6.13 and Figure 6.14. The results show that shadow regions have been correctly identified.

Contours in *Image 2* and *Image 3* are better delineated with respect to the results in Figure 6.12 thanks to edge post-processing. Thanks to the use of color information, and not only intensity, the method has correctly distinguished in *Image4* the dark object from its shadow. This would

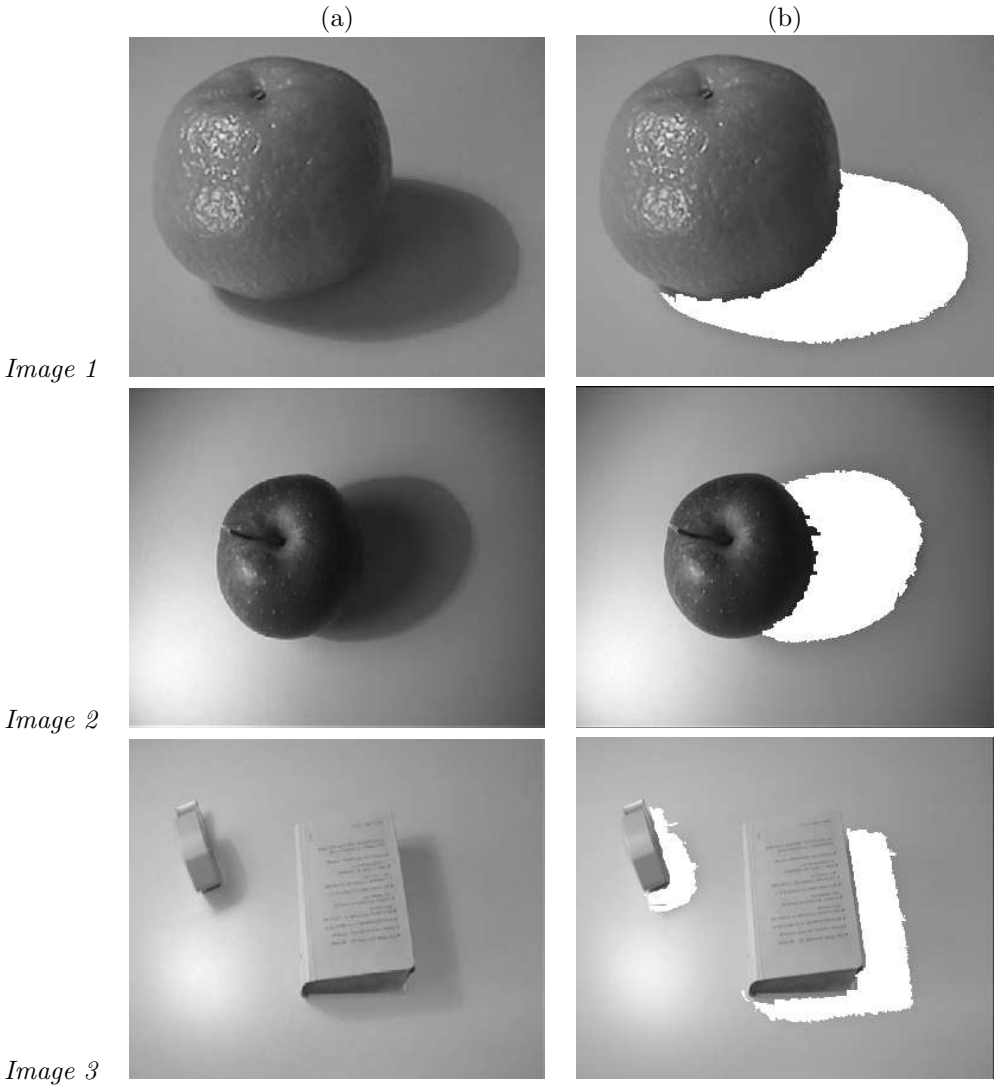

Figure 6.12: Cast shadow segmentation results for still images. (a) Original image; (b) shadow mask (white pixels) superimposed on the original image.

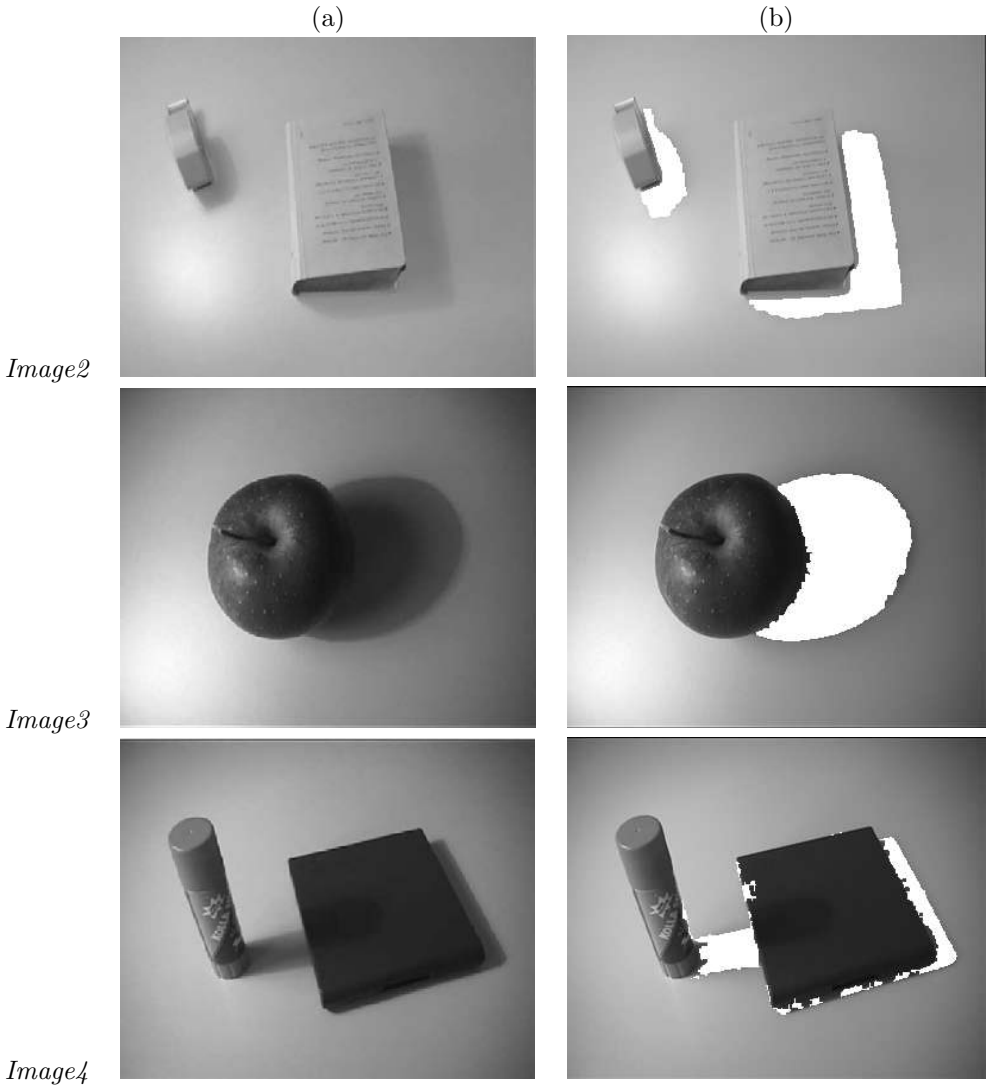

Figure 6.13: Results of cast shadow segmentation by means of color edge filling. (a) Original image; (b) shadow mask (white pixels) superimposed on the original image.

not have been possible for techniques that exploit only luminance properties of shadows for their identification.

In *Image 5*, the object on the left hand side has a pink color which is similar to the red color of the background surface. Some points on this object have been classified as shadow points because of the reduced discriminative power of photometric invariants. In *Image 6*, the gray object on the top left corner of the image has been entirely misclassified as a shadow region for the same reason. The reduced discriminative power of the invariant features has prevented the edge detection step to identify the edges of the above mentioned object. In *Image 6*, on the marker in the lower left corner, some object points have been misclassified as shadow points as well. These errors are due to the presence of highlights.

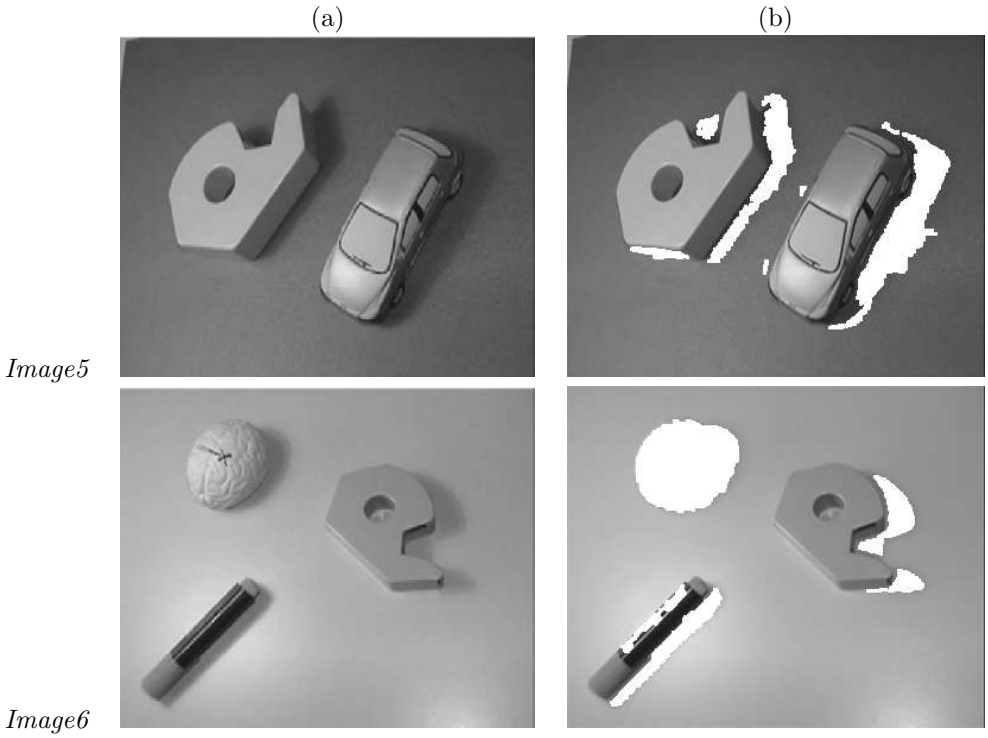

Figure 6.14: Results of cast shadow segmentation by means of color edge filling. (a) Original image; (b) shadow mask (white pixels) superimposed on the original image.

6.4 Summary

An evaluation of the proposed cast shadow segmentation method is presented in this chapter.

In order to assess the method's performance in video sequences, the results of the first two blocks of the system, that is the color analysis and the spatial analysis, were first of all analyzed. The results were evaluated subjectively, objectively, and compared to state of the art methods. Then, the improvements introduced by the temporal verification stage were assessed by means of subjective and objective comparison with the results of the first part of the segmentation algorithm.

The experimental results demonstrated the efficacy of the proposed technique in a wide range of scenes, where shadows are projected on vertical and horizontal surfaces, on surfaces of different material, in presence of different illumination conditions and with objects of different nature. This underlines the good generality of the method. The results showed moreover the improvement obtained with respect to the state of the art. The benefits introduced by shadow tracking were also demonstrated. The temporal analysis was shown to be able to eliminate the possible ambiguities of the previous analysis levels and to improve the efficiency of the overall shadow extraction algorithm.

The validity and efficiency of the proposed approach also when applied to still color images was then demonstrated through the analysis of its results on a number of typical images.

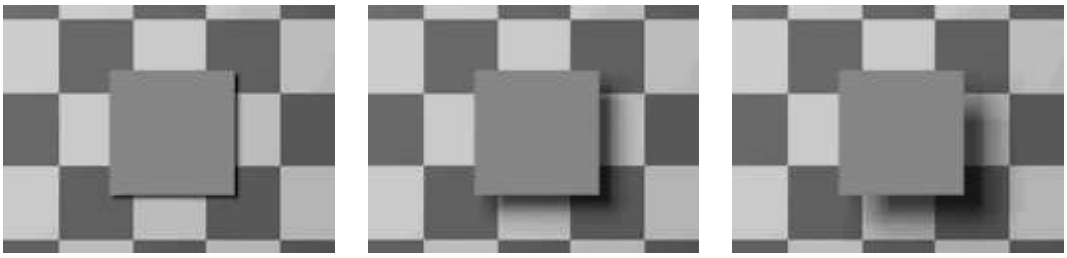

Figure 6.15: Shadow motion induces apparent motion of an object in depth (Section A.3).

7 Shadow-aware video processing

7.1 Introduction

Chapter 5 was dedicated to the description of a methodology for the segmentation of cast shadows in video sequences and still images. In order to maximize the generality of the methodology, the adopted rules, which guide the proposed method in the analysis of shadows, do not rely on models of the objects nor on particular hypotheses about the considered scenes. The developed tool is therefore flexible and can be applied to a wide range of scenes and conditions, as demonstrated in Chapter 6.

Two main uses of the developed technique can be identified: shadow segmentation can be adopted for *shadow elimination* and for *shadow manipulation*. In the former case, shadow segmentation allows to improve the performances of video object extraction, tracking and description tools. The extraction, tracking and description of video objects are fundamental steps for a wide range of object-based applications, ranging from video coding to video indexing, from video manipulation to video surveillance and immersive environments. All these applications can benefit from a flexible methodology that allows to distinguish objects from the shadows they cast. In the latter case, the identification of shadows provides information about and access to an important perceptual element of a visual scene. In applications such as object-based video editing and mixed-reality immersive environments, where new and richer visual content is created by merging objects from different sources, the ability of identifying and taking shadows into account can improve the naturalness of the merging process and have an important perceptual impact.

The objective of this chapter is to demonstrate the advantages and the possibilities offered by the proposed shadow segmentation tool in its twofold use through the discussion of some example applications. First of all, the impact of shadow segmentation on video object segmentation, tracking, and description is discussed in Section 7.2. Immersive interactive environments are presented in Section 7.3. Photorealistic video composition is finally discussed in Section 7.4.

7.2 Shadow elimination for improved video object extraction

Advances in video coding and description are driving a shift from the traditional frame-based approach to video processing, where a video sequence is composed of a set of frames, to the object-based approach, where the video sequence is composed of a set of meaningful objects. International standards, such as MPEG-4 [121, 145] and MPEG-7 [23, 146], support this type of representation and a wide variety of applications, ranging from video coding to video editing, and from video surveillance to mixed-reality, benefit from the shift.

The representation of visual information in terms of meaningful objects, that can be accessed, manipulated, coded and described separately, requires a prior decomposition of video sequences into semantically, meaningful objects. For many applications, objects of interest are moving objects and many video object extraction methods make use of motion information to automatically extract semantic objects. Moving shadow segmentation and elimination is then an important component for such methods. As discussed in Section 3.4.4, shadows cast by moving objects generate temporal changes in an image sequence and mislead both motion segmentation and motion detection approaches to automatic video object extraction.

Object segmentation – Figures 7.1-7.3 demonstrate the improvements obtained thanks to the proposed shadow segmentation method in video object segmentation results. In the second row of each figure the results of the application of the change detector described in Section 5.5.1 are shown. In the third row the refined objects obtained by eliminating the cast shadow are illustrated. Object boundaries are more accurate once shadows have been removed. Figure 7.1 illustrates a typical indoor video surveillance scenario, Figure 7.2 a smart room, and Figure 7.3 an outdoor video surveillance scene. The flexibility of the proposed approach allows its use in different applications.

Figure 7.1: Object segmentation results for some sample frames of the test sequence *Hall Monitor* without shadow segmentation (first row) and with shadow segmentation and elimination (second row).

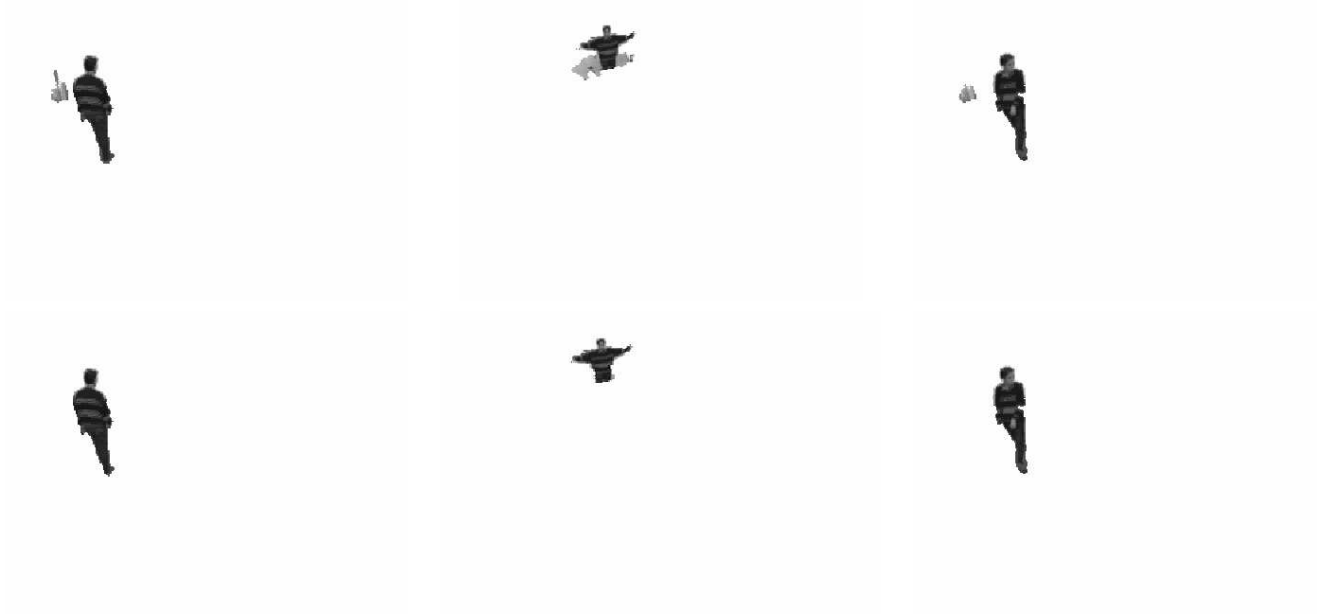

Figure 7.2: Object segmentation results for some sample frames of the test sequence *Intelligent room* without shadow segmentation (first row) and with shadow segmentation and elimination (second row).

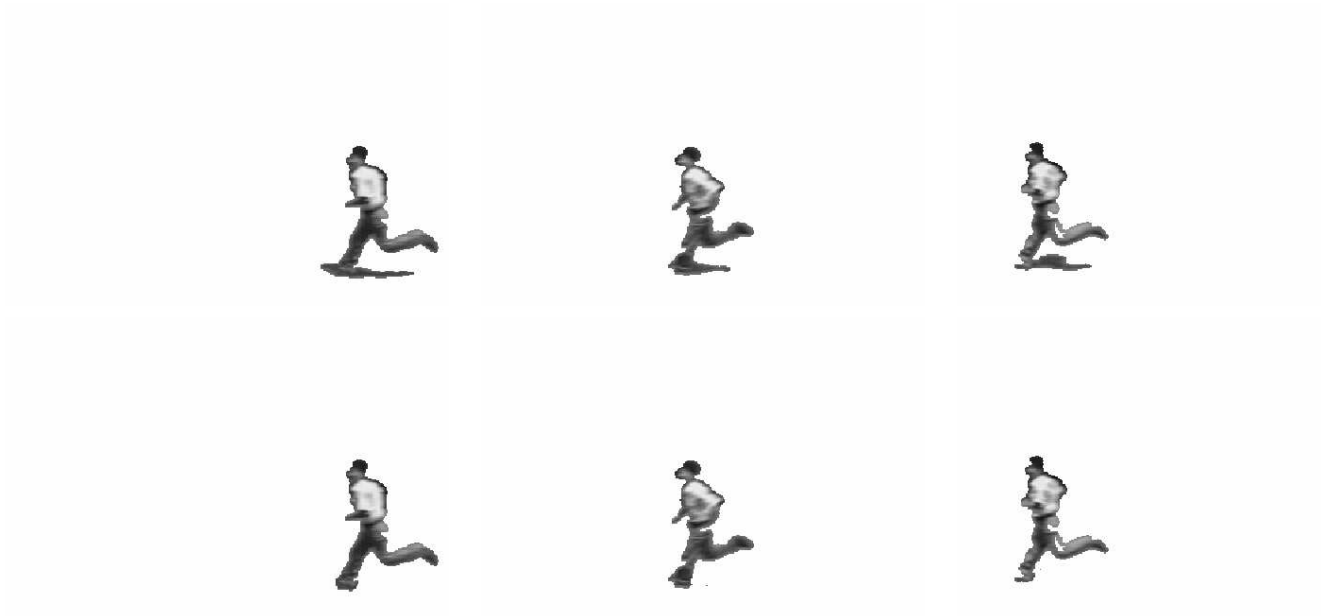

Figure 7.3: Object segmentation results for some sample frames of the test sequence *Surveillance* without shadow segmentation (first row) and with shadow segmentation and elimination (second row).

Object tracking – An accurate segmentation of foreground moving objects from the scene's background has an impact on all the subsequent video analysis operations that rely on the segmentation results as a preliminary step. A fundamental step in semantic video object extraction is given by tracking, which aims at establishing a correspondence between instances of moving objects over frames.

In order to follow an object over time, a comparison between characteristic properties of the object from frame to frame has to be performed. Spatio-temporal properties, such as color, texture and motion of object pixels, can be exploited to this end [96, 104, 158, 167]. Contours [60, 119, 123, 156], object models [53, 177] or feature points, such as corners [13], can also be used for tracking. Hybrid tracking methods [99, 162] consider first the object as an entity and then track its parts by analyzing their spatio-temporal properties.

The presence of shadows in the object segmentation results on which the tracking strategy is applied can make the computation of object features less reliable and limit the performance of the tracking algorithm. Color, texture and motion features cannot in fact be reliably computed in shadow regions since shadows change their appearance according to changes in the appearance of the surface they are cast upon. Shadows modify moreover the shape of the objects making the correspondence problem more difficult. An accurate moving object segmentation thanks to shadow elimination makes then tracking more reliable. Moreover, it can make tracking faster. Multiple hypotheses on the object's identity during time can in fact be pruned more rapidly if the object is accurately extracted.

In particular, the management of interactions between objects in the scene, which is one of the main obstacles to an effective tracking process, can benefit from the identification and elimination of shadows. Cast shadows are typically attached to the shadow-casting object and cause undersegmentation errors by creating false adjacency between objects. Two objects getting close to each other are in these cases erroneously extracted as a single object.

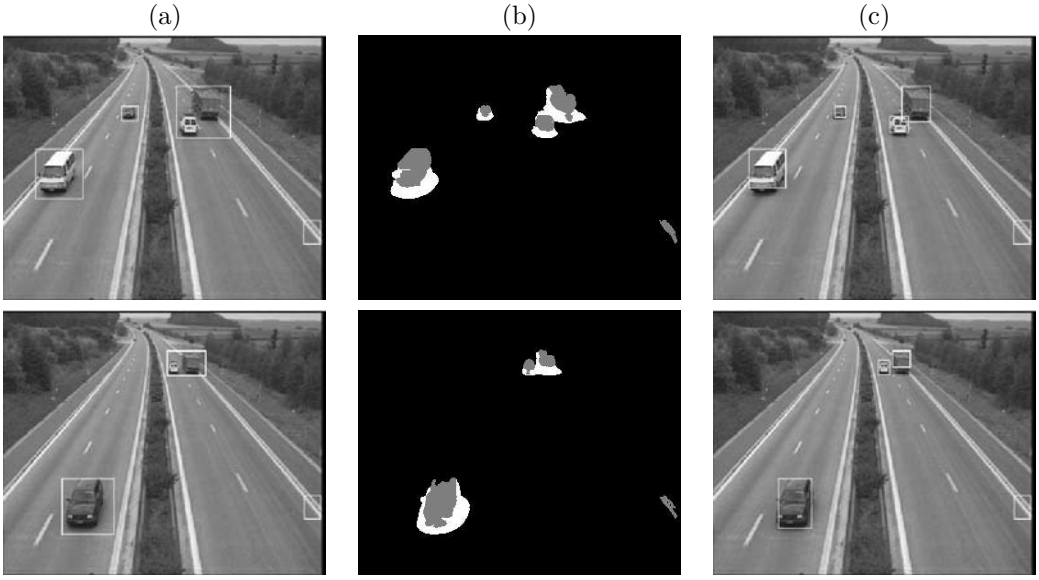

Figure 7.4: (a) Two moving objects, represented by their bounding boxes, are extracted as a single object by change detection in two sample frames of the test sequence *Highway*. (b) Object and shadow segmentation results. Objects are displayed in gray, shadows in white. (c) Thanks to shadow segmentation the undersegmentation errors have been solved.

Figure 7.4 (a) and 7.5 (a) show cases of undersegmentation due to shadow effects. The moving foreground objects extracted by means of change detection and represented by a bounding box are shown. When the proposed shadow segmentation algorithm is applied, the results of Figure 7.4 (b) and 7.5 (b) are obtained. The extracted object is color coded in gray, while the shadow pixels are color coded in white. The identification of shadow regions allows to solve the undersegmentation problem of multiple objects extracted as a single one, as illustrated by the bounding boxes in Figure 7.4 (c) and 7.5 (c).

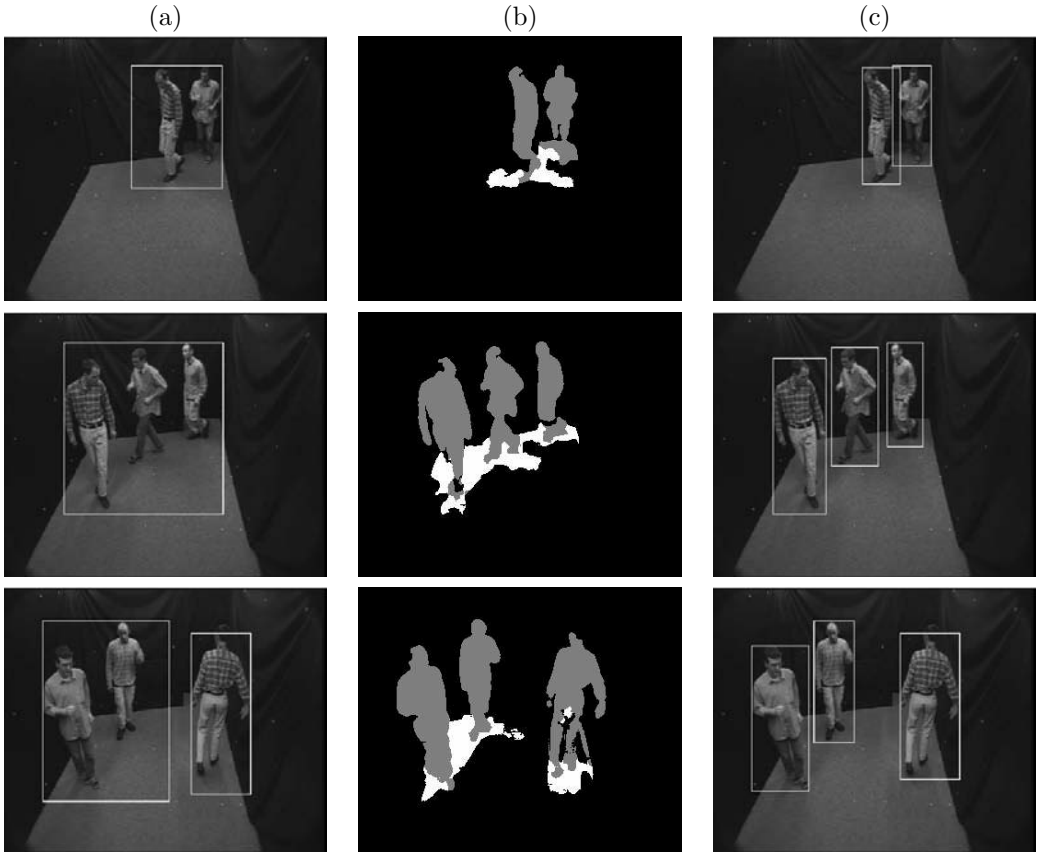

Figure 7.5: (a) Multiple moving objects, represented by their bounding boxes, are extracted as a single object by change detection in three sample frames of the test sequence *Group*. (b) Object and shadow segmentation results. Objects are displayed in gray, shadows in white. (c) Thanks to shadow segmentation the undersegmentation errors have been solved.

When the occurrence of undersegmentation errors as those discussed above is reduced thanks to shadow analysis, the tracking process can be simplified. It is indeed possible to track objects independently without the need of dedicated tracking management mechanisms. The results of object tracking by means of the simple algorithm proposed for tracking moving shadows in Section 5.6.1 on the test sequence *Highway* are shown in Figure 7.6. Each bounding box has been color coded with a different color. The correspondence between objects in successive frames is based on a simple overlap of object segmentation masks. Despite the simplicity of the tracking principle, objects have been successfully tracked over time.

The sequences considered in this subsection represent a typical traffic monitoring scenario (Figure 7.4 and Figure 7.6) and a typical video production scenario (Figure 7.5). As demonstrated, a

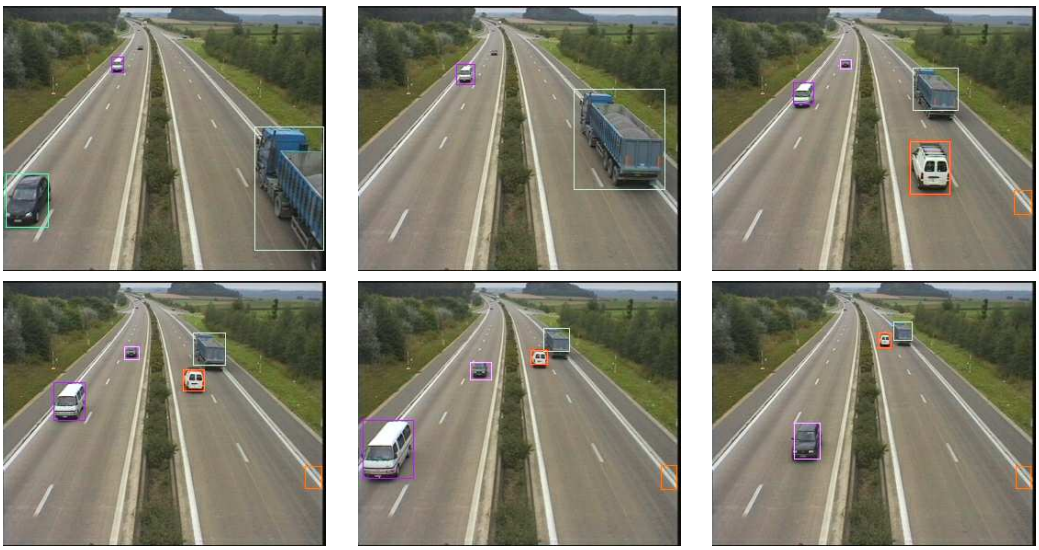

Figure 7.6: Tracking results for the test sequence *Highway*. Multiple moving objects, represented by their bounding boxes, have been successfully tracked over time with a simple nearest neighbor tracking approach.

wide variety of applications can benefit from the developed tool. In Figure 7.4 and Figure 7.6, the segmentation error visible in the lower right part of the image is due to the background image that has been reconstructed by means of a learning process. In Figure 7.6, moreover, the small car far from the camera in the first and second image of the top row is not considered by the segmentation and tracking algorithm which discards objects that are smaller than a fixed threshold.

Object description – A quantitative description of video objects can be generated once they have been extracted from a video sequence.

Low-level features, such as color, texture, and motion can be used for describing object parts at a low-level of abstraction. By attributing an identifier to each video object which describes its spatial location in the scene and by computing the object's trajectory, shape, dominant color or texture properties, a semantic description can be obtained. Low-level descriptors can be used for indexing, filtering or retrieving similar objects based on visual content in object databases. Semantic descriptors can be used for scene reconstruction [152], for video transcoding [21] and video analysis operations, such as object counting and classification. In video-based traffic surveillance applications, for instance, statistics about the vehicles passing in the field of view of cameras and traffic violations or alarming situations are of interest. The number of vehicles, their velocity, and the average distance between vehicles can be effectively computed by analyzing semantic descriptors. Moreover, descriptions provided to higher level content understanding modules can allow to monitor a scene and detect abnormal behaviors [2].

The issue of description, identification and access to multimedia information has been addressed by the MPEG-7 standard, formally known as *Multimedia Content Description Interface* [23, 146], which defines a standard set of Descriptors and Description Schemes for simple to sophisticated descriptions of a variety of multimedia content.

Shadow effects can significantly affect the description process. The case of false adjacency of multiple objects illustrated in Figure 7.4 and Figure 7.5 can for instance mislead object counting tasks. The inclusion of shadow regions in segmented objects can lead to unreliable low-level descrip-

tors computation that can limit the performance of object-based video database search. Typically, the object's shape centroid is used to describe the object's position and trajectory. Object shape is falsified by shadows and all the measured geometrical properties are then affected by an error. The explicit detection of shadows in video sequences can therefore significantly improve the accuracy of object description and support a more reliable use and interpretation.

The results of Figure 7.4 and Figure 7.5 provide some examples of how shadow segmentation can improve the description of video objects. Bounding boxes are shown which represent a simple description of the objects' shape. The bounding boxes obtained after the elimination of shadows in Figure 7.4 (c) and Figure 7.5 (c) more precisely describe the true shape, the number and the size of objects. Subsequent content understanding operations can therefore rely on a more accurate description. In case of camera calibration, 3D descriptors in the form of 3D bounding boxes could be computed from multiple views of the objects obtained from subsequent frames [128].

The discussed video object extraction, tracking and description operations are at the core of a wide range of applications. All of them can then benefit from a flexible methodology for shadow segmentation. To summarize, shadow segmentation and elimination can:

- improve the spatial accuracy of segmented objects;
- increase the reliability of object tracking;
- reduce the complexity and increase the speed of object tracking;
- increase the reliability and efficacy of object description.

Among the variety of applications of object-based video processing that can be cited, in the next section immersive interactive environments are discussed.

7.3 Immersive interactive environments

The shift from the frame-based approach to the object-based approach to visual information representation and processing allows to greatly extend the ways by which visual content can be created and manipulated. The ability of decomposing a video into a collection of meaningful objects, which can then be manipulated separately, offers in fact many novel possibilities of creating new and richer visual content. Scenes can be built by putting together objects from different sources and by mixing natural and graphical objects.

A *mixed-reality environment*, as defined by Milgram [106], is created when natural objects and synthetic objects are mixed together. Mixed-reality is also known as augmented reality and typically refers to emerging technologies that allow to insert computer-generated objects into the user's view of the real world. According to Milgram, mixed-reality consists in any combination of elements from the real, physical world, and the image capture of it, and from a virtual, completely modeled world. Here, we refer to this framework and consider the *inclusion of real objects into virtual backgrounds*. This track in the field of mixed-reality has been investigated by the European IST project *art.live*[*], whose goal was to develop an architecture and a set of tools, both generic and application dependent, for the enhancement of narrative spaces. The developed architecture [98] aimed at creating interactive stories that mix graphical elements with inputs from live cameras. The project put together partners in signal processing, artificial intelligence and multimedia authors.

The underlying concept, that is the capture of real life objects and their inclusion in a mixed-reality narrative space where they can interact with the story, is illustrated in Figure 7.7. Figure 7.8

[*]European project IST 10942 *art.live* (Architecture and authoring Tools for prototype for Living Images and new Video Experiments), http://www.tele.ucl.ac.be/PROJECTS/art.live/

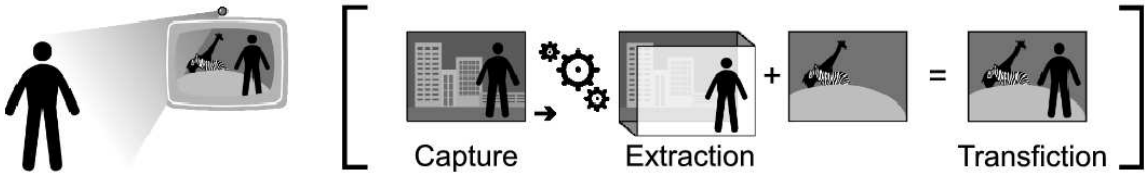

Figure 7.7: Principle of creation of a mixed-reality interactive environment (courtesy of *alterface*, www.alterface.com). The image of a person is captured by a camera and extracted from the background by means of segmentation. The background of the real scene is modified so as to create an artificial background with its perspective organization and with graphical objects. The person's image is immersed into the virtual ambiance where different events may be made happen by the person's behavior.

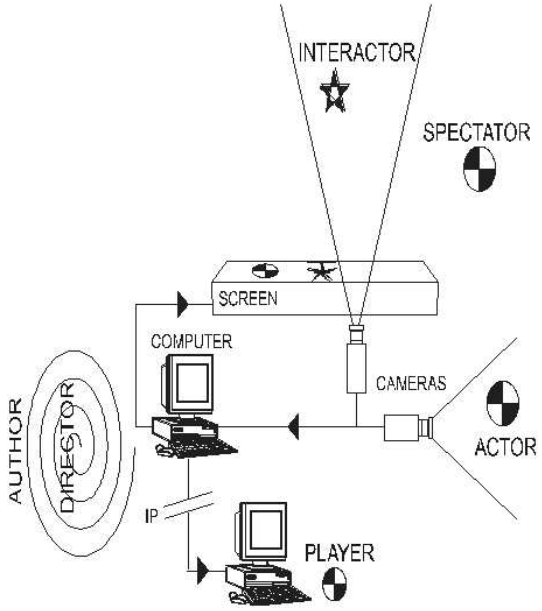

Figure 7.8: Users involved in the *Transfiction* immersive interactive narrative system [112].

provides an overview of the users involved within the designed interactive narrative architecture. The persons in the field of view of the cameras get themselves immersed within the virtual environment and are therefore involved within the narrative. An *immersive interactive mixed-reality environment* based on a *Magic Mirror* metaphor is created. The mixed-reality scene is in fact rendered on large screens facing the so called interactors who see their own images and those of other people in the field of view of the cameras embedded in the visual ambiance (Figure 7.9). The images can moreover be disseminated in real-time to the public through the Internet. Interactors as well as players behind their computer displays are offered to interact with the story. The word *Transfiction* [112] has been coined for this interactive narrative system, where users are "transported in fictional spaces".

The block diagram of the system's architecture is shown in Figure 7.10, which illustrates the different functionalities of the system's building blocks. Standards are used in order to implement an open and flexible architecture. MPEG-4 is used for the coding and transmission of the segmented objects, the author-prepared graphical material, and the descriptions of scenes associated with the narrative scenario, as well as for scene composition. MPEG-7 is used for the description of natural

Figure 7.9: Example of immersive interactive mixed-reality environment based on the *Magic Mirror* metaphor (courtesy of *alterface*). The persons in the field of view of the cameras get themselves immersed within the virtual environment and are offered to interact with the story.

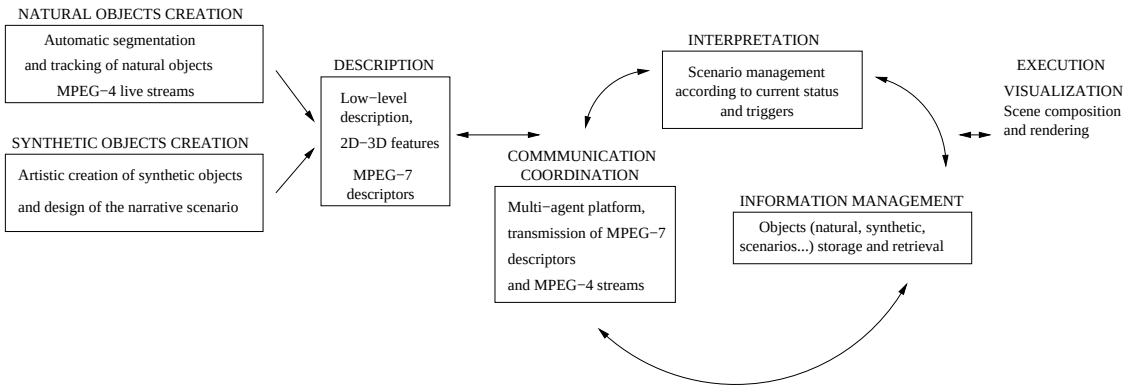

Figure 7.10: Overview of the architecture of the immersive interactive narrative system developed in the framework of the European project *art.live*.

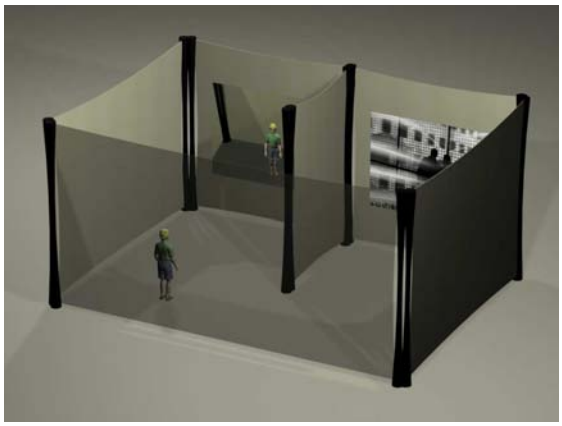

Figure 7.11: Immersive interactive gaming installation developed in the framework of the European project *art.live*. Two cameras side by side film two persons in front of two big screens. The persons see their own image and that of the neighbor immersed in a gaming scenario and are asked to play, collaborate or compete in the virtual space by means of their body movements.

and graphical objects as well as of triggers driving the scenario management. Figure 7.11 shows the *art.live* system implemented at the Royal Saltworks of Arc-te-Senans, in France, for the project's second trial*. Here, two cameras filming side by side two persons facing two big screens were used for transporting children into six gaming scenarios (an example of scenario is shown in Figure 7.12). The transfiction principle allows the players to be in the game and play by means of their body movements in a non-intrusive and seamless interaction.

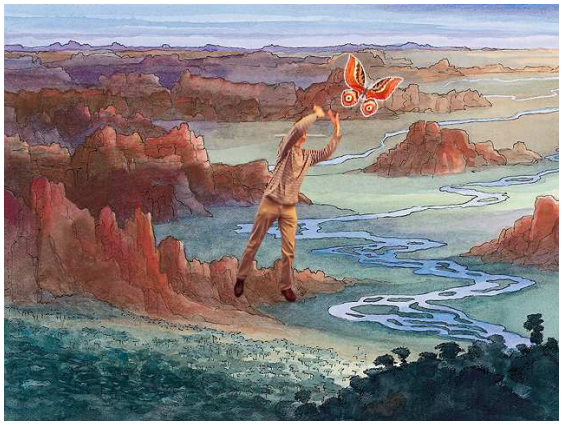

Figure 7.12: Example of mixed-reality scenario from the *art.live* interactive gaming installation. Video object segmentation allows to separate the images of the persons from the real background and to insert them in the graphical scene. Video object tracking and description allows to follow and describe persons movements which trigger events in the scenario and allow interaction. Shadow segmentation allows to improve object segmentation and helps the subsequent description, interpretation and visualization modules in the creation of the living narrative.

At the core of the creation of natural objects (see Figure 7.10), video object extraction techniques provide both immersion and interaction capabilities. Moving objects are automatically segmented

*http://www.transfiction.net/salineroyale/?lg=en

from the background and tracked in real-time. Descriptors are extracted which provide information about the position, the surface and the shape of the objects. They allow the subsequent modules to interpret users behavior and influence the narrative. In the image of Figure 7.12, for instance, the segmented image of one of the persons determines the displacements of the graphical butterfly while the image of the second person is visible and has to catch it.

The methodology for moving cast shadow segmentation proposed in this thesis has been successfully adopted and implemented in the framework of the *art.live* system. Its consideration allowed to eliminate shadows from object segmentation results and helped the subsequent description, interpretation and visualization modules in the creation of the living narrative.

7.4 Photorealistic scene composition

In the previous sections, we have shown the importance of segmenting shadows for an accurate extraction of video objects and an accurate visualization of mixed-reality scenes. Shadows were considered in those cases as a noise component to be taken into account and detected for its removal. In this section, we consider the perceptually informative role of shadows in visual scenes and demonstrate the importance of segmenting shadows for a more realistic visual content production.

As demonstrated in the previous section, one way new visual content can be created is by extracting natural objects from a scene and by composing a new scene with objects captured by different sensors and mixed with artificial elements. In television and film production, a commonly used technique for separating natural objects from the background and compositing a new, augmented scene is the *blue screen or chroma-keying* approach. While manual extraction of objects is required for high quality film production since a perfect definition of object boundaries is needed and temporal coherence has to be guaranteed, blue screens and chroma-keying allow a fully automatic extraction of natural objects thanks to a specific scene set-up. In the chroma-keying approach, the objects of interest are filmed in front of a uniformly colored background, which is usually blue or green. The extraction of objects is then performed by eliminating pixels having the known background color. Lighting is carefully controlled in order to avoid the effects of shadows cast by objects onto the background. An example of the use of chroma-keying in television studios is the production of weather bulletins. The anchorperson is filmed against the known background, its image is separated from it and placed over a background image representing the weather map.

Shadow effects are carefully avoided in the blue screen approach thanks to expensive controlled lighting or the use of specific background material and special cameras [55]. Illumination effects have however an important role in the perception of visual scenes and the fact of discarding them can limit the visual quality of the scene composition due to a lack of naturalness. Shadows cast by objects on a background surface can be in fact informative about the shape of the object, the shape of the background and the spatial arrangement of the object relative to the background. Among all these roles, it has been found [94] that cast shadows are perceptually most relevant for the recovery of spatial arrangement, especially when the shadow is in motion. Shadows are shown to be particularly salient cues to depth in dynamic scenes*. Techniques for the automatic identification of shadows cast by moving objects in real world conditions, without blue screen and ad-hoc lighting, allow then not only to cut production costs but also to increase the quality of the visual content created. When shadows cast by objects are explicitely segmented, they can be rendered in the composited scene and the overall quality can be improved by an augmented naturalness. The proposed tool for moving cast shadow segmentation can be therefore used to manipulate video content in a more perceptually relevant way.

*The perception of cast shadows is discussed in more detail in Section A.3.

(a) (b)

Figure 7.13: Sample frame from the test sequence used for video composition (a) and background image (b).

Figure 7.13 (a) shows a sample frame of a test sequence we have recorded with a digital camera in an ordinary room, where illumination is given by a table lamp and the light entering the room from the windows. We have extracted from the scene's background the moving object and moving shadow with the combination of the proposed shadow segmentation approach and the statistical change detector described in the previous chapter. We have then built a composited video emulating a weather forecast bulletin by placing the extracted object over a weather forecast map (Figure 7.13 (b)). Figure 7.14 (a) shows some sample frames of the obtained composited scene. The absence of illumination effects due to shadows gives the impression of a flat 2D scene. The human brain does not in fact receive strong cues to infer depth information in the scene.

For each point in the extracted shadow region we have then modified the color channels in the background image according to the measured decrease in RGB values due to the shadow in the original sequence. The comparison of the obtained composition in Figure 7.14 (b) with that obtained without considering the cast shadow in Figure 7.14 (a) demonstrates the perceptual importance of shadows. The hand appears now clearly positioned in a 3D space in front of the weather map and a photorealistic result is obtained, as if the hand had been filmed directly in front of the background image. The effect is more evident when the entire sequence is viewed since the object and shadow motion enhances the depth perception, in accordance with the conclusions of Mamassian in [94].

The described shadow manipulation technique can be directly applied in scenes, such as that discussed above, where the 3D geometry of the surface upon which the moving shadow is cast in the original video is the same as that of the surface upon which the shadow is cast in the artificial background. In addition to the discussed example, this is the case for rich media presentations and authoring tools such as that proposed by the *art.live* project, where the multimedia author creates the graphical background (see Figure 7.15) for the mixed-reality scene based on a 3D analysis of the visual scene obtained by means of camera calibration. The shadow segmentation technique proposed in this thesis can then be used to improve the impact of the created narratives by allowing a more realistic rendering of natural objects in the graphical backgrounds created by the multimedia authors.

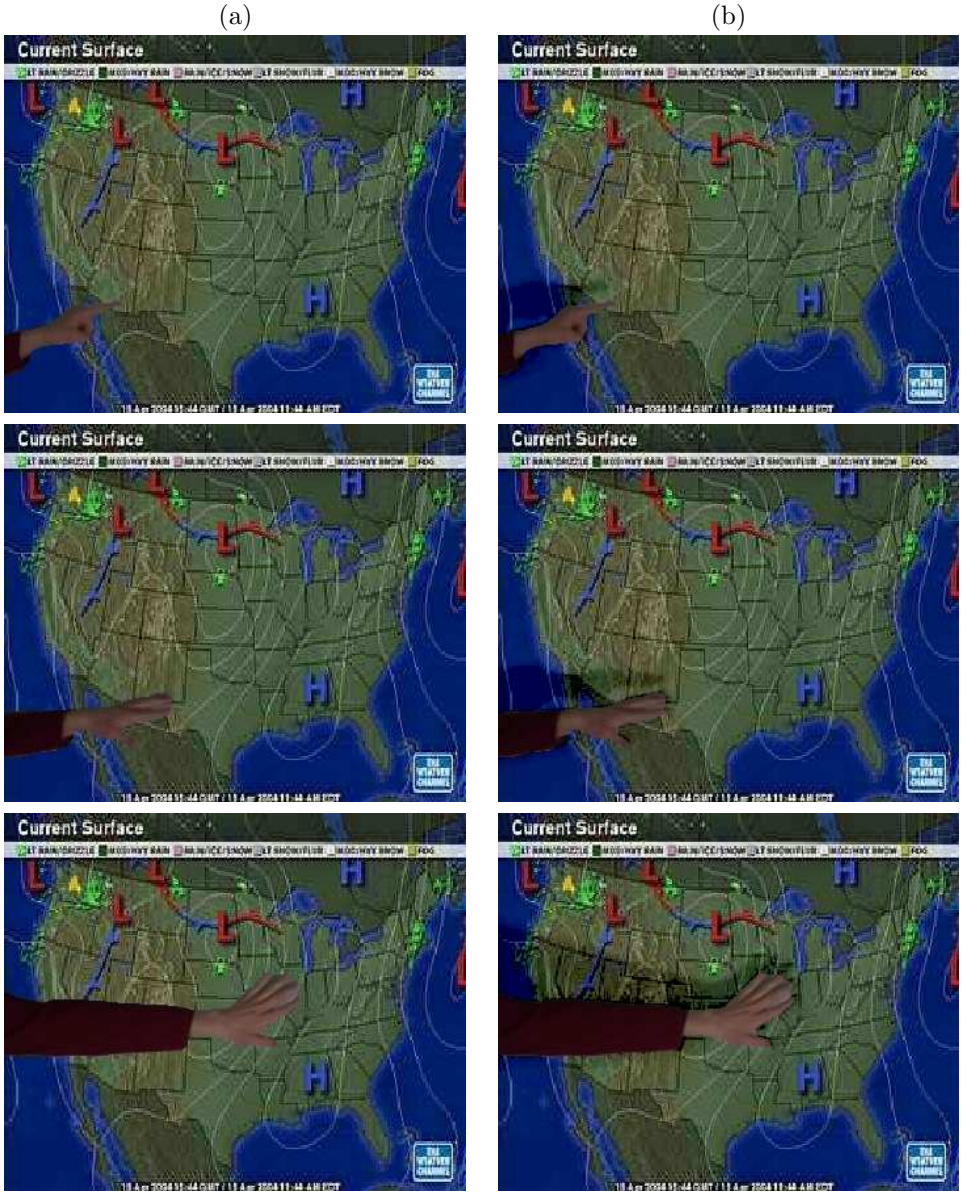

Figure 7.14: Photorealistic scene composition results thanks to shadow segmentation and manipulation. (a) Scene composition without considering the shadow; (b) with the shadow.

Figure 7.15: Graphical background fitting a 3D perspective for the creation of immersive narratives within the European project *art.live*.

7.5 Summary

In this chapter, applications of the proposed shadow segmentation method are presented in which it is possible to appreciate the benefits introduced by shadow detection in video processing. They aim at demonstrating that the tools developed in this thesis allow a more effective treatment of visual information.

The first application is the elimination of shadows for improving the accuracy of video object extraction, tracking and description operations, which are the first fundamental steps in a wide range of object-based applications. In particular, the encouraging results obtained in the framework of object segmentation for interactive immersive environments within the European project *art.live* demonstrate the reliability of the proposed methodology.

The second application of the proposed approach is the manipulation of shadows for video composition without the use of studio equipment. The rendering of shadows in a composited scene make one perceive a spatial relationship between the scene components which increases the naturalness of the result.

Figure 7.16: Shadows cast by a person's own body parts can bridge the gap between personal and extrapersonal space (Section A.4).

8 Conclusions

8.1 Summary of achievements

This thesis has discussed the problem of extracting shadows from images and image sequences. Shadows are a frequent occurrence in natural scenes and they represent an important element of visual information. In many image analysis applications, such as video surveillance and immersive gaming, shadows interfere with fundamental tasks such as object extraction, tracking and description. They have therefore to be identified and eliminated. In different applications, such as video production and mixed-reality, on the other hand, shadows, when taken into account, can increase the naturalness of visual scenes.

In literature, a list of cues has been identified on the basis of which an observer recognizes shadows in a visual scene. The list comprises spectral cues related to the brightness and color of shadows, geometric cues which relate a shadow to the shadow casting object, the casting surface and the light source whose light is occluded, and temporal cues regarding the motion of a shadow with respect to the motion of the shadow casting object and the light source. Among these cues, the most significant ones for the purpose of designing a fully automatic shadow segmentation approach have been selected and investigated. Significant, in the context of this thesis, means that they can be exploited basing only on image-derived information and with a limited number of assumptions about the scene. A shadow segmentation method that is general and flexible is in fact highly desirable for image analysis applications that deal with scenes whose content is not known a priori, as for instance video editing and visual surveillance.

The selected shadow properties which provide the major amount of information for shadow segmentation in digital images are those related to the brightness and color of shadows. Due to the nature of the problem, that is very underconstrained, it is nevertheless important to exploit also the other cues. The methodology for shadow segmentation developed in this thesis checks therefore, in addition to shadow spectral properties, the position of shadows with respect to the shadow casting objects and the temporal coherence of shadows. These properties can be defined without any knowledge of the structure of the objects and of the scene. The combination of multiple constraints based on shadows spectral, spatial and temporal properties represents an element of originality of

the proposed method with respect to the state of the art.

With regards to the use of color information for the analysis of shadows, different solutions have been proposed in literature, with regard to both the physical models of shadows adopted and to the color features exploited. This thesis has provided an original complete picture of the existing solutions with respect to this issue, having pointed out the fundamental assumptions, the adopted color models and the link with research problems such as computational color constancy and color invariance. Since the problem of shadow detection is now clearly defined with respect to such problems, the benefits of advances in those research areas could then be easily exploited by novel shadow detection methods. From the analysis, a model of shadow that is common to all shadow detection approaches that are fully automatic and do not require active processes (e.g. camera calibration and user intervention) has emerged. The model is only implicit in the majority of methods that use the invariance properties of some color transformations in presence of shadows. It has been made explicit and it has been used in the proposed method.

On the basis of the discussed theoretical background, a new analysis method for the segmentation of cast shadows has been proposed. The validity of the approach has been demonstrated through two implementations, one for the segmentation of moving cast shadows in video sequences and one for the segmentation of cast shadows in still images. As the problem of separating moving cast shadows from moving objects in image sequences is particularly relevant for an always wider range of applications, from video analysis to video coding, and from video manipulation to interactive environments, particular attention has been dedicated to the segmentation of shadows in video.

As above-stated, the proposed method exploits three sources of information, namely spectral, spatial and temporal properties of shadows. An initial shadow hypothesis is formulated on the basis of color analysis. The RGB color space as well as photometric invariant features are considered to this end. The photometric invariant features used in the proposed method have been selected basing on an extensive analysis of their behavior in presence of shadows in real images and image sequences. They are different from those typically used in literature and this represents an element of originality of the proposed approach. The initial shadow hypothesis then undergoes a spatio-temporal verification stage which allows to refine the segmentation results and improve the overall system's performance. The spatial analysis does not make any assumption about scene geometry nor about object shape. It tests the position of each hypothesized shadow with respect to the shadow casting object and allows to provide a first refinement of shadow segmentation results. The object is automatically extracted by means of a statistical model-based change detection algorithm. The temporal analysis is based on a novel shadow tracking technique. Shadow tracking was not previously addressed in literature. Based on tracking results, a temporal reliability estimation of shadows is derived which allows to discard shadows which do not present time coherence. The use of a temporal reliability estimation for improving the accuracy of shadow segmentation results is also an element of originality of the proposed approach.

The proposed approach has been evaluated by means of subjective and objective quality assessments on a wide variety of video data. It has been shown to achieve accurate segmentation results in different kind of scenes representing real world indoor and outdoor environments. The achieved generality has been demonstrated by the method's capability of dealing with different types of objects, such as vehicles and people, different types of shadows, such as strong and diffuse shadows, and different types of background geometry, such as planar and curved, horizontal and vertical surfaces. Robustness to different physically important independent variables, such as type of illumination and surface type upon which shadows are cast, has moreover been demonstrated. The integrated use of appropriate constraints derived from the analysis of the nature of shadows has allowed the proposed technique to achieve an improvement with respect to the state of the art, as demonstrated by the

comparison with different existing techniques.

Examples of application of the proposed shadow segmentation tool to the enhancement of video object segmentation, tracking and description operations, and to video composition without studio equipment, have demonstrated the advantages of a shadow-aware video processing. The application of the proposed approach in the framework of the European project *art.live,* whose achievements have been demonstrated in public trials, has moreover proved its reliability.

8.2 Perspectives

The modular structure of the proposed approach to shadow segmentation makes it particularly suitable for extending its capabilities and performances. Each level of analysis can be independently extended and improved. Some directions for further work are proposed below.

- In the developed implementation of the proposed method for shadow segmentation in video, a static camera has been assumed. This scenario is valid for many applications, such as video surveillance and immersive interactive environments. One natural extension of this work is to deal with *moving camera sequences*, by integrating the global motion information.

- The analysis of the behavior of different color invariant features has highlighted the trade-off between the degree of invariance and the inherent instability of the features in certain regions of the color space. In the color analysis stage, a *selective use of color invariants* could be envisaged. Different features could be selected according to the color content of the different parts of the image. Hue could be used in areas with highly saturated colors and normalized rgb or $c_1c_2c_3$ features in the rest of the image.

- Developments in color invariance allowing improvements with respect to stability and discriminative power of photometric invariant features could improve the performances of the system. New invariant transformation could simply replace the features used in the current implementation without the need of modifying the system. Moreover, in case the targeted framework allows it, i.e. when control on the camera is considered, different invariant derivations, such as those proposed in [35] and [95], could be considered.

- The proposed approach addresses applications that use a monocular camera. For applications that employ stereo or multiple cameras, the segmentation algorithm could be extended to exploit further constraints based on *homography and 3D geometric analyses.* This would allow to improve the performances of the spatial analysis stage.

- The proposed methods for shadow segmentation in *still images* are based on the use of *edge detection.* Due to the uncertainty in defining the shadow lines when shadows present a very diffuse penumbra, more sophisticated operators than that used in this work, such as that proposed in [176], could be used to extend the proposed method.

- In the evaluation of the proposed approach, two different sets of parameters have been used, one set for indoor and outdoor overcast scenes and one set for outdoor sunny scenes. They have been set based on extensive experiments on different test sequences. To automatically adapt the parameters to the different type of illumination conditions, it would be interesting to define a learning process based on an objective evaluation criterion.

Appendix

Shadows: from art to neurosciences

"Quam multa vident pictores
in umbris et eminentia
quae nos non videmus"

Cicero, Accademica II.20,86

As human beings, we process and interpret visual information to find our way in the world without conscious awareness of such processing. We all see the same world around us, but, as the amount of information that our eyes transmit to the brain is huge, our perception must be selective. We thus develop the ability to select and focus our perceptive attention on specific aspects of the surrounding environment of which we would otherwise be unaware. In the reported quotation, for example, Cicero underlies that artists "see in shadows and protuberances much more than we see" as they are trained to observe the world in order to reproduce it. Similarly, when carrying out a research into a certain problem we are necessarily induced to be selective and focus our attention on some among the possible aspects of the problem. Different points of view about the problem correspond to very different selectivities. Keeping an eye on such different approaches to the problem, even in domains that are far from that in which the research takes place, can be source of new ideas and intuitions.

The problem we have tackled in this work is that of defining an automatic analysis method for the identification of shadows in digital images and image sequences. At the root of such problem lies a question: *What is a shadow?* We have answered to this question by describing shadows as a physical phenomenon caused by objects which obstruct light from a source of illumination. As such, they can be characterized by means of the laws of physics and optics. This is however only one of the possible answers to the above-mentioned question. When examining them thoroughly, shadows reveal themselves as an extremely complex object. By urging our curiosity beyond the limits of our specific problem, we discovered that the investigation on the nature of shadows is becoming in recent years the central point of a very lively dialog among researchers from a number of different

domains.

Papers about the mechanisms behind shadow processing in the human visual system, published in journals such as *Nature* [120], *Perception* [81], *Trends in Cognitive Science* [94], and *Vision Research* [18], have contributed to make this topic of great impact within scientific research areas such as neuroscience, experimental psychology and vision. In parallel, prominent figures within the philosophical and history of arts domains have published interesting essays and books on shadows and illumination [10, 17, 54, 154]. Most of these works explicitly refer to the link between the artistic and scientific investigation of the nature of shadows. Experts in different fields are currently working to create opportunities for an active interaction between artists and scientists to discuss such an important aspect of our visual experience. A first international symposium* has been organized in November 2003 to provide an opportunity to share scientific results, ideas and experiences.

The aim of this appendix is to present some of the results and ideas emerging from the multidisciplinary discussion around the nature of shadows. Such results, even if not directly related to the research domain of this thesis and coming also from fields that are usually not considered in scientific investigations, allow a better understanding of the shadow phenomenon and can provide hints for further research.

The presentation is organized in four sections, which span different domains from art to neurosciences. In Section A.1 we have chosen a couple of representative examples from a book by Stoichita [154], an art historian who has investigated the representation and the significance attributed to shadows in the history of Western art from the origin of Painting to the present time. We introduce moreover the work of a contemporary artist whose research focuses on the development of shadows as a creative medium. In Section A.2, we present some findings about how the idea children have of shadows evolves over time. It is only at the age of eight, nine that our reasoning about shadows becomes purely geometric and we explain them as caused by objects which occlude light from a source of illumination. The long apprenticeship required to develop a correct cognition of shadows explains the difficulty of dealing with them for human observers and for machines. The processing of shadows in the human visual system and the ability of human observers to use information from shadows to arrive at an understanding of a visual scene are important issues related to the work presented in this thesis. Their understanding could in fact allow a better use of shadows in the creation of visual content mixing objects from different sources. In Section A.3, we provide then an overview of the works focusing on the role of shadows in visual perception. In Section A.4, we finally present an interesting study of Pavani and Castiello. In [120], they show that cast shadows could also provide cues about body position in relation to objects in the world, enhancing the ability to interact with objects in real as well as virtual environments.

A.1 Shadows and art

According to art historians, shadows have always played an important role in arts. In his book *A Short History of the Shadow* [154], Stoichita explains how the origin of Painting is related to shadows and how cinematography used shadows as a new form of expression.

A.1.1 Mythological shadows

The myth of the origin of Painting is reported by the Latin author Pliny the Elder in his *Naturalis Historia*. According to Pliny, Painting was born the first time the shadow of a man was outlined on a wall. The picture of Figure A.1 shows an illustration of Pliny's myth by the German painter

*http://www.unitn.it/convegni/neuroscienze.htm

Figure A.1: Eduard Deage, *The origin of Painting*, 1832.

Eduard Daege (1832): a woman, the daughter of Butades, potter from Sycion, was in love with the young man; when he had to leave the country, she fixed on the wall the contour of his shadow.

According to Stoichita, the first actor in Pliny's myth is nature, which, by projecting the shadow, reduces the three-dimensional world to a bidimensional image. Art does not originate therefore from the direct observation of the world but from a copy (the outlined shadow) of a copy (the shadow) of reality.

In his analysis, Stoichita relates moreover Pliny's myth to Plato's myth of the cave which marks the origin of the theory of knowledge in Western culture. In his *The Republic*, Plato relates of a cave where some prisoners have been living since their infancy in chains. They are forced by chains to look at the wall on the opposite side of the cave's entrance and they cannot turn their head around. Therefore, they consider the shadows projected by the external world on the cave's wall as real until they are allowed to turn their heads and to recognize they have made a mistake. Shadows represent for Plato the first fake semblances of reality and the starting point of the path toward real knowledge.

According to Stoichita's analysis, both art and knowledge originate from shadows.

A.1.2 Cinematographic shadows

Shadows assume during the entire Expressionist period (1919-1933) a fundamental role in cinematography. Stoichita analyzes two famous frames taken from two masterpieces of German Expressionist cinema, *The Cabinet of Dr. Caligari* (*Das Kabinett des Doktor Caligari*), realized between 1919 and 1920 by Robert Wiene and Willy Hameister, and *Nosferatu, A Symphony of Horror* (*Nosferatu, eine Symphonie des Grauens*), realized in 1922 by Friedrich Wilhelm Murnau. In an Expressionist film, each frame is conceived in such a way that it refers by analogy or by contrast to the entire film. It is then correct and indeed useful to analyze a single image as is done for paintings, something that

Figure A.2: Robert Wiene and Willy Hameister, frame from *The Cabinet of Dr. Caligari*, 1919-1920.

Figure A.3: Friedrich Murnau, frame from *Nosferatu, A Symphony of Horror*, 1922.

should not be normally done for films. All the more that Murnau and Wiene declare openly to be indebted to the past's Painting.

The chosen frame from *The Cabinet of Dr. Caligari* is shown in Figure A.2. It shows the Doctor on the left hand side of the image and the giant projection of his shadow on the right hand side. The shadow is much bigger than the character. It serves in fact to reveal his inner being, as if the movie camera was able, through the shadow, to plunge in the consciousness of the character and to project his internal mental states on the wall. The wall has then the role of a second screen and the projection of the shadow is a metaphor of the film creation. The shadow shows what is happening inside the character, that is what the character is. Stoichita makes us notice the contrast between the Doctor's attitude and his shadow. The shadow is emanating directly from the character but it is distorted and projects his psyche on the screen. The stress laid on the hand which represents action aims at conveying the idea that the shadow itself can be an active, evil instrument. What we see in this frame is the embodiment of the phantoms of the narrator of the entire story, Francis, a mad

Figure A.4: Larry Kagan, *At*.

man. It is just a "projection", an illusion, and nothing more. In the whole film Wiene is playing with the fact that the narrator is a metaphor for a film director and the projection of the shadow is, as stated above, a metaphor for the film creation which reveals its power to trick.

In the very famous frame from *Nosferatu* shown in Figure A.3, the shadow has a more delicate function. Does the silhouette of the vampire represent its shadow or the vampire himself? The deformation of hands and arms seem to support the first hypothesis but Murnau and the spectator know that, according to an old tradition, vampires do not project shadows. The only possible answer is then that the silhouette is Nosferatu himself, a kind of octopus with tentacles, translucent, almost a phantom. He lives in an underground universe, full of doors, corridors, and stairs, whose structure has been compared to that of the unconscious according to Freud. The film director then acts as "a man who shows the shadows", that is who reveals the obscure content of consciousness. The analogy between shadow and film frame is made clear to the spectator only at the end of the film, when the first ray of sunshine reaches Bremen and disintegrates Nosferatu and, most of all, when light floods into the movie theater and the screen becomes white again.

A.1.3 Wall sculptures in steel and shadow

Figure A.4 and Figure A.5 show two works by Larry Kagan*, a contemporary artist that uses shadows as an art medium. The works are created by casting light through the contours of a steel wire sculpture that protrudes from the wall. At first glance, Kagan's sculptures seem to be a mass of jumbled steel. Lit from above, however, their shadows are shaped into well-articulated sketches of everyday life.

In Kagan's art, the wires hold the critical information in an encrypted, deconstructed way that is then reassembled by light. The artist explains this concept by drawing an analogy with the delivery of information through an e-mail. The message in the e-mail gets broken up into chunks and distributed to different routes. An algorithm then reconnects the pieces to be delivered as one

*http://www.arts.rpi.edu/~kagan/

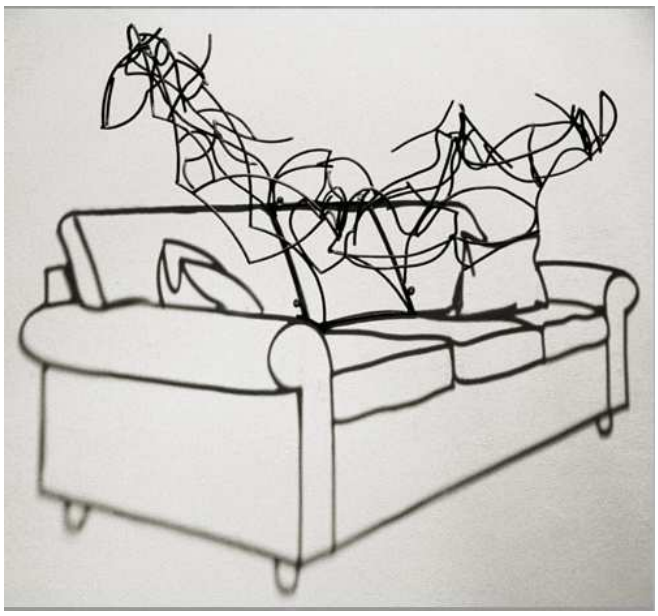

Figure A.5: Larry Kagan, *Couch*.

neatly packaged message to another's inbox. In the case of his steel-and-shadow sculptures the wires present the paths distributing chunks of information. The algorithm is the light, and all of a sudden the pieces connect and make sense.

A.2 Shadows and psychology

Between the two World Wars, the Swiss psychologist Jean Piaget (1896-1980) initiated the research about children's cognitive development. In 1927, he published in his *The Child's Conception of Physical Causality* (*La causalité physique chez l'enfant*), among other results of his studies, his findings about children's cognition of shadows. Roberto Casati in his book *The discovery of the shadow* [17] provides an overview of Piaget's and later experiments about children's ability to anticipate and explain objects' cast shadows. We summarize them here.

A.2.1 Baby shadows

To find out what children think about shadows Piaget interviewed them. Children's aged between five and ten years took part in his study. The reported answers are subtle and inventive. The shadow of a hand is dark because the hand has bones, says for instance Gall. Shadows are created by the night for Tab. For Roy, the shadow is a substance which takes up space and is impermeable to light.

From the analysis of children's answers, Piaget concluded that their understanding of shadows can be characterized by four stages. In the first stage, beginning from the age of five, the shadow of an object emanates from the environment's shadows. The night or the darkness of a room corner are some kind of "black clouds" of which shadows are made of. In the second stage, around the age of six, seven, shadows do not appear any more as emanated by the night but rather by the shadow-casting object. Consequently, the child is not able to anticipate the position of the shadow with respect to the light source. He tries to make the shadow rotate in the room by turning right

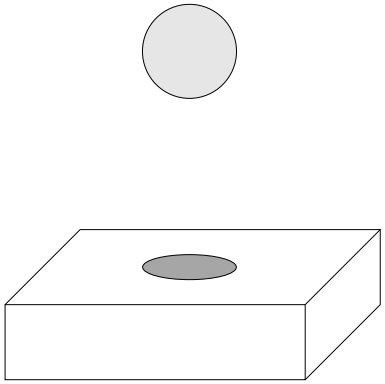

Figure A.6: The sphere, the shadow and the box can move independently. First experiment: in the unnatural situation, the shadow moves when the sphere stays still; in the natural situation, the shadow moves when the sphere moves. Second experiment: in the unnatural situation, the shadow moves when the box moves; in the natural situation, the shadow stays still below the sphere when the shadow moves.

round instead of going round the lamp. In the third stage, around the age of seven, eight years, children discover that a shadow has a geometric relationship with the light source, but they do not yet understand the cause-effect relationship between the emitted light and the shadow. They believe for instance that an object casts a shadow also during the night. In the fourth stage, around the age of eight, nine, the explanation of shadows becomes purely geometric and corresponds to that of adults: shadows are caused by objects which occlude light from a source of illumination.

Piaget's experiments were replicated in more recent years by the psychologist Rheta DeVries with a greater number of children. The experimental methodology in this case was different from the interviewing technique of Piaget. Children were involved in a game inside a room where a mobile lamp was present to create shadows on a wall and various objects were made available to children to cast shadows. Children were asked to perform different tasks, such as making their own shadow bigger than that of the toys, making their shadow touch that of the experimenter, or making their own shadow move or disappear. They were moreover asked many questions aimed at verifying if their ideas where coherent. As Piaget, DeVries found that the idea children have of shadows changes over time. However, she did not find that shadows are part of the night.

The reported experiments and results tell us something about what children say about shadows or what researches have been able to deduce from contradictions in their answers. Do these concepts derive from what children have heard from adults and what they have learned when playing, or are they related to previous phases of their cognitive development? Gretchen Van de Walle, Jayne Rubenstein and Elizabeth Spelke have conducted some experiments with children aged only some months, that is children that are not yet able to talk.

In the Sixties, a methodology has been introduced which allows to study the mental universe of children from their first weeks. By means of this methodology, researches have discovered that newborns know a lot of things. One is induced to think that they have a theory of the world or perhaps a battery of mini-theories, one for each object. What about shadows? Do newborns have a theory of shadows? Van de Walle, Rubenstein and Spelke made two experiments with respect to this question with children at ages between five and eight months. They showed to children a sphere floating in space above a box on which it casts a shadows (See Figure A.6).

The first experiment tried to understand if infants found more surprising the situation where the shadow moves while the sphere stays still or the natural situation where the shadow moves when

the sphere moves. On the contrary to adults, newborns prefer the unnatural situation where the sphere stays still. An hypothesis which could explain this preference is that the natural motion of the shadow violates one of the principles of the physical theory of children which states that objects do not act at a distance. The shadow should not move because it is attached to the box and not to the sphere. The second experiment tried to understand if newborns found more surprising the situation where the box moves and the shadow stays still below the sphere, as it is natural, or the situation where the shadow moves together with the box, as if it was attached to it. As before, the natural situation is more surprising for children.

On the basis of these experiments it can be arguably said that newborns do not have a theory of shadows. They treat shadows as objects, probably because this requires little effort, since they already have a theory for objects. A theory of shadows comes into play when children understand that shadows do not behave as objects do.

A.3 Shadows and vision

We have identified two main threads of research about shadows in visual perception and cognitive neuroscience. The first includes works on the *informational structure of cast shadows*, that is on the information they offer to the human visual system for the interpretation of a visual scene and the use that human observers make of this information. The second thread comprises studies on the role of both self and cast shadows in the *recognition of objects*. In this section, we present an overview of the main results of these two threads.

A.3.1 Shadows in the brain

The perceptual interpretation of cast shadows has been thoroughly investigated by Mamassian, Kersten and Knill [81, 94]. Cast shadows are related to two distinct surfaces, the surface of the casting object and the surface on which the shadow is cast. Cast shadows are therefore potentially informative about the shapes of both the surfaces and about the spatial layout of the scene, that is about the spatial relationship between surfaces. As the result of numerous psychophysical experiments, the above-mentioned authors have found that the human visual system does not effectively use cast shadows as cues to surface shape, despite the potential reliability of the information they provide [83]. This seems to hold for both static and moving cast shadows. Cast shadows are shown, on the other hand, to be very salient cues to the spatial layout of objects in a scene, especially in dynamic scenes.

A detailed investigation of this issue has been carried out by the authors in [81]. They show that moving cast shadows provide to the human visual system a robust source of information about depth that is resistant to conflicting cues and high-level knowledge. Shadow motion induces in fact apparent motion of an object in depth even when the object's size does not change and the object does not move. Shadow motion can override these cues which suggest the stationarity of the object. This effect is illustrated in Figure A.7. If the object's image keeps its size fixed but the object moves in the image plane, cast shadow motion still induces apparent motion of the object in depth, as illustrated in Figure A.8. The effect of apparent motion in depth is moreover resistant to changes in contrast, opacity and even some significant deformations in the shape of the shadow. Replacing the ellipsoidal shadow of the ball in the scene of Figure A.8 with a square shadow, for example, did not reduce the effect.

The motion of a cast shadow is inherently ambiguous. If Figure A.7 is considered, it is clear that the location of the shadow cast by the square on the background could be the results of an

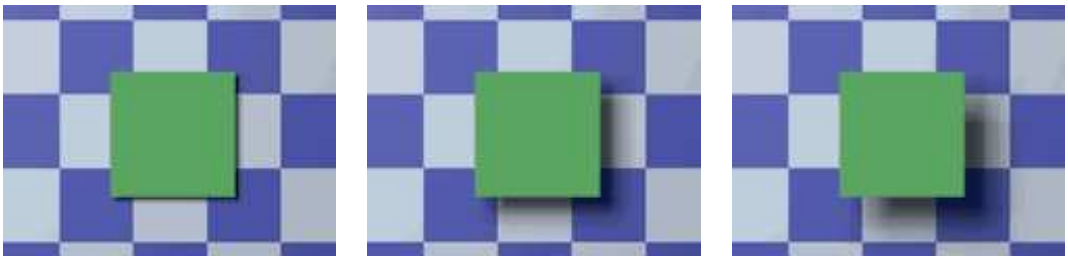

Figure A.7: Increasing the displacement between the cast shadow and the foreground object induces an impression of increasing depth relative to the background (images from [94]).

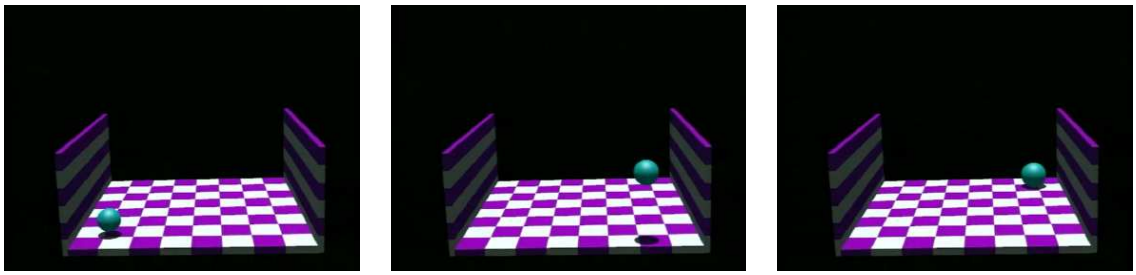

Figure A.8: When the shadow trajectory is horizontal, an impression of seeing the ball rise above the checkerboard floor is induced. When the shadow trajectory matches that of the ball, the ball appears to recede smoothly in depth along the floor (images from [94]).

infinite number of combinations of the positions of the viewpoint, the light source, the object and the background surface. The fact that a displacement of the cast shadow produces a percept of a square moving in depth indicates that the human visual system uses some a priori constraints to resolve the ambiguity in the cast shadow information. Kersten, Knill and Mamassian found that the experimental results are consistent with the hypothesis that the visual system assumes light sources and background surfaces to be stationary in the scene.

The effects of shadows on the performance of object recognition tasks have been studied by Braje, Legge, and Kersten [15] and by Castiello and colleagues [18, 19]. Braje, Legge, and Kersten have explored the effects shadows have on the recognition of natural objects such as fruits and vegetables. They found that recognition performance was not affected by the presence of shadows. In [18], Castiello investigated whether recognition performance of familiar objects other than fruits and vegetables is sensitive to different features (presence, position, and shape) of both naturally cast and artificially attached shadows. A general increase in response time was found when naming objects in incongruent shadow conditions, that is when the object was presented in conjunction with a shadow that originated from a different object. Overall, these studies indicate that humans can either marginalize the effects shadow have so that object processing is invariant across different shadow conditions, or be affected by shadows in object recognition tasks when specific shadow manipulations are performed.

The work of [18] has been extended in [19]. The first aim of this study is that of investigating to what extent the processing of shadows during an object recognition task occurs without conscious awareness. Patients with visual neglect are tested to this end. Visual neglect is a neurological phenomenon which is observed after the occurrence of lesions in various regions of the brain, but especially those involving the right parietal lobe. Visual neglect refers to the defective ability of

patients with unilateral brain damage to attend to the side of space contralateral to the lesion, and to report stimuli presented in that portion of space. Neurological evidence was found that neglect patients were still able to process shadows to optimize object shape perception, that is shadow processing is outside conscious awareness. The second aim of the study is that of determining the locus or loci of shadow processing within the human brain. This could allow to better understand the nature of the mechanisms underlying the ability to recognize objects under different illumination conditions and the residual mechanisms that allow patients with lesions in areas that may be critical for object recognition to preserve some ability to interact with environment. The results suggest that the link between object and shadow shape may occur within the brain's temporal lobe.

A.4 Shadows and neuroscience

As discussed in the previous section, shadows help our visual system decide about spatial relationships between objects and their movement and play a role in the recognition of objects. A very recent study has shown that, in addition to their effects on visual perception, cast shadows could provide cues about body position in relation to objects in the world, enhancing the ability to interact with objects in real as well as virtual environments. We present these findings here.

A.4.1 Near my shadow, near my body

Artificial body parts, such as sham arms, or repeated tool use alter the perceived position of the body in space. In relation to these situations, it has been shown that the internal representation of the body's spatial extent, the so-called *body schema*, can extend beyond the physical limit of the skin. The recent findings of Pavani and Castiello [120] indicate that body schema can also extend to incorporate shadows cast by an individual's body parts.

In their experiments, the authors tested ten individuals in a visuo-tactile interference experiment. They placed stimulators on the thumbs and index fingers of subjects and asked them to indicate via foot levers when a particular digit was being touched (see Figure A.9). When people are asked to discriminate a touch on the thumb or index finger, visual distracters, such as flashes of light, near the location of the touch, that is near the hand, are known to increase reaction times and error rates. This is what is called visuo-tactile interference. It happens because the subject is busy processing two separate inputs from the same region of the brain's body map. Pavani and Castiello's tests studied then whether visual distracters near the hand's shadow have a similar effect.

They carried different experiments where participants saw the shadow cast by one of their hands projected on a table surface (Figure A.9 (a)), the polygonal shadow cast by a shaped glove they were wearing (Figure A.9 (b)) and a line drawing silhouette of a hand (Figure A.9 (c)), respectively. In the first case, they observed that visuo-tactile interference was significantly stronger when stimulations were presented at the hand casting the shadow. This suggests that the hand shadow bound visual distractors in extrapersonal space to touches presented at the hand.

In the case of participants wearing shaped gloves projecting a polygonal shadow near the visual distracters the visuo-tactile interference was almost the same as the one with the hand not casting a shadow. In this case, the polygonal shadow movements were in accordance with the hand's movements but the shadow had no resemblance to a hand. The result then suggested that merely seeing a shadow stretching out of the body is not sufficient to produce a personal-extrapersonal binding.

In the last case, the silhouette mimicked the shape of the hand shadow while bearing no resemblance to a shadow. In addition, no real shadows of the hand were visible. Also in this case,

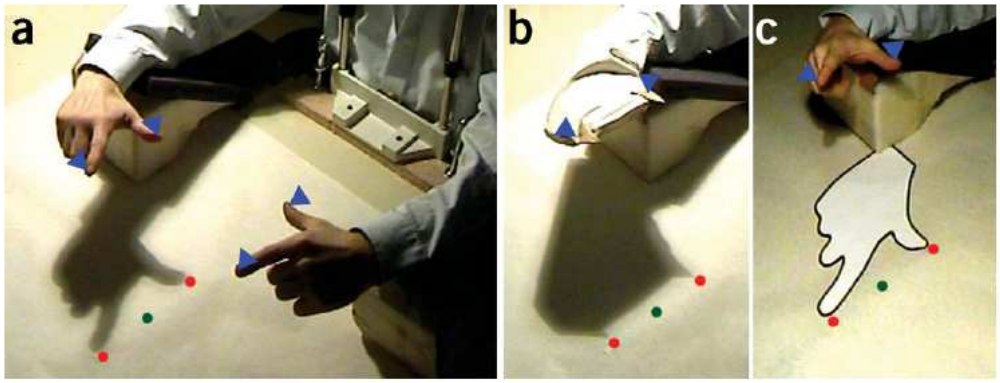

Figure A.9: Experimental setup for Pavani and Castiello's experiments (image from [120]). Participants sat with their chin on a rest, fixating on a green LED on the table surface (green central circle). The blue arrows indicate the electromagnetic stimulators attached to the fingertips. Visual distracters consisted in three successive flashes delivered by a pair of red LEDs (red side circles). In **(a)** participants saw the shadow cast by one hand; in **(b)** the polygonal shadow cast by a shaped glove; in **(c)** a line drawing silhouette of a hand.

visuo-tactile interference was not significantly different with respect to the case of the hand with no silhouette. Unlike in the first experiment, the polygonal shadow and the silhouette did not differentially affect touch discrimination performance at one hand as compared to the other.

In a final experiment, Pavani and Castiello compared the magnitude of the interference effect when the hand shadow was cast near the visual distracters with the interference observed when either the left or the right hand was physically near the distracting light, that is resting on the table surface immediately adjacent to the distracting lights. In this case, visuo-tactile interference was larger at the hand physically near the distracting lights than at the hand casting its shadow near the distracters. Although the body schema can extend to incorporate body shadows, the actual boundaries of the body remain understandably more relevant for estimating peripersonal space.

The authors suggest that body shadows may represent a new means for investigating the relationship between dynamic coding of peripersonal space and the control of action. They conclude that shadows cast by a person's own body parts can bridge the gap between personal and extrapersonal space.

Bibliography

[1] T. Aach, A. Kaup, R. Mester (1993). Statistical model-based change detection in moving video. *Signal Processing,* **31**:165–180.

[2] B. Abreu et al. (2000). Video-based multi-agent traffic surveillance system. In *Proc. of IEEE Intelligent Vehicles Symposium*, pp. 457–462, Detroit, USA.

[3] D. A. Adjeroh, M. C. Lee (2001). On ratio-based color indexing. *IEEE Transactions on Image Processing,* **10**(1):36–48.

[4] A. A. Alatan et al. (1998). Image sequence analysis for emerging interactive multimedia services–the European COST 211 framework. *IEEE Transactions on Circuits and Systems for Video Technology,* **8**(7):802–813.

[5] D. Alleysson, S. Süsstrunk, J. Hérault (2002). Color demosaicing by estimating luminance and opponent chromatic signals in the Fourier domain. In *Proc. of IS&T/SID 10th Color Imaging Conference*, pp. 331–336.

[6] N. Asada, A. Amano, M. Baba (1996). Photometric calibration of zoom lens system. In *Proc. of IEEE Int. Conf. on Pattern Recognition (ICPR)*, vol. 1, pp. 186–190.

[7] K. Barnard, G. Finlayson (2000). Shadow identification using colour ratios. In *Proc. of IS&T/SID's 8th Color Imaging Conference: Color Science, Systems and Appl.*, pp. 97–101.

[8] K. Barnard, L. Martin, A. Coath, B. Funt (2002). A comparison of computational color constancy algorithms I: methodology and experiments with synthesized data. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **11**(9):972–984.

[9] K. Barnard, L. Martin, A. Coath, B. Funt (2002). A comparison of computational color constancy algorithms II: experiments with image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **11**(9):985–996.

[10] M. Baxandall (1995). *Shadows and Enlightenment.* Yale University Press.

[11] M. Bejanin, A. Huertas, G. Medioni, R. Nevatia (1994). Model validation for change detection. In *2nd Int. IEEE Workshop on Applications of Computer Vision*, pp. 160–167.

[12] A. Bevilacqua (2003). Effective shadow detection in traffic monitoring applications. *Journal of the 11th Int. Conf. in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG),* **11**(1):57–64.

[13] D. Beymer, P. McLauchlan, B. Coifman, J. Malik (1997). A real-time computer vision system for measuring traffic parameters. In *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 495–501.

[14] J.-Y. Bouguet, P. Perona (1998). 3D photography using shadows. In *Proc. of the 1998 IEEE Int. Symposium on Circuits and Systems (ISCAS)*, vol. 5, pp. 494–497.

[15] W. L. Braje, G. E. Legge, D. Kersten (2000). Invariant recognition of natural objects in presence of shadows. *Perception,* (29):383–398.

[16] C. Bräuer-Burchardt (2001). Detection of strong shadows in monochromatic video streams. In *Proc. of 13th Scandinavian Conference on Image Analysis (SCIA)*, vol. 2749 of *Lecture Notes in Computer Science*, pp. 646–653, Springer Verlag, Berlin.

[17] R. Casati (2003). *The shadow club.* A. A. Knopf, New York.

[18] U. Castiello (2001). Implicit processing of shadows. *Vision Research,* **41**:2305–2309.

[19] U. Castiello, D. Lusher, C. Burton, P. Disler (2003). Shadows in the brain. *Journal of Cognitive Neuroscience,* **15**(6):862–872.

[20] A. Cavallaro, T. Ebrahimi (2001). Video object extraction based on adaptive background and statistical change detection. In *Proc. of SPIE Visual Communications and Image Processing (VCIP)*, pp. 465–475.

[21] A. Cavallaro, O. Steiger, T. Ebrahimi (2003). Semantic segmentation and description for video transcoding. In *Proc. of IEEE Int. Conf. on Multimedia and Expo (ICME)*, vol. 3, pp. 597–600, Baltimore, USA.

[22] M. Cavazza et al. (2003). Users acting in mixed reality interactive storytelling. In *Proc. of Second Int. Conf. on Virtual Storytelling (ICVS)*, vol. 2897 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin.

[23] S.-F. Chang, T. Sikora, A. Puri (2001). Overview of the MPEG-7 standard. *IEEE Transactions on Circuits and Systems for Video Technology,* **11**(6):688–694.

[24] CIE (1986). *Colorimetry, 2nd edition.* CIE Publ. No. 15.2, Vienna.

[25] CIE (1989). *International Lighting Vocabulary.* CIE Publ. No. 17.4, Vienna, 4 edn.

[26] R. Cook, K. Torrance (1987). A reflectance model for computer graphics. In *ARPA Image Understanding Workshop*, pp. 1–19.

[27] R. Cucchiara, C. Grana, M. Piccardi, A. Prati (2001). Detecting objects, shadows and ghosts in video streams by exploiting color and motion information. In *Proc. of 11th Int. Conf. on Image Analysis and Processing (ICIAP)*, pp. 360–365.

[28] R. Cucchiara, C. Grana, M. Piccardi, A. Prati (2003). Detecting moving objects, ghosts and shadows in video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **25**(10):1337–1342.

[29] F. Dufaux, J. Konrad (2000). Efficient, robust, and fast global motion estimation for video coding. *IEEE Transactions on Image Processing,* **9**(3):497–501.

[30] M. D. Fairchild (1997). *Color Appearance Models.* Addison-Wesley, Boston.

[31] G. Finlayson, S. Hordley, M. S. Drew (2002). Removing shadows from images. In *Proc. of European Conference on Computer Vision (ECCV)*, Lecture Notes in Computer Science, pp. 823–836, Springer Verlag, Berlin.

[32] G. D. Finlayson (1996). Color in perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **18**(10):1034–1038.

[33] G. D. Finlayson, S. S. Chatterjee, B. V. Funt (1995). Color angle invariants for object recognition. In *Proc. of IS&T/SID 3rd Color Imaging Conference*, pp. 44–47.

[34] G. D. Finlayson, M. S. Drew, B. V. Funt (1994). Color constancy - generalized diagonal transforms suffice. *J. Opt. Soc. Am. A,* **11**(11):3011–3019.

[35] G. D. Finlayson, S. D. Hordley (2001). Color constancy at a pixel. *J. Opt. Soc. Am. A,* **18**(2):253–264.

[36] G. D. Finlayson, S. D. Hordley, M. S. Drew (2002). Removing shadows from images using retinex. In *Proc. of IS&T/SID 10th Color Imaging Conference*, pp. 73–79.

[37] G. D. Finlayson, B. Schiele, J. L. Crowley (1998). Comprehensive colour image normalisation. In *Proc. of the Fifth European Conference on Computer Vision (ECCV)*, vol. 2 of *Lecture Notes in Computer Science*, pp. 475–490, Springer Verlag, Berlin.

[38] J. Foley, A. van Dam, S. Feiner, J. Hughes (1996). *Computer Graphics: Principles and Practice*. Addison-Wesley, Reading, MA, 2nd edn.

[39] D. Forsyth, A. Zisserman (1989). Mutual illumination. In *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 466–473.

[40] D. A. Forsyth, J. Ponce (2003). *Computer Vision: A Modern Approach*. Prentice Hall, New York.

[41] N. Friedman, S. Russell (1997). Image segmentation in video sequences: A probabilistic approach. In *Proc. of 13th Int. Conf. on Uncertainty in Artificial Intelligence (UAI)*, Morgan Kauffman, Providence, Rhode Island.

[42] G. S. K. Fung, N. H. C. Yung, G. K. H. Pang, A. H. S. Lai (2001). Effective moving cast shadows detection for monocular color image sequences. In *Proc. of 11th Int. Conf. on Image Analysis and Processing (ICIAP)*, pp. 404–409.

[43] G. Funka-Lea, R. Bajcsy (1995). Combining color and geometry for the active, visual recognition of shadows. In *Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 203–209.

[44] B. V. Funt, K. Barnard, L. Martin (1998). Is machine colour constancy good enough? In *Proc. of the Fifth European Conference on Computer Vision (ECCV)*, vol. 2 of *Lecture Notes in Computer Science*, pp. 445–459, Springer Verlag, Berlin.

[45] B. V. Funt, M. S. Drew (1993). Color space analysis of mutual illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **15**(12):1319–1326.

[46] B. V. Funt, G. D. Finlayson (1995). Color constant color indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **17**(5):522–529.

[47] R. Gershon, A. D. Jepson, J. K. Tsotsos (1986). Ambient illumination and the determination of material changes. *J. Opt. Soc. Am. A,* **3**(10):1700–1707.

[48] T. Gevers (2002). Adaptive image segmentation by combining photometric invariant region and edge information. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **24**(6):848–852.

[49] T. Gevers (2004). Robust segmentation and tracking of colored objects in video. *IEEE Transactions on Circuits and Systems for Video Technology,* **14**(6):776–781.

[50] T. Gevers, A. W. M. Smeulders (1999). Color-based object recognition. *Pattern Recognition,* **32**:453–464.

[51] T. Gevers, A. W. M. Smeulders (2000). PicToSeek: Combining color and shape invariant features for image retrieval. *IEEE Transactions on Image Processing,* **9**(1):102–119.

[52] T. Gevers, H. Stokman (2003). Classifying color edges in video into shadow-geometry, highlight, or material transitions. *IEEE Transactions on Multimedia,* **5**(2):237–243.

[53] B. Gnsel, A. M. Tekalp, P. J. van Beek (1998). Content-based access to video objects: Temporal segmentation, visual summarization, and feature extraction. *Signal Processing,* **66**(2):261–280.

[54] E. H. Gombrich (1995). *Shadows.* National Gallery Publications, London.

[55] O. Grau, T. Pullen, G. A. Thomas (2004). A combined studio production system for 3-D capturing of live action and immersive actor feedback. *IEEE Transactions on Circuits and Systems for Video Technology,* **14**(3):370–380.

[56] R. L. Gregory (1997). *Eye and Brain.* Princeton University Press, Princeton, New Jersey, 5th edn.

[57] D. Grest, J.-M. Frahm, R. Koch (2003). A color similarity measure for robust shadow removal in real-time. In *Proc. of 8th Fall Workshop on Vision, Modeling and Visualization (VMV).*

[58] M. D. Grossberg, S. K. Nayar (2003). Determining the camera response from images: What is knowable? *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **25**(11):1455 –1467.

[59] M. D. Grossberg, S. K. Nayar (2003). What is the space of camera response functions? In *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 602 –609.

[60] C. Gu, M.-C. Lee (1998). Semiautomatic segmentation and tracking of semantic video objects. *IEEE Transactions on Circuits and Systems for Video Technology,* **8**(5):572–584.

[61] L. N. Hambrick, M. H. Loew, R. Carroll (1987). The entry-exit method of shadow-boundary segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **9**(5):597–607.

[62] I. Haritaoglu, D. Harwood, L. S. Davis (1998). W4S: A real-time system detecting and tracking people in 2 1/2D. In *Proc. of the Fifth European Conference on Computer Vision (ECCV)*, vol. 1 of *Lecture Notes in Computer Science*, pp. 877–892, Springer Verlag, Berlin.

[63] G. Healey (1989). Using color for geometry-insensitive segmentation. *J. Opt. Soc. Am. A,* **6**(6):920–937.

[64] G. E. Healey (1992). Segmenting images using normalized color. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **22**(1):64–73.

[65] G. E. Healey, R. Kondepudy (1994). Radiometric CCD camera calibration and noise estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **16**(3):267–276.

[66] G. E. Healey, D. Slater (1994). Global color constancy: Recognition of objects by use of illumination-invariant properties of color distributions. *J. Opt. Soc. Am. A,* **11**(11):3003–3010.

[67] S. Hinz, A. Baumgartner (2001). Vehicle detection in aerial images using generic features, grouping, and context. In *Pattern Recognition 2001 (DAGM Symposium 2001)*, vol. 2191 of *Lecture Notes in Computer Science*, pp. 45–52, Springer Verlag, Berlin.

[68] S. Hinz, A. Baumgartner, P. Wasmeier (2001). The role of shadow for 3D-object reconstruction from monocular images. In *Optical 3-D Measurement Techniques (V)*, pp. 354–363, Vienna, Austria.

[69] B. K. P. Horn (1986). *Robot Vision.* MIT Press, Cambridge, MA.

[70] T. Horprasert, D. Harwood, L. S. Davis (1999). A statistical approach for real-time robust background subtraction and shadow detection. In *Proc. of IEEE Int. Conf. on Computer Vision (ICCV), FRAME-RATE Workshop.*

[71] J. Hsieh, W. Hu, C. Chang, Y. Chen (2003). Shadow elimination for effective moving object detection by gaussian shadow modeling. *Journal of Image and Vision Computing,* **21**(6):505–516.

[72] IEC (1999). *Multimedia Systems and Equipment - Colour Measurement and Management - Part 2-1: Colour Management - Default RGB Colour Space - sRGB.* IEC 61966-2-1, Geneva.

[73] K. Ikeuchi (1994). Surface reflection mechanism. In T. Y. Young (ed.), *Handbook of Pattern Recognition and Image Processing: Computer Vision*, pp. 131–160, Academic Press.

[74] R. Irvin, D. Mckeown (1989). Methods for exploiting the relationship between buildings and their shadows in aerial imagery. *IEEE Transactions on Systems, Man, Cybernetics,* **19**:1564–1575.

[75] ISO/CIE (1991). *Colorimetric Observers.* Joint ISO/CIE Standard 10527.

[76] ITU (1990). *Basic Parameter Values for the HDTV Standard for the Studio and for International Programme Exchange.* ITU-R Recommendation BT.709, Geneva.

[77] ITU (1995). *Studio Encoding Parameters of Digital Television for Standard 4:3 and Widescreen 16:9 Aspect Ratio.* ITU-R Recommendation BT.601, Geneva.

[78] Y. Ivanov, A. Bobick, J. Liu (2000). Fast lighting independent background subtraction. *International Journal of Computer Vision,* **37**(2):199–207.

[79] C. Jiang, M. O. Ward (1994). Shadow segmentation and classification in a constrained environment. *CVGIP: Image Understanding,* **59**(2):213–225.

[80] J. Kender (1976). *Saturation, Hue, and normalized colors: Calculation, digitization effects, and use.* Tech. rep., Carnegie-Mellon University.

[81] D. Kersten, P. Mamassian, D. C. Knill (1997). Moving cast shadows induce apparent motion in depth. *Perception,* (26):171–192.

[82] G. J. Klinker, S. A. Shafer, T. Kanade (1990). A physical approach to color image understanding. *International Journal of Computer Vision,* **4**:7–38.

[83] D. C. Knill, P. Mamassian, D. Kersten (1997). Geometry of shadows. *J. Opt. Soc. Am. A,* **14**(12):3216–3232.

[84] D. Koller, K. Danilidis, H.-H. Nagel (1993). Model-based object tracking in monocular image sequences of road traffic scenes. *International Journal of Computer Vision,* **10**(3):257–281.

[85] A. Koschan (1995). A comparative study on color edge detection. In *Proc. of 2nd Asian Conference on Computer Vision (ACCV)*, vol. 3, pp. 574–578, Singapore.

[86] D. J. Kriegman, P. N. Belhumeur (1998). What shadows reveal about object structure. In *Proc. of the Fifth European Conference on Computer Vision (ECCV)*, vol. 1407 of *Lecture Notes in Computer Science*, pp. 399–414, Springer Verlag, Berlin.

[87] E. H. Land (1977). The retinex theory of color vision. *Scientific American,* **237**(6):108–128.

[88] E. H. Land (1986). Recent advances in retinex theory. *Vision Research,* **26**(1):7–21.

[89] H.-C. Lee, E. J. Breneman, C. P. Schulte (1990). Modeling light reflection for computer color vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **12**(4):402–409.

[90] N. Li, J. Bu, C. Chen (2002). Real-time video segmentation using HSV space. In *Proc. of IEEE Int. Conf. on Image Processing (ICIP)*, vol. 2, pp. II–85–II–88.

[91] M. Liévin, F. Luthon (2004). Nonlinear color space and spatiotemporal MRF for hierarchical segmentation of face features in video. *IEEE Transactions on Image Processing,* **13**(1):63–71.

[92] C. Lin, R. Nevatia (1998). Building detection and description from a single intensity image. *Computer Vision and Image Understanding,* **72**(2):101–121.

[93] Y. Liow, T. Pavlidis (1991). Use of shadows for extracting buildings in aerial images. *Computer Vision Graphics Image Processing,* **49**:242–277.

[94] P. Mamassian, D. C. Knill, D. Kersten (1998). The perception of cast shadows. *Trends in Cognitive Sciences,* (2):288–295.

[95] J. A. Marchant, C. M. Onyango (2000). Shadow-invariant classification for scenes illuminated by daylight. *J. Opt. Soc. Am. A,* **17**(11):1952–1961.

[96] B. Marcotegui et al. (1998). A video object generation tool allowing friendly user interaction. In *Proc. of the IEEE International Conference on Image Processing*, vol. 2, pp. 391–395.

[97] X. Marichal, T. Umeda (2003). Real-time segmentation of video objects for mixed-reality interactive applications. In *Proc. of SPIE, Visual Communications and Image Processing (VCIP)*, vol. 5051, pp. 41–50, Lugano, Switzerland.

[98] X. Marichal et al. (2002). The ART.LIVE architecture for mixed reality. In *Proc. of Virtual Reality International Conference (VRIC)*, pp. 19–21, Laval, France.

[99] F. Marques, J. Llach (1998). Tracking of generic objects for video object generation. In *Proc. of the IEEE International Conference on Image Processing*, pp. 628–632.

[100] D. Marr (1982). *Vision.* W.H.Freeman and Company.

[101] Y. Matsushita, K. Nishino, K. Ikeuchi, M. Sakauchi (2002). Shadow elimination for robust video surveillance. In *Proc. of IEEE Workshop on Motion and Video Computing (MOTION)*, pp. 15–21.

[102] Y. Matsushita, K. Nishino, K. Ikeuchi, M. Sakauchi (2003). Illumination normalization with time-dependent intrinsic images for video surveillance. In *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. I–3–I–10.

[103] R. Mech, M. Wollborn (1997). A noise robust method for 2D shape estimation of moving objects in video sequences considering a moving camera. In *Workshop on Image Analysis for Multimedia Interactive Services*, vol. 5, pp. 494–497, Louvain la Neuve, Belgium.

[104] T. Meier, K. Ngan (1998). Automatic segmentation of moving objects for video object plane generation. *IEEE Transactions on Circuits and Systems for Video Technology,* **8**(5):525–538.

[105] I. Mikic, P. C. Cosman, G. T. Kogut, M. M. Trivedi (2000). Moving shadow and object detection in traffic scenes. In *Proc. of IEEE Int. Conf. on Pattern Recognition (ICPR)*, pp. 321–324.

[106] P. Milgram, H. Colquhoun (1999). A taxonomy of real and virtual world display integration. In Y. Otha, H. Tamura (eds.), *Mixed Reality, Merging Real and Virtual Worlds*, Ohmsha-Springer.

[107] A. Mitiche, P. Bouthemy (1996). Computation and analysis of image motion: a synopsis of current problems and methods. *International Journal of Computer Vision,* **19**(1):29–55.

[108] S. Nadimi, B. Bhanu (2001). Multistrategy fusion using mixture model for moving object detection. In *Proc. of Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems*, pp. 317–322.

[109] S. Nadimi, B. Bhanu (2002). Moving shadow detection using a physics-based approach. In *Proc. of IEEE Int. Conf. on Pattern Recognition (ICPR)*, vol. 2, pp. 701–704.

[110] K. Nagao, W. E. L. Grimson (1998). Using photometric invariants for 3D object recognition. *Computer Vision Image Understanding,* **71**(1):74–93.

[111] M. Nagao, T. Matsuyama (1979). Region extraction and shape analysis in aerial photographs. *Computer Vision Graphics Image Processing,* **10**(3):195–223.

[112] A. Nandi, X. Marichal (2000). Transfiction. In *Proc. of Virtual Reality International Conference (VRIC)*, pp. 76–88, Laval, France.

[113] S. G. Narasimhan, V. Ramesh, S. K. Nayar (2003). A class of photometric invariants: separating reflectance from shape and illumination. In *Proc. of IEEE Ninth Int. Conf. on Computer Vision (ICCV)*, vol. 2, pp. 1387–1394.

[114] S. K. Nayar, R. Bolle (1993). Computing reflectance ratios from an image. *Pattern Recognition,* **26**:1529–1542.

[115] S. K. Nayar, K. Ikeuchi, T. Kanade (1991). Surface reflection: Physical and geometrical perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **13**(7):611–634.

[116] K. Onoguchi (1998). Shadow elimination method for moving object detection. In *Proc. of IEEE Int. Conf. on Pattern Recognition (ICPR)*, pp. 583–587.

[117] Y. Otha, T. Kanade, T. Sakai (1980). Color information for region segmentation. *Computer Graphics and Image Processing,* **13**:222–241.

[118] A. Papoulis (1991). *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, 3rd edn.

[119] N. Paragios, R. Deriche (1999). Geodesic active regions for motion estimation and tracking. In *Proc. of 7th International Conference on Computer Vision (ICCV).*

[120] F. Pavani, U. Castiello (2004). Binding personal and extrapersonal space through body shadows. *Nature Neuroscience,* **7**(1):13–14.

[121] F. C. Pereira, T. Ebrahimi (2002). *The MPEG-4 Book.* Prentice Hall PTR.

[122] F. Perez, C. Koch (1994). Toward color image segmentation in analog VLSI: Algorithm and hardware. *Journal of Computer Vision,* **12**(1):17–42.

[123] N. Peterfreund (1998). Robust tracking of position and velocity with Kalman snakes. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **21**(6):564–569.

[124] B. T. Phong (1975). Illumination for computer generated pictures. *Comm. of the ACM,* **18**(6):311–317.

[125] J. Pinel, H. Nicolas (2001). Estimation of 2D illuminant direction and shadow segmentation in natural video sequences. In *Proc. of VLBV*, pp. 197–202, Athens, Greece.

[126] J. Pinel, H. Nicolas (2002). Shadows analysis and synthesis in natural video sequences. In *Proc. of IEEE Int. Conf. on Image Processing (ICIP)*, Rochester, USA.

[127] J. Pinel, H. Nicolas (2003). Cast shadows detection on Lambertian surfaces in video sequences. In *Visual Communications and Image Processing 2003*, vol. 5150 of *Proc. of SPIE*, pp. 378–384, Lugano, Switzerland.

[128] P. Piscaglia, A. Cavallaro, M. Bonnet, D. Douxchamps (1999). High level descriptors of video surveillance sequences. In *Proc. of 4th European Conference on Multimedia Applications, Services and Techniques (ECMAST)*, pp. 316–331, Madrid, Spain.

[129] K. Plataniotis, A. Venetsanopoulos (2000). *Color Image Processing and Applications,*. Springer Verlag, Berlin.

[130] A. M. Polidorio et al. (2003). Automatic shadow segmentation in aerial color images. In *Proc. of IEEE 16th Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI)*, pp. 270–277.

[131] C. Poynton (1997). Frequently asked questions about color, http://www.poynton.com/ColorFAQ.html.

[132] C. Poynton (1998). The rehabilitation of gamma. In *Proc. of SPIE, Human Vision and Electronic Imaging III*, vol. 3299, pp. 232–249.

[133] C. Poynton (2003). *Digital Video and HDTV.* Morgan Kaufmann Publishers, San Francisco.

[134] A. Prati, I. Mikic, M. Trivedi, R. Cucchiara (2003). Detecting moving shadows: Algorithms and evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **25**(7):918–923.

[135] W. Pratt (1991). *Digital Image Processing.* John Wiley & Sons, New York.

[136] V. Risson (2001). *Application de la Morphologie Mathématique à l'Analyse des conditions d'Éclairage des Images Couleur.* Ph.D. thesis, Paris School of Mines.

[137] P. L. Rosin, T. Ellis (1995). Image difference threshold strategies and shadow detection. In *Proc. of British Machine Vision Conference*, pp. 347–356.

[138] J. M. Rubin, W. A. Richards (1982). Color vision and image intensities: when are changes material? *Biol. Cybern.,* **41**:215–226.

[139] J. M. Rubin, W. A. Richards (1984). *Color Vision: Representing material categories.* Tech. rep., MIT.

[140] E. Salvador, A. Cavallaro, T. Ebrahimi (2001). Shadow identification and classification using invariant color models. In *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, pp. 1545–1548.

[141] J. M. Scanlan, D. M. Chabries, R. Christiansen (1990). A shadow detection and removal algorithm for 2D images. In *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2057–2060.

[142] O. Schreer, I. Feldmann, U. Goelz, P. Kauff (2002). Fast and robust shadow detection in videoconference applications. In *Proc. of VIPromCom 2002, 4th EURASIP IEEE Int. Symposium on Video Processing and Multimedia Communications*, pp. 371–375.

[143] S. A. Shafer (1985). Using color to separate reflection components. *COLOR Research Applications,* **10**(4):210–218.

[144] S. A. Shafer, T. Kanade (1983). Using shadows in finding surface orientations. *Computer Vision Graphics Image Processing,* **22**:145–176.

[145] T. Sikora (1997). The MPEG-4 video standard verification model. *IEEE Transactions on Circuits and Systems for Video Technology,* **7**(1):19–31.

[146] T. Sikora (2001). The MPEG-7 visual standard for content description – an overview. *IEEE Transactions on Circuits and Systems for Video Technology,* **11**(6):696–702.

[147] J. J. Simpson, Z. Jin, J. R. Stitt (2000). Cloud shadow detection under arbitrary viewing and illumination conditions. *IEEE Transactions on Geoscience and Remote Sensing,* **38**(2):972–976.

[148] D. Slater, G. Healey (1996). The illumination-invariant recognition of 3D objects using local color invariants. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **18**(2):206–210.

[149] Y. Sonoda, T. Ogata (1998). Separation of moving objects and their shadows, and application to tracking of loci in the monitoring images. In *Proc. of IEEE Int. Conf. on Signal Processing (ICSP)*, pp. 1216–1264.

[150] J. Stauder, R. Mech (1999). Detection and tracking of moving cast shadows. In *COST211 Workshop on Image Analysis for Multimedia Interactive Services*, Berlin, Germany.

[151] J. Stauder, R. Melch, J. Ostermann (1999). Detection of moving cast shadows for object segmentation. *IEEE Transactions on Multimedia,* **1**(1):65–77.

[152] O. Steiger, A. Cavallaro, T. Ebrahimi (2002). MPEG-7 description of generic video objects for scene reconstruction. In *Proc. of SPIE Electronic Imaging – Visual Communications and Image Processing*, San Jose, California, USA.

[153] A. Stockman, L. Sharpe (2000). The spectral sensitivities of the middle- and long-wavelength-sensitive cones derived from measurements in observers of known genotype. *Vision Research,* **40**(13):1711–1737.

[154] V. I. Stoichita (1997). *A short history of the shadow.* Reaktion, London.

[155] M. Stricker, M. Orengo (1995). Similarity of color images. In *Proc. of SPIE, Storage and Retrieval for Image and Video Databases III*, vol. 2420, pp. 381–392.

[156] S. Sun, D. R. Haynor, Y. Kim (2003). Semiautomatic video object segmentation using VSnakes. *IEEE Transactions on Circuits and Systems for Video Technology,* **13**(1):75–82.

[157] M. J. Swain, D. H. Ballard (1991). Color indexing. *International Journal of Computer Vision,* **7**(1):11–32.

[158] H. Tao, H. S. Sawhney, R. Kumar (2002). Object tracking with bayesian estimation of dynamic layer representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **24**(1):75–89.

[159] S. Tominaga (1991). Surface identification using the dichromatic reflection model. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **13**(7):658–670.

[160] S. Tominaga, B. A. Wandell (1989). Standard surface-reflectance model and illuminant estimation. *J. Opt. Soc. Am. A,* **6**(4):576–584.

[161] V. J. D. Tsai (2003). Automatic shadow detection and radiometric restoration on digital aerial images. In *Proc. of IEEE Int. Geoscience and Remote Sensing Symposium (IGARSS)*, vol. 2, pp. 732–733.

[162] Y. Tsaig, A. Averbuch (2002). Automatic segmentation of moving objects in video sequences: a region labeling approach. *IEEE Transactions on Circuits and Systems for Video Technology,* **12**(7):597–612.

[163] N. Vandenbroucke (2000). *Segmentation d'images couleur par classification de pixels dans des espaces d'attributs colorimétriques adaptés. Application à l'analyse d'images de football.* Ph.D. thesis, Université des sciences et technologies de Lille 1.

[164] D. Waltz (1975). *The Psychology of Computer Vision, Understanding Line Drawings of Scenes with Shadows*, pp. 19–91. McGraw-Hill, New York.

[165] B. A. Wandell (1995). *Foundations of vision.* Sinauer Associates, Sunderland, Massachusetts.

[166] C. Wang, L. Huang, A. Rosenfeld (1991). Detecting clouds and cloud shadows on aerial photographs. *Pattern Recognition Letters,* **12**:55–64.

[167] D. Wang (1998). Unsupervised video segmentation based on watersheds and temporal tracking. *IEEE Transactions on Circuits and Systems for Video Technology,* **8**(5):539–546.

[168] J. M. Wang, Y. C. Chung, C. L. Chang, S. W. Chen (2004). Shadow detection and removal for traffic images. In *Proc. of IEEE Int. Conf. on Networking, Sensing and Control (ICNSC)*, vol. 1, pp. 649–654.

[169] Y. Wang, T. Tong, K.-F. Loe (2003). A probabilistic method for foreground and shadow segmentation. In *Proc. of IEEE Int. Conf. on Image Processing (ICIP)*, vol. 3, pp. 937–940.

[170] Y. Weiss (2001). Deriving intrinsic images from image sequences. In *Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, vol. 2, pp. 68–75.

[171] S. G. Wilson (1996). *Digital modulation and coding.* Prentice Hall, New Jersey, USA.

[172] A. P. Witkin (1982). Intensity-based edge classification. In *Proc. of National Conference on Artificial Intelligence*, pp. 46–41.

[173] G. Wyszecki, W. S. Stiles (1982). *Color Science: Concepts and Methods, Quantitative Data and Formulae.* John Wiley & Sons, New York, 2nd edn.

[174] A. Yoneyama, C. H. Yeh, C.-C. J. Kuo (2003). Moving cast shadow elimination for robust vehicle extraction based on 2D joint vehicle/shadow models. In *Proc. of IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS).*

[175] J. J. Yoon, C. Koch, T. J. Ellis (2002). ShadowFlash: an approach for shadow removal in an active illumination environment. In *Proc. of British Machine Vision Conference (BMVC)*, pp. 636–645.

[176] W. Zhang, F. Bergholm (1997). Multi-scale blur estimation and edge type classification for scene analysis. *International Journal of Computer Vision,* **24**(3):219–250.

[177] J. W. Zhao, P. Wang, C. Q. Liu (2002). An object tracking algorithm based on occlusion mesh model. In *Proc. of IEEE Int. Conf. on Machine Learning and Cybernetics (ICMLC)*, pp. 288–292.

[178] S.-Y. Zhu, N. Plataniotis, A. N. Venetsanopoulos (1999). Comprehensive analysis of edge detection in color image processing. *Optical Engineering,* **38**(4):612–625.

Curriculum Vitae

Personal information

Name:	Elena Salvador
Nationality:	Italian
Date of birth:	May 20, 1975
Place of birth:	Udine, Italy
Marital status:	Single
Address:	Rue de la Tour 5 1004 Lausanne, Switzerland
Phone:	+41 78 7320720
Fax:	+41 21 693 76 00
Email:	Elena.Salvador@epfl.ch

Work experience

- ***September 2000 – present***: Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland
  Research assistant
  - Research on image and video analysis.
  - Development of shadow segmentation algorithms for still color images and video sequences.
  - Teaching: definition and supervision of student projects and responsible for exercises and lab sessions for the "Image and Video Processing" course.
  - Responsible for EPFL's contribution to the European IST project *art.live* on vision based human computer interaction in multimedia environments.
  - Participation in the European Network of Excellence VISNET (Networked Audiovisual Media Technologies).

Education

- ***September 2000 – August 2004***: Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland
  Ph. D student in Electrical and Electronic Engineering

- ***June 2000***: University of Trieste, Trieste, Italy
  Laurea (M. Sc.) in Electronic Engineering
  Recipient of the *Marisa Bellisario Award* assigned every year to young women who achieve top marks in an Engineering degree in Electronics and Telecommunications (www.fondazionebellisario.org).

- ***October 1999 – April 2000***: Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland
  Master Thesis as *Erasmus student* at the Signal Processing Institute.

Skills

Languages

Italian:	mother tongue
English:	fluent
French:	fluent
German:	intermediate

Computer literacy

Operating systems:	LINUX, Unix, Windows
Programming languages:	Pascal, C, C++
Software:	Matlab, familiarity with Spice and L-Edit, LaTeX, MS Office

Personal interests

- *Swimming*: 2002, Brevet I de natation de sauvetage, Societé Suisse de Sauvetage SSS.

- *Water–polo*: 2nd Swiss league player with Lausanne Water–polo Club (since October 2001).

- *Reading*: literature and arts.

Publications

Journal papers

- E. Salvador, A. Cavallaro and T. Ebrahimi, 'Cast shadow segmentation using invariant color features', *Computer Vision and Image Understanding*, vol. 95, n. 2, August 2004, pp. 238-259.

- A. Cavallaro, E. Salvador and T. Ebrahimi, 'Shadow-aware object-based video processing', submitted to *IEE Proceedings Vision Image & Signal Processing.*

Conference papers

- A. Cavallaro, E. Salvador and T. Ebrahimi, 'Shadow detection in image sequences', Proc. of *IEE Conference on Visual Media Production (CVMP)*, London, UK, March 2004.
- E. Drelie Gelasca, E. Salvador and T. Ebrahimi, 'Intuitive Strategy for Parameter Setting in Video Segmentation', Proc. of *SPIE, Visual Communications and Image Processing 2003*, Vol. 5150, p. 998-1008, Lugano, Switzerland, July 2003.
- E. Salvador, A. Cavallaro and T. Ebrahimi, 'Spatio-temporal Shadow Segmentation and Tracking', Proc. of *SPIE, Image and Video Communications and Processing 2003*, Vol. 5022, pp. 389-400, Santa Clara, California, USA, January 2003.
- E. Salvador and T. Ebrahimi, 'Cast Shadow Recognition in Color Images', Proc. of *11th European Conference on Signal Processing (EUSIPCO)*, Vol. 3, pp. 555-558, Toulouse, France, September 2002.
- E. Salvador, A. Cavallaro and T. Ebrahimi, 'Shadow Identification and Classification using Invariant Color Models', Proc. of *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 3, pp. 1545-1548, Salt Lake City, Utah, USA, May 2001.